All Accidents are Not Equal: Using Geographically Weighted Regressions Models to Assess and Forecast Accident Impacts

Libing Zheng
China Railway Siyuan Survey and Design Group Co., LTD
Wuhan, 430063, China, e-mail: zheng_rixin@yahoo.com

R. Michael Robinson
Assistant Research Professor, Virginia Modeling, Analysis, and Simulation Center,
Old Dominion University, Suffolk, VA, USA, e-mail: rmrobins@odu.edu

Asad Khattak
Professor of Civil and Environmental Engineering, College of Engineering,
Old Dominion University, Norfolk, VA, USA, e-mail: akhattak@odu.edu

Xin Wang
Civil and Environmental Engineering Department
Old Dominion University, Norfolk, VA, USA, e-mail: xwang008@odu.edu

Submitted to the 3rd International Conference on Road Safety and Simulation, September 14-16, 2011, Indianapolis, USA

ABSTRACT

Transportation professionals have long recognized the importance of accounting for accident and incident impacts when designing and constructing transportation networks. Studies have used statistical models of accidents to explore associations of traffic injuries/harm with driver, vehicle, roadway and environmental factors. A critical characteristic of most studies is their reliance on traditional models (referred to as global models) that assume the efficacy of a single set of estimated parameters to forecast crash impacts — an approach characterized as "one size fits all." However, in spatially diverse metropolitan regions, accident impacts and their associations with variables can vary across space, resulting in unobserved spatial heterogeneity. Overcoming this weakness has led to the use of various spatial analysis techniques. Using collision data from the Hampton Roads region of southeastern Virginia and the estimated economic costs of accidents, this paper explores spatial relationships and provides comparisons of results obtained from global models and those obtained using the technique of Geographically Weighted Regression (GWR). Estimation of collision harm (assumed to be related to the approximate monetary costs of accidents) indicates that GWR methods yield significantly more accurate results. The results provide valuable information on high-risk factors associated with collision harm and the spatial variations in these associations and suggest improved data application in dynamic traffic simulations.

Keywords: crash harm, GWR, global model, local model, spatial

INTRODUCTION

Vehicle accidents may cause property damage, traffic congestion delays, and personal injuries or death. The importance of accounting for the impacts of accidents in transportation planning and road design has long been recognized and considered in network design and operational planning. A critical characteristic of most planning studies is their reliance on traditional models (referred to as global models) that assume the efficacy of a single set of estimated parameters to forecast crash impacts throughout the modeled area — an approach characterized as "one size fits all." Severity is measured using the costs of accidents and the product of frequencies times cost is referred to as crash harm. Crash harm is generally associated with driver, vehicle, roadway, and environmental factors. Traditional Ordinary Least Square (OLS) regression (Council et al., 2003; Khattak and Targa, 2004) is usually used to model these relationships. However, in spatially diverse metropolitan regions, accident impacts and their associations with variables vary across space, causing a problem known as unobserved spatial heterogeneity. Geographically Weighted Regression (GWR) methods overcomes this deficiency and yield more accurate results.

Using collision data from the Hampton Roads region of southeastern Virginia, this paper explores spatial relationships and provides comparisons of results obtained from global and GWR models. The relative impacts of accidents are assessed by comparing the frequency and the severity of collisions at various locations throughout the region. The results provide valuable information on high-risk factors associated with collision harm and the spatial variations in these associations and suggest improved data application in dynamic traffic simulations.

LITERATURE REVIEW

Overcoming the weaknesses of "one size fits all" estimates such as global regression and ordinary least squares analysis has led researchers to use varying types of spatial analysis to provide insights on accidents impacts that might otherwise be overlooked. Such insights have provided value in better understanding the factors that contribute to accident frequencies and severities. For example Levine et al. (1995) used spatial analysis to assess vehicle crashes in Honolulu and showed the dynamic variations of crash densities with the traffic volumes and patterns associated with days of the week and time of day. Loo (2009) used spatial characteristics of road crashes to identify hot zones for crashes in Hong Kong, comparing these results with "blacksite methodology" for identifying hazardous areas.

Geographically weighted regression is a particular technique for spatially varying relationships. In essence, it uses regression parameters for each location assessed and allows evaluations of how parameter changes vary from one location to another. GWR uses the attributes considered in OLS analysis, but adds consideration of the geographic location of data points. When using GWR, users assume that points physically nearer one another, on the same road type (for this project, on Interstate highways), and sharing the same physical characteristics (such as number of lanes, pavement condition, etc.) are more alike than those further apart. It provides local parameter estimates for variables in a spatial context.

GWR is often interpreted as a smoothing function. Because variable values are weighted by the values of nearest neighbors, discontinuities and sudden changes in magnitude are minimized and GWR can create a highly accurate observed variable surface. This makes the technique attractive for various aspects of urban analysis, a benefit demonstrated by Paez and Scott (2005). Although widely used in other fields, research with an emphasis on GWR applications in transportation is a fairly recent phenomenon (Zhao and Park, 2004; Chow et al., 2006; Du and Mulley, 2003; Wang and Khattak, 2011). Relatively few studies exist using GWR in transportation safety analyses. Hadayeghi et al. (2009) utilized geographically weighted Poisson regression (GWPR) to model zonal collision counts and concluded that the local model estimation technique of GWPR can improve analysis of transportation networks. Park et al. (2010) used GWR to identify hazardous locations based on severity scores of highway crashes.

The current study differs from previous works by its use of economic impacts to assign and assess the severity of accidents. It employs GWR as described by Fotheringham et al. (2002) and makes use of the GWR analysis software provided by these authors.

METHODOLOGY

Crash Harm

Crash harm includes the estimated costs from personal and property damage. Information for accidents occurring on Interstate highways in the Hampton Roads region of Virginia in 2006, recorded by the Virginia State Police and provided by the Virginia Department of Transportation (VDOT), was used for all analyses. The severity of injuries was included in the data and was categorized as follows:

1) Dead before report;
2) Visible signs of injury requiring assistance (bleeding wounds, individual required transport from the scene):
3) Other visible injury (bruises, abrasions, swelling, etc.);
4) No visible injury, but complaint of pain or momentary loss of consciousness.

Total crash harm is crash harm multiplied by the frequency of accidents at a particular site and is expressed in units of dollars/time. This study uses estimated costs suggested by the Secretary of Transportation to the Federal Highway Administration (FHWA) which updated 1994 estimates with information from a year 2000 study of crash costs conducted by Blincoe for the National Highway Traffic Safety Administration (NHTSA) (Blincoe, 2002). The estimated costs (per occurrence) used in this study were:

- Collisions with no injury: $2250;
- Collisions with nonfatal injury: $63,000; and
- Collisions with fatal injury: $3,000,000.

Geographically Weighted Regression

GWR explores the spatial deviations of associations between dependent and explanatory variables by relaxing the assumption that estimated parameters hold globally. In this context it is useful in identifying whether the association of between harm and a particular explanatory variable is relatively stable over space or it varies substantially. This can in turn help with spatially targeted countermeasure development. In GWR, the regression model is calibrated based on data geographically proximate to a specific location. In other words, GWR assesses parameters within specified distances (called bandwidths) of one another and weights these parameters from an identified regression reference point (Fotheringhan, 2002). The basic GWR equation can be written as:

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i \text{ where}$$

y_i is the dependent variable at location i, β_{i0} is the constant at point i, β_{ik} is the coefficient at point i for variable x_{ik}, x_{ik} is the independent variable of the k^{th} parameter at location i, ε_I is the error term at location i, and p is the number of parameters being estimated. The critical difference between global and GWR analysis is that the global estimation uses one model for all observations while the GWR estimates a particular local model for each location in space. Monte Carlo significance tests for the parameter estimates can determine if estimated parameters have significant spatial variability.

DATA USED

Three datasets for the Hampton Roads region of Virginia were used in the study. Information for accidents occurring on Interstate highways in Hampton Roads in 2006, recorded by the Virginia State Police and provided by VDOT, was used for all analyses. The database contained 4517 crashes, each one with 67 variables, including crash vehicle, driver information, personal and equipment/facility damage, and environmental and roadway factors. Crash locations were recorded using Interstate milepost values. These locations were geocoded to latitude and longitude coordinates and matched with the VDOT route system. Key factors and their descriptive statistics used in the study are provided in Table 1.

VDOT, the Hampton Roads Planning District Commission (HRPDC), and the Hampton Roads Transportation Planning Organization (HRTPO) provided basic information about roadway segments in Hampton Roads. Data included segment lengths, annual average daily traffic (AADT), annual average weekday daily traffic (AAWDT), and truck contributions to total traffic (as a percentage) for each segment. HRPDC and HRTPO also provided Traffic Analysis Zone (TAZ) data, including spatial, population, and employment information for each TAZ. GIS tools were used to merge crash data, roadway data and traffic analysis zone data and to assist with identifying the key factors associated with crash harm.

Kernel density analysis can help examine accident hotspots. Here, Kernel density is used to examine the spatial distributions of secondary and non-secondary incidents. This method calculates the density of a variable in a search radius, and shows where incidents are concentrated. A kernel function K determines the shape of the bumps while the parameter h determines their width. By calculating the incident density, a surface can be created showing

Table 1 Descriptive Statistics for Variables (4,517 observations)

Variable	Mean	Std. Dev.	Min	Max
Total harm	37854	196136	100	3100000
Ln harm (log-transformed of total harm)	9.17	1.53	4.61	14.95
Percentage of truck traffic	0.770	0.422	0	2
Peak hour (6:00-9:00, 16:00-19:00)	0.302	0.459	0	1
Road Width	35.090	11.45	0	72
Bridge or not	0.079	0.27	0	1

Variable		Freq.	%
Function class	Urban Area	4,371	96.77
	Rural Area	146	3.23
Facility	Two way uncontrolled	20	0.44
	Two way full control	4,300	95.20
	One way	173	3.830
	Not stated (facility)	24	0.531
Intersection	T-leg intersection	8	0.177
	Interchange	870	19.261
	Not Intersection	3,639	80.562
Crash type	Rear End	2,481	54.93
	Angle	6	0.13
	Head on	6	0.13
	Sideswipe-same direction	602	13.33
	Sideswipe-opposite direction	5	0.11
	Fixed object- in road	13	0.29
	No collision	125	2.77
	Fix object-off road	1,211	26.81
	Deer	30	0.66
	Pedestrian	4	0.09
	Backed into	3	0.07
	Other	31	0.69

the spatial distribution of secondary and non-secondary incidents throughout the network. The kernel density function is described in flowing equations.

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

Where: n = sample size

h = bandwidth parameter (kernel radius)

X_i = Observed frequency of incidents on the segment i

$\hat{f}(x)$ = estimate of the intensity of the spatial point pattern measurement at location x

The function *K(x)* will be a symmetric probability density function, the normal density, for instance, or Gaussian function (shown in the following equation) with mean zero and variance one.

$$K\left(x\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$$

Kernel density is a relative value that can improve one's understanding of the hotspot location. Figure 1 shows the kernel density for crash frequency. Green coloring indicates areas of low density, yellow moderate density, and red high density. Figure 2 represents the same data and geographic area, but the coloring indicates relative densities of crash harm as measured in estimated dollar costs. Category descriptors are further explained in the following section.

Comparison of the two figures shows how high density locations for crash frequency may differ from areas of high crash harm; differences between spatial distribution of crash frequency density and crash harm density are clearly seen in the differing color patterns. In particular, readers should take note of the two bridge-tunnel complexes included in the graphic. Although high crash frequencies are noted for both the Hampton Roads Bridge Tunnel (HRBT) and the Monitor-Merrimac Memorial Bridge Tunnel (MMMBT), crash harm drops in both locations. Along I-64 at the bottom-center of the maps, the opposite situation exists with an area of more moderate crash frequency producing high values for crash harm. The variation of crash harm is substantially different over space, points to unobserved spatial heterogeneity.

Descriptive statistics identified in Table 1 show that the maximum total harm of these crashes is 3.1 million dollars, which represents a fatal crash with 100,000 dollars property damage and the minimum total harm is 100 dollars, which represents no one injured in that crash with property damage of 100 dollars.

The maximum percentage of truck traffic is 2 percent and minimum percentage of truck traffic is 0, as trucks are restricted from going on certain routes. Less than half of the crashes (30.2%) happened during peak hour, 6 to 9 in the morning and 4 to 7 in the afternoon.

And considering Hampton Roads area is on the East coast and has bridges and tunnels, 7.9% crashes happened in bridge or tunnel. 4,731 crashes (96%) happened in urban area and only 146 crashes(less than 4%) happened in rural area. And 4,300 crashes (95%) were involved in two-way roadway facility. Further, 217 crashes (5%) are in other kinds of facilities, including two-way uncontrolled facilities, and one way facilities.

The most common crash type is rear end crashes, which accounts for 54.93% of total crashes. Following are Fixed object (off road) crashes and sideswipe (same direction) crashes are also very common, which account for 26.81% and 13.33% of the total crashes, respectively. Other type of crash may not as common as these three, but still might be very harmful, for example, head on crashes.

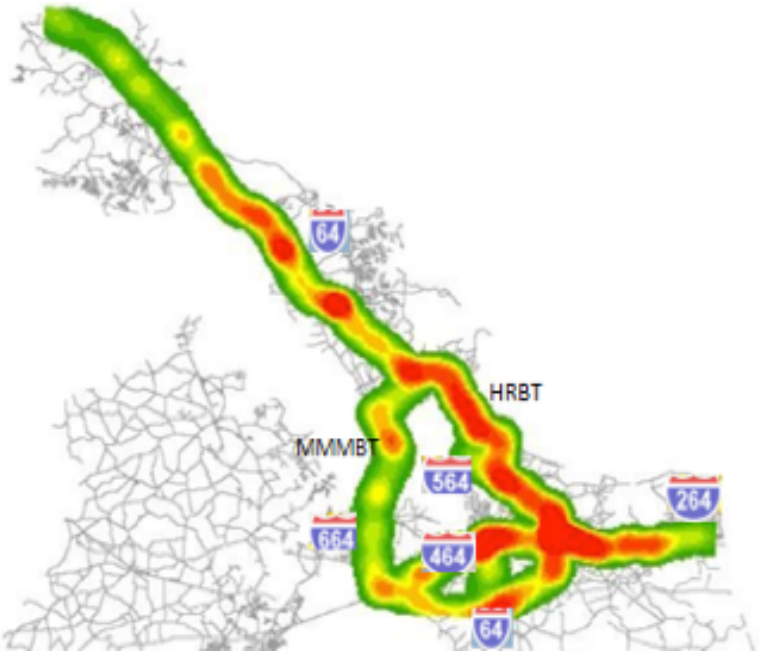

Figure 1 Crash Frequency density

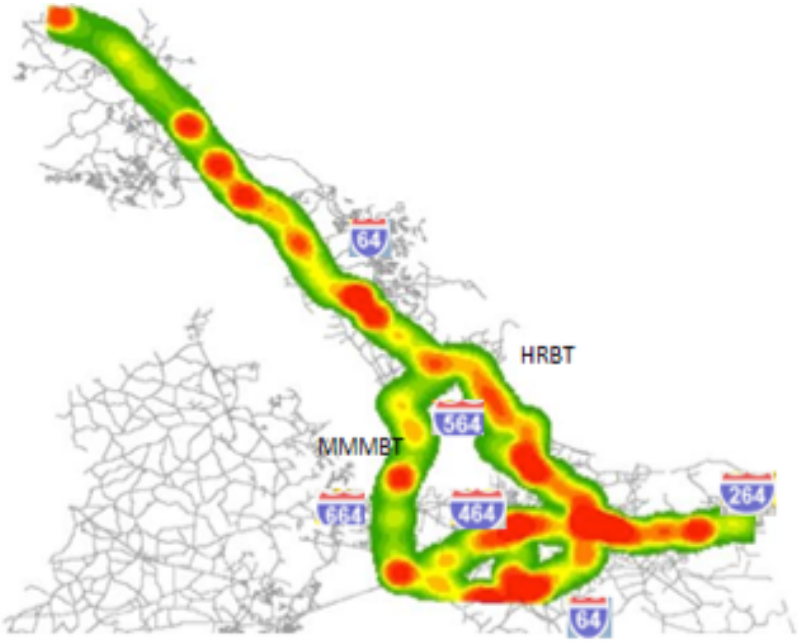

Figure 2 Crash harm density

MODELING CRASHES

The global model for the log-transformed cost of harm (ln harm) and their tests for non-stationarity are provided in Table 2. Variables with positive coefficient values contribute to higher crash harm; negative values indicate a contribution to lower crash harm. Both models are statistically significant overall. The test for bandwidth is significant, which means that the GWR model is statistically better than the global OLS model. From the global model, the factors found to be significantly associated with higher crash harm include:

- collision type, including the striking angle and direction;
- another object involved, e.g., another vehicle, a pedestrian, animals, fixed objects;
- truck involvement;
- road characteristics, especially the number of lanes and lane width;
- spatial location (on a bridge, urban or rural area, intersection, etc.); and
- time of day (peak traffic period).

Table 2 Global Model for Log Transformed Cost of Harm and Test for Non-Stationarity

Global OLS (sample size: 4,517)				
	Variable	Coef.	P>|t|	Test for non-stationary
Collision Type	Rear End	-1.144	0.000	0.140
	Angle	-1.229	0.050	0.040
	Head on	1.943	0.002	0.340
	Sideswipe-same direction	-1.233	0.000	0.200
	Sideswipe-opp. Direction	-0.691	0.313	0.250
	Fixed object- in road	-0.843	0.054	0.130
	Fix object-off road	-1.009	0.000	0.030
	Deer or other animal	-2.148	0.000	0.430
	Pedestrian	0.725	0.342	0.100
	Backed into	-2.675	0.002	0.100
	Other	-1.122	0.000	0.570
Roadway & Spatial variables	Bridge or not	0.091	0.316	0.420
	Road Width	0.006	0.008	0.090
	Urban Area	-0.187	0.054	0.960
	Two way uncontrolled	0.172	0.611	0.540
	One-way	-0.003	0.984	0.640
	Not stated (facility)	-0.25	0.432	0.998
	T-leg intersection	0.703	0.187	0.290
	Interchange	-0.08	0.197	0.730
Other variables	Percentage of truck traffic	0.15	0.008	0.000
	Peak hour (6:00-9:00, 16:00-19:00)	-0.179	0.000	0.140
	Constant	10.175	0.000	0.810
Model Info.	Number of obs. =4,517, Prob > F 0.0000 R-squared 0.0357			

Among independent variables that are significantly associated with crash harm, the test for non-stationarity shows that collision angle, fixed-object-off road collision, road width and percentage of truck traffic are statistically significant. The null hypothesis of stationarity for these variables can be rejected statistically. That is, their associations with crash harm vary significantly over space. Therefore, using a fixed global model (such as OLS) will cause misspecification and hide the detailed information on spatial distribution of the association.

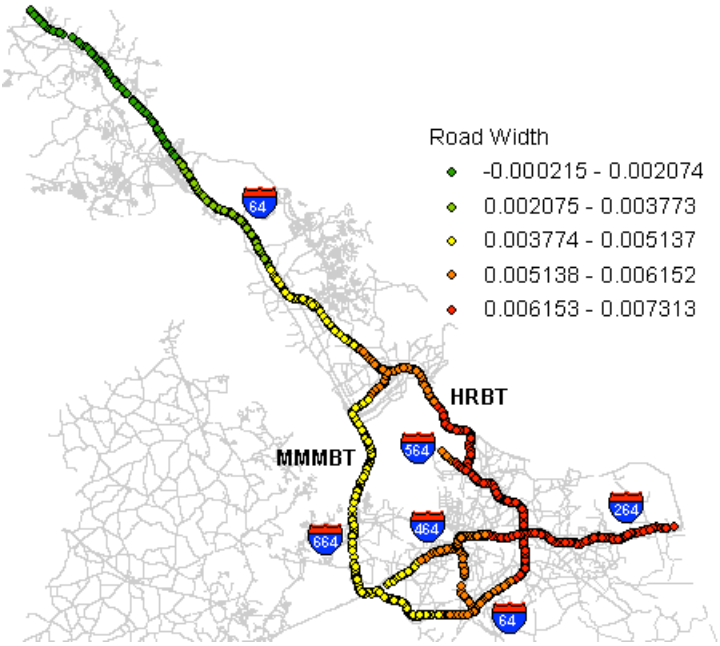

Figure 3 Local parameter estimates associated with the variable road width

Table 3 Information on Road Width for Validation of Five OLS Models

Road Width		Parameter Information		Validation Model	
		Coef.	p>t	# of obs	Prob > F
Sub-groups OLS models	1 ◆	(dropped)	na	160	0.0028
	2◆	(dropped)	na	453	0.0000
	3◆	-.0031417	0.561	833	0.0012
	4◆	.0071805	0.129	1,185	0.0000
	5◆	.0092065	0.013	1,886	0.0000
Global OLS Model		0.006	0.008	4,517	0.0000

(Note: In local GWR result, the range of each estimator subgroup for road width is: 1◆: -0.000215- 0.002074; 2◆: 0.002075- 0.003773; 3◆: 0.003774- 0.005137; 4◆: 0.005138- 0.006152, 5◆: 0.006153- 0.007313)

Spatial variation of local parameter estimations using results of geographically weighted regression models are mapped for road width and percentage of truck traffic as shown in Figures 3 and 4. Tables 3 and 4 show the estimation results. Parameter estimations are classified into five groups by the "natural break" method and mapped by point symbols with different colors. Natural break is the default classification method available in ArcGIS. It identifies natural

groupings in the data and break points, picking the breaks that best group similar values and maximize the differences between classes.

To explore the validity of GWR method, tests of the variables are provided. Several un-pooled regression models are estimated as opposed to a pooled global model. For each variable, all collisions are categorized in to five sub-groups according to quantity of the estimator and color of the crash point shown in Figures 3 and 4. The validation OLS regressions with the same dependent and independent variables are given for each crash sub-group. If the unspooled model results from different spatial areas are consistent with ones obtained from the GWR model, then that will confirm the validity of GWR.

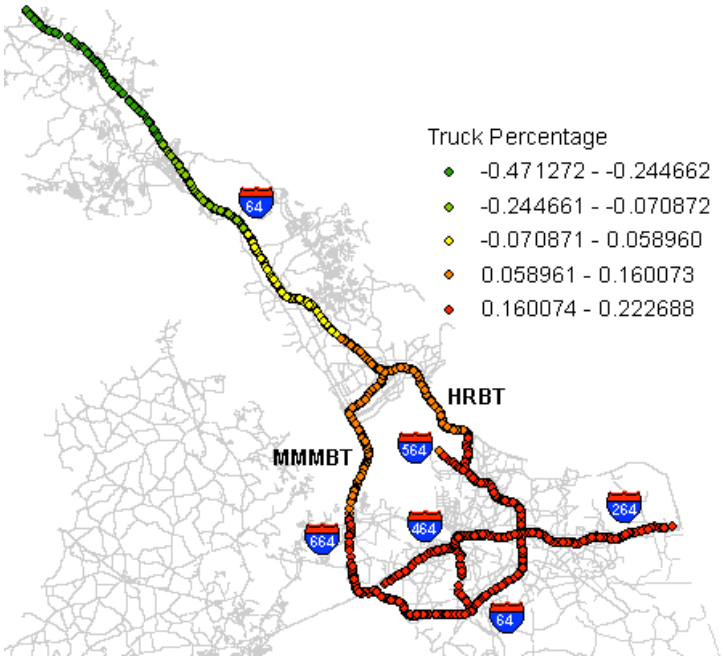

Figure 4 Local parameter estimates associated with the variable truck percentage

Table 4 Information on Truck Percentages for Validation of Five OLS Models

Truck Percentage		Parameter Information		Validation Model	
		Coef.	p>t	# of obs	Prob > F
Sub-groups of OLS models (unspooled)	1 ◆	(dropped)	na	142	0.0058
	2◆	(dropped)	na	302	0.0003
	3◆	-.2986355	0.278	479	0.020
	4◆	-1.475976	0.074	876	0.0004
	5◆	.252419	0.000	2,718	0.000
Global OLS Model (pooled)		0.150	0.008	4,517	0.0000

(Note: In the local GWR model, the range of each estimator subgroup for Truck Percentage is:
1 ◆: -0.471272- -0.244662; 2◆: -0.244661- -0.070872; 3◆: -0.070871- 0.058960; 4◆: 0.058961- 0.160073;
5◆: 0.160074- 0.22688)

An interesting finding is that the road width is significantly and positively associated with higher crash harm. One unit increase in surface width is associated with 0.6% higher harm on average. The cause of this relationship cannot be identified, given the cross-sectional nature of the dataset. The authors conjecture that the greater width may be associated with greater differences in relative speeds of vehicles and the opportunity for impacts. This finding shows the importance of specifying the criteria used to assess an area as worse in terms of accident impacts. The relationship is counter to what one would intuitively expect if the analysis used traffic delay instead of economic cost, when narrower road widths and fewer lanes contribute to significantly worse conditions as shown in Robinson (2007) and Robinson et al. (2009).

From the local model and the map of road width coefficients, one can see that the relationship between surface width and crash harm changes over space. The magnitude of road width associated with harm is larger in the eastern, more densely populated sections of the region than in the western portions of the Hampton Roads region. Similarly, the association between crash harm and percent of trucks in traffic flow also changes over space. Higher percentage of truck traffic south of HRBT and MMMBT is associated with higher collision harm compared to areas that are north of these two bridge tunnels.

CONCLUSION

While researchers have focused on understanding factors associated with crash harm, few studies have explored spatial variations in associations. This paper fills a critical gap by investigating if relationships observed in transportation safety vary across space. A critical finding is that, spatial heterogeneity exists in crash harm, as crashes on highways are often clustered. This leads to the conclusion that associations of roadway, traffic, driver and socio-demographical factors with crash harm are not identical across the space. Thus, the basic independence assumption of OLS does not hold in the situation explored. Instead, GWR, which is a local model, provides a better statistical fit than a traditional OLS model by capturing spatial heterogeneity. Although the spatial relationships uncovered through GWR in this paper are only valid for this region, due to the fact that the model itself—the weights used in the model are based on spatial locations, which are unique to the area, the methodology used in this paper can be transferred to other regions. The map of coefficients can provide a detailed picture of where (in space) certain factors are associated with higher crash harm. This can provide valuable information to help safety agencies pay more attention to critical factors in certain locations where they have the largest associations. Studies on countermeasures that might be more effective in specific locations can be facilitated by the analysis conducted. Subsequently, more efficient resource assignment, improvement and countermeasure implementation will be achieved.

ACKNOWLEDGMENTS

The authors would like to express their appreciation to the Hampton Roads Smart Traffic Center for the provision of accident and incident data and to the Old Dominion University Transportation Research Institute and Virginia Modeling, Analysis, and Simulation Center for their support of this research.

REFERENCES

Blincoe, L., A. Seay, E. Zaloshnja, T. Miller, E. Romano, S. Luchter, and R. Spicer. 2002. The Economic Impact of Motor Vehicle Crashes, 2000, National Highway Traffic Safety Administration, Washington D.C., Report No. DOT HS 809 446, May 2002.

Chow, Lee-Fang, Fang Zhao, Xuemei Liu, Min-Tang Li, and Ike Ubaka, *Transit Ridership Model Based on Geographically Weighted Regression.* Transportation Research Record: Journal of the Transportation Research Board, 2006. Vol. 1972: p. 105-114.

Council, F., Harkey, D., Nabors, D. , Khattak A., Mohamedshah Y., *Examination of Fault, Unsafe Driving Acts, and Total Harm in Car-Truck Collisions.* Transportation Research Record, 2003. 1830: p. 63-71.

Du. H. and C. Mulley, *Relationship Between Transport Accessibility and Land Value: Local Model Approach with Geographically Weighted Regression.* Transportation Research Record: Journal of the Transportation Research Board, 2006. 1977(-1): p. 197-205.

Fotheringham, A.S., Brunsdon, C., and Charlton, M.E., 2002, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Chichester: Wiley.

Hadayeghi, A., A.S. Shalaby, and B.N. Persaud, *Development of Planning Level Transportation Safety Tools Using Geographically Weighted Poisson Regression*, in *TRB Annual Meeting CD-ROM*. 2009 Washington, D.C.

Khattak, A.J., Targa, F., *Injury Severity and Total Harm in Truck-Involved Work Zone Crashes*, in *Transportation Research Record, TRB, National Research Council, Washington, D.C.* 2004.

Levin, N., K. E. Kim and L. H. Nitz, *Spatial Analysis of Honolulu Motor Vehicle Crashes: Spatial Patterns.* Accid. Anal. and Prev., 1995. Vol. 27 (No. 5): p. 663-674.

Loo, Becky P.Y., *The Identification of Hazardous Road Locations: A Comparison of the Blacksite and Hot Zone Methodologies in Hong Kong*, International Journal of Sustainable Transportation, 2009, Vol. 3, pp. 187-202.

Páez, A. and D. Scott, *Spatial statistics for urban analysis: A review of techniques with examples.* GeoJournal, 2005. 61(1): p. 53-67.

Park, S.H., D-K Kim, S-Y Kho, and S. Rhee, *Identifying Hazardous Locations Based on Severity Scores of Highway Crashes*, in *12th World Conference on Transport Research*. 2010: Lisbon.

Robinson, R.M.. *Hampton Roads Hurricane Evacuation Study*. Report Number V07-008, provided to the Virginia Department of Emergency Management, 2007, Available from the Virginia Department of Emergency Management, 10501 Trade Court, Richmond, VA 23236.

Robinson, R.M., A. Khattak, J. Sokolowski, P. Foytik, and X. Wang. What is the Role of Traffic Incidents in Hampton Roads Hurricane Evacuations? In *Transportation Research Board 2009 Annual Meeting,* No. 1339, CD-ROM. Transportation Research Board of the National Academies, Washington, D.C., 2009.

Wang, X. and Khattak, Asad, *Role of Travel Information in Supporting Travel Decision Adaption: Exploring Spatial Patterns*, Transportmetrica, pp. 1-19, 2011.

Zhao, F., N. Park, *Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic.* Transportation Research Record, 2004. 1879: p. 99-107.