University Transportation Research Center - Region 2

Final Report

Omitted Variable Bias in Crash Reduction Factors

Performing Organization: Rutgers University

September 2015

Sponsor:
University Transportation Research Center - Region 2

University Transportation Research Center - Region 2

The Region 2 University Transportation Research Center (UTRC) is one of ten original University Transportation Centers established in 1987 by the U.S. Congress. These Centers were established with the recognition that transportation plays a key role in the nation's economy and the quality of life of its citizens. University faculty members provide a critical link in resolving our national and regional transportation problems while training the professionals who address our transportation systems and their customers on a daily basis.

The UTRC was established in order to support research, education and the transfer of technology in the field of transportation. The theme of the Center is "Planning and Managing Regional Transportation Systems in a Changing World." Presently, under the direction of Dr. Camille Kamga, the UTRC represents USDOT Region II, including New York, New Jersey, Puerto Rico and the U.S. Virgin Islands. Functioning as a consortium of twelve major Universities throughout the region, UTRC is located at the CUNY Institute for Transportation Systems at The City College of New York, the lead institution of the consortium. The Center, through its consortium, an Agency-Industry Council and its Director and Staff, supports research, education, and technology transfer under its theme. UTRC's three main goals are:

Research

The research program objectives are (1) to develop a theme based transportation research program that is responsive to the needs of regional transportation organizations and stakeholders, and (2) to conduct that program in cooperation with the partners. The program includes both studies that are identified with research partners of projects targeted to the theme, and targeted, short-term projects. The program develops competitive proposals, which are evaluated to insure the mostresponsive UTRC team conducts the work. The research program is responsive to the UTRC theme: "Planning and Managing Regional Transportation Systems in a Changing World." The complex transportation system of transit and infrastructure, and the rapidly changing environment impacts the nation's largest city and metropolitan area. The New York/New Jersey Metropolitan has over 19 million people, 600,000 businesses and 9 million workers. The Region's intermodal and multimodal systems must serve all customers and stakeholders within the region and globally.Under the current grant, the new research projects and the ongoing research projects concentrate the program efforts on the categories of Transportation Systems Performance and Information Infrastructure to provide needed services to the New Jersey Department of Transportation, New York City Department of Transportation, New York Metropolitan Transportation Council , New York State Department of Transportation, and the New York State Energy and Research Development Authorityand others, all while enhancing the center's theme.

Education and Workforce Development

The modern professional must combine the technical skills of engineering and planning with knowledge of economics, environmental science, management, finance, and law as well as negotiation skills, psychology and sociology. And, she/he must be computer literate, wired to the web, and knowledgeable about advances in information technology. UTRC's education and training efforts provide a multidisciplinary program of course work and experiential learning to train students and provide advanced training or retraining of practitioners to plan and manage regional transportation systems. UTRC must meet the need to educate the undergraduate and graduate student with a foundation of transportation fundamentals that allows for solving complex problems in a world much more dynamic than even a decade ago. Simultaneously, the demand for continuing education is growing – either because of professional license requirements or because the workplace demands it – and provides the opportunity to combine State of Practice education with tailored ways of delivering content.

Technology Transfer

UTRC's Technology Transfer Program goes beyond what might be considered "traditional" technology transfer activities. Its main objectives are (1) to increase the awareness and level of information concerning transportation issues facing Region 2; (2) to improve the knowledge base and approach to problem solving of the region's transportation workforce, from those operating the systems to those at the most senior level of managing the system; and by doing so, to improve the overall professional capability of the transportation workforce; (3) to stimulate discussion and debate concerning the integration of new technologies into our culture, our work and our transportation systems; (4) to provide the more traditional but extremely important job of disseminating research and project reports, studies, analysis and use of tools to the education, research and practicing community both nationally and internationally; and (5) to provide unbiased information and testimony to decision-makers concerning regional transportation issues consistent with the UTRC theme.

Project No(s):
UTRC/RF Grant No: 49997-38-25

Project Date: September 2015

Project Title: Omitted Variable Bias in Crash Reduction Factors

Project's Website:
http://www.utrc2.org/research/projects/omitted-variable-bias

Principal Investigator(s):

Dr. Robert B. Noland
Alan M. Voorhees Transportation Center
Edward J. Bloustein School of Planning and Public Policy
Rutgers, the State University of New Jersey
33 Livingston Ave.
New Brunswick, NJ 08901
Tel: (848) 932-2859
Email: rnoland@rutgers.edu

Co author(s):
Yemi Adediji
Alan M. Voorhees Transportation Center
Edward J. Bloustein School of Planning and Public Policy
Rutgers, the State University of New Jersey
33 Livingston Ave.
New Brunswick, NJ 08901

Performing Organization:
Rutgers University

Sponsor(s):
University Transportation Research Center (UTRC)

To request a hard copy of our final reports, please send us an email at utrc@utrc2.org

Mailing Address:

University Transportation Reserch Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: 212-650-8051
Fax: 212-650-8374
Web: www.utrc2.org

Board of Directors

The UTRC Board of Directors consists of one or two members from each Consortium school (each school receives two votes regardless of the number of representatives on the board). The Center Director is an ex-officio member of the Board and The Center management team serves as staff to the Board.

City University of New York
Dr. Hongmian Gong - Geography/Hunter College
Dr. Neville A. Parker - Civil Engineering/CCNY

Clarkson University
Dr. Kerop D. Janoyan - Civil Engineering

Columbia University
Dr. Raimondo Betti - Civil Engineering
Dr. Elliott Sclar - Urban and Regional Planning

Cornell University
Dr. Huaizhu (Oliver) Gao - Civil Engineering

Hofstra University
Dr. Jean-Paul Rodrigue - Global Studies and Geography

Manhattan College
Dr. Anirban De - Civil & Environmental Engineering
Dr. Matthew Volovski - Civil & Environmental Engineering

New Jersey Institute of Technology
Dr. Steven I-Jy Chien - Civil Engineering
Dr. Joyoung Lee - Civil & Environmental Engineering

New York University
Dr. Mitchell L. Moss - Urban Policy and Planning
Dr. Rae Zimmerman - Planning and Public Administration

Polytechnic Institute of NYU
Dr. Kaan Ozbay - Civil Engineering
Dr. John C. Falcocchio - Civil Engineering
Dr. Elena Prassas - Civil Engineering

Rensselaer Polytechnic Institute
Dr. José Holguín-Veras - Civil Engineering
Dr. William "Al" Wallace - Systems Engineering

Rochester Institute of Technology
Dr. James Winebrake - Science, Technology and Society/Public Policy
Dr. J. Scott Hawker - Software Engineering

Rowan University
Dr. Yusuf Mehta - Civil Engineering
Dr. Beena Sukumaran - Civil Engineering

State University of New York
Michael M. Fancher - Nanoscience
Dr. Catherine T. Lawson - City & Regional Planning
Dr. Adel W. Sadek - Transportation Systems Engineering
Dr. Shmuel Yahalom - Economics

Stevens Institute of Technology
Dr. Sophia Hassiotis - Civil Engineering
Dr. Thomas H. Wakeman III - Civil Engineering

Syracuse University
Dr. Riyad S. Aboutaha - Civil Engineering
Dr. O. Sam Salem - Construction Engineering and Management

The College of New Jersey
Dr. Thomas M. Brennan Jr - Civil Engineering

University of Puerto Rico - Mayagüez
Dr. Ismael Pagán-Trinidad - Civil Engineering
Dr. Didier M. Valdés-Díaz - Civil Engineering

UTRC Consortium Universities

The following universities/colleges are members of the UTRC consortium.

City University of New York (CUNY)
Clarkson University (Clarkson)
Columbia University (Columbia)
Cornell University (Cornell)
Hofstra University (Hofstra)
Manhattan College (MC)
New Jersey Institute of Technology (NJIT)
New York Institute of Technology (NYIT)
New York University (NYU)
Rensselaer Polytechnic Institute (RPI)
Rochester Institute of Technology (RIT)
Rowan University (Rowan)
State University of New York (SUNY)
Stevens Institute of Technology (Stevens)
Syracuse University (SU)
The College of New Jersey (TCNJ)
University of Puerto Rico - Mayagüez (UPRM)

UTRC Key Staff

Dr. Camille Kamga: *Director, Assistant Professor of Civil Engineering*

Dr. Robert E. Paaswell: *Director Emeritus of UTRC and Distin*guished Professor of Civil Engineering, The City College of New York

Herbert Levinson: *UTRC Icon Mentor, Transportation Consultant and Professor Emeritus of Transportation*

Dr. Ellen Thorson: *Senior Research Fellow, University Transportation Research Center*

Penny Eickemeyer: *Associate Director for Research, UTRC*

Dr. Alison Conway: *Associate Director for Education*

Nadia Aslam: *Assistant Director for Technology Transfer*

Nathalie Martinez: *Research Associate/Budget Analyst*

Tierra Fisher*: Office Assistant*

Bahman Moghimi*: Research Assistant; Ph.D. Student, Transportation Program*

Wei Hao*: Research Fellow*

Andriy Blagay: *Graphic Intern*

TECHNICAL REPORT STANDARD
TITLE PAGE

1. Report No.	2.Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Omitted Variable Bias in Crash Reduction Factors		5. Report Date Sept. 2015 6. Performing Organization Code	
7. Author(s) Robert B. Noland and Yemi Adediji		8. Performing Organization Report No.	
9. Performing Organization Name and Address Alan M. Voorhees Transportation Center Edward J. Bloustein School of Planning and Public Policy Rutgers, the State University of New Jersey 33 Livingston Ave. New Brunswick, NJ 08901		10. Work Unit No. 11. Contract or Grant No. 49997-38-25	
12. Sponsoring Agency Name and Address University Transportation Research Center City College of New York Marshak Hall, 910 New York, NY 10031		13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code	
15. Supplementary Notes			

16. Abstract
Transportation planners and traffic engineers are increasingly turning to crash reduction factors to evaluate changes in road geometric and design features in order to reduce crashes. Crash reduction factors are typically estimated based on segmenting a highway and associating crashes with geometric features; this allows statistical methods to be applied to the data. Concurrently there is a stream of research that relies on spatial units of analysis to examine crashes; these typically use broad features of the road network combined with socio-economic and demographic factors that are associated with crashes. In this paper, we examine whether omission of these spatial factors in a link-based geometric model results in omitted variable bias. Our results suggest that there is no change in coefficient signs, but that there is a reduction in the magnitude of estimates. The sign of spatial variables, however, is quite different when combined into a link-based model. We also find substantial variability in coefficient estimates, and discuss the implications of these results for the use of crash reduction factors.

17. Key Words	18. Distribution Statement
traffic safety, crash reduction factors, spatial models, link-based models, specification error	

19. Security Classif (of this report)	20. Security Classif. (of this page)	21. No of Pages	22. Price
Unclassified	Unclassified	31 pages	

Form DOT F 1700.7 (8-69)

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The contents do not necessarily reflect the official views or policies of the UTRC[, (other project sponsors),] or the Federal Highway Administration. This report does not constitute a standard, specification or regulation. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government [and other project sponsors] assume[s] no liability for the contents or use thereof.

Contents

Figures

Tables

Abstract

Transportation planners and traffic engineers are increasingly turning to crash reduction factors to evaluate changes in road geometric and design features in order to reduce crashes. Crash reduction factors are typically estimated based on segmenting a highway and associating crashes with geometric features; this allows statistical methods to be applied to the data. Concurrently there is a stream of research that relies on spatial units of analysis to examine crashes; these typically use broad features of the road network combined with socio-economic and demographic factors that are associated with crashes. In this paper, we examine whether omission of these spatial factors in a link-based geometric model results in omitted variable bias. Our results suggest that there is no change in coefficient signs, but that there is a reduction in the magnitude of estimates. The sign of spatial variables, however, is quite different when combined into a link-based model. We also find substantial variability in coefficient estimates, and discuss the implications of these results for the use of crash reduction factors.

Introduction

Crash reduction factors (CRF) for the evaluation of safety interventions are being developed to provide guidance to highway engineers and planners. CRFs are developed using statistical models that link a variety of geometric design variables to police-reported crashes on road links (AASHTO, 2010). As such, the unit of analysis tends to be links or segments of a highway. This allows one to match the highway characteristics with the crashes that occur along the specific link in the road network. These models focus on the geometric design of the highway including features such as turning radius, road curvature, access points, lane widths, and number of lanes, among others (Abdel-Aty and Radwan, 2000; Caliendo et al., 2007; Malyshkina and Mannering, 2010; Milton and Mannering, 1998; Shankar et al., 1995). The output of these models is a parameter estimate associated with a specific design element which can then be used in a cost-benefit analysis to determine whether a design change should be implemented.

Another strand of research within the highway safety community has examined spatial determinants associated with road crashes (e.g., median income levels, population density, and other census-level data). These include studies of pedestrians in New Jersey (Noland et al., 2013), child and adult pedestrians in London (Graham et al., 2005), motor-vehicle crashes in England (Noland and Quddus, 2004) and Pennsylvania (Aguero-Valverde and Jovanis, 2006), and a number of studies conducted in Florida, including a spatial analysis of counties (Huang et al., 2010), analysis of pedestrians and cyclist crashes at a small spatial scale (Siddiqui et al., 2012) and analysis of the most desirable spatial unit to use (Abdel-Aty et al., 2013; Lee et al., 2014), suggesting that traffic-analysis zones (TAZ) while convenient to use may not provide the best model fit.

The issue of spatial autocorrelation is an issue that should be accounted for when modeling these effects (Aguero-Valverde and Jovanis, 2006). Conditional autoregressive models that control for spatial autocorrelation are used to derive crash estimates in (Quddus, 2008) and (Noland et al., 2013), the latter for pedestrian crashes. These studies will typically use various measures of the road network, such as network density or intersection density, but often will not include details such as lane width or number of lanes, (Noland, 2003; Noland and Oh, 2004) being two notable exceptions. These models will tend to lack the detailed context of the road geometry that a link-based model can provide.

One critique of CRF analysis and the link-based models on which these are based, is that the coefficient estimates are not transferable (Hauer et al., 2012) and often are highly variable (Elvik, 2015). The latter may be due to variation in the traffic environment in which they are measured. Put another way, there may be factors that affect safety outcomes that are not included in link-based models. The omission of variables from a statistical model can potentially bias results. This means that one may find statistically significant effects from factors that are not associated with the measured crash outcome, or non-significant effects for factors that are associated with crashes. One source of omission is the spatial context where the crashes occur. Combining the spatial context with a link-based analysis may shed light on whether this is an issue or not. (Mitra and Washington, 2012) identified issues associated with omitted variable bias when intersection-based models do not control for spatial attributes that may also affect crashes. The objective of this paper is to explore these issues using a link-based analysis that is

combined with a spatial analysis using data from New Jersey. This is done to examine whether there is omitted variable bias.

In the following sections, we first review some of the results from link-based studies, primarily to highlight the variation in results, recognizing that a difficulty with comparing results is the multitude of variables modeling approaches used. We then discuss our own data and the challenges of combining the link-based and spatial approaches. This is followed by our analysis, which includes a base spatial model and models estimated for five highways in New Jersey as well as a large database of local roads that covers the entire state. Results and discussion follows with implications for crash analysis and the use of crash reduction factors.

Link-based studies

The Poisson and Negative Binomial models have emerged as the standard in crash analysis as it best matches the distributional characteristics of crash data. Though the Poisson model is often the starting point due to its suitability for analyzing count data, the Negative Binomial model is often chosen because of the Poisson model's assumption of equi-dispersion, i.e., the mean and variance are equal. Real data usually violates this assumption, including most crash datasets.

The Negative Binomial model accounts for over-dispersion and thus provides a useful method for estimating crash models. Most of the studies we reviewed were estimated using the Negative Binomial model, however a handful employed other models, including (Chiou and Fu, 2013) who used a multinomial-generalized Poisson model with error components, and (Garnowski and Manner, 2011) who used a random parameter negative binomial model in addition to a fixed parameter model. Our review focuses on the variables included in each model and the results of the estimates, and we make no judgement about the appropriateness of various modeling techniques. Our review is by no means comprehensive, but we sought out studies that are repeatedly cited in the literature (except some that are more recent); these are listed in Table 1 with their corresponding Google Scholar citation rates.

Table 1. Google Scholar citations for papers reviewed.

	Google Scholar citations (Sept 25, 2015)	Citations per year
(Shankar et al., 1995)	490	24.5
(Milton and Mannering, 1998)	285	16.8
(Council and Stewart, 1999)	46	2.9
(Abdel-Aty and Radwan, 2000)	401	26.7
(Caliendo et al., 2007)	172	21.5
(Malyshkina and Mannering, 2010)	49	9.8

(Labi, 2011)	7	1.8
(Zeng and Huang, 2014)	4	4.0
(Bauer and Harwood, 2014)	2	2.0

In reviewing link-based studies, we found a wide variety of different variables included in the models presented. Different measures were sometimes used for the same variables, complicating efforts to compare coefficient values. Our comparison focuses on results for the following geometric design elements included in some of the studies: number of traffic lanes, lane width, average annual daily traffic (AADT), horizontal curvature, shoulder width, and median width. These are variables that we use in our analysis based on the availability of data in New Jersey. We examine each of the geometric variables in turn.

Horizontal curvature

One of the earlier studies is (Shankar et al., 1995). In this study, data for an interstate highway in Washington state was used and the focus was on horizontal curvature for different design speeds, while controlling for weather conditions. Lane and shoulder widths were virtually constant over the stretch of road analyzed given that interstate highways follow standard design guidelines. Estimates ranged from 0.046 (lower design speed) to 0.117 (higher design speed) measured using the number of horizontal curves at the two design speeds based on segmenting the interstate into ten sections. Further work in Washington state based on data for principal arterials measured curvature as the horizontal curve radius; thus a negative coefficient implies increased risk (Milton and Mannering, 1998); coefficient values for two different data sets varied between -0.0021 and -0.000221, and both were statistically significant.

An analysis of an arterial roadway in Florida also controlled for horizontal curvature, but used a different measure of "degrees/100 m arc" which is not precisely defined (Abdel-Aty and Radwan, 2000). Their coefficient estimate is positive and significant with a value of 0.124, which cannot be directly compared to the results of (Shankar et al., 1995). A study of a 4-lane motorway in Italy used a measure of km^{-1}, and found positive coefficients, for all crashes, of about 0.26 (Caliendo et al., 2007). Using data from Indiana, (Malyshkina and Mannering, 2010) used a measure of 18,000/($\pi \; x \; r$), with *r* (radius) defined in feet. The coefficient estimate is -0.0562, again difficult to compare with other studies except for the directional effect being the same.

Another study using data from Indiana focuses on rural two-lane roads (Labi, 2011). Horizontal curvature is based on "average horizontal curve radius", coefficients were estimated for different crash severity levels and for different functional road classes of rural two-lane roads; for fatal plus injury crashes, this varied from .0262 to 0.0580.

Returning to data from Washington state, (Bauer and Harwood, 2014) define horizontal curvature as 1/ln(2 x 5730/r). The coefficient estimated is statistically significant at a 95% level and is 0.19. Again, it is not clear how this can be compared with other studies. A variable was also included in the estimated model that assessed the interaction between horizontal curves and vertical grades, which was also

significant. This report, conducted for the US Federal Highway Administration, is of note partly because it was conducted to develop crash modification factors for the Highway Safety Manual and presents precise figures for both fatal/injury crashes and property-damage only crashes (this latter is surprising given the incomplete collection of data that is common in property-damage crash reporting).

Curvier roads are assumed to increase the probability of crashes, all else equal, and these results largely support that notion. However, curvier roads may also lead to reduced speeds if they are perceived as riskier (Noland, 2013). Only one study notes that the curviest stretches of roads seem to have fewer crashes and that curves tend to be riskiest following a long tangent section, i.e., a straight road leading into a curve (Milton and Mannering, 1998). While all these studies controlled for different variables (some of which we discuss below), none included spatial variables that might provide a better context for the driving population and the local area in which the crash occurred.

Shoulder width

Another commonly included geometric design variable is the shoulder width, larger shoulders are assumed to decrease the crash rate. This is presumably because a larger right-hand shoulder provides greater space for a driver to recover if a loss of control occurs. Some of the models we reviewed include parameters for shoulder width. Measurement is quite straightforward, assuming the segments correspond to constant shoulder widths.

Variation in the results is large. One study that analyzed two-lane rural roads in four different states had coefficient estimates ranging from -0.1230 to -0.4541 (Council and Stewart, 1999). Two of the models included surface width as an additional control variable; this might be correlated with shoulder width, but no mention is made of this possible confounding result. (Abdel-Aty and Radwan, 2000) in their model of Florida data estimates a coefficient of -0.12, similar to the low value in (Council and Stewart, 1999). The models estimated by (Labi, 2011) with Indiana data have estimates ranging from -0.0321 to -0.0943 for fatal and injury crashes, substantially smaller than the other estimates. Also using data from Indiana, (Malyshkina and Mannering, 2010) include a variable for interior shoulder widths and find a larger coefficient value of -1.25. All these studies used different data, models, and independent variables, thus it is not surprising to find such a wide variation in parameter estimates.

Lane width and number of lanes

Increasing the width of lanes and the number of lanes on a road is typically assumed to increase safety. Yet, theory and empirical work, much of it from spatial analysis of road safety, suggests the opposite (Dumbaugh and Rae, 2009; Noland, 2003; Noland, 2013). Three studies find that wider lanes reduce crashes with coefficient values of -0.42 (Council and Stewart, 1999), -0.364 (Abdel-Aty and Radwan, 2000), and -.09 (Labi, 2011). While these are all different road types, there is again substantial variation in coefficient estimates. An analysis conducted in Hillsborough County, Florida, included a variable for the number of lanes (Zeng and Huang, 2014) and found that more lanes increase the crash risk; coefficient values range from 0.137 to 0.167 depending on estimation method.

Traffic levels

Almost all models control for the level of traffic on each link, based on estimated annual average daily traffic (AADT) counts. Log transformations of AADT are often included in the model. At a fundamental level one would expect more traffic to lead to more crashes, although the relationship may vary in how it affects relative severity. Highly congested roads may suffer more crashes, but less severe crashes since vehicles are moving at slower speeds (Zhou and Sisiopiku, 1997). Estimated coefficient values range from 0.24 (Caliendo et al., 2007) to 1.18 (Bauer and Harwood, 2014), again showing substantial variation between estimates.

Summary of link-based analysis review

This review serves to show the variation in crash reduction factors that are estimated in the literature. For horizontal curvature this is partly due to different ways of measuring curvature, but for other simpler geometric design features this is due to the variation in the models. Our hypothesis is that much of this is due to the omission of variables that also affect crashes. While it may be naïve to think that one can estimate fully transferable crash reduction factors, the use of fixed deterministic values reported in the *Highway Safety Manual*, is equally naïve. In what follows we discuss our analysis of New Jersey safety data, including a spatial model, a link-based model, and then a combination of both. We also discuss some of the data problems we encountered that preclude a one-to-one comparison between all the models, but consider the results in that context.

Data

Crash data for New Jersey is available via Plan4Safety, maintained by the Center for Advanced Infrastructure and Transportation (CAIT) at Rutgers University (http://cait.rutgers.edu/tsrc/plan4safety). Most of the crash data is geo-coded and we downloaded data for all crashes from 2008 to 2012. This included information on the severity level of injuries, specifically fatalities, incapacitating injuries, and more minor injuries. In the analysis that follows we analyze three categories of crashes: total crashes, crashes with fatalities and incapacitating injuries, crashes with fatalities and all injuries.

For this data there were about 980,000 geocoded crash occurrences that were linked to road segments. The *Spatial Join* ArcGIS geo-process was used to assign crashes to road segments by X and Y coordinates. The result was that each segment in the NJ road network was assigned a value for number of crash occurrences and for the varying crash severity types. Approximately 46,206 road segments were assigned crash occurrences, out of a total of 104,811 segments in the NJDOT data.

Detailed information on the New Jersey road network was obtained from the New Jersey Department of Transportation (NJDOT) (New Jersey Department of Transportation, 2011; New Jersey Department of Transportation, 2013). This included the composition of the road network, including functional class, starting and ending mileposts for all road segments, AADT for selected road segments, and geometric design variables (segment length, lane counts, road width, median width and shoulder width).

Spatial data includes demographic data such as population density, employment density by employment location, median income, land-use, road density, and household vehicle ownership. Population data was

based on 2010 Census block group data, median income from 5-year (2007-2011) average American Community Survey (ACS) data, and for employment from the Census Longitudinal Employer-Household Dynamics (LEHD) database which was 2011 data. Land use data was obtained from the New Jersey Department of Environmental Protection (NJDEP) Land Use Land Cover data.

Our initial goal was to estimate a link-based model covering the entire state, as this could then be easily matched with a spatial model that covers the entire state. In processing the data, we linked the spatial data (based on block groups) to the road segments. In other words, the road segments within a block group would have the socio-economic variables of that block group attached as data attributes. This allows us to estimate three models: a link-based model with crash counts based on road segments, a spatial model with crash counts based on block groups, and a link-based model that includes the spatial attributes associated with the block group in which the road segment is located.

One constraint was the lack of AADT data on most road segments. While there are 104,811 road segments in the NJDOT data, linear estimates of AADT were only available for 1712 segments based on 5-year averages. Of these 1712 segments, only about one-tenth had exact segment matches with the segments in the road network shape file, and as such only a small portion of the entire network could be assigned AADT values using this linear AADT data. As an alternative, AADT data using 2010-2013 traffic counts were obtained from NJDOT, that could be linked using the latitudes and longitudes of these traffic count points. In total, 8336 of these points were processed into AADT values for 3863 road segments.

Given some of these data constraints, we estimated two different link-based models. In one we used all the road segments in the NJDOT shape file, in order to cover the entire state, to the best of our ability. For this analysis, the segmentation of the roads was based on how they were segmented in the NJDOT shape file (more details on the data are discussed below). The other link-based model was based on five highways that we selected (described below). Two-mile segments were created leading to a total of 587 road links, among these five highways. For both of these road link datasets, spatial variables were assigned based on the spatial attributes of block groups within 0.25 miles of each road link.

Segment length data was used to estimate VMT for each link (i.e., length x AADT). For the five model, we had AADT data for each road link. For the model covering the entire state AADT estimated based on a negative binomial regression using 680 of the 3863 data points (see model is shown in

Table 2. Predictor variables included geometric features and some socio-economic variables associated with the linked block group. Predicted AADT results were applied to those segments without data. This was done for only segments in functional class 7 (local roads) as the bulk of the data for geometric variables used in the regression were for segments in this class and other segments were largely lacking such data.

Table 2. Negative binomial regression of AADT

Dependent variable = AADT	coef.	t-value
Lane Count (ln)	-0.399	-0.38
Pavement Width (ln)	0.456	1.34
Number of 0 Vehicle Households (ln)	0.219	2.13
Median Income (ln)	0.520	2.34
Block Group Population Density (ln)	172.20	3.30
Employment Density (ln)	169.80	2.04
Sinuosity (ln)	1.926	2.43
Constant	-1.813	-0.54
Log Transformed Overdispersion	0.368	7.48
Observations	680	
Log likelihood	-5750	
Ll Constant Only	-5862	
LR Chi2	89.38	
Pseudo_R2	0.019	

We calculate horizontal curvature slightly differently than other studies, as the NJDOT data did not have turning radius. We use sinuosity, which is a measure of the deviation of a line from its shortest path and is the actual length of a segment, divided by the shortest (straight line) distance between its start and end points. This provides a differentiation between straight versus more sinuous segments, measured on a scale of 0 to 1, 0 indicating very sinuous (curvy) segments and 1 indicating perfectly straight segments. Sinuosity was measured using an ArcMap geo-processing tool (ArcGIS, 2015).

The remaining geometric design attributes include lane count, roadway width, median width, and shoulder width. These were obtained from NJDOT as tables, each with between 95,000 and 110,000 segments, for which Standard Route Identifiers (SRI) as well as the starting and ending mile posts were indicated to denote the various segments of each roadway. These were matched both with SRI and the starting and ending mile posts leading to 87,518 initial segments with lane count, pavement width, median width, median type, and shoulder width data. These successful joins did not include many of the higher functional class roads (namely class 1, 2 and 3) as we could not determine exact matches. We matched some of these segments using SRI, but not the same starting and ending mileposts. This required us to average some of the geometric design features across the mismatched segments. This worked well for the higher functional class roads due to the consistency of their geometric features across segments of roadways in these classes. This interpolation process provided us with geometric data on 87,704 segments.

Our spatial variables include total population, population density, employment density by work destination, median income, percent below poverty, percent of residential, commercial and industrial land uses, vehicle ownership by household and road density by functional class. There are 6320 block groups in New Jersey and data was processed at that level.

All data sources are listed in Table 3.

Table 3. Data sources used

CRASH DATA	http://cait.rutgers.edu/tsrc/plan4safety
ROAD DATA	
NJ Road Network Shapefile	http://www.state.nj.us/transportation/gis/data.shtm
AADT	http://www.state.nj.us/transportation/refdata/roadway/traffic.shtm
ROAD GEOMETRY DATA	
Lane count, pavement width, median width, shoulder width	(New Jersey Department of Transportation, 2011)
Sinuosity [Computation]	*Calculate Sinuosity* geoprocessing tool
SPATIAL DATA	
Block Group shapefile	https://www.census.gov/geo/maps-data/data/tiger-data.html
Geographic Data (block group area)	https://www.census.gov/geo/maps-data/data/tiger-data.html
Demographic data (age, population, income, poverty status, vehicle ownership)	https://www.census.gov/geo/maps-data/data/tiger-data.html http://www.census.gov/acs/www/data/data-tables-and-tools/index.php
Land Use Data (Residential, Commercial, Industrial)	http://www.nj.gov/dep/gis/lulc07cshp.html
Employment by employment destination	http://lehd.ces.census.gov/data/
Road Density [Computation]	http://www.state.nj.us/transportation/gis/data.shtm https://www.census.gov/geo/maps-data/data/tiger-data.html

Methods

As crash data is based on counts of crashes, whether for a spatial unit or a road link, count regression models are used to estimate models. Common practice is to use negative binomial regression models that account for overdispersion in the data. This is a generalization of the Poisson regression that assumes equidispersion, that is that the mean of the estimate is equal to the standard deviation, a condition that typically fails with empirical data. These models are estimated using maximum likelihood estimation.

Spatial data tends to also be spatially auto-correlated. Put simply, this means that the conditions in neighboring block groups likely also have an effect on the incidence of crashes in a given block group. The same can be said for road links; that is neighboring links may be spatially correlated with each other. To account for this, we also estimate negative binomial conditional autoregressive models (Levine et al., 2010) using Crimestat. These models are estimated using Markov Chain Monte Carlo estimation which is a Bayesian estimation technique.

The interpretation of Bayesian estimates differs from that of frequentist approaches (i.e., maximum likelihood estimation). In a frequentist estimate, the confidence interval represents the sampling error; that is, a 95% confidence interval implies that if a population is sampled 95 out of 100 times, then the estimates will fall within that interval. A Bayesian analysis, on the other hand, results in a credible interval. A 95% credible interval means that there is a 95% probability that the correct estimate is within

the specified range. This provides a powerful method for showing the range of coefficient estimates rather than simply a fixed coefficient estimate.

An additional issue with our Bayesian estimates is that Crimestat uses a block sampling method with large datasets (Levine et al., 2013). This is done to increase the speed of the estimates and does not affect their efficiency. The mean standard deviations, however, have substantial variation, thus the standard errors are inflated. Crimestat calculates an adjusted standard error to correct for this and we report these (for our link-based model covering the entire state). Of consequence, in some cases our 95% credible intervals may span zero; that is the estimate with an unadjusted standard error was not significant at the 95% level, but the adjustment implies that it is significant.[1]

The MCMC method is a stochastic process. What that means is that the estimation relies on a random process to converge to a stable estimate. For most of our estimates we ran 100,000 iterations with a burn-in sample of 20,000. In some cases this was increased to 300,000 iterations with a 40,000 burn-in sample. To determine whether the estimates reached convergence we examined the Gelman-Rubin (G-R) statistic.

For the full geocoded database, there were 981,483 total crashes over the 5 years of our data; 2318 of these were fatal crashes and another 8303 were incapacitating injury crashes. A total of 255,086 were less severe injury crashes (not counting the fatal and incapacitating injuries). We estimate three sets of models, each with various dependent variables (total crashes, total fatal and incapacitating injury crashes, total fatal and injury crashes).

The first model we estimate is a spatial model for the entire state. One benefit of a spatial model is that it includes the crash information across every block group in the state (of which there are 6320). While a large selection of socio-economic and demographic factors can be controlled for, the one shortcoming is the lack of VMT data at the block-group level. However, total population tends to be highly correlated with VMT and can be used in its place as an exposure variable (Noland, 2003). However, we note this one limitation of spatial analysis. Also, we used only those block groups that had roads passing through. This was important in order to include the road density variables. We estimated this model using a total of 6293 block groups.

The second set of models that are estimated are two different link-based models. We originally intended to include every road in the state, but this proved infeasible due to data limitations with some higher functional class roads. Since roads in functional class 7 (local roads) had the most robust data, we opted to use them for one of our link based models. Of a total of 94,109 links, 86,394 had data for the variables we wanted to model. For these links there are a total of 224,900 crashes, 51,450 crashes with fatalities or any injuries, and 3377 crashes with fatalities or incapacitating injuries.

[1] Crimestat uses a default of 6000 observations for block sampling. In our models using just spatial variables our number of observations is 6293; in this case we overrode the default and ran the models without block sampling (these took about one day to run). In our model with all the local road links our number of observations is 86,394 and thus we ran these with block sampling.

The final model is a more traditional link-based analysis based on five roads within functional classes 1 through 4 in New Jersey. The five roads are the Garden State Parkway (FC 4), I-80 (FC 1), I-95 (FC 4), US-46 (FC 2), and US-202 (FC 2) and these are displayed in Figure 1. There are a total of 587 links across all five roads (with a length of 471.4 miles), with 97,884 total crashes, 22,304 crashes with fatalities or injuries, and 591 crashes with fatalities or incapacitating injuries.

Summary statistics for each set of models are listed in Table 4.

Figure 1. Highways used for link-based analysis

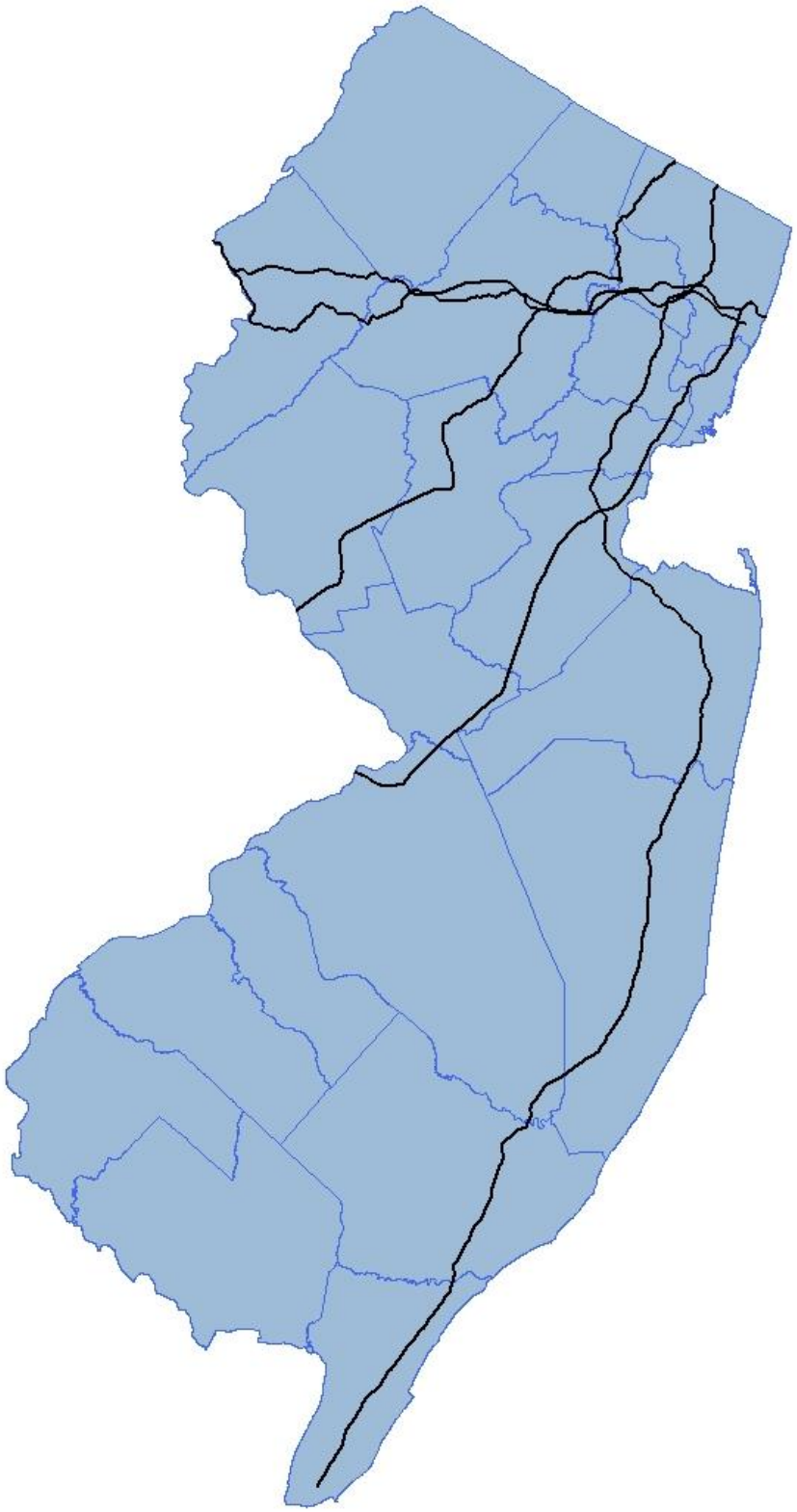


Table 4. Summary statistics

Spatial models based on block groups	N	Mean	Std. Dev.	Min	Max
Independent Variables					
Population (ln)	6293	7.14	0.51	0.00	9.07
Population Density (ln, /sq meter)	6293	0.004	0.005	0.000	0.062
Employment Density (ln)	6293	0.001	0.003	0.000	0.082
Median Income (ln)	6293	11.11	0.84	0.00	12.43
Percent Residential (ln)	6293	3.79	0.77	0.00	4.62
Percent Commercial (ln)	6293	1.87	1.10	0.00	4.60
Percent Industrial (ln)	6293	0.52	0.96	0.00	4.25
Percent Principal Arterial (ln)	6293	0.01	0.02	0.00	0.39
Percent Minor Collector (ln)	6293	0.03	0.04	0.00	0.45
Percent Local Road (ln)	6293	0.22	0.13	0.00	0.67
Dependent Variables					
Crashes	6293	155.96	199.52	0.00	3570.00
Crashes with Fatalities or Injuries	6293	40.90	51.33	0.00	853.00
Crashes with Fatalities or Incapacitating Injuries	6293	1.69	2.31	0.00	26.00
Link-based model for five highways					
Independent Variables					
Sinuosity (ln)	587	0.68	0.03	0.30	0.69
Lane Count (ln)	587	1.29	0.22	0.69	1.95
Shoulder Width (ln)	587	2.18	0.70	0.00	3.04
Lane Width (ln)	587	2.25	0.25	0.00	3.26
Vehicle Miles Traveled (ln)	587	10.48	0.99	8.10	12.36
Population Density (ln, /sq meter)	587	0.001	0.001	0.000	0.010
Employment Density (ln)	587	0.001	0.001	0.000	0.005
Median Income (ln)	587	11.28	0.75	0.00	12.14
Dependent Variables					
Crashes	587	166.75	162.79	0.00	1477.00
Crashes with Fatalities or Injuries	587	38.00	34.61	0.00	236.00
Crashes with Fatalities or Incapacitating Injuries	587	1.01	1.17	0.00	7.00
Link-based model for all local roads					
Independent Variables					
Sinuosity (ln)	86394	0.64	0.11	0.00	0.69
Lane Count (ln)	86394	1.09	0.05	0.69	1.95
Lane Width (ln)	86394	2.65	0.18	0.00	3.89
Vehicle Miles Traveled (ln)	86394	5.72	1.01	0.08	12.75
Population Density (ln, /sq meter)	86394	0.002	0.002	0.000	0.024
Employment Density (ln)	86394	0.001	0.001	0.000	0.029
Median Income (ln)	86394	11.35	0.42	0.00	12.43
Dependent Variables					
Crashes	86394	2.60	10.77	0.00	665.00
Crashes with Fatalities or Injuries	86394	0.60	3.19	0.00	210.00
Crashes with Fatalities or Incapacitating Injuries	86394	0.04	0.25	0.00	12.00

Results

Results are presented for models with spatial variables, link-based variables, and then with our effort to combine these. Both maximum likelihood and MCMC results are presented and discussed. Our focus is on the variation in coefficient estimates and the implications for developing crash reduction factors and policy. We show the problems with omitting key variables in link-based models.

Models with just spatial variables

The first set of models include only the spatial variables and are based on Census block groups. We estimated these as negative binomial regressions with both Maximum Likelihood Estimation (MLE) and as Bayesian MCMC models controlling for spatial autocorrelation. Dependent variables included total crashes, crashes with fatal and incapacitating injuries, and crashes with fatal and all reported injuries.

Results are shown in Table 5 and Table 6 (models are labeled consecutively). Each model is based on 6,293 block groups. For Model 1, all explanatory variables show statistically significant effects at the 5% alpha level, except employment density, and percent local roads. For Model 3, all explanatory variables show significant effects except median income and percent industrial land use. For Model 2, all the variables except percent local roads have significant effects. The direction of effects were consistent for all three models with population, percent commercial and industrial land uses, and percent principal arterial, minor collector and local roads being positively associated with the respective dependent variable. Population density, employment density, median income, and percent residential land use had negative effects.

As these models do not control for spatial correlation, results for our conditional autoregressive model are shown in Table 6. As these are Bayesian estimates they have the added benefit of calculating credible intervals, as discussed above. In most of the estimates we see that the 95% credible interval includes the MLE estimated coefficient. But these intervals are quite large. For example, looking at how the percent of principal arterials in a block group is associated with total crashes, the estimate has a 95% probability of falling between 5.559 and 7.296. While positive and implying that more principal arterials can cause more crashes, if used as a crash reduction factor or for cost/benefit analysis, the mean estimate (6.422) could lead to improper conclusions. Another example is the estimated coefficient for local roads which is statistically significant only in model 6. However, the credible interval for this coefficient (in models 4 and 5), ranges between a negative and a positive value. Thus, while the mean coefficient value is positive, there is a probability that it may actually be negative, which clearly would imply some difficulty in using this estimate for policy. The same holds true for the employment density variable in model 4.

Table 5. Negative binomial maximum likelihood models with just spatial variables

Dependent variables	Total crashes (model 1)		Total fatal and injury crashes (model 3)		Total fatal and incapacitating injury crashes (model 2)	
	coef.	t-value	coef.	t-value	coef.	t-value
Population (ln)	0.546	15.12	0.616	17.35	0.693	10.04
Population Density (ln)	-40.210	-12.62	-50.010	-14.46	-58.130	-10.16
Employment Density (ln)	-4.339	-1.39	-9.690	-3.03	-18.390	-3.00
Median Income (ln)	-0.100	-2.81	-0.142	-3.83	-0.081	-1.21
% Residential (ln)	-0.534	-20.23	-0.548	-21.08	-0.609	-18.74
% Commercial (ln)	0.271	21.85	0.270	20.61	0.157	8.53
% Industrial (ln)	0.071	5.56	0.062	5.03	0.016	1.08
% Principal Arterial (ln)	6.400	12.67	6.702	13.46	3.777	7.97
% Minor Collector (ln)	2.488	9.61	2.788	10.14	1.821	4.65
% Local Roads (ln)	0.041	0.24	0.213	1.23	0.875	3.47
Constant	3.531	6.43	2.206	3.99	-1.758	-1.56
ln overdispersion	-0.527	-23.11	-0.454	-20.92	-0.565	-10.11
Observations	6293		6293		6293	
Log likelihood	-36341		-28068		-10327	
Ll Constant Only	-38083		-29723		-11163	
LR Chi2	3067		2926		1435	
Pseudo_R2	0.0458		0.0557		0.0749	

Table 6. Negative binomial conditional autoregressive Bayesian models with just spatial variables

	Total crashes (model 4)				Total fatal and injury crashes (model 5)				Total fatal and incapacitating injury crashes (model 6)			
Variables	mean	t-value	2.5th	97.5th	mean	t-value	2.5th	97.5th	mean	t-value	2.5th	97.5th
Population (ln)	0.542	27.40	0.505	0.581	0.617	29.44	0.576	0.658	-1.784	-6.73	-2.300	-1.269
Population Density (ln)	-40.128	-14.76	-45.444	-34.771	-50.175	-16.58	-56.060	-44.179	0.690	23.94	0.635	0.747
Employment Density (ln)	-4.079	-1.14	-10.897	3.109	-9.363	-2.56	-16.356	-1.997	-57.977	-11.66	-67.917	-48.405
Median Income (ln)	-0.095	-6.76	-0.124	-0.068	-0.144	-9.05	-0.175	-0.113	-18.354	-3.23	-29.608	-7.315
% Residential (ln)	-0.531	-24.15	-0.575	-0.489	-0.544	-23.35	-0.590	-0.498	-0.078	-3.60	-0.119	-0.035
% Commercial (ln)	0.272	26.10	0.251	0.292	0.270	24.29	0.248	0.292	-0.605	-21.25	-0.659	-0.548
% Industrial (ln)	0.071	6.17	0.049	0.094	0.063	5.24	0.040	0.087	0.158	9.72	0.126	0.190
% Principal Arterial (ln)	6.422	14.44	5.559	7.296	6.716	14.39	5.810	7.648	0.017	1.11	-0.013	0.047
% Minor Collector (ln)	2.553	9.67	2.034	3.068	2.814	9.99	2.263	3.371	3.782	6.88	2.711	4.864
% Local Roads (ln)	0.091	0.67	-0.172	0.361	0.222	1.55	-0.059	0.501	1.807	4.48	1.022	2.598
Intercept:	3.479	28.45	3.227	3.721	2.195	16.39	1.936	2.462	-1.784	-6.73	-2.300	-1.269
Spatial correlation (phi)	-0.004	-1.79	-0.009	0.000	-0.002	-1.00	-0.006	0.001	0.000	0.14	-0.001	0.002
N	6293				6293				6293			
Df	6280				6280				6280			
Iterations	100000				100000				100000			
Burn-in	20000				20000				20000			
LL	-36344.8				-28068.3				-10326.7			

Link-based models for five highways

We first estimate a link-based model for the five highways previously mentioned. This is the method that is typically used in the development of crash reduction factors. We estimate using both MLE and MCMC. Next we add in various spatial variables that have been linked to each road segment. These models are shown in Table 7, Table 8, and Table 9.

Examining the results in Table 7 for models 7, 8, and 9, without the spatial variables, we see that coefficient estimates are statistically significant at the 95% confidence level, except for two variables in the fatal and incapacitating injury crash model (sinuosity and lane width). Results for total crashes and total fatal and injury crashes are broadly similar in the link-based models without spatial variables (model 7 and 8). In general, these results suggest that straighter roads increase crashes (sinuosity ranges from 0, very curvy, to 1, a straight road segment), more lanes increases crashes, wider shoulders reduce crashes, wider lane widths reduce crashes, and increased VMT is associated with more crashes. Other than the result on sinuosity, these results are in general agreement with much of the traffic safety literature.

We then add our spatial variables, which we know are generally associated with crashes as previously shown in Table 5 and Table 6. Based on these prior results we expect increased population and employment density to be associated with fewer crashes, and higher median incomes to also be associated with fewer crashes (with some minor variation between each model). We see that these results are quite different when linked to our road links. Population density now has a positive association with more crashes, as does employment density in model 10 and 11. Employment density is negative in model 12 for fatal and injury crashes. Median income is no longer statistically significant.

The effect on the road geometry variables of adding spatial controls is also notable, but less so. All the variables maintain the same directional effect, but almost all the coefficient values have a lower value. The implication is that geometric variables have less effect on crashes when spatial controls are added to the model. While this is indicative of omitted variable bias, the good news is that directional effects are maintained. Model fit, as measured by pseudo R^2 is higher in the models that include spatial controls.

Turning next to our Bayesian models for the five highway road link data (Table 8 and Table 9), we see a broadly similar pattern of results. Directional effects are similar for the road geometry variables. The spatial variables introduced into the model are also similar in direction to results shown in the MLE model (Table 7). The introduction of the spatial variables reduces the magnitude of the road geometry coefficients, but not by that much and in some cases the coefficient value is a bit higher. Again, while there may be some omitted variable bias, it is minor and does not distort the directional effects.

The credible intervals, however, have quite large ranges. For example, in model 17 (Table 9) we can see that the interval for sinuosity ranges from -0.100 up to -3.819, which obviously would have an impact on any consideration of how curvature affects safety. In the same model we can also see that the lane width coefficient ranges from -0.085 to -0.640, so increasing lane widths may have either virtually no safety effect or have a relatively substantive effect. It is possible that assuming a constant effect without

considering that there is a non-linear process, i.e., going from 9 ft lanes to 10 ft lanes may improve safety, while moving from 11 ft to 12 ft lanes may not. If anything, this only highlights the difficulty of model estimation to estimate the effect of geometric changes.

Table 7. Link-based negative binomial maximum likelihood models for five highways

	Total crashes (model 7)		Total fatal and injury crashes (model 8)		Total fatal and incapacitating injury crashes (model 9)		Total crashes (model 10)		Total fatal and injury crashes (model 11)		Total fatal and incapacitating injury crashes (model 12)	
Variables	coef.	t-value	coef.	t-value	coef.	t-value	coef.	t-value	coef.	t-value	coef.	t-value
Sinuosity (ln)	-3.804	-2.77	-3.286	-2.55	-1.906	-1.21	-2.279	-1.52	-2.121	-1.54	-1.853	-1.84
Lane Count (ln)	0.586	3.99	0.678	4.64	0.601	2.55	0.415	2.80	0.464	3.20	0.574	2.33
Shoulder Width (ln)	-0.440	-6.98	-0.427	-6.86	-0.199	-2.04	-0.287	-4.55	-0.294	-4.62	-0.203	-1.86
Lane Width (ln)	-0.491	-2.99	-0.504	-3.00	-0.375	-1.45	-0.343	-1.91	-0.375	-2.06	-0.430	-1.71
Vehicle Miles Traveled (ln)	0.797	20.28	0.720	18.05	0.376	5.52	0.631	12.52	0.565	11.65	0.369	4.38
Population Density (ln)							122.600	5.22	153.800	7.09	94.470	2.73
Employment Density (ln)							279.000	4.61	181.000	3.04	-174.500	-1.92
Median Income (ln)							-0.030	-1.12	-0.032	-1.28	0.011	0.42
Constant	0.429	0.43	-0.685	-0.72	-2.207	-1.79	0.718	0.65	-0.076	-0.07	-2.139	-2.10
Log Overdispersion	-0.947	-16.61	-0.998	-15.84	-1.652	-4.42	-1.130	-13.93	-1.210	-13.22	-1.778	-4.29
Observations	587		587		587		587		587		587	
Log likelihood	-3348		-2507		-767.5		-3291		-2451		-762.9	
Ll Constant Only	-3582		-2715		-801.1		-3582		-2715		-801.1	
LR Chi2	468.6		416.8		67.03		637.8		624.7		81.63	
Pseudo_R2	0.065		0.0767		0.0418		0.0813		0.0973		0.0476	

Table 8. Link-based negative binomial conditional autoregressive Bayesian models for five highways

	Total crashes (model 13)				Total fatal and injury crashes (model 14)				Total fatal and incapacitating injury crashes (model 15)			
Variables	mean	t-value	2.5th	97.5th	mean	t-value	2.5th	97.5th	mean	t-value	2.5th	97.5th
Sinuosity (ln)	-2.370	-2.03	-4.734	-0.319	-1.870	-1.99	-3.656	-0.050	-2.010	-1.59	-4.438	0.382
Lane Count (ln)	0.570	3.82	0.262	0.848	0.580	3.91	0.291	0.870	0.620	2.58	0.145	1.086
Shoulder Width (ln)	-0.360	-5.58	-0.486	-0.231	-0.350	-5.89	-0.474	-0.234	-0.200	-2.16	-0.389	-0.020
Lane Width (ln)	-0.370	-2.12	-0.725	-0.048	-0.360	-2.17	-0.692	-0.038	-0.400	-1.66	-0.873	0.073
Vehicle Miles Traveled (ln)	0.760	17.26	0.669	0.850	0.700	15.73	0.607	0.781	0.370	5.03	0.240	0.542
Constant	-0.700	-0.78	-2.416	1.236	-1.840	-2.50	-3.303	-0.293	-2.080	-1.94	-4.174	-0.107
N	587				587				587			
Df	579				579				579			
Iterations	100000				300000				300000			
Burn-in	20000				40000				40000			
LL	-3582.37				-2958.08				-769.66			

Table 9. Link-based negative binomial conditional autoregressive Bayesian models with spatially linked variables for five highways

	Total crashes (model 16)				Total fatal and injury crashes (model 17)				Total fatal and injury incapacitating crashes (model 18)			
Variables	mean	t-value	2.5th	97.5th	mean	t-value	2.5th	97.5th	mean	t-value	2.5th	97.5th
Sinuosity (ln)	-1.740	-1.72	-4.150	-0.016	-1.920	-2.00	-3.819	-0.100	-1.850	-0.96	-5.328	1.796
Lane Count (ln)	0.520	3.76	0.249	0.789	0.500	3.67	0.229	0.766	0.590	2.52	0.138	1.057
Shoulder Width (ln)	-0.270	-4.89	-0.383	-0.167	-0.290	-5.46	-0.398	-0.189	-0.190	-1.87	-0.400	0.006
Lane Width (ln)	-0.330	-2.15	-0.646	-0.043	-0.360	-2.50	-0.640	-0.085	-0.420	-1.52	-0.975	0.096
Vehicle Miles Traveled (ln)	0.640	16.34	0.552	0.708	0.580	14.69	0.498	0.653	0.360	4.89	0.227	0.525
Population Density (ln)	115.560	5.41	74.305	157.945	152.330	7.11	110.809	194.816	96.480	3.01	33.194	158.515
Employment Density (ln)	236.950	4.62	136.981	338.035	164.170	3.21	64.360	264.732	-170.710	-1.89	-350.622	2.255
Median Income (ln)	-0.030	-0.78	-0.094	0.036	-0.030	-1.01	-0.096	0.028	0.020	0.35	-0.084	0.135
Constant	0.060	0.07	-1.469	1.847	-0.420	-0.49	-2.090	1.351	-2.280	-1.56	-5.092	0.254
N	587				587				587			
Df	576				576				576			
Iterations	100000				300000				300000			
Burn-in	20000				40000				40000			
LL	-3331.99				-2459.78				-765.04			

Link-based analysis with full spatial coverage

Our original intent was to develop a link-based model that covered every road and highway of New Jersey. As noted above, we encountered major data limitations in our attempt to do alternative, we developed a dataset that includes almost a full set of geometric variables for 7 local roads. AADT data was still largely unavailable but we estimated AADT using the data available. This is obviously a major limitation, but in the end we have a model with 86,394 This in itself created additional problems of successful convergence using MCMC for our analysis. We were only able to successfully estimate models for total crashes. Results using MCMC are shown in Table 10 and

Table 11, respectively. Shoulder width is not included in these models as local roads generally have no shoulder.

Examining the MLE model (Table 10) we find that effects are quite different than in our previous analysis. Sinuosity has a positive and significant effect as does lane width. While these different effects may be because local roads have different travel characteristics than the links examined for the five highways, it is also possible that our estimates of AADT and VMT are affecting the estimates. VMT is highly significant with a coefficient value above one.

Spatial variables, shown in model 20, on the other hand, have a similar effect as in previous models. Both population and employment density are associated with more crashes, while higher median income reduces crashes. But as can be seen, introduction of the spatial variables, while not changing the sign of the geometric variables does affect the coefficient values, in some cases substantially. For example, sinuosity drops to 0.182 from 0.728.

The results for the MCMC estimation, in Table 11, show broadly the same pattern, however, coefficient values are quite different and have a large range within the credible interval. Of note, is that sinuosity loses statistical significance when spatial variables are included while lane count and lane width both have a credible interval that spans zero. The same is true of population and employment density, although the latter is not statistically significant. So these models show both different effects when the spatial variables are introduced, but also credible intervals that suggest ambiguous effects associated with all the variables except VMT and median income.

Table 10. Link-based negative binomial maximum likelihood models for all local road links

Variables	Total crashes (model 19) coef.	t-value	Total crashes (model 20) coef.	t-value
Sinuosity (ln)	0.728	13.33	0.182	2.32
Lane Count (ln)	0.503	5.44	0.897	3.77
Lane Width (ln)	0.603	16.30	0.572	9.11
Vehicle Miles Traveled (ln)	1.239	194.20	1.171	97.58
Block Group Population Density (ln)			47.540	8.53
Employment Density (ln)			47.720	3.62
Median Income (ln)			-0.739	-32.84
Constant	-9.727	-56.23	-1.107	-2.37
Log Overdispersion	0.653	73.04	0.563	34.12
Observations	86,394		86,394	
Log likelihood	-128005		-125669	
Ll Constant Only	-147993		-147993	
LR Chi2	39978		19479	
Pseudo_R2	0.135		0.151	

Table 11. Link-based negative binomial conditional autoregressive Bayesian models for all local road links

Variables	Total crashes (model 21) mean	adj. t-value	2.5th	97.5th	Total crashes (model 22) mean	adj. t-value	2.5th	97.5th
Sinuosity (ln)	0.739	12.42	-0.976	2.435	-0.072	-1.20	-1.805	1.639
Lane Count (ln)	-0.057	-0.36	-4.611	4.388	0.646	4.90	-3.181	4.385
Lane Width (ln)	0.253	6.00	-0.944	1.481	0.449	11.57	-0.661	1.569
Vehicle Miles Traveled (ln)	1.405	195.45	1.205	1.617	1.301	172.58	1.087	1.525
Block Group Population Density (ln)					24.355	6.24	-85.850	140.385
Employment Density (ln)					8.512	1.04	-215.984	257.848
Median Income (ln)					-0.836	-48.15	-1.342	-0.331
Constant	-9.387	-38.10	-16.585	-2.394	-0.124	-0.43	-8.478	8.280
N	86,394				86,394			
Df	86387				86384			
Iterations	200000				200000			
Burn-in	40000				40000			
LL	-130811.70				-127082.89			

Discussion and Conclusions

As noted previously, only one study has examined omitted variable bias in safety research (Mitra and Washington, 2012). Their analysis focused on intersection crashes in the city of Tucson, Arizona. Some 291 signalized intersections were included in the database with crashes occurring within 250 ft being attributed to the intersection location. Traffic variables that are typically included in this type of model are average daily traffic, turning movements, and infrastructure features (width of medians, posted speeds, right/left turn lanes). These were supplemented with spatial variables, specifically distance to schools of various types, number of drinking establishments within a specified distance, and total nearby population, by age. They conclude that when these variables are omitted inaccurate estimates of safety effects occurs, and this can lead to false conclusions on correlates of traffic crashes.

Our analysis provides less evidence that the omission of variables is a major issue, as in most cases directional effects are consistent with and without spatial variables being included in a link-based analysis. If anything, the spatial variables prove to be more problematic when included in a link-based analysis, for example, instead of a negative effect associated with population density, we find a positive association.

Of more interest is the wide range of the credible intervals found in our analysis. Recall that a credible interval in a Bayesian analysis represents the probability (at say 95%) that the real coefficient estimate falls within the range of the interval. This suggests at a minimum that any use of crash reduction factors should evaluate the low and high end of the range and what the impact will be on road safety. In some cases, this may even span zero, suggesting both a negative and positive effect associated with the intervention. More and better data may lead to more refined estimates, however, this means that models must be fully specified with theoretically sound relationships. Data will undoubtedly still be an issue as we found with our models.

Data is the major limitation of this work. While we sought to examine the effect of omitted variables on crash reduction factors, our models may also omit key variables due to missing data. The safety of the road network is largely determined by driver behavior and various other policies (such as vehicle regulations, drunk driving laws, and safety-belt laws). These can have large effects on safety (Noland, 2003). None of these factors are typically taken into account in this type of modeling, and would largely be dependent on how the drivers respond to these policies as well as the vehicle mix. One of the objectives of a pure spatial analysis is to control for demographic proxies for some of these variables, but even this is probably insufficient.

Our main conclusion is that those working to improve the safety of our highways should be cautious in the use of deterministic crash reduction factors. Theoretical understanding of how changes to geometric design will affect safety is useful knowledge but does not require the development of crash reduction factors (Noland, 2013). For example, highway design practice assumed that “faster, straighter, and wider” was safer, mainly because controlled access freeways follow this design practice. This ignores how applying these criteria to an urban road may actually have a negative safety outcome (Dumbaugh

and Gattis, 2005). Putting thought into design practices is more effective than blindly applying crash reduction factors.

Acknowledgements

This research was funded by the U.S. Department of Transportation through the region 2 University Transportation Research Center. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This paper is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

References

AASHTO, 2010. Highway Safety Manual, .

Abdel-Aty, M.A., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement, Accident Analysis & Prevention 32 (5), 633-642.

Abdel-Aty, M., Lee, J., Siddiqui, C., 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., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania, Accident Analysis and Prevention 38 (3), 618-625.

ArcGIS, 2015. Calculate Sinuosity, 2015 (July).

Bauer, K.M., Harwood, D.W., 2014. Safety effects of horizontal curve and grade combinations on rural two-lane highways, .

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

Chiou, Y., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components, Accident Analysis & Prevention 50, 73-82.

Council, F., Stewart, J., 1999. Safety effects of the conversion of rural two-lane to four-lane roadways based on cross-sectional models, Transportation Research Record: Journal of the Transportation Research Board (1665), 35-43.

Dumbaugh, E., Rae, R., 2009. Safe urban form: Revisiting the relationship between community design and traffic safety, Journal of the American Planning Association 75 (3), 309-329.

Dumbaugh, E., Gattis, J., 2005. Safe streets, livable streets, Journal of the American Planning Association 71 (3), 283-300.

Elvik, R., 2015. Methodological guidelines for developing accident modification functions, Accident Analysis & Prevention 80, 26-36.

Garnowski, M., Manner, H., 2011. On factors related to car accidents on German Autobahn connectors, Accident Analysis & Prevention 43 (5), 1864-1871.

Graham, D., Glaister, S., Anderson, R., 2005. The effects of area deprivation on the incidence of child and adult pedestrian casualties in England, Accident Analysis and Prevention 37 (1), 125-135.

Hauer, E., Bonneson, J., Council, F., Srinivasan, R., Zegeer, C., 2012. Crash modification factors: foundational issues, Transportation Research Record: Journal of the Transportation Research Board (2279), 67-74.

Huang, H., Abdel-Aty, M., Darwiche, A., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling, Transportation Research Record: Journal of the Transportation Research Board (2148), 27-37.

Labi, S., 2011. Efficacies of roadway safety improvements across functional subclasses of rural two-lane highways, J Saf Res 42 (4), 231-239.

Lee, J., Abdel-Aty, M., Jiang, X., 2014. Development of zone system for macro-level traffic safety analysis, J Transp Geogr 38, 13-21.

Levine, N., Lord, D., Park, B., 2010. Crimestat version 3.3 Update Notes: Part 2: Regression Modeling, .

Levine, N., Lord, D., Park, B., Geedipally, S., Teng, H., Sheng, L., Cahill, I., 2013. The CrimeStat Regression Module, Chapter 20, 215 (July 30).

Malyshkina, N.V., Mannering, F.L., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents, Accident Analysis & Prevention 42 (1), 131-139.

Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies, Transportation 25 (4), 395-413.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling, Accident Analysis & Prevention 49, 439-448.

New Jersey Department of Transportation, 2013.
NJ roadway network, .

New Jersey Department of Transportation, 2011.
Straight line diagram database, .

Noland, R.B., 2003. Traffic fatalities and injuries: the effect of changes in infrastructure and other trends, Accident Analysis & Prevention 35 (4), 599-611.

Noland, R.B., Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data, Accident Analysis and Prevention 36 (4), 525-532.

Noland, R.B., Quddus, M.A., 2004. A spatially disaggregate analysis of road casualties in England, Accident Analysis and Prevention 36 (6), 973-984.

Noland, R.B., 2013. From theory to practice in road safety policy: Understanding risk versus mobility, Research in Transportation Economics 43, 71-84.

Noland, R.B., Klein, N.J., Tulach, N.K., 2013. Do Lower Income Areas Have More Pedestrian Casualties? Accident Analysis & Prevention 59, 337-345.

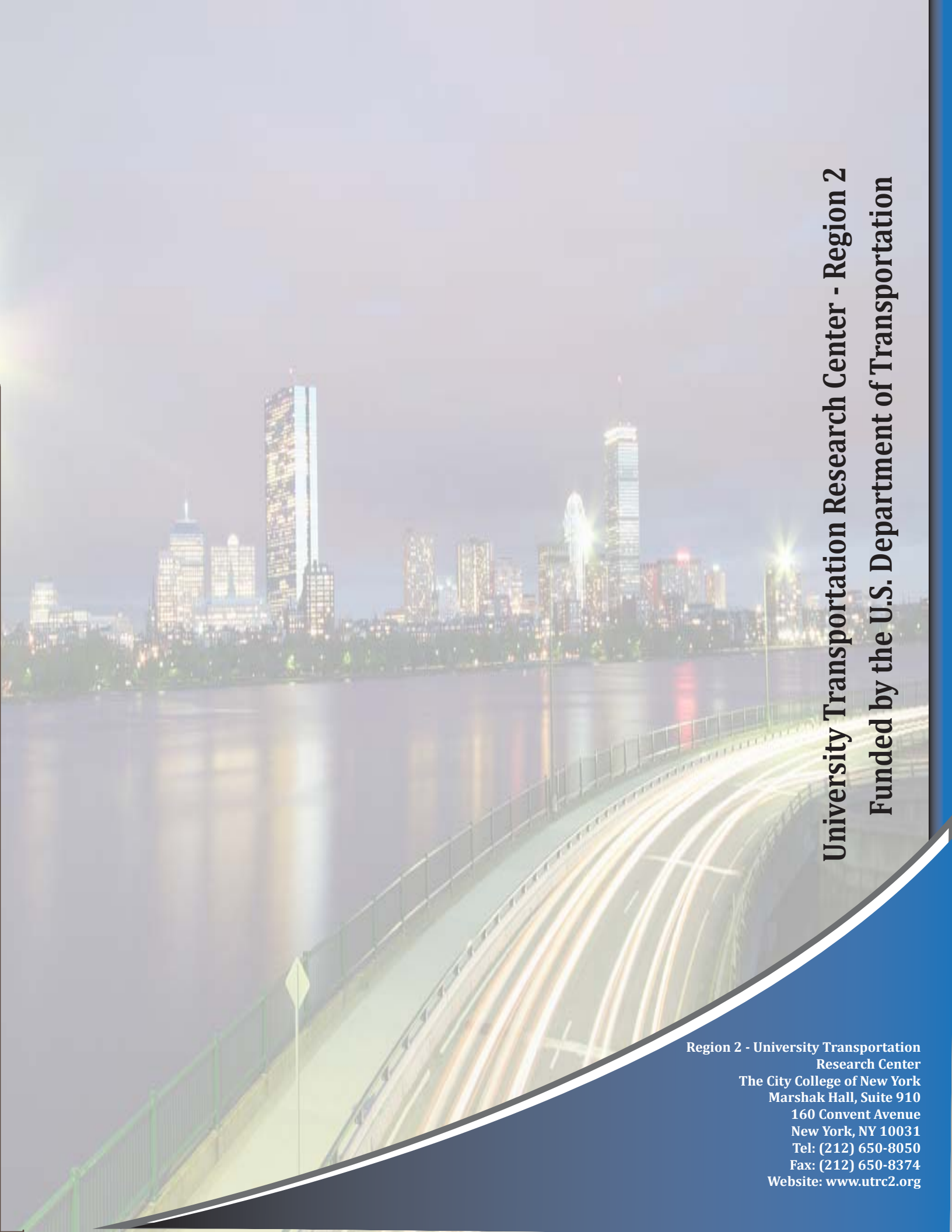
Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data, Accident Analysis & Prevention 40 (4), 1486-1497.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies, Accident Analysis & Prevention 27 (3), 371-389.

Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes, Accident Analysis & Prevention 45, 382-391.

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

Zhou, M., Sisiopiku, V., 1997. Relationship between volume-to-capacity ratios and accident rates, Transportation Research Record: Journal of the Transportation Research Board (1581), 47-52.



University Transportation Research Center - Region 2

Funded by the U.S. Department of Transportation

Region 2 - University Transportation
Research Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: (212) 650-8050
Fax: (212) 650-8374
Website: www.utrc2.org