Safety Science 58 (2013) 89–97

Contents lists available at SciVerse ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Using Geographically Weighted Poisson Regression for county-level crash modeling in California

Zhibin Li [a,b,*], Wei Wang [a,1], Pan Liu [a,2], John M. Bigham [b,3], David R. Ragland [b,3]

[a] School of Transportation, Southeast University, Si Pai Lou #2, Nanjing 210096, China
[b] Safe Transportation Research and Education Center, Institute of Transportation Studies, University of California, Berkeley, 2614 Dwight Way #7374, Berkeley, CA 94720-7374, United States

ARTICLE INFO

Article history:
Received 25 December 2012
Received in revised form 11 March 2013
Accepted 13 April 2013

Keywords:
Safety
Crash
County-level
Geographically Weighted Regression

ABSTRACT

Development of crash prediction models at the county-level has drawn the interests of state agencies for forecasting the normal level of traffic safety according to a series of countywide characteristics. A common technique for the county-level crash modeling is the generalized linear modeling (GLM) procedure. However, the GLM fails to capture the spatial heterogeneity that exists in the relationship between crash counts and explanatory variables over counties. This study aims to evaluate the use of a Geographically Weighted Poisson Regression (GWPR) to capture these spatially varying relationships in the county-level crash data. The performance of a GWPR was compared to a traditional GLM. Fatal crashes and countywide factors including traffic patterns, road network attributes, and socio-demographic characteristics were collected from the 58 counties in California. Results showed that the GWPR was useful in capturing the spatially non-stationary relationships between crashes and predicting factors at the county level. By capturing the spatial heterogeneity, the GWPR outperformed the GLM in predicting the fatal crashes in individual counties. The GWPR remarkably reduced the spatial correlation in the residuals of predictions of fatal crashes over counties.

1. Introduction

Previously, traffic safety analysis at spatially aggregated levels has drawn the interests of safety researchers to meet the needs of region-level safety inspection and emerging safety planning (Hadayeghi et al., 2003, 2006, 2010a; Noland and Quddus, 2004; Quddus, 2008; Huang et al., 2010; Zhang et al., 2012; Pirdavani et al., 2013). In recent years, state agencies have paid particular attention to the traffic safety evaluation at the county spatial level (Fridstrøm and Ingebrigtsen, 1991; Tarko et al., 1996; Karlaftis and Tarko, 1998; Amoros et al., 2003; Noland and Oh, 2004; Aguero-Valverde and Jovanis, 2006; Donaldson et al., 2006; Traynor, 2008; Darwiche, 2009; Huang et al., 2010; Chang et al., 2011; Hanna et al., 2012). In particular, crash counts are aggregated at a county level to relate traffic safety to a series of countywide factors including traffic patterns, road network attributes, as well as socio-demographic characteristics.

County-level crash risk analysis has become more popular since road safety has been increasing considered a necessary component in transportation planning process for counties (de Guevara et al., 2004; FHWA, 2005; NCHRP, 2010). Crash prediction modes are useful in predicting the expected number of crashes and estimating the normal safety situations in individual counties based on the countywide characteristics. Counties with greater-than-expected crashes can be identified and countermeasures can be implemented in these areas. For those reasons, it is desirable to develop county-level crash prediction models that have reasonably accurate predictions for crashes in individual counties.

A common technique for the county-level crash modeling is the generalized linear modeling (GLM) procedure. However, since the parameters in a GLM are assumed to be fixed, the GLM fails to capture the spatial heterogeneity in the relationships between crashes and predictors. Recently, a new methodology named Geographically Weighted Poisson Regression (GWPR) has been used by researchers for traffic safety analysis at traffic analysis zone (TAZ) levels (Hadayeghi et al., 2010a; Zhang et al., 2012; Pirdavani et al., 2013). The parameters in a GWPR are allowed to vary over space to capture the spatially varying relationships in the data. However, none of previous studies have used the GWPR for county-level crash analysis. An evaluation on the performance of a GWPR for the county-level data is necessary because the predicting variables at the county level are aggregated at a different spatial scale from

* Corresponding author at: School of Transportation, Southeast University, Si Pai Lou #2, Nanjing 210096, China. Tel.: +86 13952097374; fax: +86 25 83791816.
E-mail addresses: lizhibin@seu.edu.cn (Z. Li), wangwei@seu.edu.cn (W. Wang), liupan@seu.edu.cn (P. Liu), jbigham@berkeley.edu (J.M. Bigham), davidr@berkeley.edu (D.R. Ragland).
[1] Tel./fax: +86 25 83794101.
[2] Tel.: +86 13584057940; fax: +86 25 83791816.
[3] Tel.: +1 510 642 0655; fax: +1 510 643 9922.

0925-7535/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ssci.2013.04.005

these at the TAZ level. Besides, the sample size for the crash frequency modeling at the TAZ level is usually over hundreds. While in the analysis of county level crashes, the available sample size is strictly restricted by the number of counties in the study area. The small sample size issue could result in inaccurate estimates in statistical modeling procedures (Wood, 2002; Lord, 2006; Washington et al., 2010; Lord and Mannering, 2010).

The primary objective of this study is to evaluate the application of the GWPR modeling technique for the crash frequency modeling at the county level. More specifically, this study aims to answer the following questions: (1) if a GWPR is useful in capturing the spatial non-stationarity in the relationship between crashes and predictors over counties; and (2) if a GWPR outperforms a traditional GLM in predicting crash counts using county-level data with a small sample size. The remainder of this paper is organized as follows. The following section reviews the existing work. Section 3 introduces the methodology. Section 4 shows the data resources. Section 5 discusses the modeling results. The paper ends with concluding remarks and future work in Section 6.

2. Literature review

Previously, numerous studies have evaluated the crash risks at the county spatial level. The most commonly used technique for the county-level crash modeling is the GLM procedure with its random component follows a Poisson or Negative Binomial (NB) distribution (Fridstrøm and Ingebrigtsen, 1991; Tarko et al., 1996; Karlaftis and Tarko, 1998; Amoros et al., 2003; Noland and Oh, 2004; Traynor, 2008; Chang et al., 2011). Within a GLM framework, fixed coefficient estimates explain the associations between crash counts and explanatory variables in individual counties. For example, Noland and Oh (2004) developed the NB models to evaluate the impacts of road network infrastructure and geometric design on the county-level fatal and total crashes based on the 4-year data for 102 counties in Illinois.

Several researchers have used the Bayesian spatial models for the county-level traffic safety analysis (Aguero-Valverde and Jovanis, 2006; Darwiche, 2009; Huang et al., 2010). The advantage of Bayesian spatial models is that they can account for the spatial correlation in the county-level crash data, which refers to the fact that crashes tend to be more clustered by groups that are spatially close to each other by sharing some unobservable effects of factors. For example, Huang et al. (2010) developed the Bayesian spatial models for the 67 counties in Florida. They reported that significant spatial correlations in crashes were identified across adjacent counties. The Bayesian spatial models fitted the data better than the GLMs did.

The traditional GLMs are limited in capturing the spatial heterogeneity in the crash data. The outputs from a GLM consist of a set of fixed global parameters that do not vary over counties. The Bayesian spatial models employed previously are also limited to a CAR prior with fixed main parameters. The fixed parameters in these models represent that the impacts of countywide variables on crash counts are the same between different counties. Actually, however, the impacts of predicting variables could not be stationary over space. In other words, it is possible that some variables have large impacts in certain counties but have small impacts in other counties. Thus, the accuracy of such global models for predicting county-level crashes could be suspect.

Previously, several methods have been developed to account for the spatial heterogeneity in spatial data. In these models, the parameters of explanatory variables are allowed to vary spatially. These methods include the random parameter model (EI-Basyouny and Sayed, 2009; Anastasopoulos and Mannering, 2009), Geographically Weighted Regression (GWR) technique (Zhao and Park, 2004; Chow et al., 2006; Du and Mulley, 2006; Ibeas et al., 2011; Wang et al., 2011; Hadayeghi et al., 2010a; Zhang et al., 2012; Pirdavani et al., 2013), Full Bayesian semiparametric additive technique (Hadayeghi et al., 2010b), and Bayesian hierarchical model (Quddus, 2008). Among them, the GWR is the most commonly used modeling technique. The GWR has been reported to provide more accurate estimates than the global GLMs. Until recently, only a few studies have used the GWR for the crash frequency modeling (Hadayeghi et al., 2010a; Zhang et al., 2012; Pirdavani et al., 2013). Since crashes are presented as count data, a Poisson regression in conjunction with a GWR, i.e., a Geographically Weighted Poisson Regression (GWPR), is commonly used to fit the spatial crash data. It was reported that the calibrated GWPR captured the spatially varying relationships between crashes and predictors and outperformed the traditional GLMs in predicting the TAZ-level crashes.

A review on the literature shows that the spatial heterogeneity in the county-level crash data should be properly considered in the development of crash prediction models to improve the predicting accuracy for crashes. The GWPR technique has been used to account for the spatial heterogeneity in the crash modeling at the TAZ level. However, none of previous studies have used the GWPR for the county-level crash data analysis. The predicting variables at the county level are aggregated at a different spatial scale as compared to the TAZ level. And the county-level crash data usually suffers from a small sample size issue which does not exist in the TAZ level datasets. An evaluation on the performance of the GWPR modeling technique particularly for the county-level crash data is important to safety researchers. The findings can help state agencies select appropriate approaches in the development of county-level crash prediction models.

The GWPR in the present study is specified in the available software "GWRx3.0" developed by Charlton et al. (2003) for the GWPR calibration. It is clarified that although it would be beneficial to evaluate the Geographically Weighted Negative Binomial regression, the "GWRx3.0" does not support the calibration of GWR with a NB structure and as such these models are not calibrated in this study. In our study, the local models are fitted using a number of vicinity observations that are similar in their characteristics. Thus it is expected that the variance of crash counts will become much closer to the mean during the estimates for local parameters in a GWPR (Pirdavani et al., 2013). Besides, it is worth noting that the use of Poisson regression instead of NB does not produce much inaccurate estimates in general since the model coefficients are similar for the two error distributions (Miaou, 1994; Hadayeghi et al., 2010a). This justifies the choice of Poisson error distribution that is adopted in this study.

3. Methodology

Both the GLM and the GWPR were calibrated in the present study. The two techniques for the county-level crash frequency modeling were briefly described in this section. The goodness of fit measures for the model comparison as well as the Moran's *I* statistics for the tests on spatial correlation were also introduced.

3.1. GLM

A GLM usually consists of three components, a random component, a systematic component, and a link function that connects the random and systematic components to produce a linear predictor (Lord and Persaud, 2000). One important property of a GLM is its flexibility in specifying the probability distribution for the

random component. Thus, the GLMs have been widely used in the context of traffic safety, for which the distribution of crash counts often follows a Poisson or NB distribution (Washington et al., 2010). The difference between a Poisson model and a NB model is that the NB model can deal with the over-dispersion which indicates the variance exceeds the mean of crash counts.

The link function and linear predictor determine the functional form of the model. After a review on the model specifications for county-level crash frequency modeling in previous studies (Amoros et al., 2003; Aguero-Valverde and Jovanis, 2006; Darwiche, 2009; Huang et al., 2010), the following model form is considered in the present study:

$$\ln(\mathbf{Y}) = \ln(\beta_0) + \beta_1 \ln(\mathbf{DVMT}) + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \ldots + \beta_J \mathbf{X}_J + \varepsilon \quad (1)$$

where ln(**Y**) is the natural log of expected crash count per county per year, **DVMT** is the daily vehicle miles traveled, $\mathbf{X}_j$ is the jth explanatory variable ($j = 2, 3, \ldots, J$), β_j is the jth model parameter ($j = 0, 1, \ldots, J$), and ε is the error term.

As discussed above, the parameters **β** with explanatory variables **X** in the GLM are estimated globally and do not change over counties. The fixed parameter β_j represents the average impact of the jth variable on crash count across all counties.

3.2. GWPR

In a GWPR, the crash counts are predicted by a set of explanatory variables of which the parameters are allowed to vary over space. Similar to Eq. (1), the model specification of the GWPR in the present study is:

$$\ln(\mathbf{Y}) = \ln(\beta_0(\mathbf{u_i})) + \beta_1(\mathbf{u_i}) \ln(\mathbf{DVMT}) + \beta_2(\mathbf{u_i})\mathbf{X}_2 + \beta_3(\mathbf{u_i})\mathbf{X}_3 + \ldots + \beta_J(\mathbf{u_i})\mathbf{X}_J + \varepsilon \quad (2)$$

Note that β_j is now a function of location $\mathbf{u_i} = (u_{xi}, u_{yi})$ denoting the two dimensional coordinates of the ith point (ith county centroid in this study) in space. This means that the parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)$ estimated in Eq. (2) are allows to be different between counties. Thus, the spatial heterogeneity is addressed in the GWPR modeling framework. The parameter **β** can be expressed in the following matrix form:

$$\beta = \begin{bmatrix} \beta_0(u_{x1}, u_{y1}) & \beta_1(u_{x1}, u_{y1}) & \cdots & \beta_J(u_{x1}, u_{y1}) \\ \beta_0(u_{x2}, u_{y2}) & \beta_1(u_{x2}, u_{y2}) & \cdots & \beta_J(u_{x2}, u_{y2}) \\ \cdots & \cdots & \cdots & \cdots \\ \beta_0(u_{xn}, u_{yn}) & \beta_1(u_{xn}, u_{yn}) & \cdots & \beta_J(u_{xn}, u_{yn}) \end{bmatrix} \quad (3)$$

where n is the number of counties. The parameters for each county, which form a row in the matrix in Eq. (3), are estimated as follows (Fotheringham et al., 2002):

$$\hat{\beta}(i) = (\mathbf{X}^T \mathbf{W}(u_{xi}, u_{yi}) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(u_{xi}, u_{yi}) \mathbf{Y} \quad (4)$$

In Eq. (4), $\mathbf{W}(u_{xi}, u_{yi})$ denotes an n by n spatial weight matrix that can be conveniently expressed as $\mathbf{W}(i)$:

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & w_{in} \end{bmatrix} \quad (5)$$

where w_{ij} ($j = 1, 2, \ldots, n$) is the weight given to county j in the calibration of model for county i.

In the GWPR modeling framework, a regression equation is estimated for each county based on the observations in nearby counties. The estimation process was repeated for all regression points. Each county is weighted by its distance from the regression point. Hence, the data in counties closer to the regression point are weighted more heavily than are the data in counties farther away. In other words, the observations closer to county i have more of an influence on the estimation of i's parameter $\beta_j(\mathbf{u_i})$ than those counties farther from i. This influence around i is described by the weighting function w_{ij}. The Gaussian and bi-square functions are commonly used to produce the weighting scheme as follows:

$$\text{Gaussian}: \; w_{ij} = \exp\left(-\frac{1}{2} \times \frac{\|u_i - u_j\|}{G}\right) \quad (6)$$

$$\text{Bi-square}: \; w_{ij} = \begin{cases} [1 - (\|u_i - u_j\|/G_i)^2]^2 & \text{if } \|u_i - u_j\| < G_i \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

The parameter G_i is a quantity known as the bandwidth. When G_i approaches infinity, w_{ij} approaches 1 and the GWPR becomes a global model expressed in Eq. (1). The bandwidth is constant in the Gaussian function (fixed kernel) that sets the magnitude of the weighting function to be the same for every county. A potential problem that might arise in the GWR with fixed kernel is that for some points, where data are sparse, the local models might be calibrated on very few data points, giving rise to parameter estimates with large standard errors and unpredictable results. To reduce these problems, the bi-square function (adaptive kernel) allows the weighting scheme to vary spatially according to the density of data. The kernels have larger bandwidths where the data are sparse and have smaller bandwidths where the data are plentiful. Thus, the adaptive kernel is employed in the GWPR in this study.

The selection of bandwidth is important in the GWPR modeling procedure. With a large dataset, the bandwidth selection can be made using a sample (%) of data points in order to reduce workload and save time. However, the crash data at the county level has a small sample size. The estimates for local models may not have enough observations if a low percentage of sample is specified. Thus, during the modeling procedure in our study, all data are used in the adaptive kernel to determine the optimal bandwidth. The Corrected Akaike Information Criterion (AICc) is used for the selection of bandwidth in the adaptive kernels. The model with the lowest AICc indicates the best model performance (Fotheringham et al., 2002; Nakaya et al., 2005; Hadayeghi et al., 2010a). The AICc is also used to determine the model specification. The best GWPR is the one with the lowest AICc and is selected as the final model form.

3.3. Measures of goodness of fit

The performance of a GWPR in predicting county-level crashes is compared to that of a traditional GLM. The measures of goodness of fit used for model comparison are the mean absolute deviation (MAD), and the mean squared prediction error (MSPE). The MAD provides a measure of the average misprediction of the model. The MSPE is typically used to assess the error associated with a prediction. A smaller value of MAD or MSPE suggests that, on average, the model predicts the observed data better. These measures are described as follows:

$$\text{MAD} = \frac{\sum_{i=1}^{N} |\hat{Y}_i - Y_i|}{N} \quad (8)$$

$$\text{MSPE} = \frac{\sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2}{N} \quad (9)$$

where Y_i is the observed number of crashes in county i, $\hat{Y}_i$ is the predicted number of crashes in county i, and N is the number of counties.

3.4. Moran's I statistics

In statistics, Moran's I is a measure of spatial autocorrelation developed by Moran (1950). In this study, the Moran's I test is em-

ployed to investigate whether the residuals of predictions for county-level crashes are spatially correlated among adjacent counties. A negative (positive) value of Moran's I indicates a negative (positive) spatial autocorrelation over counties. Values of Moran's I range from −1 (indicating perfect dispersion) to +1 (perfect correlation). A zero value indicates a random spatial pattern. The Moran's I tests were conducted in the ArcGIS 10.

4. Data

4.1. Data resource

A four-year frame data, from 2007 to 2010, were collected from the 58 counties in the state of California. The data included four types of information in each individual county: counts of fatal crashes, traffic patterns, road network attributes, and socio-demographic characteristics. The reliability of the input data is important to the estimating results of the models. In this study, all the input data were collected from the authoritative agencies in the California and United States.

Fatal crash counts aggregated by county in California were obtained from the Fatal Accident Reporting System (FARS) created by the National Highway Traffic Safety Administration (NHTSA). Traffic patterns and road network attributes were collected from the Highway Performance Monitoring System (HPMS) maintained by the California Department of Transportation (Caltrans). Vehicle registration and licensed drivers information were retrieved from the Department of Motor Vehicles (DMV) in California. A variety of socio-demographic characteristics for counties in California were available from the U.S. Census Bureau. All data in each county were geocoded in the Geographic Information System (GIS).

4.2. Data description

Crash count per year in the 58 California counties ranges from 1.33 in the Alpine County to as high as 644.33 in the Los Angeles County, with a mean of 54.45 and a standard deviation of 98.25. Los Angeles County has the largest amount of fatalities, along with the largest population and DVMT. Thus, the data of Los Angeles County is not considered as an outlier. The distribution of fatal crash counts in California counties is shown in Fig. 1a. Crashes are more concentrated in the south and east regions. Regions in the northwest are associated with fewer crashes.

In the present study, the DVMT is utilized as the exposure variable as suggested by many prior studies (Amoros et al., 2003; Aguero-Valverde and Jovanis, 2006; Darwiche, 2009; Huang et al., 2010; Hadayeghi et al., 2010a; Pirdavani et al., 2013). The distribution of DVMT in California counties is shown in Fig. 1b. As expected, the distribution of DVMT is naturally consistent with that of crash count. An explanatory analysis was conducted to fit the crash counts to the exposure variable with a nonlinear regression assumption. A good-fitting relationship was obtained by taking the natural logarithm of the variables, as shown in Fig. 2. In general, the increase in DVMT results in a larger amount of fatal crashes. We also tested the correlation between the DVMT and the other explanatory variables. It was found that the DVMT positively impact most of the independent variables. It indicates the DVMT is a good measure of exposure for fatal crashes at the county level in California.

The variables used for model development and their descriptive statistics are shown in Table 1. The dependent variable is the fatal crash count per year in each county. The explanatory variables are the predictors that are commonly used in previous studies for the county-level crash analysis. Thus, the results of this study regarding the effects of predictors on county-level crashes can be conveniently compared to previous findings.

5. Results and discussion

The four-year frame data were divided into two subsets. The data from 2007 to 2009 were used to calibrate the GWPR and GLM models. The data of 2010 were used for the model validation. The estimates of the GWPR and GLM were presented and discussed in this section. The performances of the two models were then compared.

5.1. GWPR calibration

The GWPRs were calibrated based on the explanatory variables shown in Table 1. For each model, the exposure variable, i.e. DVMT, was initially considered due to its dominate predicting influence on crashes. Other candidate variables were analytically selected

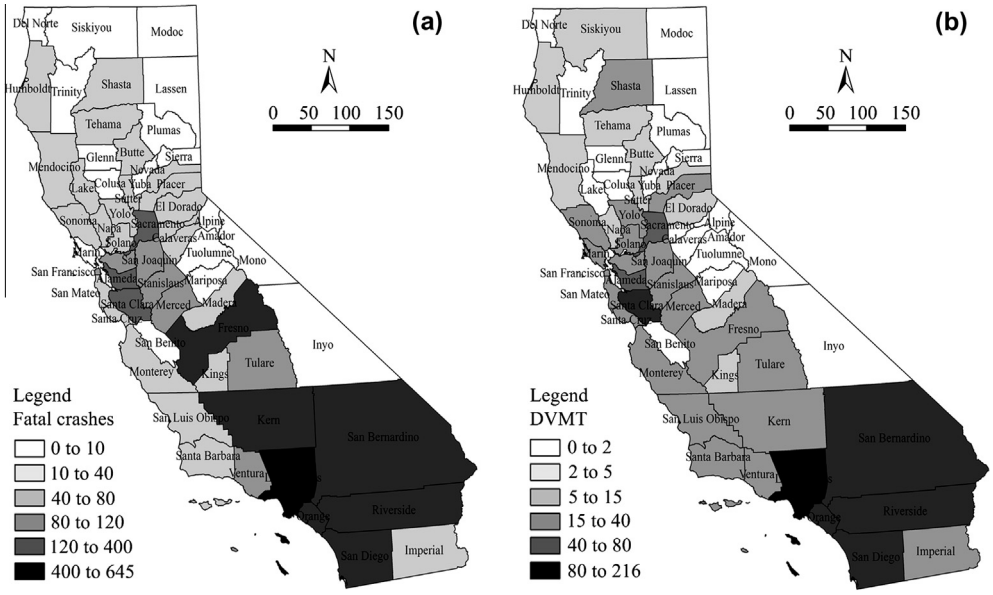

Fig. 1. (a) Yearly fatal crashes by county in California; and (b) yearly DVMT in thousand by county in California.

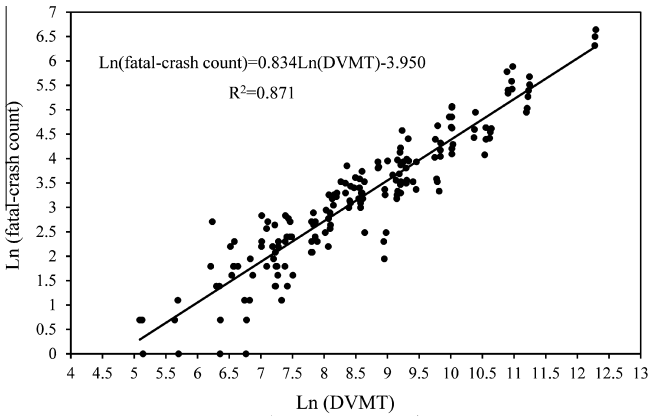

Fig. 2. Relationship between crash count and exposure variable.

into the model form. In this procedure, the variables were added into the model specification one by one, while monitoring the significance of these variables and the AICc of the model. Since one variable could be significant in several counties while insignificant in other counties, the rule that a variable was kept in the model if it was significant in more than 80% of counties was used in this study. Including insignificant variables into the model specification was found to increase the AICc of the model. The variable selection procedure was repeated several times. The GWPR with the smallest AICc was considered as the final model.

5.2. GWPR estimates

The GWPR with only DVMT as the explanatory variable was calibrated initially. The distribution of parameters of DVMT over counties is shown in Fig. 3a. It is identified that the parameters have an obvious pattern of spatial non-stationarity. The parameter of DVMT ranges from 0.65 to 0.90. All the parameter signs are positive indicating the DVMT has positive impacts on the number of fatal crashes per county. The positive impact of DVMT is consistent with most previous findings (Tarko et al., 1996; Karlaftis and Tarko, 1998; Traynor, 2008; Huang et al., 2010; Hadayeghi et al., 2010a), but contrary to a study by Aguero-Valverde and Jovanis (2006). The local *t*-statistics for the parameters of DVMT are computed to determine their significances. The results are shown in Fig. 3b. All the county-specific parameters of DVMT are significant at a 95% confidence level. The GWPR successfully captures the spatial heterogeneity in the relationship between fatal crashes and DVMT which is hidden in the global GLM.

In the preliminary analysis, several GWPRs with single-category and multiple categories of independent variables were calibrated and evaluated. The results show that in general the models with more variables produce smaller AICc as compared to these with single-category variables. It suggests that the GWPRs with more countywide predicting factors have better performances in fitting the data, though their applications may be limited due to the practical reality of data availability in some counties or regions. In this study, the GWPR with all categories of available variables was evaluated for the comparison to the GLM.

The summaries of parameter estimates in the GWPR are shown in Table 2. The local parameters are described by the 5-number summaries that present the minimum, lower quartile, median, upper quartile, and maximum of values. The distributions of parameters of predicting variables over the 58 California counties are shown in Fig. 4. It is identified that the parameters have obvious patterns of spatial variation. For most of variables, such as the freeway percentage, population density, percentage of age group under 18, traffic intensity, urban traffic percentage, and percentage of trucks/trailers, the parameters change gradually from northern counties to southern counties. For some variables, such as the logarithm of DVMT, road density, and median household income, their predicting powers on fatal crashes are more concentrated in the central counties than the other parts.

Table 1
Descriptive statistics of countywide variables.

Variable	Description	Min	Max	Mean	S.D.
Crash response variables					
Crash	Fatal crash count per year	1.25	614.25	51.62	93.25
Road network					
Road density	Road length/area (M)[a]	0.07	7.67	0.74	1.10
UR percent	Percent of urban road mileage	0.00	1.00	0.35	0.31
FW percent	Percent of freeway mileage	0.00	0.05	0.02	0.01
PA percent	Percent of principal arterial mileage	0.00	0.12	0.05	0.03
MA percent	Percent of minor arterial mileage	0.01	0.20	0.09	0.03
CR percent	Percent of collector road mileage	0.06	0.36	0.21	0.06
LR percent	Percent of local road mileage	0.47	0.76	0.63	0.06
Traffic					
DVMT	Daily vehicle miles traveled (T)	0.17	215.75	15.44	31.97
Traffic intensity	DVMT/road length (T)[a]	0.30	11.19	3.52	2.89
UT percent	Percent of urban DVMT	0.00	1.00	0.65	0.24
MVR density	Motor vehicle registration/area (T)	0.00	3.82	0.18	0.53
TRs percent	Percent of trucks and trailers	0.15	0.60	0.39	0.11
License rate	Licensed drivers/population	0.47	0.91	0.69	0.09
Socio-demographic					
Area	Area (M)	120	51,935	6964	8038
Pop density	Population/area (T)	0.00	6.67	0.25	0.90
Male	Percent of male population	0.48	0.64	0.51	0.02
Eighteen	Percent of age group under 18	0.14	0.33	0.24	0.05
Sixty-five	Percent of age group of 65 and older	0.08	0.25	0.13	0.04
MIC	Median household income (T)	34.44	86.83	53.33	13.46
Poverty	Percent of people below poverty line	0.07	0.22	0.14	0.04
Poverty18	Percent of people under 18 in poverty	0.02	0.10	0.05	0.02
UE rate	Unemployment rate	0.05	0.23	0.09	0.03
RUC	Rural-Urban Continuum	0.00	1.00	0.64	0.48

[a] M = in million, T = in thousand.

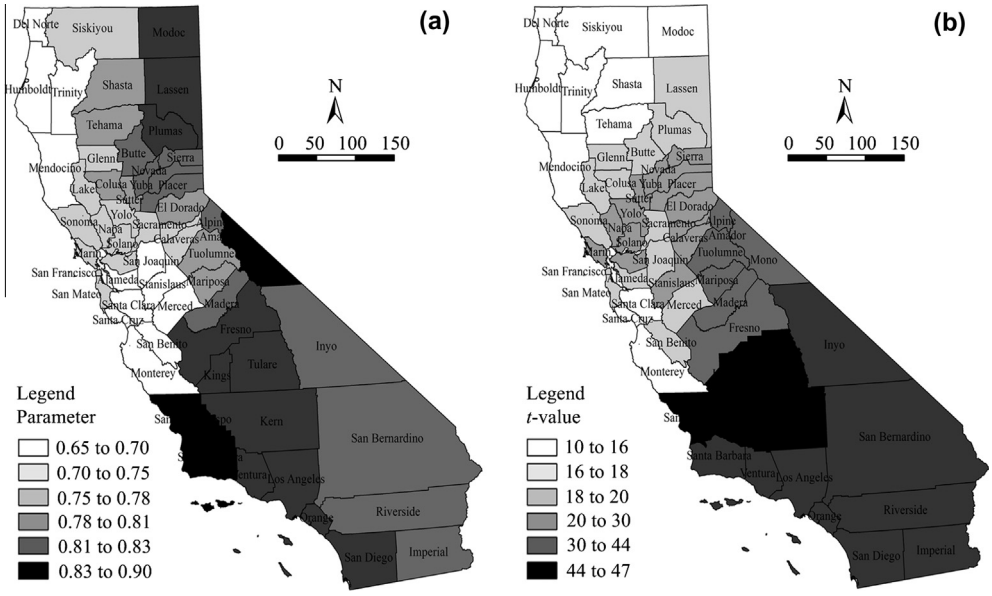

Fig. 3. (a) Parameters of DVMT by county; and (b) pseudo-*t* values for parameters of DVMT by county.

The percentage of freeway mileage in a county is negatively correlated with the fatal crashes. It suggests by controlling the DVMT, freeway produces less fatal crashes than other facilities. It could be because freeways are generally better designed and have full access control, while the other road types have numerous intersections and experience more traffic congestions which could increase the fatalities (Amoros et al., 2003; Huang et al., 2010). The road density is estimated to be negatively related to the fatal crash counts after controlling the DVMT.

The population density is positively related to the risk of fatal crash probably because more residents in an area have more activities that could result in more fatalities. Similar results were reported in several studies (Tarko et al., 1996; Hadayeghi et al., 2003; de Guevara et al., 2004), though Noland (2008) reported an opposite result in England. The percentage of age group under 18 has positive effects on fatal crashes since young population tend to take more risks in travels (Quddus, 2008; Aguero-Valverde and Jovanis, 2006; Huang et al., 2010). Higher median household income would decrease the risk of fatal crashes. It would be expected that individuals in wealthier areas seek to avoid risky activities and generally own cars with better safety performance (Huang et al., 2010).

Traffic intensity has negative coefficients suggesting the level of traffic congestion is negatively related to fatal crashes, which has also been reported in some studies (Hadayeghi et al., 2003; Noland and Oh, 2004), though a recent study reported a positive correlation (Huang et al., 2010a). The percentage of urban traffic is negatively related to fatal crashes, indicating that traffic in rural area is more likely to result in fatalities. The coefficients of percentage of trucks/trailers vary from negative to positive. The change of coefficient sign has been commonly observed in the application of the GWR or GWPR (Zhao et al., 2005; Wheeler and Calder, 2007; Hadayeghi et al., 2010a). The local *t*-statistics are computed and the results show that the negative coefficients are not significant at a 90% confidence level. Thus, more fatal crashes occur in counties with larger percentages of trucks and trailers.

5.3. GLM estimates

The GLMs with the random component follows a NB distribution were calibrated based on the same dataset. Initially, a GLM with only the DVTM as the explanatory variable was calibrated. The parameter is estimated to be 0.80 which is about the average value of parameters in the GWPR in Fig. 3a. Then a GLM with all variables contained in the GWPR were calibrated. The model estimates are shown in Table 3. Most of variables are significant at a 90% confidence level. A comparison between the two GLMs shows that the model with more explanatory variables has larger LR chi^2 and Pseudo R^2 and a smaller AIC value, which indicates a better statistical performance. The GLM with more variables was found to produce more accurate predictions of fatal crashes than the GLM with only DVMT.

In the GLM estimates in Table 3, the DVMT is the exposure variable in the GLM and is positively related to the fatal crashes. The percentage of freeway mileage has a negative impact on fatal crashes, and the total road density has a negative impact. The population density as well as the percentage of age group under 18 are positively related to crash risks. The median household income has a negative impact on crash risks. By controlling the exposure, the traffic intensity is estimated to be negatively related to crash counts. A higher percentage of urban traffic reduces the fatal crashes. The percentage of trucks/trailers is positively related to fatal crashes in a county.

Table 2
Summaries of local parameters in the GWPR.

Variable	Minimum	Lower quartile	Median	Upper quartile	Maximum
Ln(DVMT)	0.996	1.010	1.020	1.029	1.037
FW percent	−6.335	−6.041	−5.830	−4.605	−1.710
Road density (M)*	−0.027	−0.015	−0.011	−0.010	−0.008
Pop density (M)	0.247	0.262	0.276	0.293	0.393
Eighteen	1.262	1.534	1.653	1.713	1.791
MIC (T)	−0.012	−0.011	−0.010	−0.009	−0.008
Traffic intensity (T)	−0.085	−0.079	−0.076	−0.069	−0.019
UT percent	−0.457	−0.423	−0.399	−0.320	−0.050
TRs percent	−0.644	−0.089	0.285	0.814	1.178
Intercept	−5.460	−5.331	−4.997	−4.735	−4.292

* M = in million, T = in thousand.

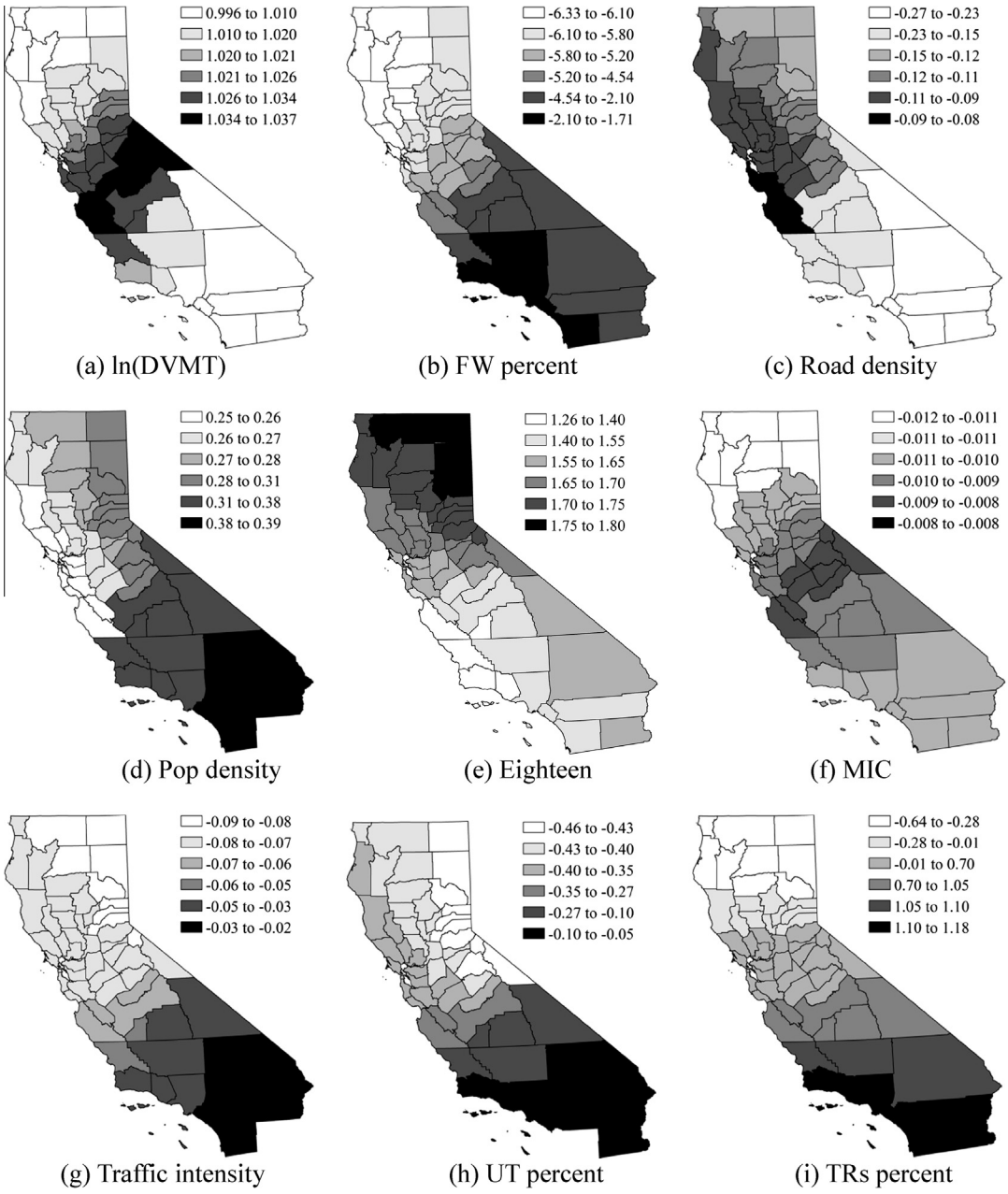

Fig. 4. Parameters of predicting variables by county in the GWPR.

5.4. Comparison between GWPR and GLM

The parameter estimates in the GLM were compared to these in the GWPR. The signs of parameters of predicting variables were found to be consistent between the two models. The difference between the two models is the GLM has a constant parameter for each variable while the GWPR has spatially varying parameters for each variable. The parameter of a variable in the GLM falls into the range of parameters of the same variable in the GWPR, indicating the parameter estimated in the GLM generally represents the average effect of the variable on fatal crashes in all counties. Thus, using the GLM, one crash prediction model was developed for all counties. While using the GWPR, different crash prediction models were developed for individual counties.

The predicting performances of the GLM and GWPR were compared. Using the GLM and GWPR calibrated in the above sections, the fatal crash counts of year 2010 in the 58 California counties were predicted for model validation. The distributions of residuals of predictions in the two models are shown in Fig. 5. It is observed that the residuals in the GWPR are obviously smaller than these in the GLM. The measures of goodness of fit introduced in Section 3.3 were computed for the two model predictions. The results are shown in Table 4. Both the MAD and MSPE in the GWPR are less than these in the GLM. The MAD and MSPE in the GWPR are reduced by 23.42% and 66.11% respectively as compared to the GLM. The results indicate that the GWPR produces more accurate predictions for fatal crash counts in individual counties than does the GLM. By capturing the spatial heterogeneity in the data, the variability of fatal crashes over counties is better predicted in the GWPR.

Both the GWPR and GLM assume the error term is independently distributed. If the spatial autocorrelation exists in the error term, the underlying model assumption is violated and biased estimates may be produced (Leung et al., 2010). In this

Table 3
Results of parameter estimates in the GLM.

Variable	Coeff.	S.E.	*t*	*p*-Value	95% Conf. interval	
Ln(DVMT)	1.022	0.054	18.960	<0.001	0.916	1.127
FW percent	−3.682	2.011	−1.782	0.083	−7.704	0.341
Road density (M)[a]	−0.025	0.009	−2.760	0.006	−0.042	−0.007
Pop density (M)	0.399	0.000	4.400	<0.001	0.000	0.001
Eighteen	1.751	0.804	2.180	0.030	0.174	3.327
MIC (T)	−0.010	0.004	−2.700	0.007	−0.016	−0.003
Traffic intensity (T)	−0.046	0.035	−1.300	0.192	−0.114	0.023
UT percent	−0.234	0.128	−1.792	0.085	−0.490	0.022
TRs percent	0.593	0.758	0.780	0.434	−0.892	2.078
Intercept	−5.250	0.704	−7.460	<0.001	−6.631	−3.870

Statistics:
Log likelihood: −166.9232.
LR chi^2(8): 241.69.
Prob > chi^2: <0.001.
Pseudo R^2: 0.4199.
[a] M = in million, T = in thousand.

Table 4
Measures of goodness of fit for the GWPR and GLM.

Measures of goodness of fit	MAD	MSPE
GWPR	13.14	516.61
GLM	17.16	858.12
Difference	23.43%	66.11%

Table 5
Moran's I statistics for residuals of predictions in the GWPR and GLM.

Model	Global Moran's I	Variance	*Z* Score	*p*-Value
GWPR	0.0113	0.0028	0.5496	0.5826
GLM	0.2029	0.0028	4.1677	<0.001

study, we computed the Moran's *I* statistics to quantify the spatial correlation in the residuals of predictions in the GWPR and GLM. The results are shown in Table 5. In the GWPR, the spatial correlation in the residuals of predictions of fatal crashes over counties is not significant at a 90% confidence level. However, in the GLM, a significant spatial correlation is found in the residuals of predictions at a 99.9% confidence level. The tests of Moran's *I* statistics suggest that because the GWPR accounts for the spatial heterogeneity in the county level data, the residuals of crash counts in the GWPR are less spatially correlated as compared to the GLM.

We also examined the spatial correlation for the fatal crashes. It is found that the crash counts are spatially correlated at a 99.9% confidence level. It suggests using the GWPR, the spatial correlation of fatal crashes is explained by the county-specific effects of predicting variables included in the model form. Thus, the residuals of predictions of fatal crashes are no longer spatially correlated. This finding is quite obvious by comparing the spatial distribution of fatal crashes in Fig. 1a and the residuals in Fig. 5a. The above analysis suggests that the GWPR is an appropriate technique for the modeling of the county-level crash data in California.

6. Summary and conclusions

This study evaluated the application of the GWPR modeling technique for the county-level crash data analysis. Based on the data collected from the 58 counties in the state of California, the GWPR was calibrated to explore the spatially varying relationships between fatal crashes and explanatory variables. A traditional GLM was also calibrated based on the same dataset. The GWPR and GLM were used to predict the fatal crashes in individual counties. The predictive performances of the two models were compared and the spatial correlations in the residuals of predictions were examined.

The results showed that the GWPR successfully captured the spatially non-stationary relationships between fatal crashes and predicting factors at the county level. The parameters of variables in the GWPR varied spatially, suggesting the effects of predictors on fatal crashes were different between counties. After considering the spatial heterogeneity in the county-level data, the GWPR outperformed the traditional GLM in predicting the fatal crashes in individual counties. The Moran's I tests showed that the GWPR remarkably reduced the level of spatial correlation in the residuals of predictions of fatal crashes over counties as compared to the GLM. This study suggested that the GWPR was more appropriate than the GLM for the crash frequency modeling based on the county level data with a small sample size.

The GWPR estimated the parameters of predicting variables for each county. Thus, the crash prediction model was developed particularly for every individual county. These crash prediction models are useful tools in evaluating the normal safety levels and

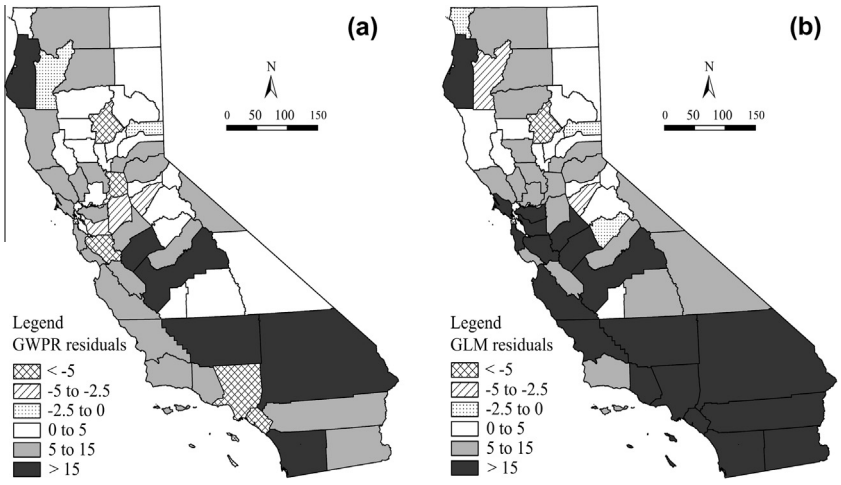

Fig. 5. (a) Residuals of fatal crashes by county in the GWPR; and (b) residuals of fatal crashes by county in the GLM.

forecasting the expected number of crashes in planning years in the counties of California. These models are also important in evaluating the effectiveness of policies or countermeasures applied in particular counties. The GWPR has an advantage that the technique has been built in the "GWRx3.0" software package. The findings of this study can help advance the progress of county-level transportation projects that incorporate safety into traditional planning process.

Although the GWPR is an excellent technique for predicting the number of county-level crashes, the calibrated models are not spatially transferable since they produce a set of local parameters for a specific geographic region. It indicates that most jurisdictions need to develop their own GWPRs for local regions. Besides, it would be interesting to compare the performance of a GWPR to other types of models such as the ransom parameter model and Bayesian spatial model which can also account for the spatial heterogeneity and spatial correlation in dataset. New findings are expected from the comparison between different models for the crash analysis at the county level. The authors recommend future studies could focus on these issues.

Acknowledgements

This research is supported by the National Key Basic Research Program (NKBRP) of China (No. 2012CB725400), the National High-tech R&D Program of China (863 Program) (No. 2012AA112304), as well as the Scientific Research Foundation of Graduate School of Southeast University (No. YBPY1211).

References

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in pennsylvania. Accident Analysis and Prevention 38 (3), 618–625.
Amoros, E., Martin, J.L., Laumon, B., 2003. Comparison of road crashes incidence and severity between some french counties. Accident Analysis and Prevention 35 (4), 537–547.
Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41, 153–159.
Chang, D.C., Eastman, B., Talamini, M.A., Osen, H.B., Tran Cao, H.S., Coimbra, R., 2011. Density of surgeons is significantly associated with reduced risk of deaths from motor vehicle crashes in us counties. Journal of the American College of Surgeons 212 (5), 862–866.
Charlton, M., Fotheringham, A.S., Brunsdon, C., 2003. Software for Geographically Weighted Regression. University of Newcastle, Newcastle, United Kingdom.
Chow, L.-F., Zhao, F., Liu, X., Li, M.-T., Ubaka, I., 2006. Transit ridership model based on geographically weighted regression. Transportation Research Record: Journal of the Transportation Research Board 1972, 105–114.
Darwiche, A., 2009. A GIS safety study and a county-level spatial analysis of crashes in the state of Florida. University of Central Florida.
de Guevara, F.L., Washington, S.P., Oh, J., 2004. Forecasting crashes at the planning level-simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board 1897, 191–199.
Donaldson, A.E., Cook, L.J., Hutchings, C.B., Dean, J.M., 2006. Crossing county lines: the impact of crash location and driver's residence on motor vehicle crash fatality. Accident Analysis and Prevention 38 (4), 723–727.
Du, H., Mulley, C., 2006. Relationship between transport accessibility and land value: local model approach with geographically weighted regression. Transportation Research Record: Journal of the Transportation Research Board 1977, 197–205.
El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41, 1118–1123.
FHWA (Federal Highway Administration), 2005. Safetea-LU: Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users. U.S. Department of Transportation.
Fotheringham, A.S., Brunsdon, C., Charlton, M.E., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.
Fridstrøm, L., Ingebrigtsen, S., 1991. An aggregate accident model based on pooled, regional time-series data. Accident Analysis and Prevention 23 (5), 363–378.
Hadayeghi, A., Shalaby, A.S., Persaud, H.N., 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record: Journal of the Transportation Research Board 1840, 87–95.
Hadayeghi, A., Shalaby, A.S., Persaud, B.N., Cheung, C., 2006. Temporal transferability and updating of zonal level accident prediction models. Accident Analysis and Prevention 38 (3), 579–589.
Hadayeghi, A., Shalaby, A., Persaud, B., 2010a. Development of planning-level transportation safety models using full Bayesian semiparametric additive techniques. Journal of Transportation Safety and Security 2, 45–68.
Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010b. Development of planning level transportation safety tools using geographically weighted poisson regression. Accident Analysis and Prevention 42 (2), 676–688.
Hanna, C.L., Laflamme, L., Bingham, C.R., 2012. Fatal crash involvement of unlicensed young drivers: County level differences according to material deprivation and urbanicity in the united states. Accident Analysis and Prevention 45, 291–295.
Huang, H.L., Abdel-Aty, M.A., Darwiche, A.L., 2010. County-level crash risk analysis in Florida Bayesian spatial modeling. Transportation Research Record: Journal of the Transportation Research Board 2148, 27–37.
Ibeas, A., Cordera, R., Dell'olio, L., Moura, J.L., 2011. Modelling demand in restricted parking zones. Transportation Research Part A – Policy and Practice 45 (6), 485–498.
Karlaftis, M.G., Tarko, A.P., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention 30 (4), 425–433.
Leung, Y., Mei, C.L., Zhang, W.X., 2010. Testing for spatial autocorrelation among the residuals of the geographically weighted regression. Environment and Planning A 32, 871–890.
Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the Estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38 (4), 751–766.
Lord, D., Mannering, F., 2010. The Statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291–305.
Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transportation Research Record: Journal of the Transportation Research Board 1717, 102–108.
Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26 (4), 471–482.
Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1), 17–23.
Nakaya, T., Fotheringham, A.S., Brunsdon, C., Charlton, M., 2005. Geographically weighted poisson regression for disease association mapping. Statistics in Medicine 24 (17), 2695–2717.
NCHRP (National Cooperative Highway Research Program), 2010. PLANSAFE: Forecasting the Safety Impacts of Socio-demographic Changes and Safety Countermeasures. Transportation Research Board, NCHRP, pp. 8–44.
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.
Pirdavani, A., Brijs, T., Bellemans, T., Wets, G., 2013. Spatial analysis of fatal and injury crashes in Flanders, Belgium: Application of geographically weighted regression technique. In: The 92th Annual Meeting of Transportation Research Board, Washington, DC.
Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. Accident Analysis and Prevention 40 (4), 1486–1497.
Tarko, A., Sinha, K., Farooq, O., 1996. Methodology for identifying highway safety problem areas. Transportation Research Record: Journal of the Transportation Research Board 1542, 49–53.
Traynor, T.L., 2008. Regional economic conditions and crash fatality rates – a cross-county analysis. Journal of Safety Research 39 (1), 33–39.
Wang, Y.Y., Kockelman, K., Wang, X.K., 2011. Anticipating land use change using geographically weighted regression models for discrete response. In: The 90th Annual Meeting of the Transportation Research Board, Washington, DC.
Washington, S., Karlaftis, M., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.
Wheeler, D., Calder, C., 2007. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. Journal of Geographical Systems 9 (2), 145–166.
Wood, G.R., 2002. Generalised linear accident models and goodness of fit testing. Accident Analysis and Prevention 34 (4), 417–427.
Zhang, Y., Bigham, J., Li, Z., Ragland, D., Chen, X., 2012. Associations between road network connectivity and pedestrian-bicyclist accidents. In: The 91th Annual Meeting of Transportation Research Board, Washington, DC.
Zhao, F., Chow, L., Li, M., Liu, X., 2005. A Transit ridership model based on geographically weighted regression and service quality variables. Report DO97591, Lehman Center for Transportation Research, Department of Civil and Environmental Engineering, Florida International University.
Zhao, F., Park, N., 2004. Using geographically weighted regression models to estimate annual average daily traffic. Transportation Research Record: Journal of the Transportation Research Board 1879, 99–107.