Exploratory Spatial Data Analysis in Traffic Safety

Amin Azimian  
*West Virginia University, amin_azimian@yahoo.com*

Dimitra Pyrialakou  
*West Virginia University*

Follow this and additional works at: [https://dc.uwm.edu/ijger](https://dc.uwm.edu/ijger)

Part of the Environmental Sciences Commons, Geographic Information Sciences Commons, Spatial Science Commons, and the Transportation Engineering Commons

**Recommended Citation**

*International Journal of Geospatial and Environmental Research: Vol. 7 : No. 1 , Article 4.*  
Available at: [https://dc.uwm.edu/ijger/vol7/iss1/4](https://dc.uwm.edu/ijger/vol7/iss1/4)

This Short Communication is brought to you for free and open access by UWM Digital Commons. It has been accepted for inclusion in International Journal of Geospatial and Environmental Research by an authorized administrator of UWM Digital Commons. For more information, please contact open-access@uwm.edu.
Exploratory Spatial Data Analysis in Traffic Safety

Abstract
This paper presents an exploratory spatial data analysis (ESDA) of road traffic crashes at different severity levels in West Virginia (WV). Although ESDA can support transportation safety decision-making by helping planners understand and summarize crash data, it is underutilized in practice. This paper describes the application of five representative easy-to-use methods to identify crash patterns and high crash-risk counties in WV. Analysis of crash data from 2010 to 2015 indicated that traffic crashes in WV were not spatially correlated. However, crash severities were found to be positively correlated.

Keywords
exploratory, spatial, data, traffic, crash, GIS
1 INTRODUCTION

Spatial data analysis is a rapidly growing area in the transportation field, allowing researchers to easily manipulate spatial data into different forms and extract additional meaning as a result (Bailey 1994). Spatial analysis techniques can be classified into two broad categories, exploratory spatial data analysis (ESDA) and confirmatory spatial data analysis (CSDA) techniques. The ESDA is an extension of the exploratory data analysis (EDA) derived from the work of Tukey (Tukey 1977) and its main objectives are to visualize and describe spatial distributions, identify standards of spatial patterns or clusters, generate hypotheses based on spatial trends, and identify cases or subsets of cases that are unusual, given their location on the map (Cressie and Wikle 2015). However, CSDA is used to perform the estimation and validation necessary for the analysis of spatial components. Hypothesis testing, spatial econometrics, and spatial regression are all CSDA techniques (Anselin and Rey 2010).

According to the Moving ahead for Progress in the 21st Century Act (MAP-21), states are required to perform transportation planning analyses to understand crash trends and identify locations with high incidence to improve road safety (FHWA 2012). Review of literature (Azimian and Eustace 2018; Bernhardt and Virkler 2002; Himes et al. 2017; Lin and Fan 2019; Maze et al. 2005; Srinivasan and Bauer 2013; Xue and Xu 2019) shows that the majority of research efforts highly rely on CSDA rather than ESDA and most transportation agencies place emphasis on developing regression models to identify high crash risk locations. However, a vast majority of existing CSDA approaches such as logit models (Rifaat et al. 2012; Tay et al. 2011) and count models (Caliendo et al. 2007; Coruh et al. 2015) operate at micro-level, that is, they can be only applied to road entities such as road segments, intersections, etc., and therefore, they are not appropriate for development of statewide road safety strategic plan. Moreover, existing macro (area)-level CSDA approaches such as Full Bayesian multivariate models (Liu and Sharma 2018; Zeng and Huang 2014) and Autoregressive models (Rhee et al. 2016; Soro et al. 2017) could take long time to converge (Nichols et al. 2011) and cannot handle large dataset (Burden et al. 2015) respectively. Compared to CSDA, ESDA is a simple statistical data analysis that provides insights into the characteristics of dataset (Karimi and Akinci 2009). Moreover, it could be a very useful tool to help statewide roadway safety strategic planning (Abdel-Aty et al. 2013; Rybarczyk and Wu 2010).

In light of the above, this paper employs five representative ESDA techniques to address the following safety concerns which are critical for transport authorities and decision makers and relate to the trend and nature of traffic crashes (Waldheim et al. 2015; Ye et al. 2013).

1. To what extent are traffic crashes prevalent in a study area?
2. Which specific regions in an area should be considered for safety improvement?
3. Is spatial autocorrelation present in the crash data?
4. Are crash severities independent?
2 METHODOLOGY

According to the National Highway Traffic Safety Administration (NHTSA 2017), West Virginia is among top US states with high fatality rates. However, few research efforts have attempted to assess the overall traffic safety in West Virginia. Hence, in this study, the ESDA techniques are illustrated using estimated average crash rates from 55 counties in West Virginia. First, county-level crash frequency at different severity levels (Fatal, injury, and property damage only crashes [PDO]) from 2010 to 2015 obtained from the West Virginia division of highways through the email communication. Thereafter, the average crash rate for each county has been estimated as follows:

\[ CR_i(s) = \frac{1}{6} \times \sum_{t=1}^{6} \frac{N_{it}(s)}{Pop_{it}} \times 10,000 \]  

where \( CR_i(s) \) represents the 6-year average crash rate of severity level \( s = \text{fatality, injury, PDO} \) in county \( i \); \( N_{it}(s) \) is the crash frequency of severity level \( s \) in county \( i \) at year \( t=2010 \) to 2015; \( Pop_{it} \) is the U.S. Census Bureau’s population estimate for county \( i \) at year \( t \). Since the estimated crash rate could be very small, it has been multiplied by 10,000 to give the crash rate per 10,000 population.

3 EXPLORATORY SPATIAL DATA ANALYSIS

This section presents the following ESDA techniques: histogram, boxmaps, Moran’s I, Pearson correlation and conditional map. These techniques are used to examine the distribution of fatal, injury, and property damage only crashes. All data analyses are based on West Virginia (WV) crash data (2010–2015) obtained from the WV Department of Transportation. The findings can be used as fundamental information for designing effective policies regarding highway safety and transportation system and driver education. The GeoDa software package was used to perform the analysis as it is an open-source software.

3.1 Histogram

Histograms provide a visual interpretation of numerical data and are used to explore the data’s underlying distribution (e.g., normal distribution), outliers, skewness, etc. To assess the crash data density and distribution, the histograms of the average crash rates per 10,000 population by severity based on six-year data (2010–2015) have been constructed (Figure 1). Figure 1 shows that the crash rate distributions at different severity levels tend to be right-skewed, with a tail on the high end and taller bins on the low end, suggesting that most counties in WV have a crash rate lower than the average. Moreover, referring to the average crash rate’s histograms, especially the injury crash rate, there is an observation (McDowell County) in red color that lies outside the overall distribution pattern. This implies that McDowell County has had a higher crash rate than any other county and should be considered as a potential hotspot for safety improvement.
Figure 1. Histograms of the county-level average crash rate per 10,000 population by severity over the years 2010–2015.

3.2 Boxmap

A boxmap is an alternative way of visualizing the distribution of a variable (Anselin 1995). It is used to represent the summary statistics (fractions of distribution) and detect potential outliers by using the Interquartile Range (Tukey 1977). Figure 2 shows the boxmaps of the average crash rates per 10,000 population by severity. Each map is a choropleth map in which a quantile classification of the data has been applied to reflect the data distribution and identify anomalous counties. Figure 2(a) shows that Pendleton County has the highest average fatal crash rate among WV counties. Moreover, eastern and southern WV counties tend to have higher fatal crashes. Referring to Figure 2(b), Raleigh and McDowell Counties experienced more in jury crashes than other counties when their population was accounted for. Finally, as shown in Figure 2(c), Ohio, Lewis,
Cabell, Kanawha, Raleigh, and McDowell Counties are among those with the highest PDO crash rates in WV.

Figure 2. Boxmaps of the county-level average (a) fatal crash rate (b) injury crash rate, and (c) PDO crash rate per 10,000 population over the years 2010–2015.
3.3 Moran’s I

Investigation of the global clustering patterns of traffic crashes across regions (e.g., a cluster of counties with a high number of crashes) is very important in traffic safety, as it can give information about underlying issues or unobserved factors in clustered regions (Kuo et al. 2018; Ouni and Belloumi 2019; Ziakopoulos and Yannis 2020). Moran’s I statistics is a powerful tool that can be used to detect such global spatial patterns (Li et al. 2014; Xie et al. 2019). It is defined as equation (2):

\[
I = \frac{N}{\sum_{i} \sum_{j} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i} (y_i - \bar{y})^2}
\]  

(2)

where \(I\) is Moran’s Index value, \(N\) is the number of spatial units (counties), \(y_i\) and \(y_j\) are the crash rates related to targeted county \(i\) and neighboring county \(j\), respectively, is the average crashes, and \(w_{ij}\) is an element of a matrix of spatial weights. A Moran’s Index value near +1.0 indicates clustering, an index value near −1.0 indicates dispersion, and a value close to zero indicates a random spatial pattern. To construct the neighboring structure, the Queen Contiguity was considered; that is, counties that share an edge or have coincident boundaries are neighbors. The global Moran’s I statistics in each year from 2010 to 2015 are calculated using GeoDa, and the results are summarized in Table 1. The results show that little spatial autocorrelation is present in the crash data. These results contradict the findings of previous studies (Aguero-Valverde and Jovanis 2006; Huang et al. 2010) that have shown that a strong spatial autocorrelation is present in the crash data. It should be noted that both studies used multivariate conditional autoregressive model to capture spatial autocorrelation. Moreover, the differences in research results could be justified in this way that pattern of traffic crashes and their contributing factors vary from one state to another because of differences in infrastructure characteristics and other underlying factors (Aguero-Valverde and Jovanis 2006).

### Table 1. Global Moran’s I statistics.

<table>
<thead>
<tr>
<th>Year</th>
<th>Fatal Index</th>
<th>P-value</th>
<th>Injury Index</th>
<th>P-value</th>
<th>PDO Index</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>-0.004</td>
<td>0.40</td>
<td>-0.07</td>
<td>0.30</td>
<td>-0.07</td>
<td>0.26</td>
</tr>
<tr>
<td>2014</td>
<td>-0.01</td>
<td>0.40</td>
<td>-0.06</td>
<td>0.30</td>
<td>-0.05</td>
<td>0.40</td>
</tr>
<tr>
<td>2013</td>
<td>0.04</td>
<td>0.20</td>
<td>-0.05</td>
<td>0.30</td>
<td>0.04</td>
<td>0.20</td>
</tr>
<tr>
<td>2012</td>
<td>0.05</td>
<td>0.20</td>
<td>0.08</td>
<td>0.13</td>
<td>-0.03</td>
<td>0.40</td>
</tr>
<tr>
<td>2011</td>
<td>0.03</td>
<td>0.30</td>
<td>-0.04</td>
<td>0.40</td>
<td>-0.01</td>
<td>0.40</td>
</tr>
<tr>
<td>2010</td>
<td>0.10</td>
<td>0.10</td>
<td>0.02</td>
<td>0.30</td>
<td>0.003</td>
<td>0.40</td>
</tr>
</tbody>
</table>

3.4 Pearson Correlation

Some research efforts (Barua et al. 2014; Boulieri et al. 2017) have reported that crash severities are correlated and cannot be modeled independently. That is, there may be shared factors that simultaneously affect fatal, injury, and PDO crashes. One possible
way to assess the dependency among crash severities is to estimate the Pearson correlation (Liu 2018).

Table 2. Pearson correlation among crash severities.

<table>
<thead>
<tr>
<th>Crashes</th>
<th>Fatal</th>
<th>Injury</th>
<th>PDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatal</td>
<td>1</td>
<td>0.71 ($p &lt; 0.0001$)</td>
<td>0.78 ($p &lt; 0.0001$)</td>
</tr>
<tr>
<td>Injury</td>
<td>1</td>
<td>0.87 ($p &lt; 0.0001$)</td>
<td></td>
</tr>
<tr>
<td>PDO</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2 gives the Pearson correlation coefficients of fatal injury and PDO crashes. From the results, all crash severities are positively associated in WV. This implies that an increase in one WV county’s fatal crash rate will likely increase the injury and PDO crash rate in that county and vice versa.

3.5 Conditional Map

As discussed in the previous section, the findings of the Pearson correlation indicated positive associations among crash severities. However, this finding may not be true for all counties, as the correlation coefficients only show the overall trends across the study area.

Figure 3. Conditional map of the average fatal crash rate in comparison with the average PDO and injury crash rates.
To better understand how crash severities change in relation to one another in any counties, it is necessary to draw a conditional map (Anselin 2005). As outlined in figure 3, each column indicates that how PDO and fatal crash rate changes across counties with the same injury crash rate range (< 42.789, 42.78 to 54.885, ≥ 54.885). Whereas each row represents the variations in injury and fatal crash rate among counties with the same PDO crash rate range (< 98.966, 98.96 to 132.982, ≥ 132.982). Referring to the counties associated with the highest injury and PDO crash rate ranges, it can be seen that Cabell, Doddridge, and Ohio have had the lowest fatal crash rates. However, when it comes to the group of counties associated with the lowest injury and PDO crash rate ranges, Mingo and Lincoln have had the highest fatal crash rates. Such differences could be due to variation in demographic, environmental, transportation and sociological factors (Merlin et al. 2020).

4 CONCLUSIONS

In this paper, five representative ESDA tools are introduced, and applications to crash data sets are presented. Such tools could help traffic safety practitioners and transportation agencies assess crash data beyond formal statistical modeling and identify crash patterns and high crash risk areas across the study. Exploratory analysis of traffic crashes at different severity levels in WV indicated that their distributions are right skewed, suggesting that most counties have a crash rate lower than the average. From the results, Pendleton County was found to have the highest fatal crash rates, whereas Raleigh and McDowell Counties have extreme (i.e., beyond the cut-off) injury crash rates. Moreover, counties located in eastern and southern WV tend to have higher fatal crash rates, while injury and PDO crashes tend to be more randomly distributed across WV. The results of the Moran’s I test demonstrated that the crash rates of different severities in neighboring counties are not significantly correlated, while the Pearson correlation indicated that crashes of different severities tend to be positively correlated across the study area. Such trend, as shown in the conditional map is not present in Cabell, Doddridge, Ohio, Mingo and Lincoln counties.

The findings can be used by state agencies and corresponding decision-makers to effectively allocate limited resources and funds to mitigate traffic safety issues in high crash-rate counties. This could be done by adopting effective speed management strategies such as reduction of posted speed limits, enhancement of road delineation and increasing sobriety check points in rural roadways in high crash rate counties. In terms of future work, conditional maps can be used to incorporate crash and socioeconomic data to discover potential factors contributing to traffic crashes. In terms of future work, conditional maps can be used to incorporate crash and socioeconomic data to discover potential factors contributing to traffic crashes.

REFERENCES


