Utilization of Methods of Spatial Analysis in Road Safety Evaluation

Paper # 15-2482

Johnny Tse
Graduate Student
Department of Civil Engineering
Carleton University
1125 Colonel By Drive
Ottawa, Ontario, Canada, K1S 5B6
Tel: (613) 291-5805
E-mail: johnnytse@cmail.carleton.ca

Corresponding Author:
Yasser Hassan
Professor
Department of Civil and Environmental Engineering
Carleton University
1125 Colonel By Drive
Ottawa, Ontario, Canada, K1S 5B6
Tel: (613) 520-2600X8625
Fax: (613) 520-3951
E-mail: yasser_hassan@carleton.ca

Dan Patterson
Instructor
Department of Geography and Environmental Studies
Carleton University
1125 Colonel By Drive
Ottawa, Ontario, Canada, K1S 5B6
E-mail: Dan.Patterson@carleton.ca

Word count:
Manuscript 6,787
Tables (2*250) 500
Figures (2*250) 500
Total 7,787

Paper submitted for possible presentation at the 94th Annual Meeting of the Transportation Research Board, Washington, D.C., January 11-15, 2015 and possible publication in Transportation Research Record

July 2014
Abstract

Network screening is an important step in road safety analysis. Statistical techniques are currently available and have been integrated in the Highway Safety Manual and the SafetyAnalyst software. However, analysis can still be complex to conduct. Several alternative approaches that use spatial analysis methods are identified and applied to collision data from the Ontario Provincial highway network. The results are compared to those of SafetyAnalyst. The results show the potential of the spatial analysis methods, particularly the Local Moran I index method in road network screening. In addition, with the ability of the spatial analysis methods to identify clusters of a specific type of collisions, they can be easily used in speed management efforts in order to identify clusters of speed-related collisions.
INTRODUCTION

Road safety is a concern for many countries around the world. The World Health Organization lists vehicle collisions as a leading cause of death amongst deceases \((1)\). In 2010, there were 215,533 vehicle collisions on Ontario roadways. Of these collisions, there were 534 fatal collisions and 44,430 collisions involving injuries \((2)\).

Based on published statistics, speed is a contributing factor in many vehicle collisions in Ontario, Canada. In 2010 “excessive speed for road conditions” and “speed over the posted limit” are listed as contributing factors for 13% of all fatal collisions in Ontario \((2)\). Speed management is a range of measures designed to provide a balance between the safety of all road users and the efficiency of a road network \((3)\). Speed management measures are a combination of engineering design, public education, and law enforcement.

The engineering design portion of speed management may be implemented at the design stage or during the operational stage. The design stage includes the initial design of horizontal curves, tangent sections, and other geometric features. The other aspect of engineering design in speed management is the implementation of countermeasures once an existing road segment is identified to experience more speed-related collisions than normal through network screenings.

To identify such road segments, statistical techniques, which have evolved in recent years, are typically used for road safety analysis. However, Azam et. al. \((4)\) surveyed 48 state transportation agencies and found that only few state transportation agencies used advanced statistical techniques for road safety analysis as a standard procedure. On the other hand, many agencies continue to use basic statistical analysis methods. This may be due to the complexity of the statistical techniques and the lack of statistical expertise.

To assist transportation agencies in adopting advanced statistical techniques, the Federal Highway Administration (FHWA) offers software solutions for road safety analysis: IHSDM for project level analysis and SafetyAnalyst for network level analysis. Both of these software applications can help reduce the expertise required for the application of advanced road safety analysis techniques. They use graphical user interfaces and step-by-step analytical processes as opposed to general statistical software packages, which require a certain level of programming knowledge. Still, a user with a strong statistical background should operate such software to ensure that statistical techniques are applied properly and correct inferences are drawn from the results.

Geographical information systems (GIS) are also commonly used in the field of transportation engineering and planning. The most common application is digitizing and inventoring existing data such as road segment classification and posted speed limits. However, GIS software packages also offer a range of in-depth statistical analysis methods based on the location of an event. These methods have been widely developed and used in geography and biology where events such as animal road kills are studied \((5, 6)\). However, such analysis methods have been sparsely applied in road safety analysis.

With advances in data collection, storage, and archival methods by transportation agencies, the amounts of data available for analysis have increased significantly. Given the amounts of data collected and with the increasing utilization of GIS software by transportation agencies, additional insights may be gained by applying spatial analysis techniques to collision data. With the ability to analyze collision data in combination with a wide range of data on potential...
contributing factors such as speed, weather, and socioeconomic data; spatial analysis techniques can reveal collision patterns and causative factors that may not be easily realized using traditional statistical analysis methods.

OBJECTIVE AND SCOPE

The objective of this paper is to evaluate the potential of using methods of spatial analysis for the purpose of network screening as an alternative to the advanced statistical analysis commonly used in safety analysis. These network screening methods must have the ability to: use collision data to identify highway segments that experience more collisions than normal and thus have a potential for safety improvement, provide statistically sound analysis results at both segment and network level, be implemented in freely available and/or commonly used commercial geographical information system software packages. Finally, the analysis methods need to have the ability to easily consider a specific contributing factor so that they can be easily applied in such efforts as speed management.

The proposed network screening techniques are evaluated by applying them to the road segments that make up the Ontario provincial road network. The application is limited to road segments as the spatial analysis methods used in this paper are developed to examine events along a path. Ramp and intersection collisions are not considered because they are always recorded at the ramp or intersection. Therefore, the spatial analysis methods would always flag all ramps and intersections as locations for potential of improvement.

In applying the proposed network screening methods, the objective is set as to determine locations where speed is a contributing factor to higher collision occurrence or collision severity. Thus, safety at these locations can be improved through speed management countermeasures, whether they may be education, law enforcement, or engineering.

SAFETY ANALYSIS TOOLS

Background

A review of available spatial analysis techniques showed that they may be used in network screening and to identify road segments with potential for improvement through speed management. In general, there are two types of spatial analysis techniques: global and local. Global analysis techniques reveal whether there is a clustering of events in the area under study. Local analysis techniques on the other hand, can identify the locations at which these clustering occur, also known as “hotspots”. The latter type of spatial analysis is particularly useful as it has the potential to locate road segments with potential for improvement based on the clustering of events, which are vehicle collisions in this case. Given the general goal of network screening, only local spatial analysis methods are used in this paper.

The physical distances between events are a crucial part of spatial analyses. Many existing spatial analysis techniques are based on the use of Euclidean (direct) distances between events. In this approach, the study area is treated as a two-dimensional planar space, and the distance between two events is measured along the straight line between the two events. The use of Euclidean-distance-based methods is particularly useful as spatial analyses tend to look for the clustering of events that are spread out over an area of interest. However, the use of these
methods may lead to overestimation in collision hotspots. This is due to the fact that the location of automobile collisions takes place only on the road network, which is a one-dimensional subset of the two-dimensional study area. Therefore, the use of Euclidean distance based methods violates this fundamental assumption (7, 8). However, the capability of identifying clusters of collisions over an area, even when they are on different road segments, can be useful in identifying areas where a specific type of collisions (e.g. speed-related) is prominent. Such areas may be targeted by education and enforcement countermeasures to improve road safety. Spatial analysis methods that will be applied in this paper involve both Euclidean distance and network distance based methods. The effect of the use of Euclidean distance based methods is evaluated by analyzing the highway network as a whole and then analyzing the highways on an individual basis.

The complete spatial randomness (CSR) hypothesis is a fundamental hypothesis in statistical spatial analysis, where point events (collisions in this case) within a study area are hypothesized to occur in a completely random manner. When testing the CSR hypothesis, the random point events are generated according to the homogeneous binomial point process for network methods (methods based on network distances) or the Poisson point process for non-network methods (methods based on Euclidean distances). The details of these methods are available in the literature (9, 10) When the CSR hypothesis is rejected at a particular level of significance, the events along the network are believed to be nonrandom and that there are clusters of events along the network. The CSR hypothesis is applied in all spatial analysis methods in this paper.

**Spatial Analysis Methods**

**Local Moran’s I**

The Local Moran’s I (LMI) index is a non-network method (based on Euclidean distances) that indicates the presence of statistically significant spatial clusters for each location. The index may also indicate the absence of spatial clustering, also known as dispersion. It belongs to a class of spatial association indicators known as Local Indicators of Spatial Association (LISA) (11-13). By comparing the local index at a particular location to a global one for the whole study area, the presence or absence of a cluster at this location is identified. The Local Moran’s index, \( I_L \), is given as:

\[
I_L = \frac{x_i - \bar{X}}{S^2} \sum_{j=1, j \neq i}^n w_{ij} (x_j - \bar{X})
\]  

Where \( x_i \) = an attribute (e.g. the level of injury severity) for feature \( i \) (e.g. a collision); \( \bar{X} = \frac{\sum x_i}{n} \) = mean of the corresponding attribute; \( w_{ij} \) = spatial weight between features (collisions) \( i \) and \( j \); \( n \) = total number of features; and \( S^2 = \frac{\sum_{j=1}^n (x_j - \bar{X})^2}{n-1} \) = sample variance of the attribute.

The z-score \( z_{I_L} \) of the index \( I_L \) is computed as:

\[
z_{I_L} = \frac{I_L - E[I_L]}{\sqrt{V[I_L]}}
\]  

Where: \( E[I_L] \) and \( V[I_L] \) = expected value and population variance of the index \( I_L \), respectively, calculated as:
As mentioned earlier, the null hypothesis for the LMI method is the complete spatial randomness (CSR) hypothesis. If the null hypothesis is rejected for a specific feature \( i \), other features (collisions) are said to be spatially clustered around this feature. In existing transportation engineering literature, LMI index has been preferred over other local spatial autocorrelation measures as it produces more consistent results \((14)\). A disadvantage of this method is that some authors have noted that LMI index only identifies the presence of a cluster or lack thereof, but it does not offer feedback on whether it is a cluster of high or low attributes associated with the features in the cluster. With current spatial analysis packages (such as ArcGIS, GeoDa), this is not an issue as the final step of the analysis procedure will sort through the identified clusters and label each cluster as high or low based on the mean attribute value \((\bar{X})\) for the features in the cluster. Lastly, the analysis requires testing the CSR hypothesis for all feature locations, which is referred to in the literature as multiple comparisons \((13)\). The problems with multiple comparisons are present in many spatial autocorrelation methods, and can be significant in analyzing collision data, where there are hundreds or thousands of records to be compared. If a single hypothesis test is conducted at 5\% level of significance, the probability of false rejection of the null hypothesis is only 5\%. When a large number of tests are conducted simultaneously, for example if 100 tests were conducted simultaneously at 5\% level of significance and all null hypotheses were rejected, the expected number of false rejections is 5. The problem is amplified for collision analysis using spatial methods as thousands of tests are conducted simultaneously; therefore a high number of tests are expected to be falsely rejected. Some software utilize error control techniques to minimize this problem as explained later.

**Getis-Ord** \( G_i^* \)

The Getis-Ord \( G_i^* \) (GOG) statistic uses Euclidean distances to measure the degree of association resulting from the concentration or dispersion of the values of an attribute \( x \) (e.g. whether the collision is speed-related) associated with features (collisions) in the study area. The \( G_i^* \) statistic allows for tests of the CSR hypothesis about the spatial concentration of the sum of \( x \) values associated with all features within distance \( d \) of the \( i \)th feature \((15, 16)\).

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

(5)

Where \( x_j \) = attribute value for feature \( j \); \( w_{ij} \) = spatial weight between features \( i \) and \( j \); \( n \) = total number of features; \( \bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \) = mean value of \( x_j \); and \( S = \frac{\sum_{j=1}^{n} (x_j - \bar{X})^2}{n-1} \) = standard deviation of \( x_j \) values.

The \( G_i^* \) statistic has a similar formulation to the LMI index. However, the \( G_i^* \) statistic is a \( z \)-score, and therefore no additional calculations are needed for the \( z \)-score. Whilst LMI statistic will identify locations with positive spatial autocorrelation or clustering, it does not directly identify whether it is a cluster of high values or a cluster of low values. Whereas the \( G_i^* \) statistic will locate hot spots (clusters of high values) and cold spots (clusters of low values). As
mentioned earlier, further data processing is required for LMI statistic in order to identify which clusters are of high values and vice versa. Overall, Getis and Ord (17) recommended both $G_t^*$ and $I$ statistics to be used to verify the results.

**Network Constrained Global Cross Nearest-Neighbor**

The Network Constrained Global Cross Nearest-Neighbor (NCGCNN) method uses the network distances to examine the spatial association between two attributes of observed events (collisions) in terms of the shortest path distance from every type $A$ event to the nearest type $B$ event. In other words, NCGCNN attempts to answer the question “Given two sets of points on a network, how can we test whether the shortest-path distance from each point of type $A$ to the nearest point of type $B$ is significantly short (or long)?” For the purpose of collision analysis, it can be used to examine the association between two collision attributes such as the level of severity and whether the collision is speed-related. In such example, we attempt to answer the question: “Do speed-related collisions tend to occur near collisions resulting in injuries or fatalities?” (10).

To answer this question, the CSR hypothesis is tested. The observed events (collision data) are compared against the expected events, which are generated by simulation. This is done by first estimating the number of type $A$ points (e.g. injury collisions) that are on average significantly close to the nearest type $B$ points (e.g. speed-related collisions). For a specific distance $t$, the number of the number of type $A$ points within this distance, $n_A(t)$, that is too close to a type $B$ point needs to satisfy the following condition:

$$d_s(p_{A_j}, p_{B_i}^*) \leq t \forall j = [1, n_A]$$  

(6)

Where $d_s$ = nearest network distance between type $A$ point ($p_{A_j}$) and the closest type $B$ point ($p_{B_i}^*$); $n_A$ = number of all type $A$ points.

The probability that the distance $d_s$ is less than or equal to the distance $t$ in the simulated data is given by the probability distribution function, $F_B(t)$:

$$F_B(t) = \frac{1}{|L|} \sum_{i=1}^{n_B} |L|(t|p_{B_i})|$$  

(7)

Where $L$ = length of analysis segment and $n_B$ = number of all type $B$ points.

For a given $t$, the random number $n_A(t)$ under the CSR hypothesis follows a binomial distribution whose mean, $E(n(t))$, and variance, $Var(n(t))$, are given by:

$$E(n(t)) = n_A \frac{1}{|L|} \sum_{i=1}^{n_B} |L|(t|p_{B_i})|$$  

(8)

$$Var(n(t)) = n_A \left( \frac{1}{|L|} \sum_{i=1}^{n_B} |L|(t|p_{B_i})| \right) \left( 1 - \frac{1}{|L|} \sum_{i=1}^{n_B} |L|(t|p_{B_i})| \right)$$  

(9)

**Other Spatial Analysis Methods Considered**

The network constrained Kernel Density Estimation (KDE) and network constrained K-Function (simply referred to as K-Function) methods were also considered but were not used for analysis. The reason for the exclusion for these two methods will be explained in further detail in the preliminary analysis section.
The K-Function method is a distance-based point pattern analysis method similar to the aforementioned NCGCNN method. Both the K-Function and NCGCNN methods can be used for clustering analysis, but they do so differently. As mentioned earlier, the NCGCNN method attempts to answer the question whether the shortest distance between points are significantly short (or long). The K-Function method on the other hand addresses the question whether the number of points within a distance from each point is significantly many (or few) (10).

The network constrained KDE method (hereby KDE only), is a network point density estimation method. It attempts to answer the question: “For a given set of points on a network, how can we estimate the density of points along the network and detect significantly high-density areas on the network?” (10).

Software Applications for Spatial Analysis

A number of software applications were investigated for use in spatial analysis. The two main applications used here are GeoDa and ArcGIS. GeoDa is a free, open source application that “serves as an introduction to spatial data analysis” (18). ArcGIS is a commercial GIS application that offers a range of tools including those for spatial analysis (19). Both of these applications offer the LMI and GOG methods, but differ in the implementation of these methods. Specifically, ArcGIS’ implementation provides the option of using False Discovery Rate (FDR) error control, to attempt to control the expected portion of falsely rejected null hypotheses; whereas GeoDa’s implementation does not. The results from both applications will be compared and evaluated against each other.

Part of the analysis is to examine the differences or similarities between the locations of clusters identified by spatial analysis methods and the locations flagged by SafetyAnalyst. To do this, NCGCNN method is used. This method is available in SANET, which is a third-party add-on toolset for analyzing network events and is designed to be used in conjunction with ArcGIS (20). The KDE and K-Function methods are also part of the SANET toolset, but these two methods will not be used in this paper as mentioned earlier.

Aside from analysis methods offered by the aforementioned software packages, GeoDaNet (21) and PySal (22) also offer several network constrained spatial analysis methods. Unfortunately, GeoDaNet’s development has largely ceased and its spatial analysis tools were not functional. PySal on the other hand have not fully integrated network constrained analysis methods into their code library and therefore cannot be used.

SafetyAnalyst

As mentioned earlier, SafetyAnalyst is a tool developed by the FHWA to allow transportation agencies to use advanced statistical methods in safety analysis (23). It implements several proven advanced statistical analysis methods adopted by the Highway Safety Manual (HSM) (24). It is currently an AASTHOware licensed software which includes a set of tools to aid the decision making process in implementing road safety improvement projects. The tools include: network screening and diagnosis, countermeasure selection and evaluation, collision data management, and administration. The network screening and diagnosis tool in the SafetyAnalyst is used here for comparison against spatial analysis methods. SafetyAnalyst was chosen as it is widely used by North American transportation agencies.
CASE STUDY

This section presents the evaluation of identified spatial analysis tools by applying them to the Ontario provincial road network as a case study. The data, scope of analysis, analysis procedure, and results are discussed.

Data

The collision and road data for the entire Ontario provincial highway network were provided by the Ministry of Transportation Ontario (MTO). Road data include the linear highway referencing system (LHRS) data, traffic volume data based on LHRS segments (e.g. AADT), and highway specifications data. The highway specifications data include information such as whether the highway is divided, pavement material, and shoulder width.

The collision data were originally collected by the Ontario Provincial Police (OPP) and were provided in three databases: MVAB, MVAD, and MVAI. The three databases are part of the Accident Information System, which is a digital collision database managed by MTO. The MVAB database includes all collisions recorded by OPP and includes basic information such as collision location, number of injuries involved, and collision severity. The MVAD database includes more details on all collisions such as the road condition, driver’s age, and the speed limit of the road. The MVAI database only includes collision records with injuries and/or fatalities. Information such as a more detailed injury classification, vehicle damage level, and the vehicle type are included in the MVAI database.

GIS road data and shapefiles were created by Digital Mapping Technologies Inc. Spatial (DMTI Spatial) and were available through Carleton University library. The GIS dataset included the entire Ontario road network but only data related to the Ontario provincial highways were used. In addition to the collision and road data provided by MTO, a formatted collision and road information database was also provided by MTO for use with SafetyAnalyst. Included with the database are modified safety performance functions (SPFs) to be used by SafetyAnalyst. The SPFs are specifically modified for Ontario provincial highway collision data and are different from the SafetyAnalyst’s default SPFs.

Scope of Analysis

Data analysis was conducted on Ontario provincial highways, which make up a network of freeways and undivided arterials managed by MTO. The collision records of only year 2008 were used in this paper, mainly due to the computational resources required for spatial analysis methods. Different sets of analysis were conducted at the network level for two sub-networks of freeways and arterial highways as well as for each road separately. Finally, the analysis focused on examining two collision attributes: severity and whether the collision is speed-related. Therefore, the results should address the feasibility of spatial analysis methods as general and speed management network screening methods.

Variables

Two variables from the collision database were analyzed. The first variable, DRACT, contains information of the drivers’ apparent action immediate prior to the collision excluding any evasive action. The list of actions includes “driving properly”, “following too close”, “speed exceed limit”, “speed too fast for condition”, “speed too slow”, “improper turn”, “disobey traffic
controls”, “fail to yield”, “improper passing”, “lost control”, “wrong way”, and “lane change”.

The specific driver actions of concern are “speed exceed limit”, and “speed too fast for
condition”. For the purpose of analysis, these two categories were merged into one, and
 collisions with these two apparent driver actions are considered as speed-related collisions. In
order to find locations where speed-related collisions may occur, the variable DRACT was re-
coded from fourteen categories into a binary variable containing two categories: speed-related
collisions (code = 1) and all other collisions (code = 0). Thus, local spatial analysis methods
should identify locations where the clustering of events with high and/or low attributes occurs,
where the high attribute value is 1, and the low value is 0. Subsequently, identified clusters of 1s
would correspond to clusters of speed-related collisions.

The second variable, CLASAC, indicates the class of the collision according to severity. It
includes 5 categories: fatal, injury, property damage only (PDO), non-reportable, and non-
vehicle related events on the highways. The CLASAC variable was also re-coded form multiple
categories into a binary variable, where “1” corresponds to fatal and injury collisions (denoted
here as severe collisions) and “0” corresponds to all other type of collisions (PDO collisions).
Both DRACAT and CLASAC variables were considered separately in the analysis. Although the
NCGCNN method can consider both variables together, the results are not presented here
because of paper size limitation.

Pilot Tests

A pilot test was conducted using the 2008 collision data for one arterial road (Highway 7) to
screen all identified spatial analysis methods. There were 295 speed-related collisions (DRACT
= 1) and 279 severe collisions (CLASAC = 1). Out of these collisions, there were 64 collisions
that both were speed-related and severe (DRACT = 1 and CLASAC = 1).

Visual inspection of the results of ArcGIS, GeoDa, and SANET showed that locations with
clustering identified by ArcGIS and GeoDa using LMI and GOG methods are similar to those
identified using KDE. Peak search analysis was applied to the collisions involving injury and/or
fatality on the entire Ontario Provincial highway network using SafetyAnalyst. The locations
identified by these spatial analysis methods were in line with LHRS points flagged by
SafetyAnalyst. Cohen’s Kappa test for agreement (25) was used to compare the results between
SafetyAnalyst and spatial analysis methods. The Kappa coefficient between the results of
SafetyAnalyst and ArcGIS was 0.382 for the DRACT variable, which represents a fair
agreement. On the other hand, the coefficient for the CLASAC was close or equal to zero
indicating no agreement.

A second pilot test was conducted on the freeway network to test the functionality of spatial
analysis methods at the network level. The main finding from this pilot test is that the K-
Function and KDE methods were not feasible when applied to collisions at a network level. This
is due in part to the extensive computational time required by the K-Function methods when
applied to a dataset of over 20,000 collisions. For the KDE method, the analysis is not feasible
because of the software’s limitation. Therefore, the main finding of the two pilot tests is that both
the K-Function and KDE methods can produce sound results when applied to a separate highway
but may not be applied to a full network.
Analysis Procedure

A number of spatial analysis methods and software were applied to the freeways and undivided arterial highways that make up the Ontario Provincial highway network. Because of the differences between the two classes of highways, they were analyzed separately. In addition, the analysis was carried out at the network level (all roads at the same class were analyzed as one network) and highway level (each road is analyzed separately). Table 1 shows a summary of all analysis tests that were carried out including the software used, the analysis method, and the variable(s) considered in the analysis. As shown in the table, these combinations produced 34 different tests, among which 32 tests utilized spatial analysis methods. Because there were no differences between the network and highway levels of analysis in SafetyAnalyst, only the network level is considered. Finally, only severe collisions (involving injury and/or fatality) were analyzed in the two SafetyAnalyst tests to reduce the manual effort in matching the results to those of spatial analysis methods.

Highway and collision data needed to be in a proper format to be used for spatial analysis. Fortunately, MTO provided GIS shapefiles for the Ontario Provincial highway network. Data were prepared using the following procedures:

- A “route” is created using MTO highway data and ArcGIS’ linear referencing tool box.
- MVAB collision dataset imported into GIS environment using a Collision Import Tool created by the authors.
- MVAD collision dataset is then merged with the already imported MVAB collision dataset; this is done to simplify the importing procedure.
- Both highway and collision data that are imported into the GIS environment are split into freeway and undivided arterial data.

The results of the spatial analysis methods could be compared to those of SafetyAnalyst visually by overlaying the spatial analysis identified clusters with the SafetyAnalyst flagged LHRS points. A more quantitative comparison can be carried out using the NCGCNN method to

Table 1. Summary of All Analysis Tests Conducted.

<table>
<thead>
<tr>
<th>Software</th>
<th>Analysis Method</th>
<th>Variables</th>
<th>Freeways(^1)</th>
<th>Arterials(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Network</td>
<td>Highway</td>
</tr>
<tr>
<td>ArcGIS</td>
<td>LMI</td>
<td>DRACT</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>LMI</td>
<td>CLASAC</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>LMI</td>
<td>DRACT</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>LMI</td>
<td>CLASAC</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>GOG</td>
<td>DRACT</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>GOG</td>
<td>CLASAC</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>GOG</td>
<td>DRACT</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>GOG</td>
<td>CLASAC</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>GeoDa</td>
<td>LMI</td>
<td>DRACT</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>LMI</td>
<td>CLASAC</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>SafetyAnalyst</td>
<td>Peak Search</td>
<td>All</td>
<td>33</td>
<td>---(^2)</td>
</tr>
</tbody>
</table>

\(^1\) The numbers displayed are identification numbers for the analysis tests.
\(^2\) Because there were no differences between the network and highway levels of analysis in SafetyAnalyst, only the network level is considered.
examine whether the clusters identified by the spatial analysis methods are close to the SafetyAnalyst’s flagged points. Sixteen analysis tests were conducted to compare freeway spatial analysis results and SafetyAnalyst results. NCGCNN method was not applied to arterial results because of the dispersed nature of the network as explained later. The procedure for comparison of results is as follows:

- Select and export LHRS points flagged by SafetyAnalyst from GIS shapefile.
- Select and export collisions that have been identified as a cluster from GIS shapefile.
- Analyze both sets of data using the NCGCNN method.

Results and Discussion

Comparison of Spatial Analysis Methods

Table 2 shows the number of clusters of collisions identified in the 32 spatial analysis tests using a 1000 m search radius. As expected, more clusters are generally identified for the variable DRACT compared to CLASAC as the former collisions include all severity levels while the latter searches for clusters with high attribute, i.e. injuries and fatalities. The table also shows that more clusters were identified in all cases in the network level of analysis compared to the highway level. This is expected as the planar spatial analysis searches for clusters within a two-dimensional area instead of a distance along the highway. As a result, clusters will be identified involving collisions nearby highways within the search radius.

For freeways and variable DRACT, the differences in the number of clusters identified between network and highway analyses are between 8 and 13% with the exception of ArcGIS’ GOG analysis, the difference reaches 28.8%. The tendency of ArcGIS’ GOG analysis to yield more clusters in the network level is also noted in comparing the results of clusters related to variable CLASAC. In this case, the highest difference between the two levels of analysis was 83.6% and corresponded to ArcGIS’ GOG analysis, whereas the differences in the other analysis methods ranged from 8 to 24%. Similar trends were also observed for the clusters on arterials but the numbers of clusters were generally lower and percentages of differences between the network and highway analyses were generally higher than the corresponding values for

<table>
<thead>
<tr>
<th>Software</th>
<th>Variable</th>
<th>Method</th>
<th>Freeways</th>
<th>Arterials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Network</td>
<td>Highway</td>
</tr>
<tr>
<td>ArcGIS</td>
<td>DRACT</td>
<td>LMI</td>
<td>2225</td>
<td>1944</td>
</tr>
<tr>
<td>GeoDa</td>
<td>DRACT</td>
<td>GeoDa</td>
<td>1787</td>
<td>1643</td>
</tr>
<tr>
<td>ArcGIS</td>
<td>GOG</td>
<td></td>
<td>3626</td>
<td>2581</td>
</tr>
<tr>
<td>GeoDa</td>
<td>GOG</td>
<td></td>
<td>4510</td>
<td>4038</td>
</tr>
<tr>
<td>ArcGIS</td>
<td>CLASAC</td>
<td>LMI</td>
<td>643</td>
<td>593</td>
</tr>
<tr>
<td>GeoDa</td>
<td>CLASAC</td>
<td>GeoDa</td>
<td>524</td>
<td>399</td>
</tr>
<tr>
<td>ArcGIS</td>
<td>GOG</td>
<td></td>
<td>627</td>
<td>103</td>
</tr>
<tr>
<td>GeoDa</td>
<td>GOG</td>
<td></td>
<td>2111</td>
<td>1685</td>
</tr>
</tbody>
</table>

¹ Difference is equal to difference between numbers of clusters in network and highway analyses divided by number of clusters in network analysis in percentage.
freeways. This can be attributed to the fact that the arterials are spread throughout the province with some relatively long segments of highways. The extra clusters identified in the network level were mainly around the Greater Toronto Area (GTA) where there is a dense road network. These extra clusters were generally small in number but large in percentage.

Based on the Kappa test results, there was a substantial level of agreement between network and highway analyses of freeways. In this case, the Kappa-statistic was higher than 0.7 for all comparisons except for ArcGIS GOG method using the variable CLASAC, where the Kappa-statistic was equal to 0.398. Lower agreement levels, as indicated by lower values of the Kappa-statistic, were noted in analyzing collisions on arterials.

The GOG method was also observed to identify more clusters than the LMI method regardless of the software application used. This may be explained by the fact that the GOG method only categorizes clusters into three categories: not significant, clustering of low values, and clustering of high values. With the LMI method, clusters are sorted into five categories, the two additional categories are: a high value surrounded by low values, and a low value surrounded by high values. However, only clusters of high values surrounded by high values are taken as clusters of collisions.

In terms of the differences between the results produced by ArcGIS and GeoDa, ArcGIS identified more clusters than GeoDa when LMI method was used, but GeoDa identified more clusters when GOG method was applied. The results of the Kappa test showed good agreement between ArcGIS and GeoDa where the values of Kappa statistic were mostly over 0.7. The notable exception was the GOG method when applied to CLASAC where Kappa statistics was very low.

Comparison between Spatial Analysis and SafetyAnalyst Results

As mentioned earlier, the results of spatial analysis methods and SafetyAnalyst can be compared visually and using NCGCNN method. The following question is asked: “Do spatial analysis methods identify collision clusters near the LHRS points flagged by SafetyAnalyst?” To answer this question, clusters identified using the CLASAC and DRACT variables were compared separately to the SafetyAnalyst’s flagged LHRS points.

As a sample of the visual inspection of the spatial analysis and SafetyAnalyst results, Figure 1 and Figure 2 show the clusters identified by ArcGIS LMI and the LHRS points flagged by SafetyAnalyst for freeways and arterials, respectively. The figures show some of the points mentioned earlier: (i) more clusters are identified for the variables DRACT compared to CLASACT; (ii) more clusters are identified in network analysis compared to highway analysis; and (iii) the arterial network is dispersed throughout Ontario. Additional observations are: (i) almost all SafetyAnalyst’s flagged LHRS points are along the highly trafficked corridors in the GTA and the TransCanada Highway as it passes through the national capital (Ottawa); and (ii) LMI identified clusters around the flagged LHRS points as well as at other locations which have considerably less traffic volumes.

By comparing the results on freeways using the NCGCNN method, the distances between clusters of speed-related collisions (variable DRACT) and SafetyAnalyst’s flagged LHRS points showed no statistically significant grouping of collisions near flagged LHRS points. On the other hand, the clusters of severe collisions (variable CLASAC) were found to be significantly close to the flagged LHRS points at distances up to 1500 m. This implies that clusters of collisions
involving injuries and/or fatalities tend to be near LHRS points flagged by SafetyAnlayst. The findings were true for all but one of the comparisons (clusters identified by ArcGIS’ GOG).

A similar NCGCNN-based comparison between clusters identified along the arterials and LHRS points flagged by SafetyAnlayst was not possible as the arterials were highly fragmented. SANET software could not analyze the data, as NCGCNN analysis is generally better suited for an individual highway segment that is not segmented.

Finally, the agreement between the clustering locations as identified by the spatial analysis methods and the LHRS points flagged by SafetyAnalyst was tested using the Kappa test. The results showed a general lack of agreement, with most Kappa-statistic values less than 0.1 and the highest being 0.215 for clusters identified by ArcGIS LMI method for highway level of analysis on arterials using the variable CLASAC. The reason may mostly be attributed to the large number of clusters identified outside the high-trafficked corridors where almost all SafetyAnalyst’s flagged LHRS points were located. Secondly, whilst spatial analysis methods attempted to identify clusters of specific collisions (speed-related or injury related), this was not possible in SafetyAnalyst where it was not possible to analyze only speed-related collisions in the database available for analysis in this study.

CONCLUDING REMARKS

This paper examined the capability of utilizing spatial analysis methods in road safety evaluation for the purpose of network screening. Several software platforms and analysis methods were
examined. ArcGIS, GeoDa, and SANET were used for spatial analysis using actual data on Ontario provincial road network, and the results were compared to those produced by SafetyAnalyst as a known network screening tool. Three spatial analysis methods were used: Local Moran’s I (LMI) index and Getis-Ord $G^*_I$ (GOG) were used for analysis, and Network Constrained Global Cross Nearest-Neighbor (NCGCNN) was used for results comparison. Spatial analysis were conducted on both network and highway level. The DRACT and CLASAC variables were analyzed in order to locate clusters of speed-related and severe collisions. Results from spatial analysis were compared to network screening results from SafetyAnalyst using Cohen’s kappa coefficient.

Spatial analysis methods should be considered by road safety researches as it has shown to produce sound network screening results. The results showed the capability of spatial analyses to identify clusters of collisions on road network, with LMI producing relatively consistent results. The spatial analysis methods was shown to identify clusters of collisions all over the road network, while SafetyAnalyst identified locations almost exclusively at the high-trafficked corridors. The NCGCNN analysis showed that clusters of collisions involving injury and/or fatality tend to be located within 1,500 m or less of LHRS points flagged by SafetyAnalyst. This indicates is an indication of the reliability of spatial analysis methods. No such finding was noted for speed-related collisions. It was also shown that individual highway analyses produced more sound results when compared to full network analyses and would be the better analysis method for engineering assessment. That said, network analyses still have its merits as it may be used to

Figure 2. Graphical Display of SafetyAnalyst and ArcGIS LMI Results on Arterials (SafetyAnalyst’s LHRS points are the dark dots).
identify general areas where safety is a concern for reasons most likely not related to road conditions. Safety in such areas may be improved through increased driver education and enforcement. In addition, spatial analysis method can be used to analyze different variables and methods such as NCGCNN method can be used to study the relationship between two types of collisions and can be used to discover the underlying factors that may cause collisions.

It should be noted that the spatial analysis methods do not consider the traffic volume in identifying clusters of collisions, which is a major contrast with the statistical techniques adopted in SafetyAnalyst. The effect of this was minimized to some extent by dividing the Ontario provincial road network to two sub-networks of freeways and undivided arterials. In addition, the effect is minimized further in the highway level of analysis. Finally, this paper used collision data of only one year due to the computational resources required for all analysis methods. Different locations of collision clusters may be identified and/or different LHRS segments may be flagged by SafetyAnalyst if multiple years of data were analyzed. Therefore, the results of this paper should be used only as illustration of the capabilities of the spatial analysis methods.

ACKNOWLEDGEMENT

The author would like to thank Susan Nichol and Michael Pardo from the Ministry of Transportation Ontario for providing collision and road data and assistance in running the analysis in SafetyAnalyst. Financial support by the Natural Sciences and Engineering Council (NSERC) is gratefully acknowledged.

REFERENCES


