International Journal of Geospatial and Environmental Research

Volume 7 | Number 1 Article 4

May 2020

Exploratory Spatial Data Analysis in Traffic Safety

Amin Azimian
West Virginia University, amin_azimian@yahoo.com

Dimitra Pyrialakou
West Virginia University

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 method 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:

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

where $CR_i(s)$ represents the 6-year average crash rate of severity level s = *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.

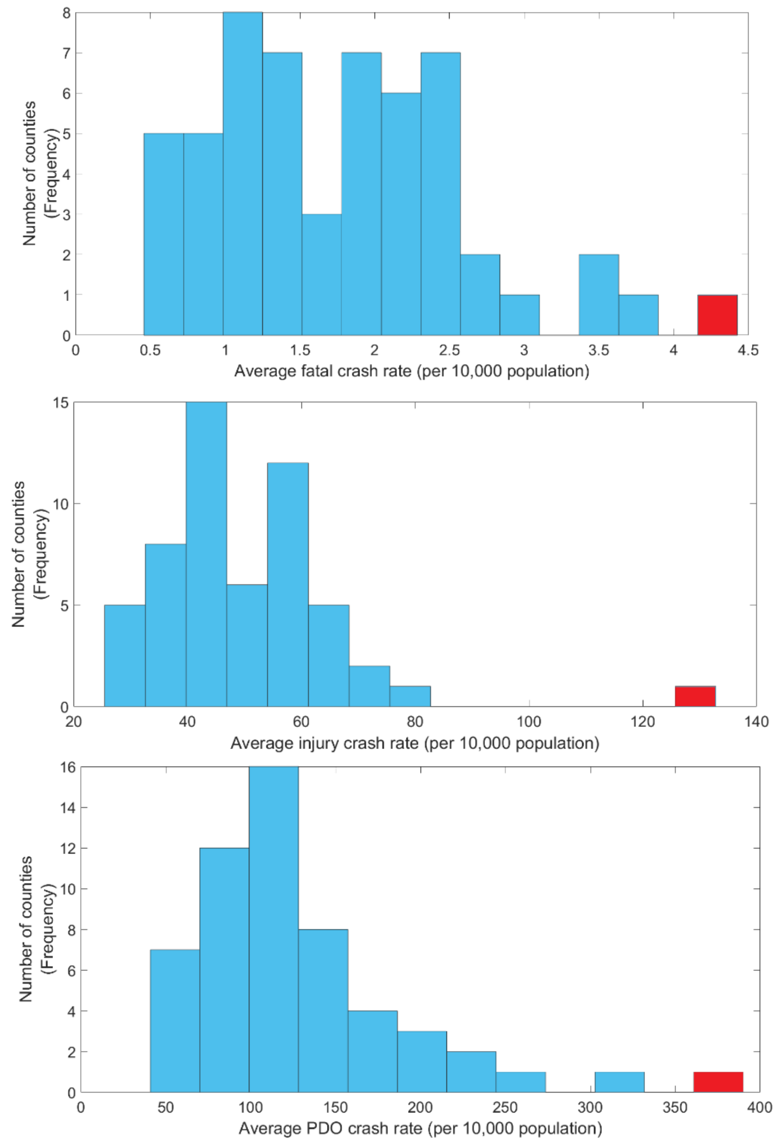
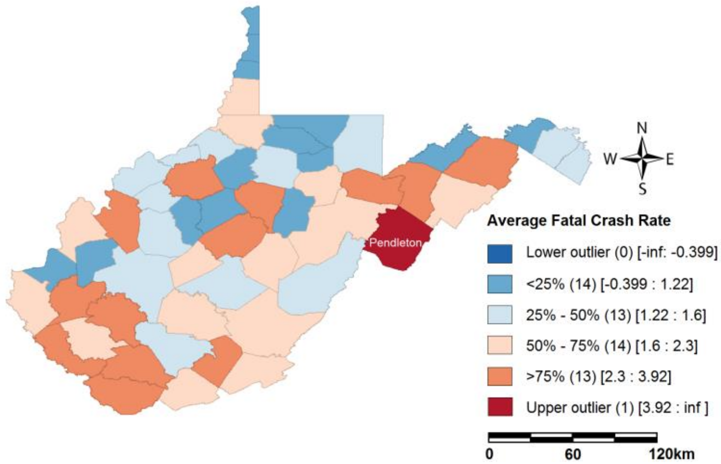


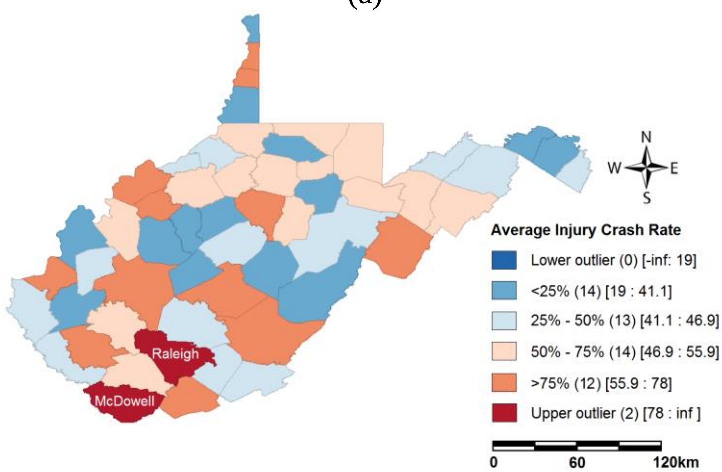
Figure 1. Histograms of the county-level average crash rate per 10,000 population by severity over the years 2010–2015.

3.2 Boxmap

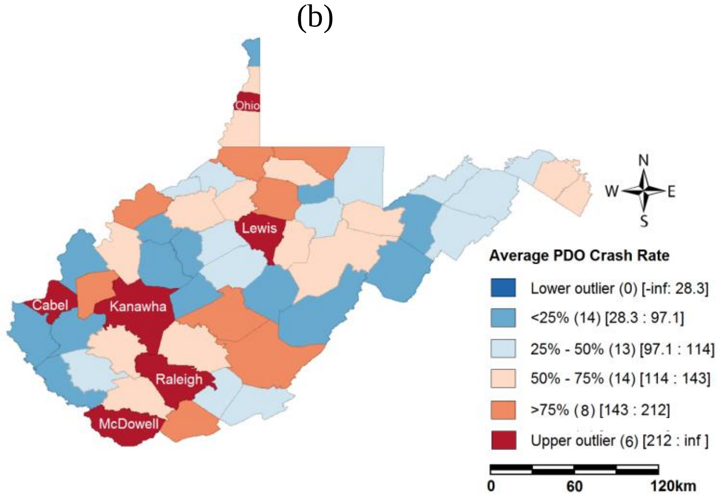
A boxmap is an alternative way of visualizing the distribution of a variable (Anselin 1995). It is used to represent the summary statistics (fractions of distribution) and detect potential outliers by using the Interquartile Range (Tukey 1977). Figure 2 shows the boxmaps of the average crash rates per 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.



(a)



(b)



(c)

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):

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{2}$$

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.

	Fatal		Injury		PDO	
Year	Index	P-value	Index	P-value	Index	P-value
2015	−0.004	0.40	−0.07	0.30	−0.07	0.26
2014	−0.01	0.40	−0.06	0.30	−0.05	0.40
2013	0.04	0.20	−0.05	0.30	0.04	0.20
2012	0.05	0.20	0.08	0.13	−0.03	0.40
2011	0.03	0.30	−0.04	0.40	−0.01	0.40
2010	0.10	0.10	0.02	0.30	0.003	0.40

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.

Crashes	Fatal	Injury	PDO
Fatal	1	0.71 ($p < 0.0001$)	0.78 ($p < 0.0001$)
Injury		1	0.87 ($p < 0.0001$)
PDO			1

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.

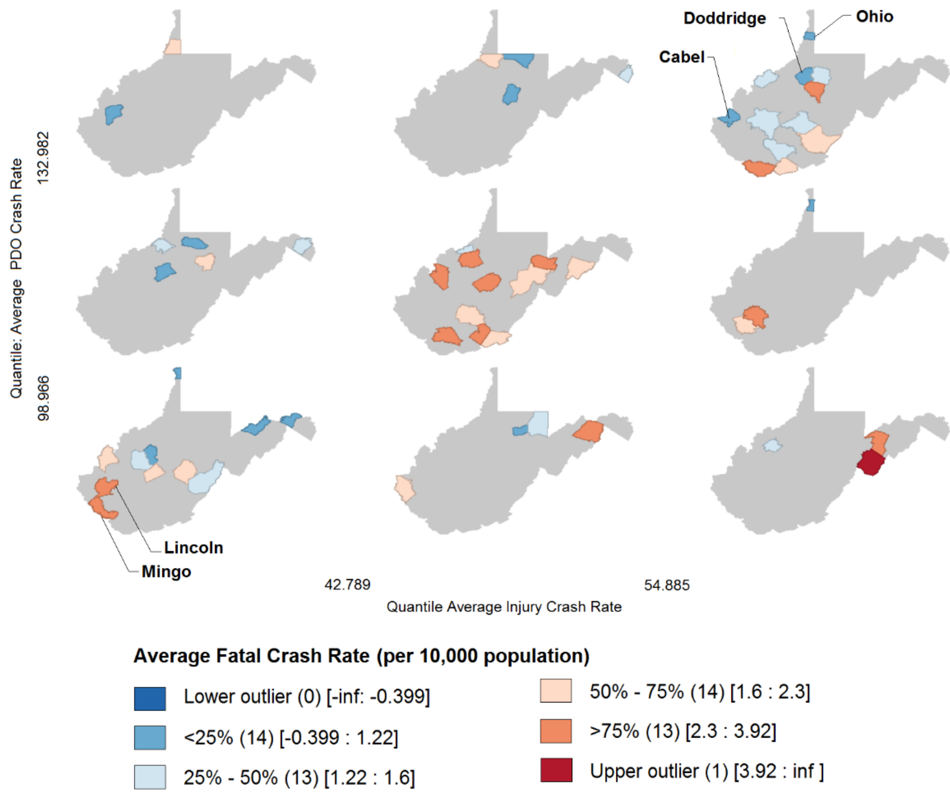


Figure 3. Conditional map of the average fatal crash rate in comparison with the average PDO and injury crash rates.

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, 98.966 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

Abdel-Aty, M., Lee, J., Siddiqui, C. and Choi, K. (2013) Geographical unit based analysis in the context of transportation safety planning. *Transportation Research Part A: Policy and Practice*, 49, 62-75.

Aguero-Valverde, J. and Jovanis, P.P. (2006) Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis and Prevention*, 38(3), 618-625.

Anselin, L. (2005) *Exploring Spatial Data with GeoDaTM: A Workbook*. Center for spatially integrated social science-University of Illinois at Urbana Champaign: Urbana, Champaign.

Anselin, L. and Rey, S.J. (2010) Perspectives on spatial data analysis. In *Perspectives on Spatial Data Analysis* (pp. 1-20): Springer.

Azimian, A. and Eustace, D. (2018) Modeling Socio-Economic Determinants of Traffic Fatalities. *International Journal of Engineering Research and Management*, 5(11).

Bailey, T. C. (1994) A review of statistical spatial analysis in geographical information systems. In S. Fotheringham & P. Rogerson (Eds.). *Spatial Analysis and GIS* (pp. 13-44). Bristol, PA: Taylor & Francis.

Barua, S., El-Basyouny, K. and Islam, M.T. (2014) A full Bayesian multivariate count data model of collision severity with spatial correlation. *Analytic Methods in Accident Research*, 3, 28-43.

Bernhardt, K.L.S. and Virkler, M.R. (2002) Improving the identification, analysis and correction of high-crash locations. *ITE Journal*, 72(1), 38-42.

Boulieri, A., Liverani, S., de Hoogh, K. and Blangiardo, M. (2017) A space–time multivariate Bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1), 119-139.

Burden, S., Cressie, N. and Steel, D.G. (2015) The SAR model for very large datasets: a reduced rank approach. *Econometrics*, 3(2), 317-338.

Caliendo, C., Guida, M. and Parisi, A. (2007) A crash-prediction model for multilane roads. *Accident Analysis and Prevention*, 39(4), 657-670.

Coruh, E., Bilgic, A. and Tortum, A. (2015) Accident analysis with aggregated data: The random parameters negative binomial panel count data model. *Analytic Methods in Accident Research*, 7, 37-49.

Cressie, N. and Wikle, C.K. (2015) *Statistics for Spatio-Temporal Data*. John Wiley and Sons: Hoboken, New Jersey.

Federal Highway Administration (FHWA). (2012) *Moving Ahead for Progress in the 21st Century Act (MAP-21): A Summary of Highway Provisions*. U.S. Deptartment of Transportation.

Himes, S., Gross, F.B., Persaud, B. and Eccles, K.A. (2017) *Safety Evaluation of Edge-Line Rumble Stripes on Rural Two-Lane Horizontal Curves*. Federal Highway Administration: McLean,Virgina.

Huang, H., Abdel-Aty, M.A. and Darwiche, A.L. (2010) County-level crash risk analysis in Florida: Bayesian spatial modeling. *Transportation Research Record*, 2148(1), 27-37.

Karimi, H.A. and Akinci, B. (2009) *CAD and GIS integration*. CRC Press: Boca Raton, Florida.

Kuo, P.-f., Hsu, T.P., Putra, I.G.B., Ilmy, H.F., Chiu, C.S. and Wu, C.Y. (2018) Defining the effects of traffic violations on crash frequency by applying a spatial panel model. *Proceedings of the 39th Asian Conference on Remote Sensing: Remote Sensing Enabling Prosperity*.

Li, Q., Song, J., Wang, E., Hu, H., Zhang, J. and Wang, Y. (2014) Economic growth and pollutant emissions in China: A spatial econometric analysis. *Stochastic Environmental Research and Risk Assessment*, 28(2), 429-442.

Lin, Z. and Fan, W. (2019) Cyclist injury severity analysis with mixed-logit models at intersections and nonintersection locations. *Journal of Transportation Safety and Security*, 1-23.

Liu, C. (2018) *Three Essays on Crash Frequency Analysis*. Ph.D. thesis in Civil Engineering, Iowa State University.

Liu, C. and Sharma, A. (2018) Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. *Analytic Methods in Accident Research*, 17, 14-31.

Maze, T.H., Preston, H., Storm, R.J., Hawkins, N. and Burchett, G. (2005) Safety performance of divided expressways. *ITE Journal*, 75(5), 48.

Merlin, L.A., Guerra, E. and Dumbaugh, E. (2020) Crash risk, crash exposure, and the built environment: A conceptual review. *Accident Analysis and Prevention*, 134, 105244.

National Highway Traffic Safety Administration (NHTSA). (2017) *Traffic Safety Facts 2017*. U.S. Department of Transportation: Washington, DC.

Nichols, J., Moore, E. and Murphy, K. (2011) Bayesian identification of a cracked plate using a population-based Markov Chain Monte Carlo method. *Computers and Structures*, 89(13-14), 1323-1332.

Ouni, F. and Belloumi, M. (2019) Pattern of road traffic crash hot zones versus probable hot zones in Tunisia: A geospatial analysis. *Accident Analysis and Prevention*, 128, 185-196.

Rhee, K.-A., Kim, J.-K., Lee, Y.-I. and Ulfarsson, G.F. (2016) Spatial regression analysis of traffic crashes in Seoul. *Accident Analysis and Prevention*, 91, 190-199.

Rifaat, S.M., Tay, R. and De Barros, A. (2012) Severity of motorcycle crashes in Calgary. *Accident Analysis and Prevention*, 49, 44-49.

Rybarczyk, G. and Wu, C. (2010) Bicycle facility planning using GIS and multi-criteria decision analysis. *Applied Geography*, 30(2), 282-293.

Soro, W.L., Zhou, Y. and Wayoro, D. (2017) Crash rates analysis in China using a spatial panel model. *IATSS Research*, 41(3), 123-128.

Srinivasan, R. and Bauer, K. (2013) *Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs*. Federal Highway Administration Report FHWA-SA-14-005: Washington, DC.

Tay, R., Choi, J., Kattan, L. and Khan, A. (2011) A multinomial logit model of pedestrian–vehicle crash severity. *International Journal of Sustainable Transportation*, 5(4), 233-249.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley: Reading, Massachusetts.

Waldheim, N., Wemple, E.A. and Fish, J.K. (2015) *Applying Safety Data and Analysis to Performance-Based Transportation Planning*. Federal Highway Administration Report FHWA-SA-15-089: Washington, DC.

Xie, K., Ozbay, K. and Yang, H. (2019) A multivariate spatial approach to model crash counts by injury severity. *Accident Analysis and Prevention*, 122, 189-198.

Xue, C. and Xu, D. (2019) Factors Influencing Crash Severity at Rural Horizontal Curves in Maine. *ITE Journal*, 89(5), 36-41.

Ye, X., Pendyala, R.M., Shankar, V. and Konduri, K.C. (2013) A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention*, 57, 140-149.

Zeng, Q. and Huang, H. (2014) Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis and Prevention*, 67, 105-112.

Ziakopoulos, A. and Yannis, G. (2020) A review of spatial approaches in road safety. *Accident Analysis and Prevention*, 135, 105323.