County-Level Crash Risk Analysis in Florida

Bayesian Spatial Modeling

Helai Huang, Mohamed A. Abdel-Aty, and Ali Lotfi Darwiche

An increasing research effort has been made on spatially disaggregated safety analysis models to meet the needs of region-level safety inspection and recently emerging transportation safety planning techniques. However, without explicitly differentiating exposure variables and risk factors, most existing studies alternate the use of crash frequency, crash rate, and crash risk to interpret model coefficients. This procedure may have resulted in the inconsistent findings in relevant studies. This study proposes a Bayesian spatial model to account for county-level variations of crash risk in Florida by explicitly controlling for exposure variables of daily vehicle miles traveled and population. A conditional autoregressive prior is specified to accommodate for the spatial autocorrelations of adjacent counties. The results show no significant difference in safety effects of risk factors on all crashes and severe crashes. Counties with higher traffic intensity and population density and a higher level of urbanization are associated with higher crash risk. Unlike arterials, freeways seem to be safer with respect to crash risk given either vehicle miles traveled or population. Increase in truck traffic volume tends to result in more severe crashes. The average travel time to work is negatively correlated with all types of crash risk. Regarding the population age cohort, the results suggest that young drivers tend to be involved in more crashes, whereas the increase in elderly population leads to fewer casualties. Finally, it is confirmed that the safety status is worse for more deprived areas with lower income and educational level and higher unemployment rate in comparison with relatively affluent areas.

In the past two decades, crash prediction models have been widely applied to examine the safety effect associated with road elements of different types such as highway segments and intersections. Recently, an increasing research effort is being shifted to a higher aggregated level of crash analyses. Specifically, traffic crashes are aggregated by a certain specific spatial scale to relate safety with zone-level factors such as socioeconomic status, demographic characteristics, land use, and traffic pattern. Although the exploration of this topic is still in its infancy, two essential incentives may ensure its further development in the near future: estimation and forecasting of regional road safety.

First, there is a need for state agencies to regularly monitor region-level safety and provide incentives to reduce the number of traffic casualties in a region's safety program (*1*). Therefore, a reliable assessment of safety is indispensable by estimating the aggregating crash potentials associated with the target road network on different spatial scales. It can be used to estimate the normal level of safety as a means for examining regions that have greater-than-expected levels, and as such, region-level safety interventions could be implemented effectively. Furthermore, it allows cross-sectional comparisons between regions as well as examination of variations over time of safety effects associated with certain factors.

Second, road safety is increasingly considered a necessary component in the transportation planning process. In the United States, the Transportation Equity Act for the 21st Century (TEA-21) and the more recent SAFETEA-LU (*2*) create a positive agenda for increased safety on the highways and mandate the agencies of transportation planning—metropolitan planning organizations and state departments of transportation—to engage in proactive safety planning. As such, safety-conscious planning is under rapid development. Of vital importance in incorporating safety into planning is a reliable tool to forecast safety at the regional planning scale. This calculation requires the forecast of crash potentials for alternative transportation planning schemes given a number of zone-level characteristics.

In response to those needs, there are a good number of studies on developing zone-level safety analysis models. However, as reviewed in the next section, several inadequacies are identified, among which confusion on exposure variables and risk factors is the most noticeable. Without explicitly differentiating these variables, most of the existing studies alternate the use of crash frequency, crash rate, and crash risk to interpret parameter coefficients. This method may have resulted in the inconsistent findings in relevant studies, as discussed later. The current study proposes a Bayesian spatial model to account for county-level variations of crash risk by explicitly controlling for exposure variables. Safety effects are investigated by using data from the Florida state-maintained road network for various socioeconomic factors, demographic characteristics, and aggregate features of different types of road elements.

LITERATURE REVIEW

The unit of analysis in previous spatially disaggregate studies varies extensively, ranging from states (*3*), counties (*4–9*), traffic analysis zones (*1, 10–12*), census wards (*13–15*), local health areas (*16*), and grid-based structures (*17*). In most of these studies, aggregate crash prediction models were developed to relate the road crashes to a variety of explanatory factors including road network composition (disaggregated mileages of different road types, road density, intersection density, etc.), traffic patterns [posted speed, vehicle miles traveled (VMT), inflow and outflow from subject zones, volume-to-capacity ratios, etc.], and area-level demographic and socioeconomic characteristics (area, population, households, age cohorts, land use, employment, income, deprivation, improvement of medical technology, etc.).

H. Huang and A. L. Darwiche, 301-A, and M. A. Abdel-Aty, 301-H, Engineering Building II, Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL 32816-2450. Corresponding author: H. Huang, huanghelai@alumni.nus.edu.sg.

Transportation Research Record: Journal of the Transportation Research Board, No. 2148, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 27–37.
DOI: 10.3141/2148-04

The effects of various factors associated with road and traffic characteristics have been investigated to understand safety at different levels of spatial zones. An early study (*13*) at a small geographical-area level in Honolulu, Hawaii, found that more road mileage was associated with more crashes. Tarko et al. (*5*) examined the effects of VMT and found that higher VMT, especially on urban roads, is associated with an increased number of crashes. These results are consistent with a number of subsequent studies (*1, 6, 9, 11, 12, 15*). Furthermore, traffic congestion has been examined by means of proxy measures such as level of urbanization (*14*) and ratio of volume to capacity (*11*). The results seemingly show that more congested urban areas would be less likely to result in fatalities. This finding implies that policies of reducing congestion in urbanized areas may have unanticipated safety consequences (*18*).

A series of studies by Noland and his colleagues (*3, 8, 14*) extensively investigated the effect of various infrastructure changes on traffic-related fatalities and crashes at different levels of disaggregate spatial units, specifically, 50 states in the United States, 102 counties in Illinois, and 8,414 census wards in England. In general, they found that some improvements in road infrastructure have actually led to increased crashes and fatalities.

In contrast, a number of demographic and socioeconomic factors have been confirmed to be important predictors to account for zonal crash variations while controlling for road and traffic factors. Among a variety of relevant factors, of vital importance are land use, population density, age cohorts, income, deprivation, and employment. Specifically, it was found that there are more crashes in areas with higher population density (*1, 5, 11, 13*). However, Noland and Quddus (*14*) reported a conflicting result: that lower population density experiences relatively more casualties in England. However, employment density was found to have a positive effect on the likelihood of casualties in their study. Therefore, they explained that traffic within commercial areas may increase the risk of casualties, whereas those areas with high residential population density have relatively fewer casualties. In addition, it was found that total employment may be positively associated with crashes (*1, 15*) and different types of employment may result in different effects on crash patterns (*13*).

Although population age cohort is well known as an important risk factor of casualties, inconsistent results have been reported in previous literature. Noland and Quddus (*14*) and Aguero-Valverde and Jovanis (*9*) found a positive effect for younger population (under 16 and 15 years of age, respectively), but a negative effect was identified by Guevara et al. (*1*) for fatalities (under 18 years of age). With regard to the adolescent age cohort encompassing young drivers (generally 18 to 25 years of age), Noland (*3*) found that the percent of the population between 15 and 24 years of age significantly increases both fatalities and injuries, which conforms with the findings by Aguero-Valverde and Jovanis (*9*). But Noland and Oh (*8*) failed to confirm the significant association between age cohort variables and fatalities and crashes at the county-level analysis for the state of Illinois. Further, it was reported (*8, 15*) that an increase in the percent of the population over age 75 leads to fewer fatalities and injuries. However, conflicting results were found by Noland and Quddus (*14*) and Aguero-Valverde and Jovanis (*9*), in which higher percentages of elderly population are associated with higher traffic casualties.

Another safety predictor that has been generally studied in prior research is area-level socioeconomic deprivation. This factor has been measured by several different forms or proxy variables such as indexes of deprivation (*14, 19, 20*), percentage of households with no cars (*15*); per capita income (*3, 8*), and unemployment rate (*21*). In general, higher casualty rates were found to be associated with more deprived areas in comparison with relatively affluent areas (*14, 15, 19, 20*). However, in the highly aggregate analysis of 50 of the United States, Noland (*3*) identified a large statistically significant positive effect of per capita income on fatalities and injuries. Most important, the way to explain how deprivation affects safety remains unclear. It would be generally expected that wealthier areas seek to avoid riskier activities and thus are associated with a better safety situation. A significant positive correlation between pedestrian casualties and areawide deprivation was reported by Graham and Glaister (*20*). Since most pedestrian casualties normally occur in areas where they live, this fact sheds light on the understanding of a higher casualty rate in more deprived areas. Noland and Quddus (*14*) also examined the potential influence of unmeasured differences in road characteristics between more deprived and less deprived areas, but no significant pattern was found.

As discussed earlier, inconsistencies exist as to what factors could be used in predicting area-level safety, and their effects on traffic casualties remain indeterminate. This situation is most probably due to the variable accuracy and extent in measurement of the data under investigation in the literature. Moreover, inconsistent definitions of many relevant factors are also noticeable, especially when it comes to their total amount and density, for example, total road length and road density (road length/area), population and population density (population/area), and total employment and density of employment (total employment/area). This sort of problem is also related to a lack of explicit discrimination between crash risk and crash frequency, which essentially reflect different aspects of safety. Particularly in the context of spatial disaggregate analysis, crash risk refers to the crash potential given a unit of traffic exposure, whereas crash frequency is an aggregate crash count associated with the unit of analysis during a particular time period. As indicated by Quddus (*15*), the ideal exposure variable would be annual VMT in each area of analysis, by which the estimated crash risk would reflect the crash potential given vehicle mileage traveled. Unfortunately, because of data unavailability, only a few studies have appropriately considered the exposure by using some proxy variables. For example, in the work by Quddus (*15*), a gravity model was constructed to measure the exposure to risk of a census ward by use of total number of registered cars in each ward.

Furthermore, some prior studies are limited in their scope in handling data of spatiotemporal context where unmeasured confounders and spatiotemporal autocorrelation are evident (*16*). The negative binomial model as generally used is not able to account for any spatial correlation and structured heterogeneities between adjacent units of analysis. By use of spatial lag models, Levine et al. (*13*) examined spatial variations in crashes at a small geographical area level and found that spatial autocorrelation significantly exists; that is, crashes tend to be more clustered by block group than what would be expected by a random distribution. Recently, Bayesian hierarchical models have been successfully developed by several studies to systematically account for the spatial autocorrelation in aggregate crash prediction models (*9, 15, 22*).

With the aforementioned issues revealed by prior research in mind, this study attempts to conduct a reliable spatial aggregate analysis of relative crash risk associated with all 67 counties in the state of Florida. The spatial autocorrelation is accommodated by adopting a Bayesian spatial model. Moreover, instead of the prediction of crash frequency, an innovative model formation is specified to directly

examine the crash risk by controlling for aggregate average daily VMT (DVMT) and population. The proposed model is calibrated with a 5-year data set (2003 to 2007).

METHODOLOGY

Explanatory Analysis

A preliminary analysis is first conducted to find out whether observed crashes are spatially correlated among adjacent counties by use of Moran's *I,* which takes the following form:

$$\text{Moran's } I = \frac{n\sum_i\sum_j \omega_{ij}(Y_i-\bar{Y})(Y_j-\bar{Y})}{\left(\sum_{i\neq j}\omega_{ij}\right)\sum_i (Y_i-\bar{Y})^2} \quad (1)$$

where

n = total number of observations,
Y_i, Y_j = respective number of crash rates in counties i and j,
$\bar{Y}$ = average crash rate of all observations in analysis, and
ω_{ij} = entries of proximity matrix $\boldsymbol{\omega}$, which generally reflects spatial association of two units.

In the current study, binary specification is employed; that is, if counties i and j share a common border, they are considered neighbors and as such ω_{ij} equals 1; otherwise it would be 0 (*23*). Apparently the Moran's *I* range from −1 to +1. A positive value of Moran's *I* indicates positive spatial correlation, or clustering, within the study area. If the value is negative, it indicates negative spatial autocorrelation, or dispersion.

If there is no spatial correlation among Y_i's, that is, they are independent and identically distributed (iid), *I* is asymptotically normally distributed with a mean of $-1/(n-1)$ and a standard deviation $S(I)$:

$$S^2(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2} \quad (2)$$

where

$$S_0 = \sum_{i\neq j}\omega_{ij}$$

$$S_1 = \frac{1}{2}\sum_{i\neq j}(\omega_{ij}+\omega_{ji})^2$$

and

$$S_2 = \sum_i\left(\sum_j \omega_{ij} + \sum_i \omega_{ji}\right)^2$$

Thus, the significance of the *I*-value could be evaluated by a *Z* score:

$$Z(I) = \frac{I - E(I)}{S(I)} \quad (3)$$

Values of *Z* greater than +1.68 or less than −1.68 indicate significant positive and negative spatial autocorrelation, respectively, at the 10% level.

Bayesian Spatial Analysis

As shown earlier, the Moran's *I* is asymptotically normally distributed and as such it is limited to a continuous approximation for the crash count in the current study. Hence, it may be thought of as just an initial indication of spatial correlation. Furthermore, although the Moran's *I* statistic provides an exploratory measure of spatial autocorrelation, it is not able to estimate and test its magnitude and significance by controlling for a variety of observable county-specific risk factors (*23*). In this study, a Bayesian spatial model is developed to relate various county-level socioeconomic and traffic-related factors to crash occurrence while accounting for the possible spatial autocorrelation among adjacent counties.

The yearly crash rate for each county is treated as an observation unit for analysis. It is assumed that Y_{it} is the observed number of crashes at county i in year t, $i = 1, \ldots, I$, and $t = 1, \ldots, T$; $\mathbf{X_{it}}$ denotes various covariates having parameter coefficients $\boldsymbol{\beta}$. The spatial autocorrelation is realized by specifying a conditional autoregressive prior (CAR) model to the residual term of the link function in an ordinary Poisson regression:

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(\mu_{it}) \\ \mu_{it} &= E_{it} \times R_{it} \\ \log(R_{it}) &= \beta_0 + \boldsymbol{\beta_1}\mathbf{X_{it}} + \theta_i + \phi_i \end{aligned} \quad (4)$$

where μ_{it} is the parameter of the Poisson model, whereas E_{it} is the expected number of crashes at county i in year t, calculated as follows:

$$E_{it} = \text{expo}_{it}\frac{\sum_{i,t} Y_{it}}{\sum_{i,t}\text{expo}_{it}} \quad (5)$$

where expo_{it} is the traffic exposures associated with the target county, specifically DVMT or population in the current study. Thus it is clear that the R_{it} reflects the relative crash risk of county i in year t given a unit of exposure; that is, >1, higher risk; = 1, average; and <1, lower risk. θ_i is a county-specific random effect, which is assumed as iid among different counties. In this study, this statewide heterogeneity component θ_i is specified via an ordinary, exchangeable normal prior:

$$\theta_i \sim N\left(\frac{0,1}{\tau_h}\right) \quad (6)$$

where τ_h is a precision term (reciprocal of the variance) that controls the magnitude of the θ_i. These county-specific random effects capture extra-Poisson variability in the log-relative risk that varies globally, that is, over the entire state. Moreover, ϕ_i is the spatial correlation residual, or in other words, the correlated heterogeneity. That is, it models extra-Poisson variability in the log-relative risk that varies locally, so that nearby counties will have more similar rates. In this study, ϕ_i is assigned a CAR prior as recommended by Besag (*24*):

$$\phi_i \sim N\left(\frac{\bar{\phi}_i, 1}{\tau_i}\right) \qquad \bar{\phi}_i = \frac{\sum_{i\neq j}\phi_j\omega_{ij}}{\sum_{i\neq j}\omega_{ij}} \quad \text{and} \quad \tau_i = \frac{\tau_c}{\sum_{i\neq j}\boldsymbol{\omega}_{ij}} \quad (7)$$

where ω_{ij} is the binary entries of the proximity matrix as described earlier and τ_c is the precision parameter in the CAR prior. Clearly, the values of τ_h and τ_c control the amount of extra-Poisson variability

allocated to statewide heterogeneity and clustering effects among adjacent counties. Thus, it may also be interesting to estimate the proportion of variability in the random effects that is due to spatial clustering as follows:

$$\alpha = \frac{\mathrm{sd}(\phi)}{\mathrm{sd}(\theta) + \mathrm{sd}(\phi)} \tag{8}$$

where sd is the empirical marginal standard deviation function. Likewise, $1 - \alpha$ is the proportion of extra-Poisson variability accounted for by the statewide heterogeneity.

In this study, it should be noted that the longitudinal observations for a specific county share the same variation in terms of θ_i and ϕ_i. Nevertheless, temporal effects could also be introduced in the model via either fixed effects or random effects (globally iid or some county-specific autoregressive structure). However, no temporal effect appeared to be statistically significant for the data used in this study. This finding may be because of the high aggregation of crash data, and as such the time variation pattern is not obvious for the relatively short-term period (5 years only). Hence, for brevity, specification of temporal effects will not be elaborated on. For explicit model specification methods of temporal effects, interested readers are referred to work by Banerjee et al. (*23*), Aguero-Valverde and Jovanis (*9*), and Quddus (*15*).

DATA

Data Collection

In this county-level analysis, four different data sets were collected for all state-maintained roadways in Florida's 67 counties during a period of 5 years (2003 to 2007). These data sets include crash data, road and traffic characteristics, demographic and socioeconomic factors, and spatial features reflecting the geographic proximity of those counties.

Crash data were obtained from the Florida Department of Transportation (FDOT) Crash Analysis Reporting System. For this study, the most important information provided by the crash data is the county in which the crash occurred and its injury severity levels 1 through 5: 1, no injury; 2, possible injury; 3, nonincapacitating injury; 4, incapacitating injury; and 5, traffic fatality.

Traffic-related data were collected mainly from two sources: FDOT's Roadway Characteristics Inventory and geographic information system maps with Florida road characteristics. All these data were aggregated into the county level:

1. Roadway length: total centerline miles of roads per county;
2. Highway classification: centerline miles of roads for each type of highway functional classification per county (urban: principal arterial, minor arterial, urban collector, and local; rural: principal arterial, minor arterial, collector, and local); the fractional percent of each road class within a given county was calculated;
3. DVMT: calculated by multiplying each road section's average daily traffic (ADT) by its centerline mile length and then summing all DVMT values for each county;
4. Interchanges: number of interchanges per county;
5. Intersection: number of intersections per county;
6. Truck annual ADT (AADT): ADT of trucks per county;
7. Traffic-monitoring site: number of traffic-monitoring sites per county; and
8. Travel time to work: average travel time to work for residents in a county.

Besides the traffic-related data, a variety of demographic and socioeconomic factors were investigated, which were obtained from the U.S. Census Bureau. These include geographical area of each county, population segregated by gender and age cohort, income (yearly median income per household), poverty level (percent of people living below the poverty line), bachelor's degree ratio (percent of people older than 25 with a bachelor's degree or higher), and unemployment rate.

Finally, the proximity matrix representing the geographic neighboring structure of counties was automatically generated by using the GeoBUGS software.

Data Parameterization

To appropriately reflect the effects of various factors on crash risk, the raw data were carefully rescaled and parameterized. For simplicity, all the variables used in the analysis and the descriptive statistics are given in Table 1. Specifically, all crashes and severe crashes associated with the counties were investigated separately. Crashes with Severity Levels 4 (incapacitating injury) and 5 (traffic fatality) were considered severe crashes. DVMT was utilized as the exposure variable as suggested by many studies (*9*). In addition, the population was considered as an alternative exposure variable. Therefore, all coefficients later estimated from the models reflect the effects of covariates on crash risk by accounting for DVMT or population. It is rational to expect that both DVMT and population represent a good estimation of travel activities and crash exposure on the road network. By controlling for the exposure variables, a number of factors potentially affecting crash risk could be constructed. It is important to note that all the factors considered here are rationally supposed to be effective on crash risk given a unit of exposure.

As shown in Table 1, several road traffic variables were coded in the analysis. In particular, aggregate road density and road densities segregated by road functions were considered potential factors influencing crash risk. Further, the ratio of DVMT to overall road length was constructed to reflect the traffic intensity on those road networks. It is well known that intersections are generally associated with more traffic conflicts, and as such the intersection density (no. of intersections/road length) was investigated to understand the variation of crash risk across different counties. Since truck-related safety problems are continuously of concern as one of the top priorities in safety campaigns, the proportion of truck AADT over the total AADT was related to explain the varying crash risk. Moreover, the average travel time to work was also investigated since past studies revealed that the travel distance from the residence is correlated with crash occurrence (*19*). Although the travel time to work is the only proxy variable available in the current study, it should be noted that travel time to work alone may not be a good indicator of travel distance when one takes into account the use of different transportation means and traffic congestion.

Various demographic and socioeconomic variables were also properly constructed to account for variations in crash risk. Population density was the first important demographic factor considered in the analysis as demonstrated by many prior studies. Moreover, several interesting age cohorts were considered: the proportion of the population under 5, under 18, between 15 and 24 (surrogate for young drivers), and 65 years or older (surrogate for senior drivers).

TABLE 1 Summary of Variables and Descriptive Statistics

Variable	Description	Min.	Max.	Mean	SD
Crash Response Variables					
All crash	Frequency of all types of crashes (per year)	25	32,810	2,352.50	4,821.47
Severe crash	Frequency of severe crashes (per year)	3	1,617	203.44	304.45
Exposure Variables					
DVMT	Daily vehicle miles traveled (in thousands)	159	31,098.4	4,491.97	6,198.62
Population	Population (in thousands)	7.4	2,427.02	264.42	423.21
Road and Traffic-Related Factors					
Road density	Road length by area ∗100	5.6	96	25.23	13.52
Traffic intensity	DVMT/road length	1.9	56.7	16.59	12.22
Urban over rural	Urban road length/rural road length	0	268.82	5.28	31.14
FW density	Freeway length/area ∗100	0.01	18	3.47	4.01
PA density	Principal arterial length/area ∗100	0.02	51.55	14.18	9.97
MA density	Minor arterial length/area ∗100	0.01	35.55	8.06	5.54
CR density	Collector road length/area ∗100	0.01	12.11	2.71	2.75
Intersection density	No. of intersections/road length	0.08	27.25	10.97	5.01
Truck AADT	Truck AADT/total AADT ∗100	3.98	40.22	11.16	5.62
TTTW	Avg. travel time to work (min)	18.4	35.5	26.55	3.67
TMS density	No. of traffic monitoring sites/road length	0	2.45	0.55	0.38
Demographic and Socioeconomic Factors					
Area	Area	240.29	2,025.34	804.87	385.67
Population density	Population/area	8.86	3,304.17	316.47	501.71
Five	Percent of age group under 5	3.7	8.6	5.87	1.05
Eighteen	Percent of age group under 18	15.6	28.9	21.27	2.78
Young	Percent of population between 15 and 24	13	36.8	19.70	4.18
Sixty-five	Percent of population of 65 and older	8	31.2	16.93	5.78
Female	Percent of female population	34.4	52.5	48.71	3.65
White	Percent of white population	42.7	94.5	82.32	9.57
MIC	Median household income (in thousands)	26.41	55.71	37.19	6.88
Poverty	Percent of population below poverty line	7.1	20.9	12.79	3.39
Bachelor	Percent of population above 25 with bachelor's degree	6.8	41.7	16.73	8.03
UE rate	Unemployment rate	2.1	6.2	3.30	0.63

To reflect area deprivation level, three surrogate indicators were employed: median household income, percent of population below poverty line, and unemployment rate. The percent of white population and population above age 25 with a bachelor's degree were also considered to reflect the race composition and education level, respectively.

RESULTS AND DISCUSSION

Crashes and Exposure Variables

As shown in Figure 1, the overall state-road crash frequencies during the 5-year period (2003 to 2007) in 67 Florida counties range from 150 in Lafayette County to as high as 154,694 in Miami–Dade County with a standard deviation of 24,199. Within these, 68,151 crashes resulted in incapacitating injuries and fatalities, ranging from 31 to 7,514, with a standard deviation of 1,523. Results in Figure 1 also imply that the overall crash distribution is spatially consistent with the severe-crash distribution. Crashes are more concentrated in the southeast and central regions, whereas Duval County peaks in north Florida. Most central regions in north Florida are associated with the least crashes.

As expected, the crash spatial distribution is naturally associated with land use features, which are relevant to varying population density and subsequent travel activities, partially reflected by transportation infrastructure such as highway density. Thus, an illustration of crash rate by controlling for traffic exposure is helpful to understanding crash risk associated with different regions.

Figure 2*a, b* shows the yearly rate of crashes by million DVMT and Figure 2*c, d* by population. Apparently the former represents the crash risk for travelers on road networks residing in different zones, whereas the latter reveals the average risk that residents in a target zone will be involved in road crashes. Specifically, by controlling for DVMT, Holmes and Clay Counties are associated with the highest risk, and Miami–Dade and Escambia Counties with the lowest for the all-crash rate and the severe-crash rate, respectively.

Overall, with the average risk of 358 crashes and 47 severe crashes per million DVMT, the standard deviations amount to 186 and 16, respectively. Moreover, in terms of crash risk given population,

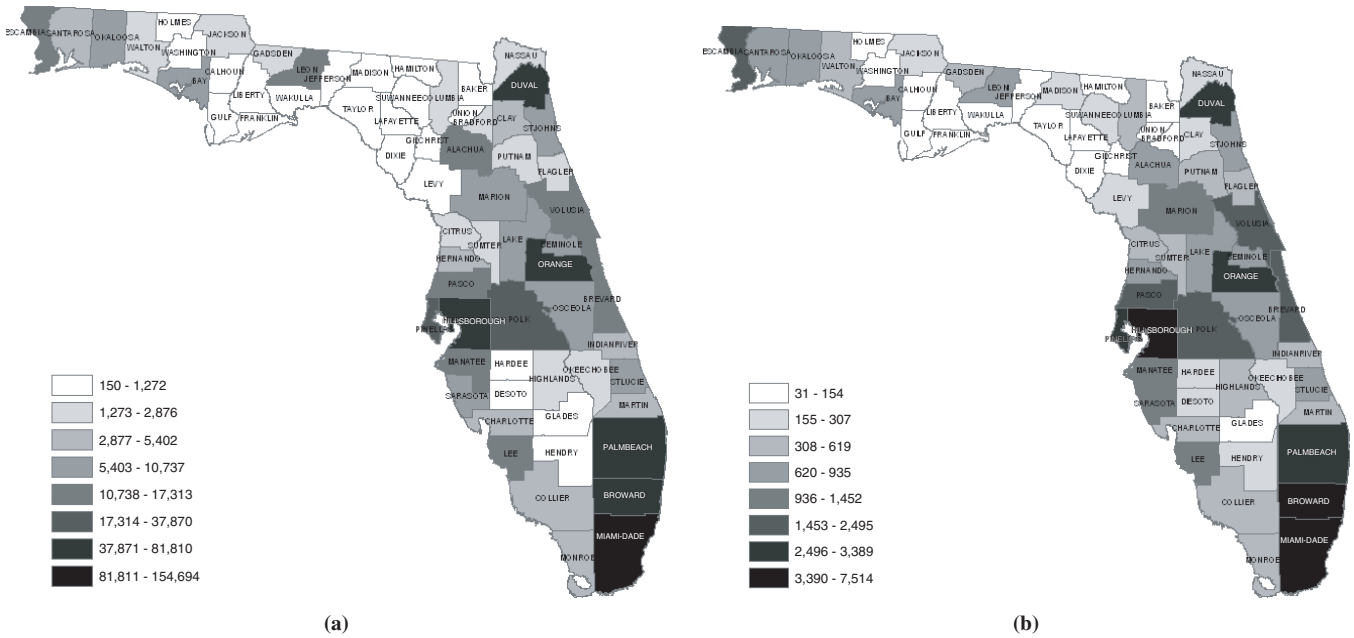

FIGURE 1 Crash frequencies by county in Florida (2003–2007): (*a*) number of all crashes and (*b*) number of severe crashes.

Collier and Clay Counties are identified as the safest, and Alachua and Madison Counties are the most dangerous for all-crash risk and severe-crash risk. On average, 71 crashes and 10 severe crashes per 10,000 population occurred yearly with standard deviations of 27 and 4, respectively. As can be seen in Figure 2, crash risk is more moderately distributed across different counties in comparison with the overall crash distribution. Nevertheless, it is still surprising to find substantial variation with regard to crash risk across different counties. These variations significantly negate the hypothesis of a linear relationship between crash frequencies and exposure variables, which would be expected if crash risk were approximately equal across different regions.

With the variation in crash risk in mind, an exploratory analysis was conducted to fit crash frequencies to exposure variables with a variety of potential nonlinear regression assumptions. Fortunately, good-fitting models were obtained by taking the natural logarithm to the variables, as shown in Figure 3. Specifically, four equations are obtained as follows:

$$\text{Log}(\text{all-crash count}) = 1.2202\,\text{Log}(\text{DVMT}) - 1.2279 \qquad R^2 = .943$$

$$\text{Log}(\text{severe-crash count}) = 0.9665\,\text{Log}(\text{DVMT}) - 1.2501 \qquad R^2 = .9223$$

$$\text{Log}(\text{all-crash count}) = 1.1001\,\text{Log}(\text{population}) + 0.6224 \qquad R^2 = .9547$$

$$\text{Log}(\text{severe-crash count}) = 0.8606\,\text{Log}(\text{population}) + 0.2369 \qquad R^2 = .9108$$

Judging by the coefficients, it seems that, in general, the increases in DVMT and population result in a higher increasing rate of all-crash frequencies but a lower increasing rate of severe-crash frequencies. In essence, the nonlinear relationships of crash rate and exposure are anticipated since different counties are associated with different features in terms of socioeconomic and demographic characteristics, road infrastructure, as well as traffic patterns. Hence, for the purpose of safety monitoring and improvement in safety-conscious planning, it is necessary to understand the effects of a variety of zone-level factors potentially affecting crash risk.

Preliminary Spatial Analysis

As has been proved in prior studies (*9, 15, 22*), spatial correlation exists widely across spatial zones and there would be significant impacts on the accuracy of the estimated effect of crash risk factors without explicitly taking into account the spatial autocorrelation. In Figures 1 and 2, it can be observed that counties in the northern portion of Florida share similar low crash rates, whereas the southern ones have higher rates in terms of overall crashes and crashes given DVMT.

However, by controlling for population, most of the northern counties turn out to be at higher risk, whereas southern ones tend to be more moderate. To explore the scope of spatial autocorrelation, Moran's *I*-statistics were calculated by using ArcMap 9.2, as shown in Table 2. The results indicate that substantial positive spatial correlation does exist among the counties with all *I*-values being positive. Except for crash rate by population, all other *Z*-scores are higher than 1.68, which indicates that there is less than 10% chance that spatial association is due to random chance. Therefore, the results justify the incorporation of spatial autocorrelation specification into the following crash risk models.

Model Calibration

To account for the cross-county variations observed in crash risk given exposure, a variety of zone-level risk factors are investigated by calibrating the proposed Bayesian models. A CAR prior is used to model the potential spatial patterns across counties, which are assumed to be able to account for various spatially correlated factors that are not observed or are unobservable in the analysis. In the

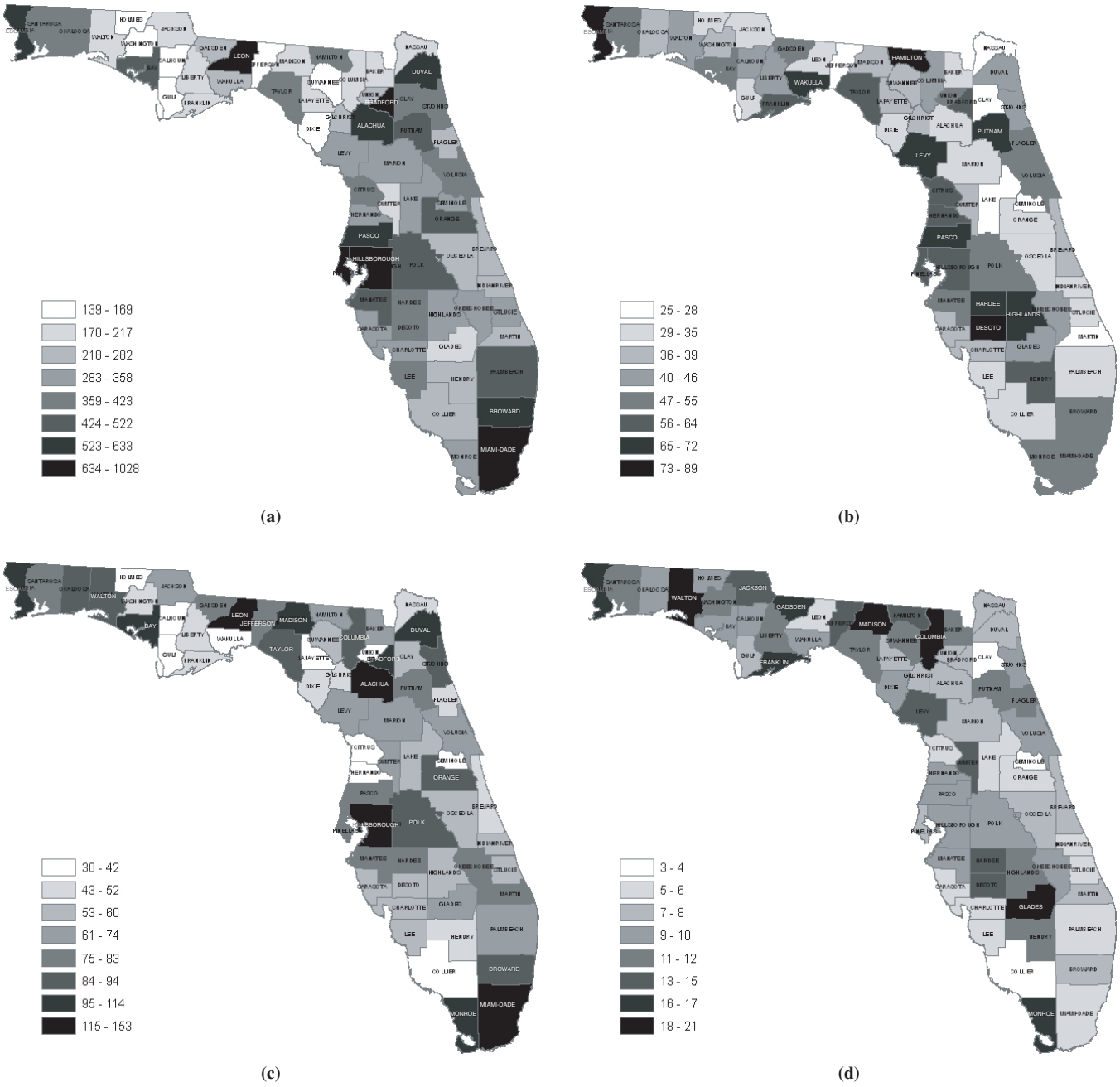

FIGURE 2 Yearly rate of crashes by million DVMT: (*a*) all crashes, (*b*) severe crashes; yearly rate of crashes by population in 10,000s: (*c*) all crashes, and (*d*) severe crashes.

context of the CAR model, those confounding factors are further supposed to be spatially correlated among adjacent counties and their effects on crash risk are homogeneous.

Because of a lack of consistent prior information, uninformative priors were specified to the model parameters. Diffused normal distributions are used for priors of regression parameters, $\beta \sim \text{norm}(0.0, 1.0\text{E-}3)$. However, in the context of the Bayesian approach, the hyperpriors on τ_h and τ_c cannot be arbitrarily vague for θ_i, and ϕ_i would be unidentifiable as noted by Banerjee et al. (*23*). In this study, the fair priors suggested by Best et al. (*25*) are specified on τ_h and τ_c:

$$\tau_h \sim \text{gamma}(1.0\text{E}-3, 1.0\text{E}-3) \qquad \tau_c \sim \text{gamma}(0.1, 0.1)$$

The CAR models are very convenient computationally, using a Gibbs sampler in Bayesian inference, which operates by successively sampling from the full conditional distribution of each parameter given the data. The proposed models were estimated by using the WinBUGS package (*26, 27*), which provides a flexible and simplified platform for calibrating Bayesian models with the BUGS programs. The convergence of multiple Markov chains was evaluated by using the built-in Brooks–Gelman–Rubin (BGR) diagnostic statistic (*28*).

Despite the convenience of model calibration, appropriate selection of variables included in the final models has been challenging. Some general criteria exist in subset selection, such as statistical significance of covariate coefficients, overall model goodness of fit, and model parsimony. Although a number of automatic variable

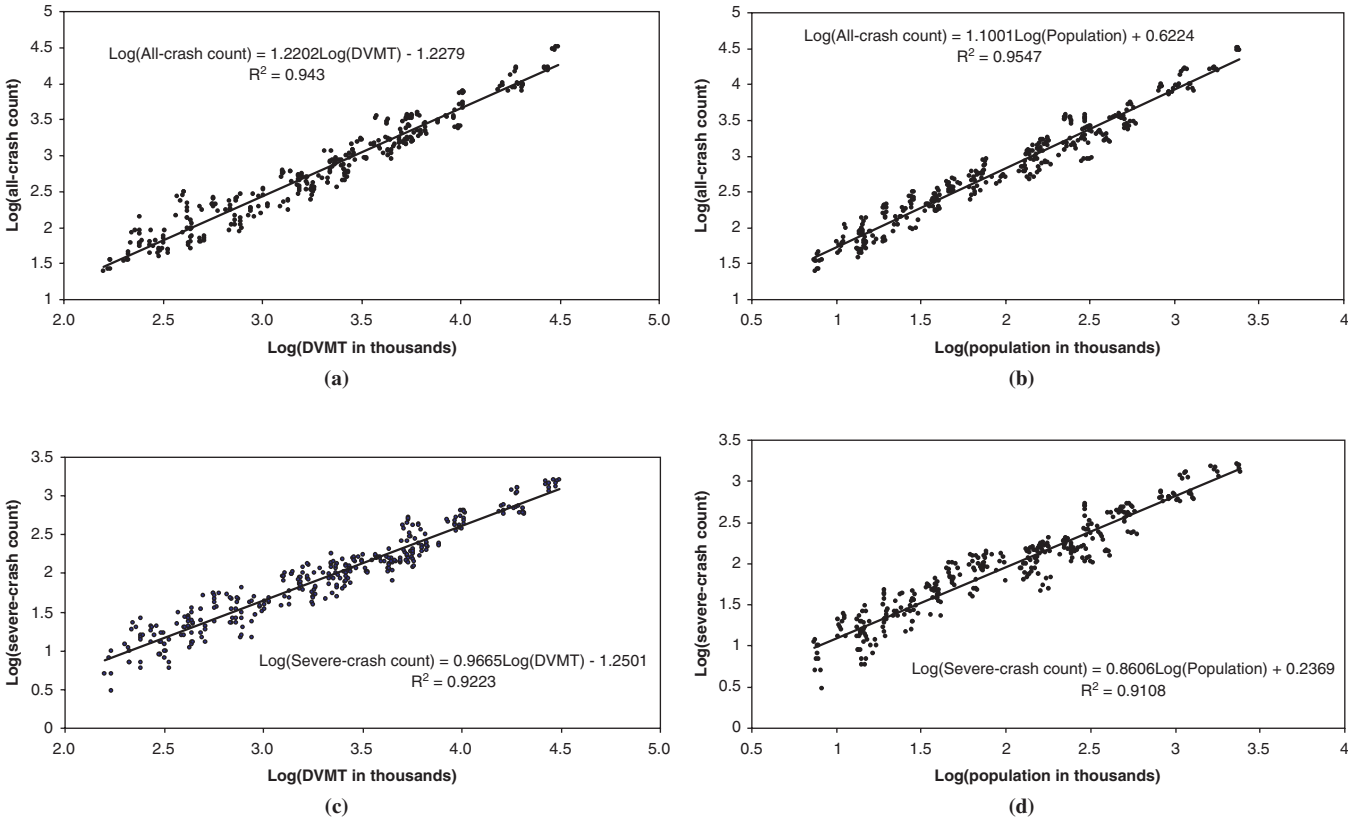

FIGURE 3 Crash frequency and exposure variables: (*a*) all-crash count versus DVMT, (*b*) all-crash count versus population, (*c*) severe-crash count versus DVMT, and (*d*) severe-crash count versus population.

selection methods are available, conscious selection is always desirable for optimal subsets from a relatively large number of available variables. This challenge becomes more critical in the current study, in which risk factors are widely intercorrelated. By exploratory examination of the data, the multicollinearities mostly arise from two sources. In particular, some variables are correlated because they intrinsically reflect identical latent characteristics in nature. For example, the correlation among median household income (MIC) and percentage of population below the poverty line is not surprising since they represent economic deprivation status, although probably from different aspects. The second source of multicollinearity is relevant to the nature of land use and transportation planning in reality. For example, more urbanized zones are generally associated with higher population density, denser road infrastructure, and most probably higher traffic intensity.

TABLE 2 Results of Moran's *I*-Statistics

	Global Moran's *I*	*Z*-Score	*P*-Value
All crashes	0.2	3.12	0.003
Severe crashes	0.29	4.03	0.000
All crashes by DVMT	0.17	2.35	0.025
Severe crashes by DVMT	0.13	1.75	0.086
All crashes by population	0.02	0.43	0.36
Severe crashes by population	0.12	1.62	0.11

To deal with the multicollinearity problem, by use of a scatterplot matrix and the variance inflation factor, manual logical inspection was conducted to predetermine the candidate subsets of covariates. Significantly redundant variables were properly transformed or carefully selected to ensure the robustness of model estimation. Meanwhile, some variables that seem very useful in explaining the variation of crash risk are retained although they are to some extent correlated with other variables in the model. Moreover, the deviance information criterion (DIC), a Bayesian measure of model complexity and fit (*29*), was employed to compare alternative models with different subsets of covariates, and the ones producing the lowest DICs were selected in the final models.

Interpretation of Model Results

Table 3 presents the final model estimation results for crash risk controlled by DVMT and by population. The BGR convergence diagnostics indicated that all final models converge well. Estimation of the marginal standard deviations of statewide heterogeneity SD(θ_i) and clustering effects among adjacent counties SD(ϕ_i) is used to calculate the proportion of the variability in the random effects that is due to clustering (α). The results show that variations accounted for by spatial clustering are substantial for all the all-crash and severe-crash risk models, specifically, 51.7% and 42.4% for models controlled by DVMT and 25.9% and 26.4% for models with population given. In general, the estimated values for the two models by DVMT are more significant than those by population, which is consistent with the preliminary results using Moran's *I* spatial diagnostics.

TABLE 3 Model Estimation Results

Risk Factors	All-Crash Risk by DVMT		All-Crash Risk by Population		Severe-Crash Risk by DVMT		Severe-Crash Risk by Population	
Traffic intensity	0.897	(0.812, 0.981)	0.756	(0.675, 0.837)	0.487	(0.403, 0.571)	0.333	(0.243, 0.424)
FW density	–0.385	(–0.476, –0.295)	–0.264	(–0.354, –0.175)	–0.223	(–0.312, –0.133)	–0.089	(–0.168, –0.010)
PA density	0.270	(0.171, 0.368)	0.216	(0.121, 0.310)	0.264	(0.174, 0.354)	0.205	(0.115, 0.295)
MA density	0.156	(0.098, 0.213)	0.107	(0.052, 0.163)	0.056	(0.004, 0.109)	—	—
Intersection density	—	—	—	—	—	—	0.081	(0.019, 0.142)
Truck AADT	—	—	—	—	—	—	0.094	(0.024, 0.163)
TTTW	–0.093	(–0.141, –0.046)	–0.128	(–0.176, –0.080)	–0.059	(–0.107, –0.011)	–0.101	(–0.150, –0.052)
Young	0.170	(0.123, 0.218)	0.091	(0.026, 0.155)	—	—	–0.108	(–0.177, –0.040)
Sixty five	—	—	–0.125	(–0.199, –0.051)	–0.128	(–0.207, –0.048)	–0.160	(–0.241, –0.079)
MIC	—	—	—	—	–0.087	(–0.146, –0.028)	–0.096	(–0.154, –0.039)
UE rate	—	—	—	—	0.057	(0.010, 0.105)	—	—
DIC	1,401.3	—	1,359.1	—	1,457.5	—	1,374.2	—
SD(θ_i)	0.182	(0.055, 0.281)	0.091	(0.012, 0.153)	0.156	(0.078, 0.235)	0.126	(0.098, 0.281)
SD(ϕ_i)	0.17	(0.041, 0.322)	0.260	(0.186, 0.301)	0.212	(0.112, 0.350)	0.351	(0.254, 0.465)
Alpha	0.517		0.259		0.424		0.264	

NOTE: Coefficient mean (95% Bayes credible interval). Variables are defined in Table 1.

A number of covariates were identified as significantly affecting the crash risk, as shown in Table 3. Most of the effects of those significant variables are consistent. This finding implies that, in general, there is no substantial difference for the risk factors on all crashes and severe crashes given DVMT and population as exposure.

Traffic intensity (i.e., the ratio of DVMT to overall road length) has a positive coefficient. Since it is highly correlated with population density (0.79) and level of urbanization (0.84), only the traffic intensity is included in the model. The results imply that elevated crash risk is associated with counties with a higher concentration of road traffic and population, as well as a higher level of urbanization. It is not surprising that more concentrated traffic leads to more interactions and conflicts, the increase rate of which should be larger than the rate of traffic increase. Hence, given a mile of vehicle travel, the average crash risk would increase because of more conflicts with other vehicles on the roads. This justification could also be applied to the cases of population as exposure. This finding means that compared with sparse population distribution, more-concentrated population (or a higher level of urbanization) will in general increase crash risk associated with each resident. However, several previous studies found traffic congestion level to be negatively related to crashes (*11, 14*). This result is most probably due to significant speed reduction in more congested urban areas as explained in prior literature. Moreover, the lack of explicit exposure variables and the multicollinearity examination as in previous studies may have led to these inconsistent results.

In addition to overall road density, the effects of disaggregate road densities were calculated for different highway functional classifications. It is not surprising that roads with different functional classifications show different effects with regard to safety. Freeways and principal arterials are designed for movement of large traffic volumes over relatively long distances and carry many trips not destined for or originating within a county. Unlike freeways, principal arterials do provide access management; however, it is controlled to the maximum extent possible in comparison with minor arterials and local roads. Minor arterials, in contrast, carry moderate volumes of traffic and provide an intermediate connection between principal arterials, collectors, and local roads. The results show that crash risk is negatively correlated with freeway density but positively correlated with the densities of principal and minor arterials. This finding could be attributed to the fact that freeways are generally better designed and have full access control and low speed variance, whereas arterials have intersections and experience more traffic congestion, which increases crash risk. This justification is consistent with the positive coefficient of intersection density, which implies that more intersections on equal lengths of road increase crash risk. Given the exposure of DVMT or population, densities of roads or of certain elements such as intersections rather than road length or element number such as intersection number should be logically used as the risk factors. Unfortunately, in many existing studies, the road length or element number and densities were alternatively used without explicit differentiation between the exposure variable and the risk variable. This approach makes the comparison between relevant studies extremely difficult.

Among the traffic factors, the proportion of truck AADT to overall AADT is positively correlated with severe-crash risk given population. This finding is consistent with the well-known fact that trucks lead to a substantial portion of road casualties. In the United States, almost 5,000 people are killed each year in truck-related crashes. Because of their size and often dangerous payloads, commercial truck crashes are devastating to pedestrians and occupants of other vehicles.

The average travel time to work was negatively correlated with all types of crash risk. One might think this finding to be somewhat surprising since it would seem logical that the probability of crash involvement increases as the average time spent traveling on the

road increases. Recent research found that most people tend to live close to their workplace, which would result in lower average travel times to work (*30*). Most crashes have been found to occur within an area close to home. Strillacci (*30*) attributed this finding to the fact that people tend to be less attentive when driving short distances because of a false sense of security that arises from proximity. This result is also consistent with work by Abdalla et al. (*19*) in which more accidents were found to more likely occur near home.

Population age cohort has been generally recognized as a significant factor in crash occurrence. The results show that the percentage of young population (age 15 to 24), a surrogate for young drivers, has significantly positive effects on all-crash risk but is negative on severe-crash risk given population. Generally, as found in many prior studies (*9, 15*), higher risk is expected for young drivers since the young population tends to have a higher level of mobility, whether as driver, passenger, cyclist, or pedestrian. They also take more risks and are aggressive in driving. With respect to the elderly population, the results conform with those of Noland and Oh (*8*) and Quddus (*15*) that an increase in the percent of elderly population leads to fewer casualties.

Among the variables indicating area deprivation (i.e., median household income and percent of population below the poverty line), the median household income is included in the model because of a coefficient with higher significance. As shown in Table 3, it has negative effects for severe-crash risk but is insignificant for all-crash risk. In addition, the coefficient of unemployment rate for the severe-crash risk model by DVMT is significant and positive. These findings imply that counties with higher median household income and lower unemployment rate are relatively safer in terms of severe-crash risk. This result confirms most of the prior studies in which higher casualty rates were found to be associated with more deprived areas in comparison with relatively affluent areas (*14, 15, 19, 20*). The proportion of the population older than 25 years with a bachelor's degree was also excluded from the final models, since it is highly correlated with median household income. The large positive correlation (0.74) may imply a significant causal relation between educational level and area deprivation.

CONCLUSION

This study presents a county-level road safety analysis for the state of Florida. Good-fitting nonlinear relations were obtained to relate crash rates with exposure variables such as DVMT and population. Significant spatial correlations in crash occurrence were identified across adjacent counties. To account for the variations in crash risk associated with different counties, a variety of aggregate road features, traffic patterns, and demographic and socioeconomic characteristics were investigated. The comprehensive literature review and preliminary research show that the development of zone-level safety prediction models could be challenging in terms of explicit discrimination of crash rate and crash risk, exposure variables and risk factors with consistent definitions, and accommodation for spatially structured heterogeneities.

A Bayesian spatial model was successfully applied to investigate crash risk given the two exposure variables. The results imply that there is no significant difference in safety effects of risk factors on all crashes and severe crashes by controlling for DVMT and population. Counties with higher traffic intensity and population density and a higher level of urbanization are associated with higher crash risk. As opposed to arterials, freeways seem to be safer with respect to crash risk given either VMT or population. It was also found that the increase in truck traffic volume tends to result in more-severe crashes. The average travel time to work is negatively correlated with all types of crash risk investigated in this study, which confirms the prior finding that more crashes occur within an area close to home. Regarding the population age cohort, the results show that young drivers tend to be involved in more crashes whereas the increase in elderly population leads to fewer casualties. Finally, it is confirmed that safety status is worse for more deprived areas with lower income and educational level and higher unemployment rate in comparison with relatively affluent areas.

REFERENCES

1. Ladron de Guevara, F., S. P. Washington, and J. Oh. Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1897,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 191–199.
2. *SAFETEA-LU: Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users*. FHWA, U.S. Department of Transportation, 2005. http://www.fhwa.dot.gov/safetealu.
3. Noland, R. B. Traffic Fatalities and Injuries: The Effect of Changes in Infrastructure and Other Trends. *Accident Analysis and Prevention,* Vol. 35, No. 4, 2003, pp. 599–611.
4. Fridstrom, L., and S. Ingebrigtsen. An Aggregate Accident Model Based on Pooled, Regional Time-Series Data. *Accident Analysis and Prevention,* Vol. 23, No. 5, 1991, pp. 363–378.
5. Tarko, A. P., K. C. Sinha, and O. Farooq. Methodology for Identifying Highway Safety Problem Areas. In *Transportation Research Record 1542,* TRB, National Research Council, Washington, D.C., 1996, pp. 49–53.
6. Karlaftis, M. G., and A. P. Tarko. Heterogeneity Considerations in Accident Modeling. *Accident Analysis and Prevention,* Vol. 30, No. 4, 1998, pp. 425–433.
7. Amoros, E., J. L. Martin, and B. Laumon. Comparison of Road Crashes Incidence and Severity Between Some French Counties. *Accident Analysis and Prevention,* Vol. 35, No. 4, 2003, pp. 537–547.
8. Noland, R. B., and L. Oh. 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,* Vol. 36, No. 4, 2004, pp. 525–532.
9. Aguero-Valverde, J., and P. P. Jovanis. Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. *Accident Analysis and Prevention,* Vol. 38, No. 3, 2006, pp. 618–625.
10. Ng, K., W. Hung, and W. Wong. An Algorithm for Assessing the Risk of Traffic Accidents. *Journal of Safety Research,* Vol. 33, No. 3, 2002, pp. 387–410.
11. Hadayeghi, A., A. S. Shalaby, and B. N. Persaud. Microlevel Accident Prediction Models for Evaluating Safety of Urban Transportation Systems. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 87–95.
12. Hadayeghi, A., A. S. Shalaby, B. N. Persaud, and C. Cheung. Temporal Transferability and Updating of Zonal Level Accident Prediction Models. *Accident Analysis and Prevention,* Vol. 38, No. 3, 2006, pp. 579–589.
13. Levine, N., K. Kim, and L. Nitz. Spatial Analysis of Honolulu Motor Vehicle Crashes. II. Zonal Generators. *Accident Analysis and Prevention,* Vol. 27, No. 5, 1995, pp. 675–685.
14. Noland, R. B., and M. A. Quddus. A Spatially Disaggregate Analysis of Road Casualties in England. *Accident Analysis and Prevention,* Vol. 36, No. 6, 2004, pp. 973–984.
15. Quddus, M. A. Modeling Area-Wide Count Outcomes with Spatial Correlation and Heterogeneity: An Analysis of London Crash Data. *Accident Analysis and Prevention,* Vol. 40, No. 4, 2008, pp. 1486–1497.
16. MacNab, Y. C. Bayesian Spatial and Ecological Models for Small-Area Accident and Injury Analysis. *Accident Analysis and Prevention,* Vol. 36, No. 6, 2004, pp. 1028–1091.

17. Kim, K., I. M. Brunner, and E. Y. Yamashita. Influence of Land Use, Population, Employment, and Economic Activity on Accidents. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1953,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 56–64.
18. Shefer, D., and P. Rietveld. Congestion and Safety on Highways: Towards an Analytical Model. *Urban Studies,* Vol. 34, No. 4, 1997, pp. 679–692.
19. Abdalla, I. M., R. Raeside, D. Barker, and R. D. McGuigan. An Investigation into the Relationships Between Area Social Characteristics and Road Accident Casualties. *Accident Analysis and Prevention,* Vol. 29, No. 5, 1997, pp. 583–593.
20. Graham, D. J., and S. Glaister. Spatial Variation in Road Pedestrian Casualties: The Role of Urban Scale, Density and Land-Use Mix. *Urban Studies,* Vol. 40, No. 8, 2003, pp. 1591–1607.
21. Roberts, I., R. Marshall, R. Norton, and B. Borman. An Area Analysis of Child Injury Morbidity in Auckland. *Journal of Paediatrics and Child Health,* Vol. 28, No. 6, 1992, pp. 438–441.
22. Miaou, S., J. J. Song, and B. K. Mallick. Roadway Traffic Crash Mapping: A Space–Time Modeling Approach. *Journal of Transportation and Statistics,* Vol. 6, No. 1, 2003, pp. 33–57.
23. Banerjee, S., B. Carlin, and A. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall/CRC Press, Boca Raton, Fla., 2003.
24. Besag, J. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society,* Vol. 36B, No. 2, 1974, pp. 192–236.
25. Best, N. G., L. A. Waller, A. Thomas, E. M. Conlon, and R. A. Arnold. Bayesian Models for Spatially Correlated Diseases and Exposure Data. In *Bayesian Statistics* (J. M. Bernardo et al., eds.), *Proc. of the Sixth Valencia International Meeting,* Vol. 6, 1999, pp. 131–156.
26. Lunn, D. J., A. Thomas, N. G. Best, and D. Spiegelhalter. WinBUGS—A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing,* Vol. 10, No. 4, 2000, pp. 325–337.
27. Spiegelhalter, D. J., A. Thomas, N. G. Best, and D. J. Lunn. *WinBUGS version 1.4.1 User Manual.* MRC Biostatistics Unit, Cambridge University, United Kingdom, 2003.
28. Brooks, S. P., and A. Gelman. Alternative Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics,* Vol. 7, 1998, pp. 434–455.
29. Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and V. D. Linde. Bayesian Measures of Model Complexity and Fit (with discussion). *Journal of the Royal Statistical Society,* Vol. 64B, No. 4, 2003, pp. 583–616.
30. Strillacci, L. Car Accidents Tend to Occur Close to Home. http://info.insure.com/auto/collision/accidentlocation502.html. Accessed February 2009.

The Safety Data, Analysis, and Evaluation Committee peer-reviewed this paper.