5-1-2016

A Spatial Analytical Framework for Examining Road Traffic Crashes

Grace O. Korter

UNIVERSITY OF IBADAN, kortergrace@gmail.com

A Spatial Analytical Framework For Examining Road Traffic Crashes

Grace O. Korter
University of Ibadan
Ibadan, Nigeria

A number of different modeling techniques have been used to examine road traffic crashes for analytic and predictive purposes. Map-based spatial analysis is introduced. Applications are given which show the power in a combination of existing exploratory and statistical methods.

Keywords: Box plots, Cartograms, Maps, Regression, Spatial autocorrelation

Introduction

Road Traffic Crashes (RTC) data are obtained for observations which are ordered in space or in space and time. In these situations, the observations can be characterized by their absolute location, using a coordinate system, or by their relative location, based on a particular distance metric. In other words, RTC data are organized by spatial units of observation across geographical space in the most general sense.

In many instances where regression analysis is applied to spatial data, various measurement problems, spatial externalities, spatial dependence and other spatial effects which may lead to dependent error terms could arise. As such, the standard econometric modeling techniques become invalid. The main methodological problems stem from the existence of spatial effects in the form of spatial dependence (also known as spatial autocorrelation). For instance, spatial autocorrelation impacts mildly on regression coefficient estimates for geographically distributed non distance based covariates, regardless of whether the statistical distribution of a variable is Normal, Poisson or Binomial. In contrast spatial autocorrelation impacts significantly upon the variance of these variables, with positive autocorrelation inducing variance inflation. Also, this increased

Dr. Korter is a Lecturer at Federal Polytechnic Offa. Email her at:
kortergrace@gmail.com.

variance tends to flatten a bell shaped curve and create over dispersion in Poisson and Binomial variables. Consequently, frequency distributions are distorted when spatial autocorrelation is present.

As positive spatial autocorrelation increases, for a normal curve, the center is squashed and the tails are thickened. Whereas, the shape of a Poisson distribution increasingly resembles that of an exponential distribution, with the frequency of zero and/or near-zero values increasing concomitantly with the appearance of marked outliers. Also, the shape of a Binomial distribution first approaches that of a Uniform distribution and then that of a Sinusoidal distribution.

These metamorphoses are the only way variance can increase for these types of variables. In addition, spatial autocorrelation compromise the descriptive ability of conventional versions of the above probability models, as well as the statistical inferences associated with them.

The majority case of positive spatial autocorrelation results in variance inflation, which for a normal frequency distribution means the central frequencies (i.e., numbers of values near the mean) decrease and the tail frequencies increase - peakedness (i.e., kurtosis) is altered because the curve is flattened. The most common case of positive spatial autocorrelation also results in variance inflation for a Poisson frequency distribution. Here, the chances of smaller and of extremely large values increase at the expense of intermediate values. Again, positive spatial autocorrelation also, results in variance inflation for a binomial frequency distribution, which means the chances of the extreme probabilities increase at the expense of more centrally located intermediate probabilities.

Exploratory data analysis and spatial autoregressive models have brought new power to the analysis of spatial data, capturing effects of redundant locational information contained in these data. Their descriptive abilities better portray observed data and they offer more fully developed nomothetic approach that in turn will support an increasing emphasis on the geographically varying local statistics that fluctuate around their global parameter counterparts (see Anselin & Griffith (1988), Anselin (1988), Anselin (1990), Griffith (2009), LeSage & Pace (2009)).

A great amount of effort has gone into assessing the different factors that contribute to RTCs. Previous studies have utilized a number of different modeling techniques for both analytical and predictive purposes. Jovanis and Chang (1986), Maher and Mountain (1988), Jones, Jannsen and Mannering (1991) and Roque and Cardoso (2014) applied Poisson regression models to study aggregate RTC

data, while, Ye et al. (2009) and Yu and Abdel-Aty (2013) investigated the frequencies of RTC using multivariate Poisson regression model.

Again, the Poisson based models; Poisson-Lognormal, Poisson-Gamma and Poisson-Lognormal with conditional autoregressive priors have been found suitable for RTC studies and these have been widely utilized in literature. Examples include the works of Shanker, Mannering and Barfield (1995), Sheffer and Riet Veld (1997), Milton and Mannering (1998), Abdel-Aty and Radwan (2000), Amoros, Martin and Laumon (2003), Miaou and Lord (2003), Noland and Quddus (2005), Kim, Brunner and Yamashita (2006), Lord and Miranda-Moreno (2008), Quddus (2008) and Wang, Quddus and Ison (2009).

Also, the negative binomial model was used by Jovanis and Chang (1986), Fridstrøm and Ingebrigtsen (1991), Miaou and Lord (2003), Guo, Wang and Abdel Aty (2010), Ahmed et al.(2011) , Yu, Abdel Aty and Ahmed (2013) and Goh, Currie, Sarvi and Logan (2014) to model RTC.

Using the Bayes approach, Aguero-Valverde and Jovanis (2006) and Jiang, Abdel Aty and Alamili (2014) employed random effect Poisson Log-Normal models for crash risk hotspot identification. While, Yu and Abdel-Aty (2013) utilized the random effects logistic regression models to investigate the dynamics of RTC.

Similarly, Jegede (1988), Okamoto and Koshi (1989) and Aderamo (2012) used the multiple regression model to study the incidence of RTC. However, Anastasopoulos, Tarko and Mannering (2008) identified the tobit regression model as an alternative method for studying RTC. In addition, Erdogan (2009) studied RTC using the geographically weighted regression model, while, Abdullah and Zamri (2012) applied the fuzzy regression model using two threshold levels. Taking cognizance of spatial autocorrelation, Levin, Kim, and Nitz (1995a, 1995b) and Korter, Olubusoye and Salisu (2014) modeled RTC using the spatial autoregressive model.

Methodology

Data on number of licensed drivers, registered vehicles and 2011 RTC was obtained from the Federal Road Safety Commission (FRSC). The land area encompassing each state and the Federal Capital Territory (FCT) was obtained, while, the corresponding population was obtained from National Population Commission.

Cartograms

To illustrate the relative sizes of the RTC for the states/FCT within the country, the area of each state/FCT is scaled in proportion to its RTC.

Cartograms are unique representations of geographical space that depict spatial transformations and attributes of geographic objects as the object's area. Examined more closely, the value-by-area mapping technique encodes the mapped data in a simple and efficient manner with no data generalization or loss of detail. Cartograms vary on the degree in which geographic space is changed.

A cartogram is yet another way to highlight extreme values on a map. Here the original spatial units are replaced by circles. The area of the circle is proportional to the value of a selected variable. The circles themselves are aligned as closely as possible to the original location of the matching spatial units by means of a nonlinear optimization routine.

A Dorling cartogram maintains neither shape, topology nor object centroids, though it has proven to be a very effective cartogram method. To create a Dorling cartogram, instead of enlarging or shrinking the objects themselves, the cartographer will replace the objects with a uniform shape, usually a circle, of the appropriate size. Circles are distributed to resemble the original topology. This method suggests that the shapes do not overlap but rather be moved so that the full area of each shape can be seen. The shape and relative location of each state and FCT is retained to as large an extent as possible (see Dorling, 1996; Anselin 2005).

Box Plots

Box plots are good at portraying extreme values and are especially good at showing differences between distributions. It is a type of graph which is used to show the shape of the distribution, its central value, and variability. The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median.

In descriptive statistics, a box plot is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.

Box plots display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution: box plots are non- parametric. The spacings between the different parts of the box indicate the

degree of dispersion (spread) and skewness in the data, and show outliers. In addition to the points themselves, they allow one to visually estimate the interquartile range, midhinge, range, mid-range and trimean. Boxplots can be drawn either horizontally or vertically.

Here, the rectangle represents the cumulative distribution of RTC sorted by value. The red bar in the middle corresponds to the median; the dark part shows the interquartile range (going from the 25th percentile to the 75th percentile). The individual observations in the first and fourth quartile are shown as blue dots. The thin line is the hinge which corresponds to the default criterion of 1.5. The hinge criterion determines how extreme observations need to be before they are classified as outliers (see McGill, Tukey & Larsen, 1978; Anselin, 2005).

The Spatial Autoregressive Model Specification

$$y = \rho W(y) + X\beta + \varepsilon \tag{1}$$

where, y is an $N \times 1$ vector of the dependent variable (RTC cases) for all 36 states and the FCT. This is formed by observing the total incidence of RTC for each of the 36 states and the FCT. ρ is the coefficient of the spatial lag. W is a known $N \times N$ spatial weights matrix whose diagonal elements are zero. Two states are said to be contiguous to one another when they share a common border. A spatial weights matrix is defined by using an interaction between the geographical location of each spatial unit and every other location. The neighbourhood contiguity distance between each location and every other location was used.

To develop the spatial weights (contiguity based matrix), the value of 1 was assigned to a state when it shares common border with another, and otherwise the value 0 was assigned. Using the common border, the rook contiguity matrix was used to form a weights matrix. Taking into cognizance the common border and the common vertex border, the queen contiguity matrix was formed. W satisfies the condition that $(I_N - \rho W)$ is non singular for all $|\rho| < 1$, I_N is an identity matrix of dimension N. $W(y)$ is a weighted matrix of $N \times 1$ vector of values for the RTC cases summed over all states.

X is an $N \times K$ matrix of observations on the K explanatory variables, namely population, drivers, area and vehicles. X is assumed to be of full column rank and its elements are assumed to be asymptotically bounded in absolute value. β is a $K \times 1$ vector of K parameters, representing the 4 regression coefficients for this

study. ε is an $N \times 1$ vector of normally distributed random error terms, with mean and constant variance, while, $E(X / \varepsilon) = 0$, strict exogeneity.

The ordinary least squares estimators are obtained from a regression of the X on y and on Wy respectively. Clearly, the maximum likelihood estimate for β is a function of these auxiliary regression coefficients as well as of ρ. Therefore, the estimate for β can be found directly, once the value for ρ has been determined. These steps can be carried out by standardized regression package. This study employed the GeoDa software (Anselin, 2005).

Predicted Value and Residual Maps

The predicted value and regression residuals are variables to any exploratory function in GeoDa, including mapping. Such maps, referred to as predicted value maps and residual maps are useful for a visual inspection of patterns. The predicted value map can be thought of as a smoothed map, in the sense that random variability, due to factors other than those included in the model has been smoothed out. A residual map may give an indication of systematic over- or under prediction in particular regions, which could be evidence of spatial autocorrelation, although this needs to be assessed more vigorously by means of a hypothesis test.

The residual map does suggest that similarly colored areas tend to be in similar locations, which could indicate positive spatial autocorrelation. Also, it indicates a tendency to over predict (negative residuals) in the outlying areas and a tendency to under predict (positive residuals) in the core, suggesting the possible presence of spatial heterogeneity in the form of spatial regimes (Anselin, 2005).

Results

From Figure 1, extreme value of RTC was found in Abuja (red circle). The numbers of RTC across the 36 states was normal (green circles). The size of each circle is proportional to the number of RTC in the spatial units (the states).
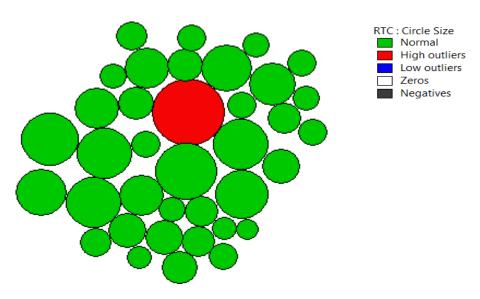
Figure 1. Dorling Cartogram of RTC (2011) in Nigeria

The upper outlier is Abuja. Niger state is the central position of RTC in Nigeria. 50 percent of the states have RTC number greater/lower than the number of RTC (88) that occurred in Niger state. 25 percent of the states had above 222 RTC cases in 2011, while the remaining 75 percent had RTC cases ranging from 16-180. Nine states had between 16-52 RTC cases, these include Akwa Ibom, Anambra, Bayelsa, Borno, Cross River, Ebonyi, Jigawa, Kebbi and Yobe states. In ten other states RTC cases ranged between 54 and 88. This includes Abia, Adamawa, Ekiti, Enugu, Gombe, Katsina, Niger, Plateau, Sokoto and Lagos states. Nine other states, Bauchi, Zamfara, Delta, Edo, Imo, Kaduna, Rivers and Taraba had between 90-180 RTC cases. At least 222 number of RTC cases where recorded in Benue, Kano, Kogi, Nassarawa, Ogun, Ondo, Osun and Oyo states (see Figure 2). Figure 3 allows a virtual inspection of the distributions across the spatial units.
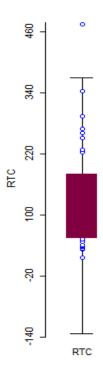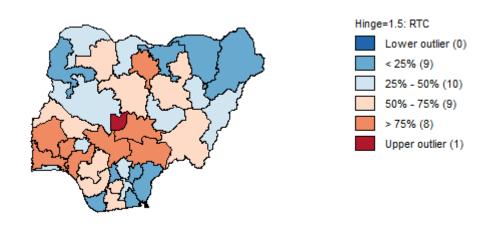
Figure 2. Box plot of RTC (2011) in Nigeria

Figure 3. Box plot Map of RTC (2011) in Nigeria

Spatial Autoregressive Model

Adopting the queen contiguity based spatial weights matrix, the spatial autoregressive model was built (Table 1). The number of observations equal 37, number of variables 6, while the degrees of freedom equal 31. The R-squared equal 50%. The mean of the dependent variable (RTC) is 128.78 and the standard deviation equal 104.74. The residual variance (sigma-square) is 5460.34, while, the standard error estimate (standard error regression) is 73.89.

A limited number of diagnostics are provided with the maximum likelihood lag estimation. First is the Breusch-Pagan test for heteroskedasticity in the error terms. The highly insignificant value of 0.88 (p-value 0.93) suggests that heteroskedasticity is not a serious problem. The second test is an alternative to the asymptotic significance test on the spatial autoregressive coefficient; it is not a test on remaining spatial autocorrelation. The value 5.88 (p-value 0.01) confirms strong significance of the spatial autoregressive coefficient.

The estimated ρ is positive and significant, indicating moderate spatial autoregressive dependence in RTC. In other words, RTC tend to be more clustered by states than what would be expected by a random distribution. The number of RTC cases for a given state is affected by the RTC cases of the neighbouring states.

Table 1. Spatial Autoregressive Model Estimates

Variable	Coefficient	Std. Error	z-value	Probability
Spatial Lag	0.45	0.17	2.64	0.01
Constant	1.44	39.29	0.04	0.97
Drivers	0.01	0.02	4.54	0.05
Vehicles	0.65	0.01	0.04	0.96
Area	-0.01	0.01	-0.21	0.83
Population	0.87	0.89	1.08	0.28

The parameter estimates for population, vehicles and area are not significant. This means that each of these exogenous variables does not contribute significantly to the incidence of RTC in Nigeria. However, the sign of the coefficient suggests the following:

Controlling for the spatial lag, population is positively related to the number of RTC occurring within the localities. Indicating that population generate a certain level of RTC. Number of registered vehicles characteristic produced a positive sign. The land area encompassing each state has a negative sign. All other things being equal, states with larger residential populations tend to have

more RTC and states with larger number of registered vehicles tend to have more RTC. Although, increase in the area of administration suggests less RTC.

Spatial units with higher than expected future likelihood of RTC include the FCT-Abuja, Kaduna, Kano, Oyo, Edo and Rivers states (Figure 4). States that seem to have high spatial dependence and where the RTC tend to be over predicted include Zamfara, Nassarawa, Benue, Kogi, Ondo and Osun states (Figure 5).
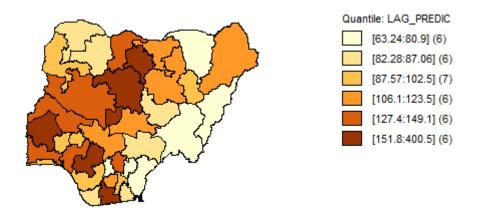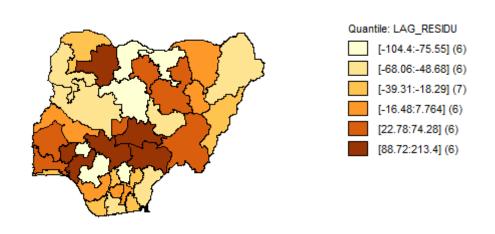

Figure 4. Predicted Value Map

Figure 5. Residual Map

Conclusion

In this paper, a combination of statistics and geographic information system is proposed. The applications have shown the power in the descriptive ability in statistical analysis. While, the proposed model uses a combination of statistics and geo visual representation of parameter estimates, tests for spatial autocorrelation is another option to increase emphasis on spatial dependencies. However, the proposed model has the advantage of giving a better understanding of parameter estimates.

References

Abdel-Aty, M. A. & Radwan, A. E. (2000). Modelling traffic accident occurrence and involvement. *Accident Analysis & Prevention, 32*(5), 633-642. doi:10.1016/S0001-4575(99)00094-9

Abdullah L. & Zamri N. (2012). Road accident models with two threshold levels of fuzzy linear regression. *Journal of Emerging Trends in Computing and Information Sciences, 3*(2), 225-230.

Aderamo A. J. (2012). Spatial pattern of road traffic accident casualties in Nigeria. *Mediterranean Journal of Social Sciences, 3*(2), 61-72. doi:10.5901/mjss.2012.v3n2.61

Aguero-Valverde, J. & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention, 38*(3), 618-625. doi:10.1016/j.aap.2005.12.006

Ahmed, M., Huang, H., Abdel-Aty, M. & Guevara, B. (2011). Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis & Prevention, 43*(4), 1581-1589. doi:10.1016/j.aap.2011.03.021

Amoros, E., Martin, J. L. & Laumon, B. (2003). Comparison of road crashes incidence and severity between some French countries. *Accident Analysis & Prevention, 35*(4), 537-547. doi:10.1016/S0001-4575(02)00031-3

Anastasopoulos, P. Ch., Tarko, A. P. & Mannering, F. L. (2008). Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis & Prevention, 40*(2), 768-775. doi:10.1016/j.aap.2007.09.006

Anselin, L. (1988). *Spatial econometrics: methods and models*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Anselin, L. (1990). Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics, 20*(2), 141-163. doi:10.1016/0166-0462(90)90001-J

Anselin, L. (2005). *Exploring spatial data with GeoDa: A workbook.* 165-212. Center for Spatially Integrated Social Science. University of Illinois, Urbana-Champaign. Retrieved January 20, 2013 from http://www.csiss.org/ or http://sal.agecon.uiuc.edu/

Anselin, L & Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association, 65*(1), 19-33. doi:10.1111/j.1435-5597.1988.tb01155.x

Dorling, D. (1996). *Area cartograms: Their use and creation.* Concepts and Techniques in Modern Geography series no. 59. Norwich: University of East Anglia.

Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research, 40*(5), 341-351. doi:10.1016/j.jsr.2009.07.006

Fridstrøm, L. & Ingebrigtsen, S. (1991). An aggregate accident model based on pooled regional time series data. *Accident Analysis & Prevention, 23*(5), 363-378. doi:10.1016/0001-4575(91)90057-C

Goh, K. C. K., Currie, G., Sarvi, M. & Logan, D. (2014). Bus accident analysis of routes with/without bus priority. *Accident Analysis & Prevention, 65,* 18-27. doi:10.1016/j.aap.2013.12.002

Griffith, D. A. (2009). Spatially autoregressive models. In R. Kitchin & N. Thrift (Eds.) *International Encyclopedia of Human Geography* (pp. 396-402). Oxford: Elsevier.

Guo, F., Wang, X. & Abdel-Aty, M. A. (2010). Modelling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention, 42*(1), 84-92. doi:10.1016/j.aap.2009.07.005

Jegede, F. J. (1988). Spatio-temporal analysis of road traffic accidents in Oyo State, Nigeria. *Accident Analysis & Prevention, 20*(3), 227-243. doi:10.1016/0001-4575(88)90007-3

Jiang, X., Abdel-Aty, M. & Alamili, S. (2014). Application of Poisson random effect models for highway network screening. *Accident Analysis & Prevention, 63,* 74-82. doi:10.1016/j.aap.2013.10.029

Jones, B., Janssen, L. & Mannering, F. (1991). Analysis of the frequency and duration of freeway accidents in Seattle. *Accident Analysis & Prevention, 23*(4), 239-255. doi:10.1016/0001-4575(91)90003-N

Jovanis, P. & Chang, H. (1986). Modelling the relationship of accidents to miles traveled. *Transportation Research Record, 1068*, 42-51.

Kim, K., Brunner, I. & Yamashita, E. (2006). Influence of land use, population, employment, and economic activity on accidents. *Transportation Research Record: Journal of Transport Research Board, 1953*, 56-64. doi:10.3141/1953-07

Korter, G. O., Olubusoye, O. E. & Salisu, A. A. (2014). Spatial modelling of road traffic accidents in Oyo State, Nigeria. *African Journal of Sustainable Development (AJSD), 4* (2), 99-118.

LeSage, J. & Pace, R. K. (2009). *Introduction to spatial econometrics.* Boca Raton: Chapman and Hall/CRC.

Levine, N., Kim, K. E. & Nitz, L. H. (1995a). Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis & Prevention, 27*(5), 663-674. doi:10.1016/0001-4575(95)00017-T

Levine, N., Kim, K. E. & Nitz, L. H. (1995b). Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. *Accident Analysis & Prevention, 27*(5), 675-685. doi:10.1016/0001-4575(95)00018-U

Lord, D. & Miranda-Moreno, L. F. (2008). Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson–gamma models for modeling motor vehicle crashes: A Bayesian perspective. *Safety Science, 46*(5), 751-770. doi:10.1016/j.ssci.2007.03.005

Maher, M. J. & Mountain, L. J. (1988). The identification of accident blackspots: A comparison of current methods. *Accident Analysis & Prevention, 20*(2), 143-151. doi:10.1016/0001-4575(88)90031-0

McGill, R., Tukey, J. W. & Larsen, W. A. (1978). Variations of box plots. *The American Statistician, 32*(1), 12-16.doi:10.1080/00031305.1978.10479236

Miaou, S.-P. & Lord, D. (2003). Modelling traffic crash-flow relationship for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. *Transportation Research Record: Journal of Transport Research Board 1840,* 31-40. doi:10.3141/1840-04

Milton, J. & Mannering, F. (1998). The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation, 25*(4), 395-413. doi:10.1023/A:1005095725001

Noland, R.B. & Quddus, M.A. (2005). Congestion and safety: A spatial analysis of London. *Transportation Research Part A: Policy and Practice, 39*(7/9), 737-754. doi:10.1016/j.tra.2005.02.022

Okamoto, H. & Koshi, M. (1989). A method to cope with the random errors of observed accident rates in regression analysis. *Accident Analysis & Prevention, 21*(4), 317-332. doi:10.1016/0001-4575(89)90023-7

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis & Prevention, 40*(4), 1486-1497.doi:10.1016/j.aap.2008.03.009

Roque, C. & Cardoso, J. L. (2014). Investigating the relationship between run-off the-road crash frequency and traffic flow through different functional forms. *Accident Analysis and Prevention, 63,* 121-132. doi:10.1016/j.aap.2013.10.034

Shanker, V., Mannering, F. & Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural accident frequencies. *Accident Analysis & Prevention, 27*(3), 371-389. doi:10.1016/0001-4575(94)00078-Z

Shefer, D. & Rietveld, P. (1997). Congestion and safety on highways: Towards an analytical model. *Urban Studies, 34* (4), 679-692. doi:10.1080/0042098975970

Wang, C., Quddus, M. A. & Ison, S. G. (2009). Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention, 41*(4), 798-808. doi:10.1016/j.aap.2009.04.002

Ye, X., Pendyala, R. M., Washington, S. P., Konduri, K., & Oh, J. (2009). A simultaneous equations model of crash frequency by collision type for rural intersections. *Safety Science, 47*(3), 443-452. doi:10.1016/j.ssci.2008.06.007

Yu, R. & Abdel-Aty, M. (2013). Investigating different approaches to develop information priors in hierarchical Bayesian safety performance functions. *Accident Analysis & Prevention, 56,* 51-58.doi:10.1016/j.aap.2013.03.023

Yu, R., Abdel-Aty, M., & Ahmed, M. (2013). Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis & Prevention, 50,* 371-376. doi:10.1016/j.aap.2012.05.011