CHALMERS
UNIVERSITY OF TECHNOLOGY

Geographically weighted machine learning for modeling spatial heterogeneity in traffic crash frequency and determinants in US

Accident Analysis and Prevention 199 (2024) 107528

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Geographically weighted machine learning for modeling spatial heterogeneity in traffic crash frequency and determinants in US

Shuli Wang [a,b], Kun Gao [b,*], Lanfang Zhang [a,*], Bo Yu [a], Said M. Easa [c]

[a] Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, CN-201804, China
[b] Department of Architecture and Civil Engineering, Chalmers University of Technology, Goteburg SE-412 96, Sweden
[c] Department of Civil Engineering, Toronto Metropolitan University, Toronto M5B 2K3, Canada

ARTICLE INFO

Keywords:
Traffic crash frequency
Spatial machine learning
Spatial heterogeneity
Interpretability

ABSTRACT

Spatial analyses of traffic crashes have drawn much interest due to the nature of the spatial dependence and spatial heterogeneity in the crash data. This study makes the best of Geographically Weighted Random Forest (GW-RF) model to explore the local associations between crash frequency and various influencing factors in the US, including road network attributes, socio-economic characteristics, and land use factors collected from multiple data sources. Special emphasis is put on modeling the spatial heterogeneity in the effects of a factor on crash frequency in different geographical areas in a data-driven way. The GW-RF model outperforms global models (e.g. Random Forest) and conventional geographically weighted regression, demonstrating superior predictive accuracy and elucidating spatial variations. The GW-RF model reveals spatial distinctions in the effects of certain factors on crash frequency. For example, the importance of intersection density varies significantly across regions, with high significance in the southern and northeastern areas. Low-grade road density emerges as influential in specific cities. The findings highlight the significance of different factors in influencing crash frequency across zones. Road network factors, particularly intersection density, exhibit high importance universally, while socioeconomic variables demonstrate moderate effects. Interestingly, land use variables show relatively lower importance. The outcomes could help to allocate resources and implement tailored interventions to reduce the likelihood of crashes.

1. Introduction

Road traffic accidents annually cause a substantial toll on human lives and well-being, resulting in considerable societal costs (Ziakopoulos and Yannis, 2020). The United States (US) witnesses a sobering statistic of over 30,000 fatalities attributable to traffic crashes, reflecting a mortality rate of 23 per 1 million individuals (Abdel-Aty et al., 2013). Despite a noticeable decrease in the incidence of traffic crashes in the US, the substantial number of casualties underscores the imperative for continuous efforts in comprehensively analyzing the underlying factors or determinants of these incidents. This is aimed to formulate and implement effective countermeasures for mitigating the deleterious consequences associated with traffic crashes.

Traffic crash frequency varies substantially between and within states in the US. Existing spatial analyses of traffic crashes predominantly involve the investigation of crash counts or frequencies across spatial units (Lord and Mannering, 2010). Existing research acknowledges the spatial dependence and heterogeneity in crash occurrences (Huang and Abdel-Aty, 2010). Spatial dependence, in this context, denotes the phenomenon that crashes at a location are highly influenced by events at neighboring locations, often quantified through spatial auto-correlation metrics. Some existing studies analyzed spatial dependence of traffic crash frequencies, effectively representing local conditions (Lord and Mannering, 2010; Mannering and Bhat, 2014). However, as far as we are concerned, few studies have not revealed the spatial heterogeneity in the effects of influence factors (e.g. road environment and land use) on the traffic crash frequencies. Herein, the spatial heterogeneity in the effects of a factor denotes that the effect of the same factor on traffic crashes varies in different areas (or spatially). Modelling and analyzing the potential spatial heterogeneity in the associations between crash frequency and various factors could help to capture unobserved trends and particularities of each area spatially. The results would allow for a more precise crash frequency analysis and help policy-makers for implementing tailored safety measures effectively to reduce the likelihood of crashes in different spatial areas.

There is research on spatial modeling of traffic crashes to demonstrate associations between spatial-level crash frequency and

* Corresponding authors.
E-mail addresses: gkun@chalmers.se (K. Gao), zlf2276@tongji.edu.cn (L. Zhang).

https://doi.org/10.1016/j.aap.2024.107528
Received 21 December 2023; Received in revised form 5 February 2024; Accepted 25 February 2024
Available online 5 March 2024

influencing factors such as land use factors (Cai et al., 2016), road environment factors (Bao et al., 2018; Wang et al., 2016), and demographics and socioeconomic factors (Pervaz et al., 2022). Some studies have suggested various statistical regression models, such as generalized linear models (GLMs) (Chiou and Fu, 2013) and random-parameter models (Amoh-Gyimah et al., 2017; Zeng et al., 2017). These models address spatial heterogeneity by introducing random effects or allowing the parameters to vary across different spatial zones or groups. To consider spatial autocorrelation issues, some studies have employed autoregressive (AR) models, e.g., conditional autoregressive (CAR) models and spatial autoregressive (SAR) models, to consider the correlation of crash data in different spatial units to introduce the spatial structure random effect and spatial lag effect into the traditional statistical model, respectively (Chiou et al., 2014; Jonathan et al., 2016; Wen et al., 2019; Luo et al., 2023). Recently, the process of Bayesian inference has been employed for modeling spatially aggregated crash data. Bayesian spatial models can simultaneously account for the spatial correlation and heterogeneity in crash data by integrating spatial random effects and spatial covariance functions (Ma et al., 2017; Wang et al., 2022). However, Bayesian spatial models require prior knowledge to select appropriate prior distributions for random parameters; otherwise, it may lead to inaccurate modeling results.

Another widely used method is the geographically weighted regression (GWR) model, which extends traditional statistical regression methods to incorporate spatial effects of factors in their structure. The method that accounts for spatial variation is the simultaneous development of several localized regression models considering spatial distance weights (Liu et al., 2017). These GWR models have good theoretical interpretability, allowing for a direct and clear understanding of the relationship between crash frequency and the analyzed influencing factors for different areas. However, their main disadvantage is that the linear models are susceptible to outliers and require strong assumptions about the linear relationship between exploratory variables and the dependent variable, as well as the multicollinearity among exploratory variables. Nowadays, the prevalence of machine learning technologies have enabled the combination of geographically weighted structures and machine learning models (Quiñones et al., 2021; Santos et al., 2019). The geographically weighted Random Forest (GW-RF) model has the potential to address the limitations of the GWR model, as it allows for modeling nonlinear relationships between crash frequency and influencing factors and exhibits robustness to outliers in the crash data (Wu et al., 2024). Meanwhile, the GW-RF endeavors to improve predictive performance over a non-geographically weighted RF model by accounting for spatial heterogeneity in the effects of influencing factors.

To the best of our knowledge, extant literature lacks endeavors utilizing the Geographically Weighted Random Forest (GW-RF) model to investigate non-stationarity in the relationships between zone-level crash frequency and diverse influencing factors. This study addresses this gap by deploying a GW-RF model, designed to account for spatial correlation and heterogeneity within crash data. The objective is to meticulously examine the local associations between zone-level crash frequency and a spectrum of influencing factors, discerning spatial variations in these associations. This study draws upon multiple datasets from open data platforms in the United States, encompassing crash data and influencing factors including road network attributes, socio-demographic characteristics, and land use variables, aggregated at the granularity of ZIP Code Tabulation Areas (ZCTAs). Firstly, the calculation of Moran's I statistic index is undertaken to scrutinize local relationships between crash frequency and safety influencing factors. Subsequently, the investigation of associations between influencing factors and zone-level crash frequency is conducted through both local geographically weighted models (GW-RF and GWR) and global models (RF and Ordinary Least Squares regression). The comparative assessment aims to compare the predictive performance of GW-RF relative to conventional local and global models. Lastly, the GW-RF model is employed to elucidate variations in the local effects of influencing factors on crash frequency, employing interpretive techniques. The culmination of these analyses yields nuanced insights into the underlying determinants of crash frequency in different geographical areas. The outcomes hold promise for effectively mitigating the likelihood of crashes and enhancing road safety within specific spatial domains through tailored prevention and interventions.

The remainder of this paper is structured as follows. Section 2 overviews relevant studies about safety influencing factors and spatial modeling approaches on traffic crash frequency. Section 3 introduces the study area, multiple data sources used, and candidate influencing factors. The analysis and model specifications of RF, GWR, and GW-RF are introduced in Section 4. The results and conclusions are presented in Section 5 and Section 6, respectively.

2. Literature review

2.1. Safety influencing factors

Traffic crashes result from a multifaceted process influenced by various factors such as road network attributes and driver characteristics. From a spatial perspective, a thorough analysis of the influencing factors on zonal traffic crashes becomes imperative, aiding researchers in identifying factors impacting traffic crashes in diverse areas and subsequently enabling the targeted implementation of improvements specific to each location. Previous studies have investigated the contributory factors to crashes at spatially aggregated levels. The safety influencing factors can be broadly categorized into four general categories: road network attributes, traffic states, socio-demographic characteristics, and land use characteristics. Notably, road network attributes stand out as pivotal factors in the spatial modeling of traffic crashes (Xu et al., 2017). In the traffic crash frequency models, road network attributes are often characterized by variables such as the length, density, and proportion of different road segment types. Extant literature suggested that factors such as freeway-segment length and intersection density exert a significant positive influence on the occurrence of traffic crashes (Wang et al., 2022). Moreover, various traffic states are relevant to the modeling of traffic crashes, including average hourly traffic volume (AHTV), annual average daily traffic (AADT), average speed, speed variance, and speed limit (Abdel-Aty et al., 2013). While the relationship between traffic volume and crash frequency is widely acknowledged, the influence of average speed and speed limits remains a point of contention. Elevated traffic volumes are observed to correlate with an increased frequency of crashes increase crash frequency, but the impact of average speed and speed limits remains inconclusive (Bao et al., 2017). Moreover, spatial modeling has explored various socio-demographic characteristic (Bao et al., 2017; Lee and Abdel-Aty, 2018), such as population density, age, gender, education attainment, unemployment rate, income condition, etc. Other noteworthy findings include a positive correlation between employment density and crash frequency, while other studies have reported associations indicating that higher levels of education and favorable income conditions are related to a reduction in crash occurrences (Cai et al., 2017).

Concerning land use factors, some studies have revealed that areas characterized by intensified commercial activities exhibit higher traffic crash frequencies (Gomes et al., 2017). Due to the difficulty in data collection for the areas of diverse land use categories, the traditional method directly utilizes the number of Points of Interests (POIs) in distinct categories to represent the land use ratios of these categories. However, it is crucial to note that this method lacks precision, as the accurate determination of land use ratios is defined as the area allocated to a specific land use divided by the total space in the study area (Gao et al., 2021). The amount of POIs in different categories provided by the online map is often highly imbalanced, which may lead to biases in calculating land use ratios. To address the imbalanced

nature of online map POI data, it is necessary to propose a more accurate method for computing the ratios of diverse land use categories within each study area.

However, numerous existing studies have primarily concentrated on the effect of influencing factors on the dependent variable (crash frequency) without fully accounting for the nonlinear and interactive effects of these factors on crash frequency. Therefore, there is a critical need to introduce an innovative method to comprehensively address these intricate relationships of influencing factors on crash frequency to gain a more precise understanding of how influencing factors affect traffic crashes and, consequently, develop more effective strategies for addressing traffic crash issues.

2.2. Spatial modeling approaches

Various methods have been developed to account for spatial dependence and heterogeneity, incorporating these spatial characteristics when modeling spatially aggregated crash data. Spatial autocorrelation of crashes is often initially assessed in spatial analyses by selecting the appropriate scale of spatial units for analysis. To accurately examine autocorrelation phenomena, researchers have employed various geographic spatial statistical methods, such as Moran's I, Local Moran's I, and Getis–Ord-Gi* statistics (Wen et al., 2019). Moreover, to incorporate spatial characteristics, numerous spatial models have been proposed, primarily falling into five categories, including generalized linear models (GLMs), autoregressive models, random parameter models, Bayesian spatial models, GWR models, and Machine learning.

Some early studies sought to explain the unobserved spatial heterogeneity in crash data using GLMs (Chiou and Fu, 2013), such as random effects Poisson or negative binomial model. These models introduced random effects into traditional statistical models to represent random variations between different observed spatial units, often denoted as random intercepts. Random effects allow the model to accommodate heterogeneity between different spatial units, thus providing a better explanation for spatial heterogeneity in crash data. However, it is crucial to acknowledge that these models rest on the assumption that crashes are independent, random events. This assumption poses a potential challenge, especially in the case of spatially correlated crash data, thereby raising concerns about the suitability of these models in accurately capturing the complex spatial relationships among crash events. To address spatial autocorrelation issues, researchers have turned to conditional autoregressive (CAR) models, incorporating spatial structure random effects to enhance the random effects defined to follow a normal distribution in the basic Poisson model (Chiou et al., 2014; Jonathan et al., 2016). In the CAR model, the spatial structure random effect of each spatial unit follows a CAR prior distribution, which is computed based on the adjacency matrix and globally smoothed over the space. CAR models have demonstrated superior performance compared to Poisson models, particularly in processing the discrete nature of spatially aggregated crash data (El-Basyouny and Sayed, 2009). On the other hand, some studies have employed spatial autoregressive (SAR) models, considering that the crash frequency (dependent variable) of a spatial unit exhibits an interactive effect not only with the explanatory variable of the same spatial unit but also with the crash frequency of adjacent spatial units (spatial lag), which aims to account for the spatial autocorrelation inherent in crash data (Wen et al., 2019). Nevertheless, when dealing with a substantial number of geographic zones, both CAR and SAR models require estimating a significant number of spatial autocorrelation parameters, which can lead to an escalation in the complexity of the model. Differing from the aforementioned the fixed-parameter models, random-parameter models tackle heterogeneities from unobserved factors by allowing the parameters to vary across distinct spatial zones or groups (Amoh-Gyimah et al., 2017). Numerous studies indicated that random-parameter models outperform traditional fixed-parameter models in both goodness of fit and practical guidance (Zeng et al., 2017). The estimation of parameters in random-parameter models typically involves estimating them independently for each observational unit, which may demand considerable data and computational resources.

In recent years, Bayesian spatial models have been employed for modeling spatially aggregated crash data. Relevant studies have indicated that models with Bayesian approaches consistently outperform their non-Bayesian counterparts due to their advantages in handling complex spatial structures of spatial data, particularly when considering random effects and spatial autocorrelation (Wang and Huang, 2016). Some works developed CAR models within a Bayesian framework, which have been proven to be effective in simultaneously accounting for spatial correlation and unobserved heterogeneity in aggregated crash data. This capability enables a thorough investigation of influential factors associated with crash frequency (Quddus, 2008). Furthermore, some studies have integrated spatial random effects and spatial covariance functions into Bayesian spatial models, which can achieve a better understanding and modeling of the correlation and heterogeneity in spatially aggregated crash data (Ma et al., 2017). In such models, spatial covariance functions are employed to quantify the correlation between different spatial units. However, both random-parameter models and Bayesian spatial models are relatively intricate and require the specification of the prior distribution for the random parameters.

Another commonly used method is the geographically weighted regression (GWR) model, which extends traditional statistical regression methods to incorporate spatial effects of influencing factors in their structure. Several localized regression models utilizing GWR have been developed to account for spatial variation, such as geographically weighted ordinary least squares regression (GW-OLS) (Pirdavani et al., 2014) and geographically weighted negative binomial regression (GWNBR) (Gomes et al., 2017; Li et al., 2010). To enhance the spatial transferability of the GWR model, researchers have endeavored to extend GWR to semiparametric GWR (SGWR) (Xu and Huang, 2015), which combines geographically varying parameters with geographically constant parameters. These GWR models, appreciated for their strong theoretical interpretability, offer a direct and clear understanding of the relationship between crash frequency and analyzed safety influencing factors. However, a main drawback lies in their assumption of a linear relationship between influencing factors and crash frequency, which may not accurately reflect the complexity of real-world scenarios. There is a urgent need for further exploration to model the complex nonlinear relationships of influencing factors with crash frequency, considering the spatial heterogeneity in the effects of these factors.

Machine learning (ML) methods, recognized for their potency and popularity as data-driven prediction tools, have been increasingly utilized in crash analysis. ML methods exhibit greater flexibility and robustness in handling data, unburdened by the assumptions and constraints of traditional statistical methods (Qu et al., 2023; Gao et al., 2023). The prevalence of machine learning technologies has facilitated the combination of geographically weighted structures and machine learning models, presenting a promising direction for advancing the field of crash analysis. Indeed, a noteworthy advancement in the field is the recent development of geographically weighted random forest (GW-RF), specifically applied to explore and visualize relationships between exploratory variables and target variables at the spatial level (Quiñones et al., 2021; Wu et al., 2024). The pioneering work of Wu et al. (2024) utilized the GW-RF model to predict crash number from London, showing that the GW-RF model has good predictive performance when selecting the appropriate bandwidth. The GW-RF model, a tree-based non-parametric ensemble model, offers a valuable solution to the limitations associated with linear GWR models. The GW-RF model not only improves modeling and predictive performance but also enables a comprehensive investigation of spatial heterogeneity in the effects of exploratory variables on target variables. This dual capability makes the GW-RF model a potential tool for advancing our understanding of how various factors contribute to crash frequencies across different spatial zones.

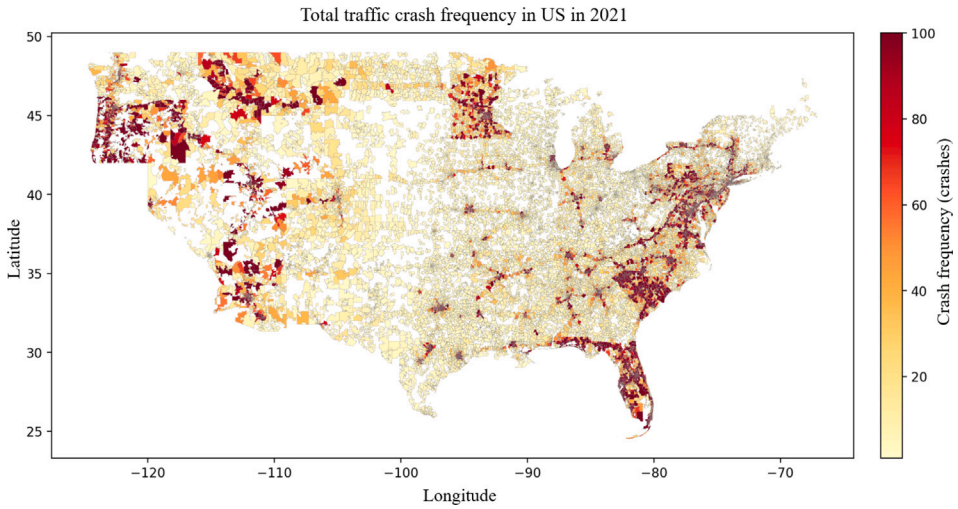


Fig. 1. Spatial distribution of crashes across different ZCTAs in US in 2021.

3. Study areas and data description

The study area encompasses the entire US, and the study period spans from January 1st to December 31st, 2021. In this study, the ZIP Code Tabulation Area (ZCTA) serves as the basic zone unit of analysis, which has been widely adopted in previous studies and is recommended as a reasonable zoning scale for spatial analysis involving crashes and human activities (Qian and Ukkusuri, 2015; Bao et al., 2021). ZCTAs are built by the Census Bureau through the aggregation of census blocks with common postal addresses assigned to streets. The average coverage area of ZCTA is 451 m^2, which is generally larger than a Census Tract yet smaller than a Census County Division. This size provides a stable and moderately sized geographic unit for spatial modeling of traffic crashes. The study area excludes Alaska and Hawaii due to the limited number of crash observations in these regions. The final dataset included 18,411 ZCTAs across the US.

The following four types of data are utilized: traffic crash data, road network attributes, social-demographic information, and land use features. The crash data were collected from the Kaggle platform using two APIs (MapQuest and Bing) with traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. Each crash record in the data consists of a unique crash ID, the starting and ending time, crash severity, and location coordinates. To facilitate analysis and modeling at the ZCTA level, a reverse geocoding method (Nominatim) is employed to obtain the ZIP code of the crash occurrence location based on its coordinates. Subsequently, the crash data can be aggregated into the respective ZCTAs for further analysis. The used dataset comprises a total of 1,144,4991 crashes during the selected period in the study area (18,411 ZCTAs). Fig. 1 illustrates the distribution of total crashes across various ZCTAs. It is observed that the number of crashes in a zone is strongly correlated with the area of the zone; larger areas tend to exhibit higher crash counts. However, the study units (ZCTAs) do not have a fixed area. To address this variability, an area-adjusted crash frequency is proposed as the dependent variable, defined as the number of crashes per 100 m^2 within each ZCTA. Consequently, the dependent variable in the analysis is comparable among zones with different areas, making it more suitable for investigating the effects of influencing factors.

To explore the effects of zone-level factors on traffic crash frequency, various road network attributes, demographics and socioeconomic factors, and land use factors are collected from multiple data sources. The adopted factors are determined based on our available datasets and cover most factors that may influence crash frequency, as reported in the literature (Ziakopoulos and Yannis, 2020). The definitions of candidate factors are summarized in Table 1.

The road network data and the land use data are collected from the Open Street Map (OSM). The road network data are obtained from the ArcGIS shape files depicting the road network attributes. The information provided by OSM includes the length and road type of each road segment, with a total of 28 road types defined by OSM. For analysis, each road segment is categorized into six road types based on the road classification standard in the US. The utilized road types consist of motorways, primary roads, secondary roads, tertiary roads, residential roads, and service roads. The matchups between road types in this study and road types from OSM are detailed in Table A.1. To calculate road network characteristics within each ZIP Code Tabulation Area (ZCTA), different types of road segments are delineated by the boundaries of the selected ZCTAs using tools provided by ArcGIS (Bi et al., 2022).

As for the land use data, we utilize the point of interest (POI) data from OSM. The information of each POI consists of the element name, element type, and location coordinates (longitude and latitude). There are 147 categories of POI defined by OSM to represent different utilization purposes, such as residence, commerce, entertainment, education, and industry. Further, we categorize each POI into four land-use categories for analysis, referring to the classification standards of the major land-use classification system in the United States Geological Survey (USGS). The investigated land-use types comprise residential, industrial, communication and utility, and commercial and service land use. The matching between the land use categories and the POI categories from OSM is presented in Table A.2. By mapping the POIs into the respective ZCTAs, the number of POIs of different categories in each ZCTA can be extracted and utilized to represent the land use characteristics in each ZCTA. Specifically, the ratios of different land use categories are calculated using the term frequency-inverse document frequency (TFIDF) method, as proposed by Gao et al. (2021). This method draws inspiration from numerical statistical models commonly used to reflect the importance of a word in a set of documents or a corpus. In the context of this study, each ZCTA is treated as a document, and each POI category is regarded as a word within that document. Thus, the TFIDF method allows for the determination of the significance of a certain category of POI, analogous to the importance of a specific word in a document, thereby addressing the imbalanced nature of POI data. The degree of a category of POI in a ZCTA can be obtained by:

$$tdidf_{ki} = td_{ki} \times idf_k \tag{1a}$$

$$td_{ki} = \frac{N_{ki}}{\sum_{k=1}^{K} N_{ki}} \tag{1b}$$

Table 1
The candidate dependent variables and influencing factors and their definitions.

Variables Name	Symbols	Definitions	Unit
Dependent variables			
Total crash frequency	TCF	Total number of crashes in each ZCTA	crashes
Area-adjusted crash frequency	ACF	The number of crashes per 100 km^2 in each ZCTA	crash/100 km^2
Road network variables			
Motorway density	MD	The length of the motorway divided by area in each ZCTA	km/km^2
Primary road density	PRD	The length of the primary road divided by area in each ZCTA	km/km^2
Secondary road density	SDRD	The length of the secondary road divided by area in each ZCTA	km/km^2
Tertiary road density	TRD	The length of the tertiary road divided by area in each ZCTA	km/km^2
Residential road density	RRD	The length of residential road divided by area in each ZCTA	km/km^2
Service road density	SVRD	The length of the service road divided by area in each ZCTA	km/km^2
Intersection density	ISD	The number of total intersections divided by area in each ZCTA	number/km^2
Signal intersection rate	SIR	The number of signalized intersections divided by the number of total intersections in each ZCTA	%
Demographics and Socioeconomic variables			
Total population	TP	The total number of people in each ZCTA	persons
Population density	PD	The number of people divided by area in each ZCTA	persons/km^2
Bachelor proportion	BP	The proportion of people with bachelor's degrees in each ZCTA	%
High school proportion	HSP	The proportion of people who graduated from high school in each ZCTA	%
Labor force proportion	LFP	The proportion of labor force in each ZCTA	%
Poverty rate	PR	The proportion of people under poverty in each ZCTA	%
Unemployment rate	UER	The rate of unemployment people in the labor force in each ZCTA	%
Median household income	MHI	Median household income in each ZCTA	dollars
Commuting proportion	CP	The proportion of commute people in each ZCTA	%
Average travel time to work	ATTW	Average travel time to work in each ZCTA	minutes
White proportion	WP	The proportion of White people in each ZCTA	%
Black and African American proportion	BAAP	The proportion of Black and African American people in each ZCTA	%
American Indian proportion	AIP	The proportion of American Indian people in each ZCTA	%
Asian proportion	AP	The proportion of Asian people in each ZCTA	%
Young proportion	YP	The proportion of young people (15-24 years old) in each ZCTA	%
Prime adult proportion	PAP	The proportion of prime adults (25-44 years old) in each ZCTA	%
Middle-aged adult proportion	MAP	The proportion of middle-aged adults (45-64 years old) in each ZCTA	%
Elder proportion	EP	The proportion of elder people over 65 years old in each ZCTA	%
Land use variables			
Residential land use ratio	RLUR	The ratio of the area allocated for residential land use in each ZCTA	%
Commercial and services land use ratio	CSLUR	The ratio of the area allocated for commercial and services land use in each ZCTA	%
Industry land use ratio	ILUR	The ratio of the area allocated for industrial land use in each ZCTA	%
Transportation, communications, and utilities land use ratio	TCULUR	The ratio of the area allocated for transportation, communications, and utilities land use in each ZCTA	%

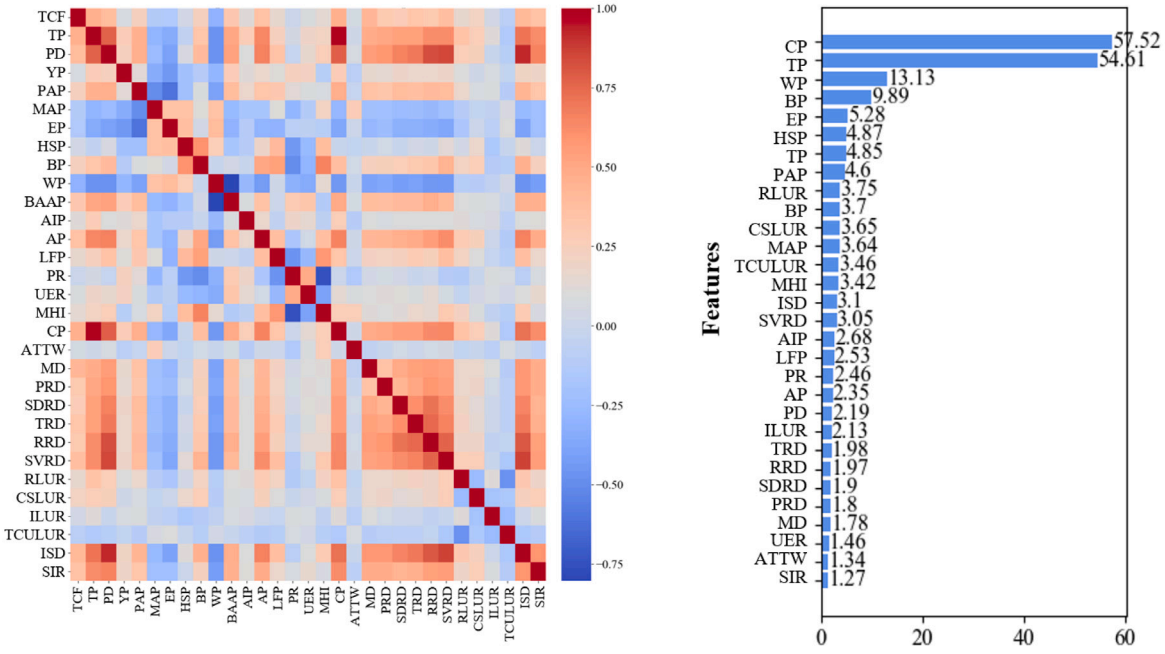
$$idf_k = log\frac{\sum_{i=1}^{D} N_{ki}}{\sum_{i=1}^{D}\sum_{k=1}^{K} N_{ki}} \tag{1c}$$

where td_{ki} and idf_k is the occurrence frequency and weight of POI category k, respectively; N_{ki} is the number of POIs belonging to POI category k in ZCTA i, K is the amount of POI categories in ZCTA i, D is the number of ZCTAs in the study area. The ratio of the land use category k can be calculated by:

$$R_{ki} = \frac{tdidf_{ki}}{\sum_{k=1}^{K} tdidf_{ki}} \tag{2}$$

where R_{ki} is the ratio of land use category k in ZCTA i. Based on this method, the ratios of the four land use categories studied are obtained, as shown in Table 1. Furthermore, the social-demographic data are acquired from the U.S. Census Bureau. The used information encompasses the number of people segregated by age, education attainment, labor force participation, poverty level, unemployment population, median household income, and the average travel time to work. The social-demographic data are aggregated into the corresponding ZCTAs for analysis.

The variance inflation factors (VIF) are calculated for all factors to execute a multicollinearity check. As depicted in Fig. 2(b), the VIF values for total population, white proportion and commuting proportion are larger than 10, indicating multicollinearity with other variables (Bao et al., 2018; Azimian et al., 2021). Considering that only the VIF value of white proportion is slightly greater than 10 in this study, we relax the threshold constraint and retain the variable. Consequently, two variables are excluded from the analysis. Besides, variables exhibiting a high correlation (Pearson correlation values exceeding 0.7) are also removed. The results of the correlation test, as shown in Fig. 2(a), reveal a strong negative correlation between the poverty rate and median household income. Despite this correlation, the poverty



(a) Estimated correlated coefficients among variables (b) VIF

Fig. 2. Correlation and collinearity test results of variables.

rate is retained for analysis as it is a variable of interest in many related studies. After addressing multicollinearity and high correlations, a set of 27 variables remains for further analysis.

4. Methodology

We applied two local spatial models (GW-RF and GWR) and two global models (Random Forest and OLS). The predictive performance of GW-RF is evaluated by comparing it with traditional local and global models. Moreover, for both the RF and GW-RF models, the importance ranking of each variable is calculated to explore the association between crash frequency and influencing factors and discern how this association varies across ZCTAs.

4.1. Moran's I statistic

To explore the intrinsic spatial autocorrelation of factors, Moran's I statistic is employed, including global Moran's I (Wang et al., 2019; Bao et al., 2017) and local Moran's I (Yuan et al., 2018). The global Moran's I index is adopted to determine whether the explanatory variables in the ZCTA-level traffic crash frequency model are spatially correlated. Global Moran's I index can be calculated as

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(z_i - \bar{z})(z_j - \bar{z})}{\sigma^2 S_0} \tag{3a}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \tag{3b}$$

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} \tag{3c}$$

where n denotes the total number of ZCTAs, z_i represents the observed value of a variable at ZCTA i, $\bar{Z}$ means the global average value of the variable at all ZCTAs, σ^2 denotes the variance of z, $w_{i,j}$ is the spatial proximity weight between ZCTA i and j. Herein, we used the most prevalent structure, 0–1 first-order neighbor, to obtain the spatial proximity weight. Specifically, if ZCTA i and j are connected to each other directly, $w_{i,j} = 1$; otherwise, $w_{i,j} = 0$. S_0 means the aggregation of all spatial weights.

The score of the statistic z_I is calculated as

$$z_I = \frac{I - E[I]}{\sqrt{V[I]}} \tag{4}$$

where

$$E[I] = -1/n - 1 \tag{5a}$$

$$V[I] = E[I^2] - E[I]^2 \tag{5b}$$

The value of Moran's I ranges from −1 to 1. A positive Moran's I indicates positive spatial correlation, and the larger the value, the more pronounced the spatial correlation, while a negative Moran's I indicates negative spatial correlation.

Moreover, we calculate bivariate local Moran's I (BLMI) (Yuan et al., 2018) to explore the degree of spatial correlation (positive or negative) between crash frequency in a given ZCTA and independent variables in its neighboring ZCTAs. The results of BLMI are categorized into four categories: low-low (LL), low-high (LH), high-low (HL) and high-high (HH). For example, ZCTAs with HH refer to significant clusters of high crash frequency that also have high values of the independent variable in neighboring ZCTAs. BLMI is expressed as:

$$I = \frac{(z_i - \bar{z})}{\sigma^2}\sum_{j=1, j\neq i}^{n} w_{i,j}(x_i - \bar{x}) \tag{6}$$

where x_j represents the observed value of an independent variable at ZCTA j, $\bar{x}$ means the global average value of the independent variable at all ZCTAs.

4.2. Random forest (global model)

Herein, a global model refers to a model that treats data points in different spatial areas uniformly and does not consider the spatial heterogeneity in the effects of a factor on the dependent variable, when learning and making predictions. Among various types of supervised machine learning, we employ Random Forest (RF), a widely applied

non-parametric machine-learning method for regression analysis (Luo et al., 2021). The RF is an ensemble of decision trees constructed through random feature selection and random sample sampling, improving the robustness and generalization ability of the model. Since each decision tree is built on random samples and only selects a subset of features for splitting, RF can efficiently manage the challenges posed by large datasets and mitigate the impact of feature correlations and noise in high-dimensional data. Additionally, RF operates without assuming specific statistical distributions of the data or predefined relationships between the dependent variable and explanatory variables, making the method well-suited for modeling nonlinear effects of factors.

Specifically, each decision tree in RF is generated and trained independently based on a subset from a given training dataset. First, this subset is formed by randomly selecting samples with replacement from the original training dataset, typically constituting around 2/3 of the training dataset. Meanwhile, the remaining data (usually the other 1/3) comprise the out-of-bag (OOB) set, which is excluded from training and reserved for testing purposes. Subsequently, a subset of variables (denoted as "m") is created by randomly selecting from each sample with k variables. Each decision tree grows with the selected subset of variables to its maximum extent without pruning until it cannot be split. In addition, the prediction error for each tree is calculated. The same process is iterated for hundreds or thousands of trees, resulting in the creation of a forest of random trees. Finally, the principle of averaging is utilized (for regression) to make predictions and create the final output. Particularly, the OOB set is also employed to assess the feature importance of each independent variable. The Permutation Feature Importance (PFI) approach is applied to estimate the importance of each variable, whose core thought is to investigate how much the accuracy of the model decreases when a particular variable is randomly permuted. A higher decrease in accuracy indicates that the feature is more crucial for predicting the dependent variable. This technique provides valuable insights into the relative importance of each variable in contributing to the model's predictive performance (Li et al., 2024). Beyond modeling, an interpretation approach known as the Partial Dependency Profile (PDP) is utilized to interpret the effects of exploratory variables on the dependent variable in a data-driven manner, based on trained global model. PDP relies on a trained global model and allows for a nuanced understanding of how individual variables influence the predicted outcomes.

4.3. Geographically weighted models

Global models may not have good performances in geographical analysis as they fail to account for spatial dependence or heterogeneity in the effects of factors. The principal idea of geographically weighted models is to take into account spatial non-stationarity by establishing a local equation at each ZCTA within the study area, namely considering the spatial heterogeneity in the effects of a factor on the dependent variable. In this study, two geographically weighted models, i.e., GWR and GW-RF, are employed to explore spatial heterogeneity in the relationships between traffic crash frequency and various factors across different ZCTAs.

4.3.1. Geographically weighted regression (GWR)

Geographically weighted regression (GWR) is a local linear regression method that considers spatial heterogeneity by allowing regression coefficients to vary across different locations. To this end, GWR fits a regression equation for each ZCTA using neighboring observations specific to that location. The general expression for GWR models is given by:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{n} \beta_k(u_i, v_i) x_{ki} + \varepsilon_i \tag{7}$$

where y_i denotes the dependent variable for ZCTA i, u_i and v_i stands for the coordinates for each ZCTA i, ε_i means the residual, and the parameter $\beta_k(u_i, v_i)$ represents the local coefficient estimate for the independent variable x_k at ZCTA i, which can be different between ZCTAs to address the spatial heterogeneity.

The spatially varying coefficients $\beta_k(u_i, v_i)$ are estimated by employing the ordinary least squares (OLS) method based on neighboring observations within the selected bandwidth for each ZCTA. The mechanism of this approach is that the model at ZCTA i is influenced more by nearby observations compared to those observations that are farther away. To minimize the sum of squared residuals, the coefficients estimated using the OLS method can be calculated in matrix form as follows:

$$\beta_e(u_i, v_i) = (\mathbf{X}^T \mathbf{W}(u_i, v_i)\mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i)\mathbf{y} \tag{8}$$

where $\beta_e(u_i, v_i)$ denotes the estimate of the location-specific parameter, T means the matrix transpose operation, $\mathbf{X}$ and $\mathbf{y}$ are matrixes of independent variables and dependent variable for neighboring study area determined by the kernel bandwidth, respectively, $\mathbf{W}(u_i, v_i)$ represents the spatial weight matrix.

Besides, selecting an appropriate kernel bandwidth is crucial in GWR, as the model's performance is highly sensitive to it. A large kernel bandwidth includes more neighboring observations in the regression model, which may lead to overlooking local features and heterogeneity. Conversely, a smaller kernel bandwidth potentially hinder the model's ability to capture spatial smoothing effects (Li et al., 2010). The bandwidth's size usually depends on the kernel type, bandwidth method, distance, and the number of neighbors. In this study, the optimum bandwidth (the best neighbor size) is determined with the lowest Akaike Information Criterion (AIC) values by applying "bi-square kernels" with adaptive distance. The weight between ZCTAs is calculated using a bi-square kernel function, expressed as follows:

$$W_{ij} = \begin{cases} [1 - (d_{ij}/h)^2]^2, & d_{ij} < h \\ 0, & otherwise \end{cases} \tag{9}$$

where W_{ij} represents the geographical weights between the observed data at ZCTA i and j, h denotes the bandwidth or the distance threshold beyond which observations are not considered in the weighting.

4.3.2. Geographically weighted random forest (GW-RF)

The GW-RF, which combines the concepts of GWR and traditional RF, is employed for analysis due to its potential to address the limitations of the GWR model and improve predictive performance over a non-geographically weighted RF model. Similar to the local regression analysis framework of GWR, GW-RF consists of multiple sub-models calibrated locally using RF instead of linear regression (Quiñones et al., 2021). The mechanism of GW-RF could simultaneously account for spatial heterogeneity and spatial correlation since a local model for each ZCTA i is calibrated locally using RF, which could address the issue of spatial heterogeneity. On the other hand, a local RF is constructed using only neighboring observations within the defined bandwidth of the target location to consider the spatial correlation with adjacent areas. The appropriate kernel bandwidth in the GW-RF is selected using the same method as the GWR model. The local RF at ZCTA i in GW-RF is

$$y_i = RF_i(\mathbf{X}) \tag{10}$$

where $RF_i(\mathbf{X})$ represents the trained sub-model using RF for ZCTA i.

In the training process of the GW-RF model, the optimal hyper-parameters are fine-tuned using the K-fold cross-validation method. The hyper-parameters of the GW-RF ("ntree": the number of trees, and "mtry": the number of variables randomly sampled) are determined using Random Grid Search (RGS). Then, these hyper-parameters are kept fixed to train the local GW-RF model. The bandwidth is also determined using the ten-fold cross-validation method. During the parameter

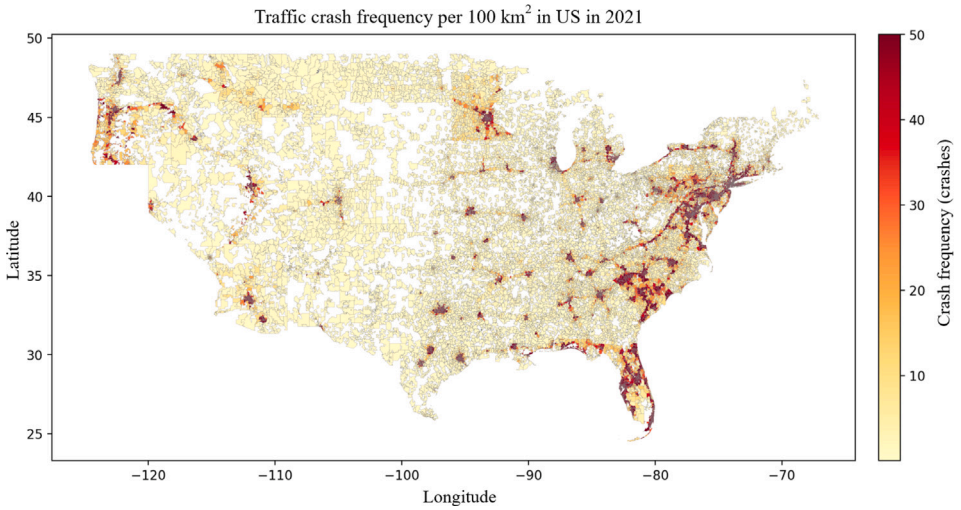


Fig. 3. Spatial distribution maps of ZCTA-level area-adjusted crash frequency in 2021.

tuning process, both the global RF and local GW-RF models are trained with the aforementioned US traffic crash data.

To assess the predictive performance of the GW-RF model and other models, several evaluation metrics are used, including Mean Square Error (MSE), Akaike Information Criterion (AIC) and goodness of fit (R^2):

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \tag{11}$$

$$AIC = 2K - 2\ln(L) \tag{12}$$

$$R^2 = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2} \tag{13}$$

where y_i is the true value for observation i, $\hat{y}_i$ denotes the predicted value of observation i, $\bar{y}_i$ means the average value of the dependent variable, n is the total sample size, k is the number of factors, L means the maximum likelihood estimate of the model.

The GW-RF model facilitates the assessment of feature importance for explanatory variables at each location, which aids in exploring the spatial heterogeneity in the effects of a factor across different zones. Local permutation feature importance is calculated for each zone based on the local RF, providing feature importance values for a factor in different zones.

5. Results

5.1. Descriptive analysis

Figs. 1 and 3 show the spatial distribution of ZCTA-level crash frequency and area-adjusted crash frequency based on the used data. The spatial distribution in Fig. 3 indicates that ZCTAs with higher crash frequencies are concentrated on the east and west coasts of the US, while crashes exhibit a clustered point distribution around major cities in the mid-western areas of the US. Regarding the area-adjusted crash frequency, they are primarily concentrated in major cities located in the southern and northeastern regions of the US.

The results of the global Moran's I index reveal that the Moran's indexes for most variables exceed 0.3, indicating that most factors at the ZCTA level are spatially correlated. Furthermore, variables such as intersection density, residential road density, the proportions of White people, the proportion of Black and African-American people, and the proportion of Asian people exhibit high spatial correlation, with Moran's indexes larger than 0.6. This implies the feasibility and necessity of the spatial analysis of crash frequency at the ZCTA level in this study.

From a local perspective, LMI is employed to test the correlation between area-adjusted crash frequency and influencing factors, exploring a significant correlation exists. Four representative LMI results depicting crash frequency and influencing factors are presented in Fig. 4. As shown in Fig. 4(a), the red-colored areas (HH) primarily concentrated in the southern and northeastern regions correspond to significant clusters of high crash frequency that also have high density of motorway in neighboring ZCTAs. Conversely, the orange-colored areas (HL) represent significant clusters of high crash frequency with low motorway density. Additionally, the majority of the areas in Fig. 4(a) are deep blue (LL), indicating significant clusters of low crash frequency with low motorway density. This suggests that the correlation between crash frequency and motorway density changes spatially (i.e. in different areas), which may be positive or negative. Moreover, spatial variations in the correlation also exist between crash frequency and other influencing factors. In Fig. 4(b), the red-colored areas (HH), representing significant clusters of high crash frequency with high intersection density, are mainly distributed near large cities. While the orange-colored areas (HL), representing significant clusters of high crash frequency with low intersection density, are concentrated in the southern and northeastern regions. For the proportions of White people, the red-colored areas (HH) are distributed in the western, mid-western, and northeastern regions, while the orange-colored areas (HL) predominantly concentrate in the southern regions (Fig. 4(c)). Fig. 4(d) illustrates spatial clusters of crash frequency and poverty rate, where the red-colored areas (HH) concentrate in the southern regions, while the orange-colored areas (HL) are mainly present in the northeastern regions. These spatial variations in the correlation between a factor and crash frequency underscore the necessity of considering geographical modeling and local estimation of effects of influencing factors on crash frequency.

5.2. Comparisons of predictive performance

5.2.1. Linear models

The performance of GW-RF and other benchmarks (OLS, GWR, RF) in modeling the relationship between influencing factors and crash frequency is evaluated using five-fold cross-validation. MSE, AIC, and R^2 are used as metrics to assess the predictive performance of the models. As shown in Table 3, from an overall perspective, models using area-adjusted crash frequency as the dependent variable perform better in their predictions compared to models using total crash frequency as

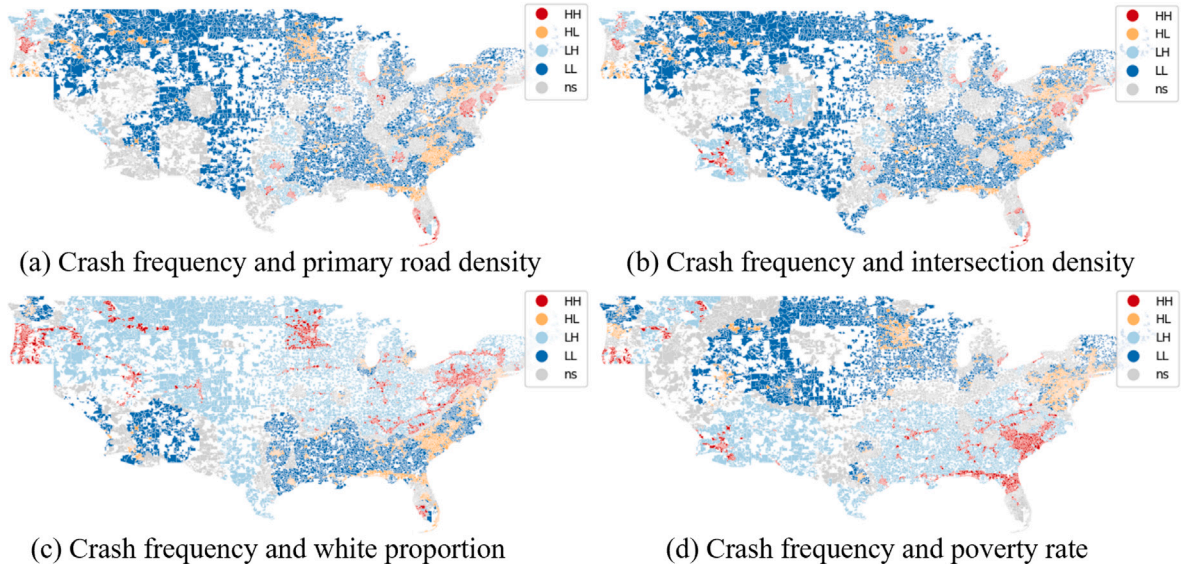


Fig. 4. BLMI cluster of area-adjusted crash frequency and four influencing factors. Figures are generated by the GeoDaSpace package for geodata analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The results of the predictive performance of the models.

Model		Total crash frequency				Area-adjusted crash frequency			
		AIC	MSE	R^2	OBB R^2	AIC	MSE	R^2	OBB R^2
OLS		48 121	0.8933	0.2021	NA[a]	47 716	0.7793	0.2206	NA[a]
GWR	Min	17 327	0.0	0.0085	NA[a]	16 714	0.0	0.0095	NA[a]
	Max		14.0509	**0.7761**	NA[a]		12.2934	**0.8584**	NA[a]
	Mean[b]		0.6923	**0.3095**	NA[a]		0.5509	**0.3285**	NA[a]
RF		/	0.6251	0.3748	0.3664	/	0.6085	0.3914	0.3835
GW-RF	Min	/	0.0	0.0091	0.0118	/	0.0	0.0101	0.0118
	Max		12.3115	**0.7964**	0.7784		10.8891	**0.8673**	0.8574
	Mean[b]		0.6792	**0.3468**	/		0.6259	**0.3704**	/

[a] Not applicable.
[b] The mean value of R^2 for GWR and GW-RF is the arithmetic mean.

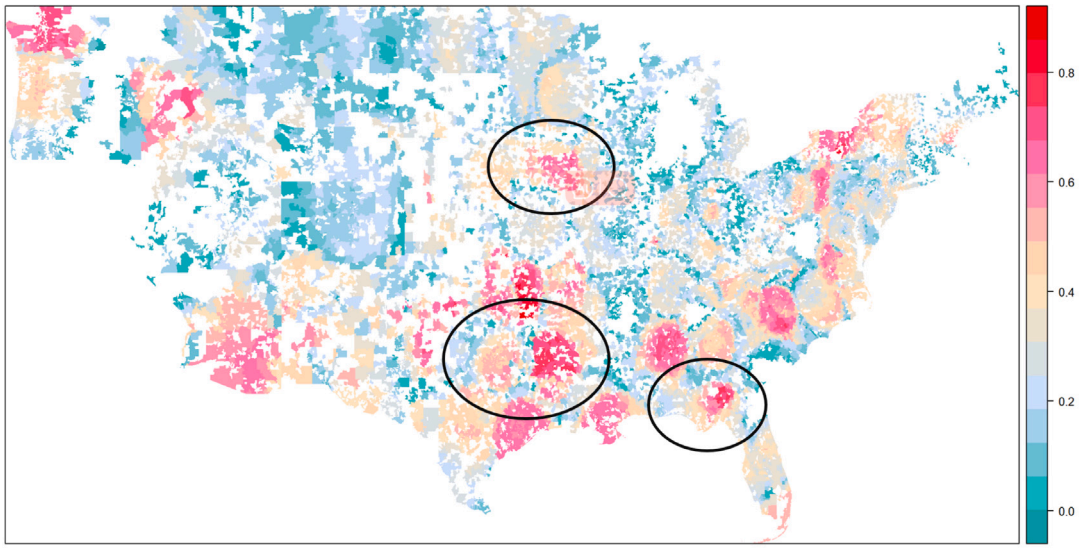
the dependent variable. Therefore, we exclusively discuss the results of models in predicting area-adjusted crash frequency.

The R^2 of the global OLS model is 0.22, which serves as a baseline for comparison. The GWR model exhibits a higher average adjusted R^2 values (0.3285) and a lower AIC value (see Table 2), indicating superior predictive performance over the global OLS model. In comparison to conventional models, the RF model shows a significant improvement, reflected by a higher R^2 value (see Table 2). The GW-RF model shows a lower MSE value than that of the global RF model. What is more, the GW-RF model shows a slight improvement in the overall performances compared to the GWR model with a higher average R^2 value (see Table 2).

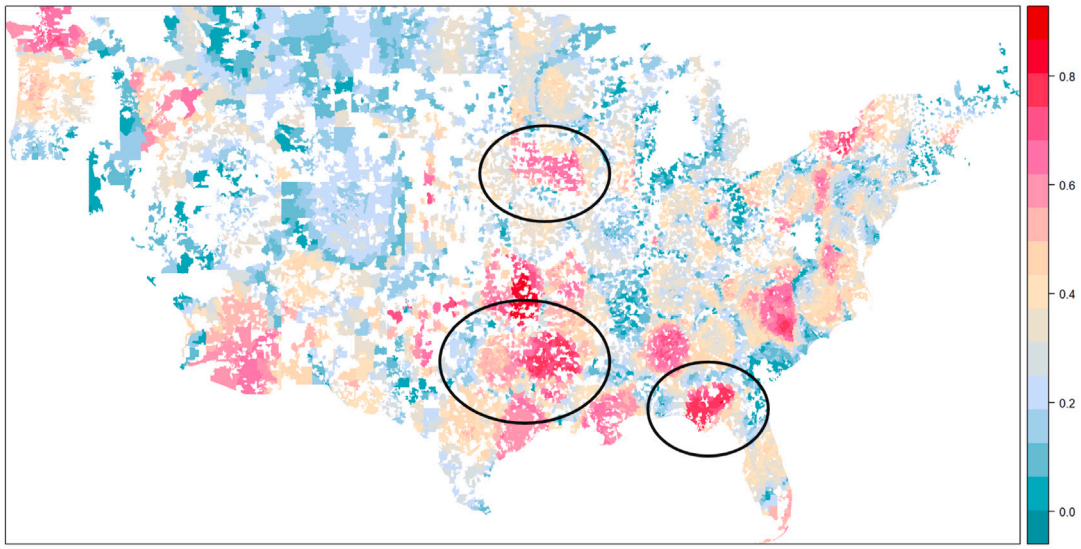
We further evaluate the local R^2 values for both the GWR model and GW-RF model pertaining to area-adjusted crash frequency, as shown in Fig. 5. In the GWR model, the local R^2 ranges from 0 to 0.858, with a mean value of 0.328. Notably, the local R^2 values are relatively high in the majority of ZCTAs in the southern regions and some ZCTAs in the western and northeast regions. Additionally, the local R^2 values for the GW-RF model range from 0.01 to 0.867, exhibiting an average value of 0.370, which is 12.8% higher than the corresponding mean R^2 value for the GWR model (0.328). The distribution of local R^2 values for the GW-RF model is similar to that of the GWR model, yet it demonstrates an improvement in model performances, as indicated by the circled areas in Fig. 5. The local GW-RF models exhibit strong robustness ($R^2 > 0.6$) in 10% of ZCTAs, whereas the local GWR models demonstrate satisfactory predictive performance ($R^2 > 0.6$) in only 8.6% of ZCTAs. A plausible explanation for these observed patterns lies in the flexibility and strength of the GW-RF model, which avoids assuming a linear relationship between crash frequency and influencing factors, and has the advantage of dealing with the complex nonlinear and interactive effects among the predictors. The local R^2 values are relatively high in most ZCTAs within the southern regions and some ZCTAs in the western and northeast regions. Conversely, the local R^2 values for most regions in the western and mid-western parts of the US are relatively low. One potential reason for this phenomenon is that the data in these regions are relatively sparse and have smaller sample size under the specified bandwidth, leading to a deficiency of explanatory power for the RF. Upon comparisons, the GW-RF model emerges as the optimal model to investigate the associations between crash frequency and influencing factors with consideration of spatial variations in the effects of a factor. Therefore, the subsequent local effect analysis of influencing factors is mainly based on the results of the GW-RF model.

5.3. Spatial variations in the local effect of influencing factors

The above results indicate the potential correlations of factors and the crash frequency. In this section, we utilize the results of OLS and RF global models to understand the relationship between influencing factors and crash frequency from a global perspective. More importantly, we further employ the GWR and GW-RF models to analyze the local relationships between crash frequency and influencing factors,



(a) Local R^2 for the GWR model



(b) Local R^2 for the GW-RF model

Fig. 5. Local R^2 for the GWR and GW-RF models of area-adjusted crash frequency. The improvement of GW-RF is shown in the circled areas.

providing an understanding concerning spatial variations in the local effects of influencing factors on crash frequency.

5.3.1. GWR model

As indicated in Table 3, the global OLS model reveals that the majority of influencing factors are positively correlated with crash frequency ($p < 0.05$), except for the residential land use ratio, industry land use ratio, labor force proportion, poverty rate, and average travel time to work. However, the residential land use ratio and poverty rate, despite showing no statistically significant correlation in the global OLS model, have been acknowledged as impactful in some existing studies (Xie et al., 2019; Wang and Kockelman, 2013). These studies have mainly concentrated on a specific city or state, while the scope of this study encompasses the whole US. Different study scopes may affect the influence of a specific factor on crash frequency from a global perspective. The expansion of the study area tends to average the effects of influencing factors across the entire study area, potentially diminishing the significance of these effects.

Further, we analyze the local effects of the influencing factors using the GWR model. The results of GWR are summarized in Table 3, presenting descriptive statistics of the estimated coefficients for influencing factors across various ZCTAs. These statistics provide general views on the variances in the effects of influencing factors. Across all variables, local coefficients exhibit both positive and negative values. For example, the local coefficients for intersection density range from −0.897 to 5.005, with a median value of 0.004. Notably, the coefficients for intersection density are significant and positive in more than 50% of ZCTAs in the US, which is significantly different from the results of the global model (0.002). The coefficients for tertiary road density range from −1.111 to 1.877, which are narrower. Besides, we found that the median coefficients for residential land use ratio and commercial and services land use ratio are zero, indicating a positive impact on crash frequency in 50% of ZCTAs and a negative impact in the remaining half. Similar patterns can be found from the results of other factors. The results indicate the existence of notable spatial variation in the effects of a factor on crash frequency, which cannot be appropriately captured by global methods without considering spatial

Table 3
Summary results of OLS and GWR models.

Variables	OLS		GWR				
	Estimate	$Pr(>\|t\|)$	Min	Median	Max	Mean	Std
Road network variables							
Motorway density	0.063	<2e-16***	-0.060	0.011	0.531	0.038	0.081
Primary road density	0.056	9.48e-11***	-0.674	0.007	1.224	0.050	0.223
Secondary road density	0.073	<2e-16***	-1.261	0.014	2.523	0.088	0.346
Tertiary road density	0.214	<2e-16***	-1.111	0.053	1.877	0.144	0.288
Residential road density	0.067	5.98e-14***	-1.113	0.008	2.154	-0.037	0.196
Service road density	-0.090	2.76e-15***	-3.058	0.008	1.226	-0.014	0.310
Intersection density	0.002	9.25e-13***	-0.897	0.004	5.005	0.162	0.555
Signal intersection rate	0.026	0.0005***	-0.186	-0.001	0.423	0.002	0.039
Demographics and Socioeconomic variables							
Population density	0.168	<2e-16***	-0.430	0.013	0.364	0.011	0.055
Bachelor proportion	-0.050	1.02e-06***	-0.182	0.005	0.502	0.007	0.050
High school proportion	-0.040	0.0059**	-0.059	-0.005	0.581	0.020	0.071
Labor force proportion	-0.014	0.1417NS	-1.934	0.002	2.548	-0.024	0.317
Poverty rate	-0.001	0.9767NS	-1.686	0.002	2.868	-0.005	0.300
Unemployment rate	-0.020	0.0104*	-1.355	0.000	0.848	-0.021	0.166
Average Travel Time to Work	0.004	0.5043NS	-0.819	0.002	0.477	-0.031	0.117
White proportion	-0.160	4.67e-12***	-0.397	-0.002	0.752	0.020	0.099
Black and African-American proportion	-0.107	1.36e-07***	-0.336	-0.004	0.544	0.047	0.095
American Indian proportion	-0.054	2.41e-07***	-0.375	-0.008	0.366	0.013	0.064
Asian proportion	-0.084	<2e-16***	-0.367	-0.009	0.500	0.035	0.082
Young proportion	0.045	0.0018**	-0.843	0.007	0.378	0.001	0.076
Prime adult proportion	0.053	0.0001***	-0.529	0.016	1.043	0.007	0.108
Middle-aged adult proportion	0.030	0.0159*	-0.273	0.007	0.254	0.009	0.038
Elder proportion	0.036	0.0131*	-0.330	0.016	0.992	0.015	0.112
Land use variables							
Residential land use ratio	0.012	0.3298NS	-2.243	0.000	1.244	0.090	0.252
Commercial and services land use ratio	0.137	0.0006***	-0.503	0.000	1.330	0.030	0.123
Industry land use ratio	0.002	0.8011NS	-1.044	0.006	0.189	-0.019	0.100
Transportation, communications, and utilities land use ratio	0.045	0.0002***	-0.624	0.001	0.126	-0.008	0.073
Intercept	-1.02e-16	1.000NS	-0.387	-0.055	0.863	-0.029	0.113

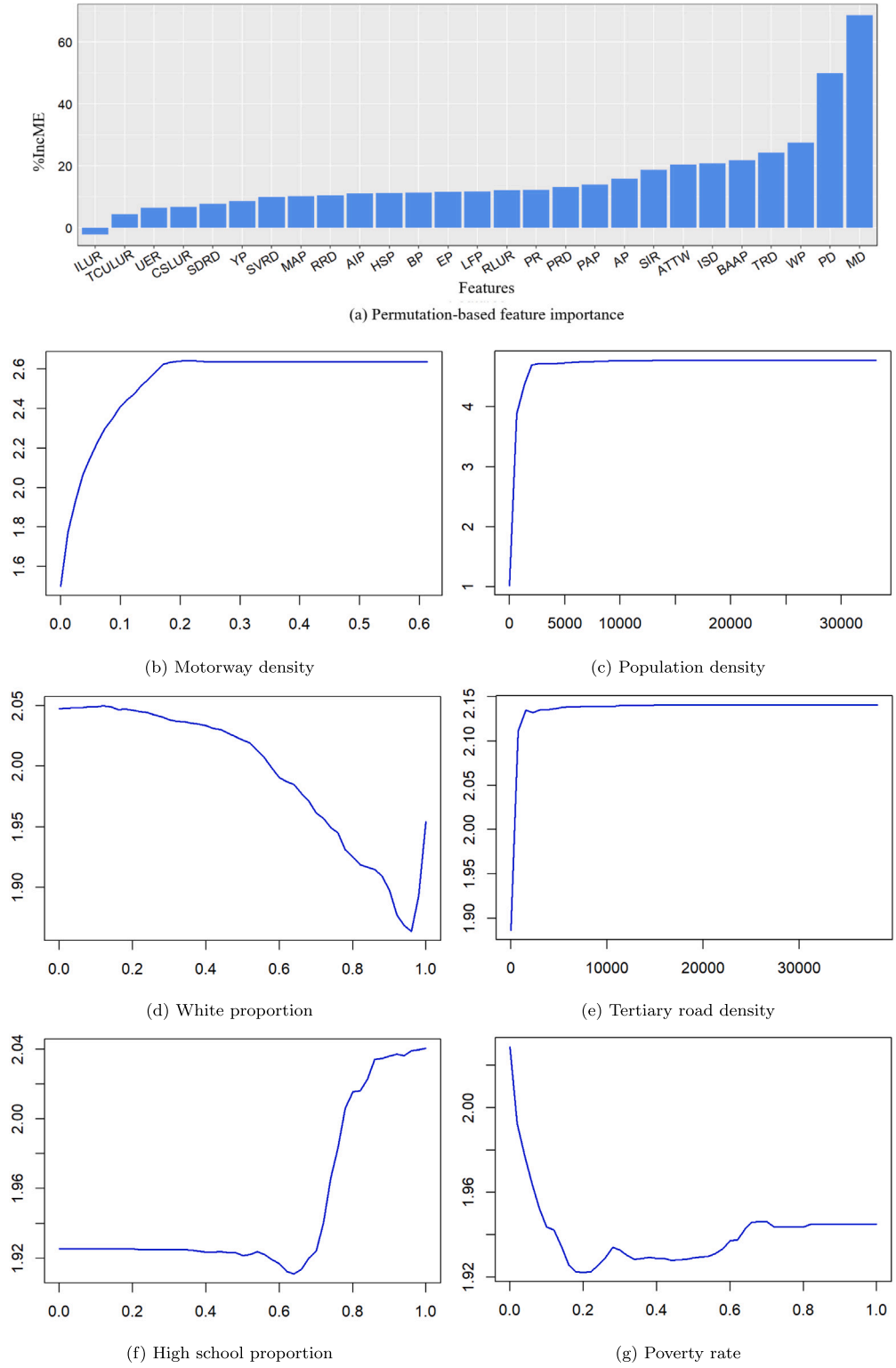
*** means $p<0.001$, ** means $p<0.01$, * means $p<0.1$, and NS means not significant.

heterogeneity in the effects of the factors. Therefore, global models may result in biases in estimating the effects of a factor within a specific geographical area, which highlights the necessity and superiority of geographically weighted modeling. Herein, we do not elaborate the effects of each factor from OLS and GWR, as more detailed discussions will be provided in the following section, combining the results from RF and GW-RF that have the best predictive performances.

5.3.2. RF and GW-RF model

According to the PFI values from the trained RF model in Fig. 6, the factor "white proportion" has the highest importance, followed by motorway density, primary road density, secondary road density, and utilities land use ratio. From an overall perspective, this ranking is significantly different from the coefficient ranking of the OLS model (Please note that the values of factors are normalized, so the coefficients of factors are comparable in OLS model). In the RF model, the motorway density, population density, and white proportion rank as the top three important features, whereas in the OLS model, the tertiary road density, population density, and white proportion are the top three factors contributing most significantly to the prediction of crash frequency. This disparity could be attributed to the data-driven modeling mechanism of machine learning such as RF to depict complex nonlinear and interactive effects of several factors, which cannot be modeled by OLS. Noting that the feature importance value of the "Industry land use ratio" is negative. This may be due to this variable originally having a low importance value and the presence of data noise, because when some variables do not provide valuable information for predicting the dependent variable, a small perturbation in the training sample may completely change the importance ranking of the variables (Louppe et al., 2013; Li et al., 2020). More importantly, it indicates that the "Industry land use ratio" is not relevant for predicting crash frequency based on our datasets.

Moreover, the partial dependency analysis is used to further interpret how a factor will affect the area-adjusted crash frequency in a more quantitative manner. The interesting results of six important



(a) Permutation-based feature importance

(b) Motorway density

(c) Population density

(d) White proportion

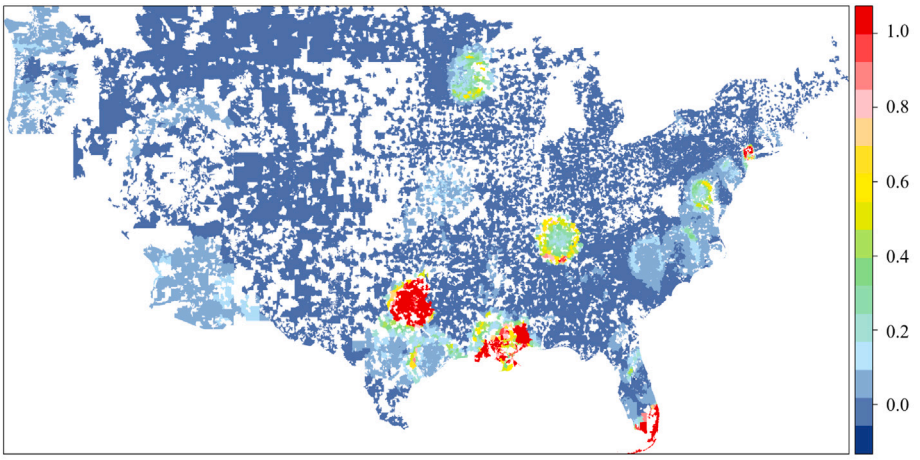
(e) Tertiary road density

(f) High school proportion

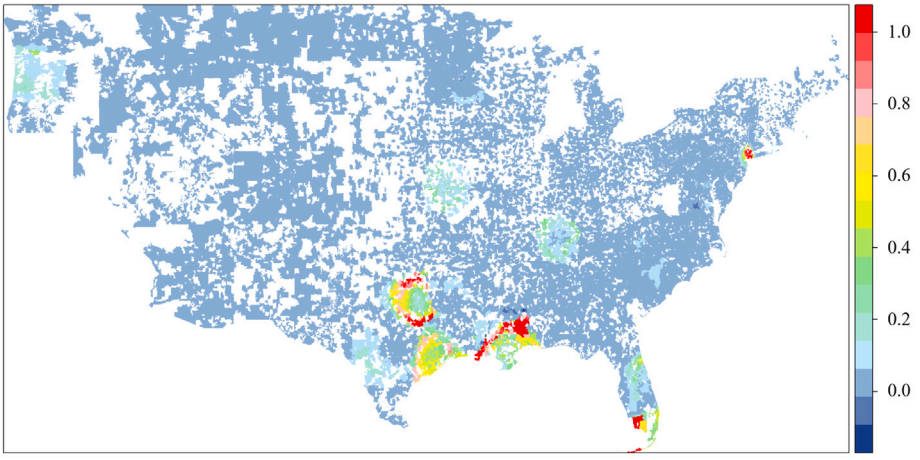
(g) Poverty rate

Fig. 6. Feature importance from global RF. (a) Permutation-based feature importance; (b–g) partial dependency profiles of the six important variables of the global random forest model.

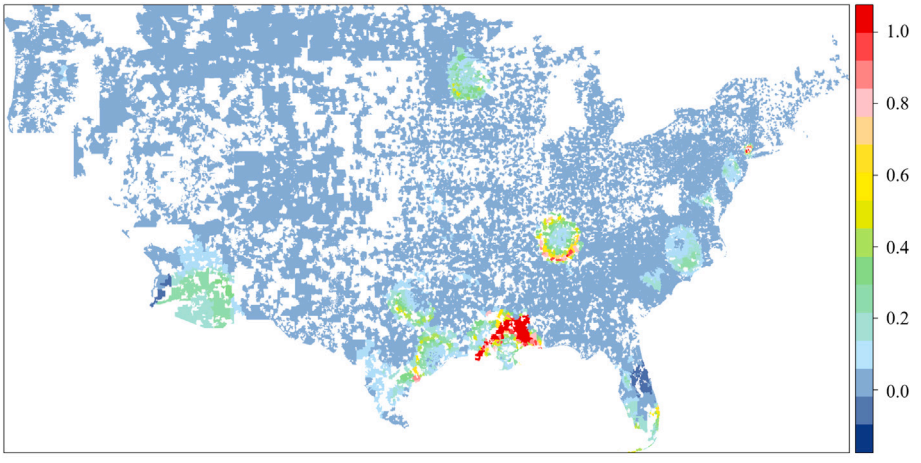
variables are demonstrated in Fig. 6 for discussions. When controlling for the influence of other factors, the motorway density (Fig. 6(b)), population density (Fig. 6(c)), tertiary road density (Fig. 6(e)), and high school proportion (Fig. 6(f)) are all positively related to area-adjusted crash frequency. More interestingly, the effect of a factor presents nonlinear and threshold patterns. Taking Fig. 6(b) as an example, the area-adjusted crash frequency increases pretty significantly when "motorway density" is less than 0.2, but its effect on crash frequency is trivial when "motorway density" is larger than 0.2. The effect of "high school proportion" is negligible when its value is below 0.65, but its effect exhibits a pronounced increase when "high school proportion" is larger than 0.65. It is worth noting that such nonlinear and threshold effects cannot be modeled by conventional models with prior assumptions (e.g. linear relations). For the effect of "white proportion" and



(a) Intersection density



(b) Service road density

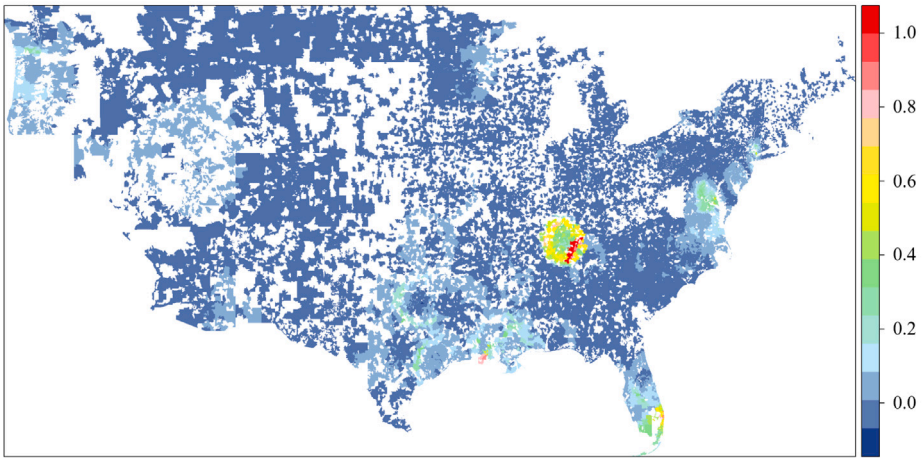


(c) Secondary road density

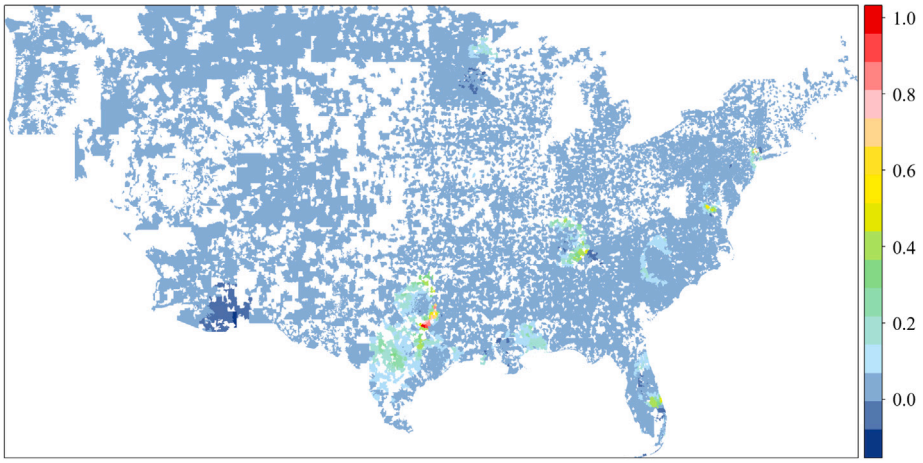
Fig. 7. Spatial heterogeneity distribution of local effects of first ten influencing factors on crash frequency. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

"poverty rate", the results also show nonlinear and threshold patterns. As for the effects of white proportion in Fig. 6(d), it decreases as "white proportion" increases when "white proportion" is less than 0.95, but the it sharply increases when "white proportion" is larger than 0.95. Additionally, as shown in Fig. 6(g), the area-adjusted crash frequency decreases rapidly when "poverty rate" is less than 0.2, and its effect slightly increases when "poverty rate" is larger than 0.6. Partial dependency profiles of the remaining variables are summarized in Fig. B.2 in the Appendix in case of redundancies.

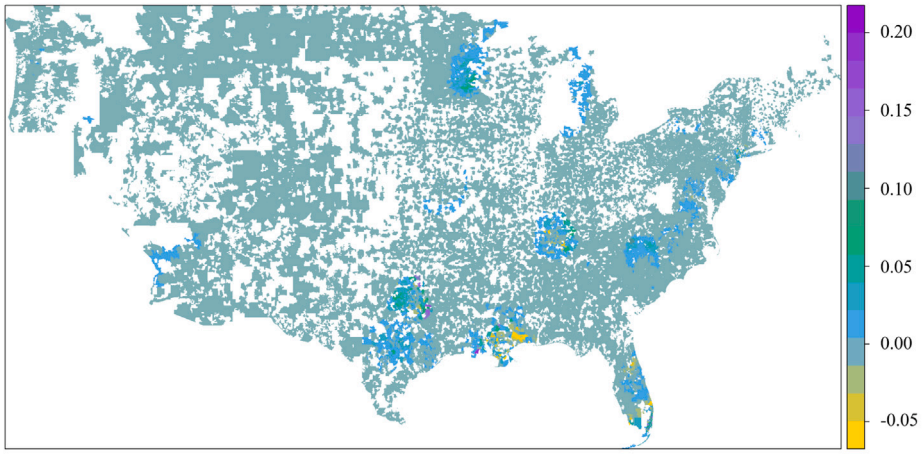
To clearly illustrate the spatial heterogeneity in the effects of a factor on area-adjust traffic crashes, the results of GW-RF are projected into the ZCTAs for detailed discussions. This mainly demonstrate and discuss factors with high importance from GW-RF, the results of remaining factors are provided in Fig. B.3 in the Appendix to avoid



(d) Primary road density



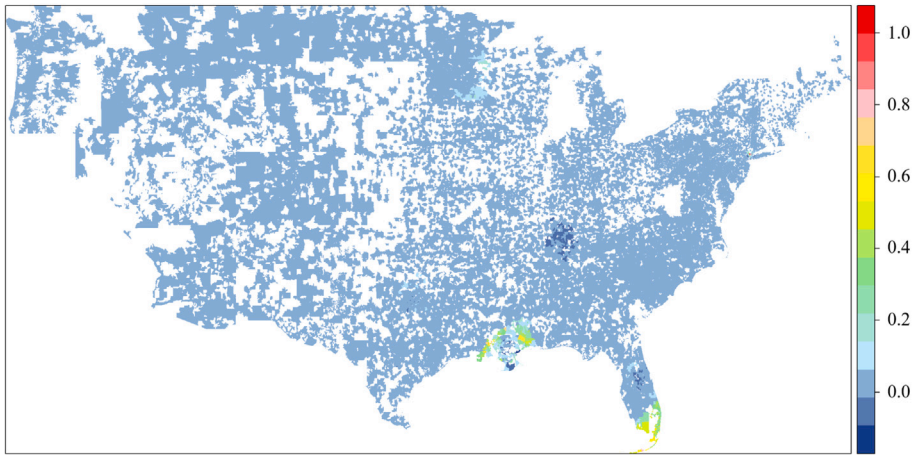
(e) Young (15-24 years old) proportion



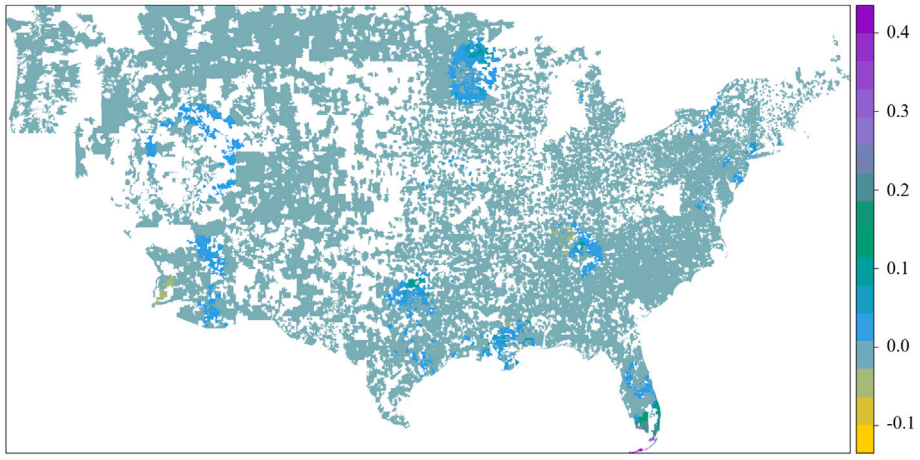
(f) Elder (> 65 years old) proportion

Fig. 7. (*continued*).

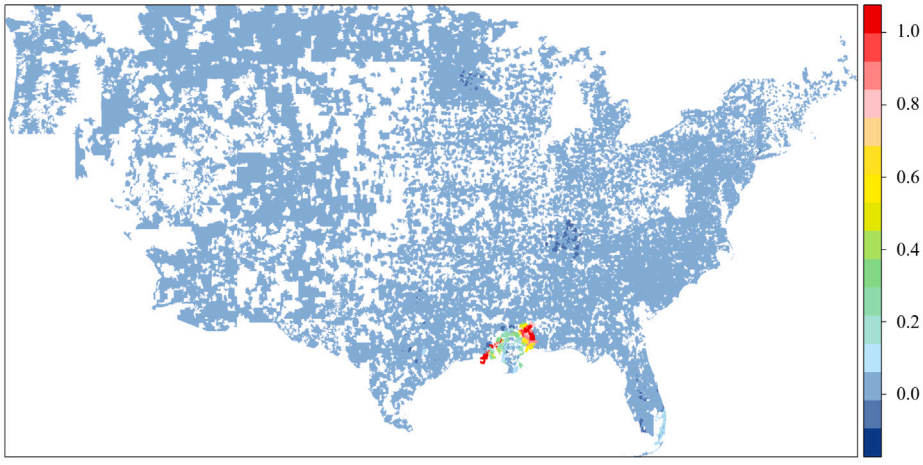
redundancy. For the road network factors, Fig. 7(a) shows that importance of intersection density on crash frequency varies significantly in different zones. The high PFI values (>0.6) of "intersection density" are mainly concentrated in some ZCTAs in the southern and northeastern regions, indicating the intersection density has higher impacts on crash frequency in these areas than other areas. These ZCTAs are mainly in or around major cities with high intersection density. Meanwhile, Table 4 summarizes that the importance of intersection density ranks within the top three and top five factors in 58.6% and 74% of the ZCTAs, respectively. Combined with the spatial clustering results of BLMI in Fig. 4, the "intersection density" is found to be positively associated with crash frequency in some areas but negatively associated in others. This finding is consistent with that of Azimian et al. (2021). Two potential reasons may explain the distinct effects of intersection density



(g) White proportion



(h) Black and African-American proportion

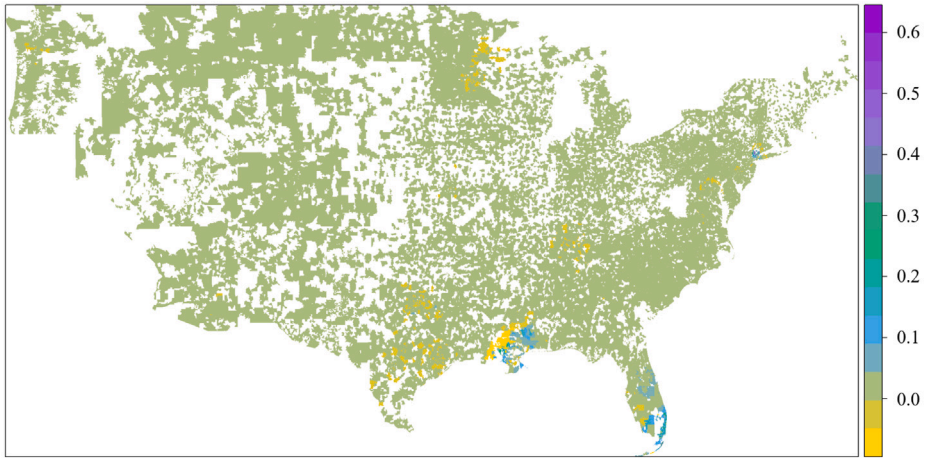


(i) Commercial and services land use ratio

Fig. 7. (*continued*).

in different areas. First, the types of intersections and the portion of traffic control policy may differ from state to state. Second, driver behaviors can vary from region to region. For instance, drivers need to drive carefully and at a lower speed near intersections because several states have strict penalties for reckless drivers, such as West Virginia. More importantly, the results reveal the spatial variation in the effects of intersection density on crash frequency, which would be covered up by methods ignoring geographical or spatial differences (e.g. the global models aforementioned).

Fig. 7(b–c) demonstrate that the importance of service road density and secondary road density is high in some cities in the southern and northeastern regions such as Houston, New Orleans, and Boston. Service road density is the top three influential factors in 47% of the ZCTAs (Table 4). One possible reason is that service roads primarily include cycleways and sidewalks, which are more likely to have traffic



(j) Residential land use ratio

Fig. 7. (*continued*).

crashes due to the randomness of pedestrian and cyclist behavior. Similarly, the importance of secondary road density is noteworthy, given that these roads serve as vital connectors between main roads and residential or recreational zones. Consequently, they play a principal role in accommodating commuter traffic, experiencing large traffic volumes during peak commuting periods, thereby heightening the risk of traffic accidents. Moreover, primary road density within the top three in about 30% of the ZCTAs, and ranks within the top five in 45% of the ZCTAs, respectively (Table 4). The ZCTAs with high importance value of primary road density are distributed in the southern regions, and in some areas of the western and northeastern regions (Fig. 7(d)), which are consistent with the red-colored areas (HH) in the BLMI cluster results (Fig. 4(a)). This implies primary road density is positively correlated with crash frequency, consistently with previous studies' findings, which potentially because of their higher traffic speed and speed limits in primary roadways (Bao et al., 2018). More importantly, the importance or effects of all the three factors show obvious differences in different zones, unraveling the spatial variation in the effects and highlighting the superiority and necessity of using geographical modeling such as GW-RF.

As for the results regarding the socioeconomic factors explored, the importance of the "young proportion" in Fig. 7(e) is not notable in the majority of ZCTAs, and ranks as the top three and top five important factors only in 9.4% and 18.9% of the ZCTAs. The importance of "young proportion" in a ZCTA is low, with an average of 0.15, and has no significant spatial clustering characteristics. Thus, in general, the proportion of young people has no significant correlation with crash frequency in most regions, but locally, it is positively correlated with crash frequency in some regions where the proportion of young people is particularly high as shown in Fig. 7(e). This is consistent with the results of previous studies, which have suggested that younger people tend to drive at higher speeds and are more likely to be engaged in traffic crashes, but studies did not taking into account the spatial variation in the effects of young people (Bao et al., 2018; Zhang et al., 2023). By contrast, the importance of the proportion of the elder is not notable in the whole study area. Interestingly, the "Black and African-American proportion" is positive associated with crash frequency and has pretty high importance in ZCTAs in the southern regions (Fig. 7(h)). This may be because the road infrastructure in these black communities in the southern regions is in disrepair and prone to traffic crashes. Conversely, the importance of "White proportion" is not significant in the vast majority of regions, which ranks the top three, and top five highest value of local variable importance only in 3.1% and 8.6% of the ZCTAs, respectively.

Table 4
The proportion of ZCTAs where a factor has the top three or top five highest local importance on the area-adjusted crash frequency.

Top 12 important factors	Proportion of ZCTAs	
	Top three	Top five
Intersection density	58.6	74.0
Service road density	47.1	65.7
Secondary road density	35.5	54.6
Primary road density	29.6	44.7
Tertiary road density	27.3	47.0
Residential road density	24.7	43.0
Bachelor proportion	13.6	25.4
Young proportion	9.4	18.9
Black and African-American proportion	3.6	7.5
White proportion	3.1	8.6

Note: "Top three" and "Top five" mean that the importance of a factor is among top three and five highest importance among all factors in a ZCTA, respectively.

The local feature importance distribution of the remaining variables does not exhibit significant spatial heterogeneity. Among them, the "average Travel Time to Work", "bachelor proportion", and "population density" have pretty high importance for crash frequency in the whole study area, but do not show spatial heterogeneity. Literature have reported that the population with a bachelor's degree in a ZCTA is negatively correlated with traffic crashes, and average travel time to work and population density are positively correlated with traffic crashes (Bao et al., 2018). The importance of all investigated land-use factors does not show significant spatial heterogeneity. Compared with the other three land use factors explored, the importance of "commercial and services land use ratio" is relatively high (Fig. 7(i)). We find that the ratio of the area allocated for commercial purpose is positively correlated with crash frequency. The finding is intuitive and consistent with the results of previous studies (Rhee et al., 2016) that areas allocated for commercial purpose tend to have dense traffic networks and large traffic volumes, which are prone to traffic crashes. In addition, the importance of "residential land use ratio" is lower than 0.01 from the perspective of the whole study area (Fig. 7(j)). This finding implies that the residential land use has noticeable effect on traffic crashes, which is consistent with previous study (Ouyang and Bejleri, 2014). An interesting finding reported by Ouyang and Bejleri (2014) is that the distance between the residential land use and commercial land use is negatively related to traffic crashes, and the explanation is that long distance may limit the frequency of residents traveling to commercial areas, thereby reducing the interaction between transport participants. Therefore, the distance between areas

Table 5
Local importance results of the GW-RF model.

Variables	GW-RF				
	Min	Median	Max	Mean	Std
Road network variables					
Motorway density	−0.0712	0.0004	0.9106	0.0351	0.0987
Primary road density	−0.0567	0.0010	0.9190	0.0515	0.1415
Secondary road density	−0.0727	0.0017	0.9478	0.0677	0.1856
Tertiary road density	−0.0825	0.0010	0.9609	0.0493	0.1630
Residential road density	−0.0728	0.0013	0.9123	0.0387	0.0873
Service road density	−0.0765	0.0033	0.9242	0.0513	0.1278
Intersection density	−0.0594	0.0038	0.9511	**0.0909**	0.2282
Signal intersection rate	−0.0912	8e−5	0.5896	0.0082	0.0447
Demographics and socioeconomic variables					
Population density	−0.0756	0.0005	0.8746	0.0583	0.0662
Bachelor proportion	−0.0518	0.0001	0.8437	0.0492	0.1025
High school proportion	−0.0456	4e−5	0.5943	0.0192	0.0549
Labor force proportion	−0.0721	4e−5	0.8213	0.0088	0.0367
Poverty rate	−0.0681	3e−5	0.8175	0.0113	0.0649
Unemployment rate	−0.0858	0.0001	0.8470	0.0060	0.0501
Average Travel Time to Work	−0.0888	0.0003	0.8885	0.0357	0.0832
White proportion	−0.0879	0.0001	0.9179	0.0310	0.0934
Black and African-American proportion	−0.0614	6e−5	0.3639	0.0248	0.0544
American Indian proportion	−0.0860	0.0000	0.1803	0.0006	0.0124
Asian proportion	−0.0661	6e−5	0.6907	0.0065	0.0409
Young proportion	−0.0764	0.0002	0.8668	0.0387	0.0652
Prime adult proportion	−0.0803	0.0000	0.2398	0.0025	0.0181
Middle-aged adult proportion	−0.0811	1e−5	0.2349	0.0095	0.0261
Elder proportion	−0.0478	2e−5	0.1926	0.0126	0.0329
Land use variables					
Residential land use ratio	−0.0494	0.0000	0.6160	0.0091	0.0387
Commercial and services land use ratio	−0.0671	2e−5	0.9030	0.0103	0.0518
Industry land use ratio	−0.0434	0.0000	0.1960	0.0061	0.0262
Transportation, communications, and utilities land use ratio	−0.0779	0.0000	0.3233	0.0068	0.0316

with different land use purposes should be considered to explore its effect on crash frequency in subsequent studies (see Table 5).

6. Conclusions

This study leverages a geographically weighted machine learning model, the GW-RF model, as both a predictive and exploratory methods, to explore spatial heterogeneity of local associations between zone-level crash frequency and selected influencing factors. This targets to provide insights about how different factors affect the crash frequency in different urban contexts. Considering potential influencing factors such as road network attributes, socio-demographic characteristics, and land use factors, the GW-RF model outperforms global (OLS and RF) and conventional geographically weighted regression models in predicting crash frequency and explains spatial heterogeneity in the associations between crash frequency and exploratory variables. The results offer insight into the underlying reasons for crash frequency in various areas, which would help crash-prone areas for tailored prevention and interventions to reduce the likelihood of crashes. The main contributions and findings of this study can be summarized as follows.

The predictive efficacy of the GW-RF model surpasses that of GWR model. In contrast to the widely adopted GWR model, which assumes linear effects of explanatory variables, the GW-RF model employs the Random Forest (RF) model to scrutinize local associations. This is particularly pertinent when nonlinear and interactive effects of factors on crash frequency are considered. Moreover, the GW-RF model exhibits superior predictive accuracy when compared to the global RF model. Capitalizing on the merits of local modeling, the GW-RF model accommodates spatial heterogeneity in the effects of a factor by training sub-models based on neighboring observations. This approach enhances the model's predictive performance by capturing and incorporating localized nuances in the relationship between factors and crash frequency.

The GW-RF is further applied to explore the spatial variations in the effects of exploratory factors on crash frequency. First, the results of the GW-RF model demonstrate that all road network factors (except the variable "Signal intersection rate") are of pretty high importance in all zones, especially intersection density, which are positively correlated with crash frequency. As for demographics and socioeconomic variables, population density, bachelor proportion, high school proportion, poverty rate, average travel time to work, White proportion, Black and African-American proportion, young proportion and elder proportion have moderate effects on crash frequency (average value of feature importance > 0.01). The results also illustrate that the importance of land use variables is relatively low, only the importance of commercial and services land use ratio is larger than 0.01, which has a significantly positive correlation with crash frequency. Our findings are practically instructive for the planning of road networks and intersection, the arrangement of distance between residential areas and companies, and the improvement of road infrastructures at zonal level.

The GW-RF model shows that the effects of some factors on crash frequency are significantly distinct in different areas. The importance of intersection density has significant spatial heterogeneity. The importance intersection density ranks within the top three in 58.6% of the ZCTAs, which are distributed in the southern and northeastern (high-high cluster) regions and some areas in the western regions. The importance of low-grade road density (service roads, secondary roads, and tertiary roads) is high in some cities in the southern and northeastern regions. The importance of young people (15–24 years old) proportion is weaker than that of low-grade road density. It ranks as the top three most important variables only in a few ZCTAs (9.4%). These results provide an understanding of spatial distinctions of the effects of a factor on zone-level crash frequency.

While this study considers a selection of widely recognized influencing factors and one year of crash data to explore the spatial heterogeneity of crash frequency, the primary focus of this study is to demonstrate that the GW-RF model can serve as a powerful method

Table A.1
Road types from OSM and corresponding road types.

Road types	Code	Road type from OSM	Road types	Code	Road type from OSM
Motorway	5111	motorway	Service	5124	pedestrian
	5131	motorway_link		5141	service
	5112	trunk		5142	track
	5132	trunk_link		5143	track_grade1
Primary	5113	primary		5144	track_grade2
	5133	primary_link		5145	track_grade3
Secondary	5114	secondary		5146	track_grade4
	5134	secondary_link		5147	track_grade5
Tertiary	5115	tertiary		5151	bridleway
	5135	tertiary_link		5152	cycleway
Residential	5121	unclassified		5153	footway
	5122	residential		5154	path
	5123	living_street		5155	steps
	5125	busway		5199	unknown

Table A.2
POI category from OSM and corresponding land use categories.

Land use categories	Code	POI category from OSM
Residential land use	1–6	House, clothes, apartments, farmyard, alpine_hut, chalet, shelter
Industrial land use	7–8	Industrial, quarry
Communication and utility land use	9–38	Police, fire_station, post_box, post_office, telephone, library, town_hall, courthouse, prison, embassy, community_centre, nursing_home, arts_centre, graveyard, market_place, university, school, kindergarten, college, public_building, pharmacy, hospital, clinic, doctors, dentist, veterinary, toilet, bench, drinking_water, fountain
Commercial and services	39–129	Clothes, florist, chemist, bookshop, butcher, shoe_shop, beverages, optician, jeweller, gift_shop, sports_shop, stationery, outdoor_shop, mobile_phone_shop, toy_shop, newsagent, greengrocer, beauty_shop, video_shop, car_dealership, bicycle_shop, doityourself, furniture_shop, computer_shop, garden_centre, hairdresser, car_rental, car_wash, car_sharing, bicycle_rental, travel_agent, laundry, vending_machine, vending_cigarette, vending_parking, vending_any, bank, atm, tourist_info, attraction, museum, monument, memorial, artwork, castle, ruins, archaeological, wayside_cross, wayside_shrine, battlefield, fort, picnic_site, viewpoint, zoo, theme_park
Other	130–147	Ecycling, recycling_glass, recycling_paper, recycling_clothes, recycling_metal, hunting_stand, waste_basket, camera_surveillance, tower, comms_tower, water_tower, observation_tower, windmill, lighthouse, wastewater_plant, water_well, water_mill, water_works

for investigating the spatial heterogeneity of crash frequency and addressing spatial variations in the effects of influencing factors on crash frequency in a data-driven way. The GW-RF model is applicable in spatial modeling problems for traffic safety analysis at various geographical locations. Nonetheless, several aspects of this study can be further improved. Firstly, due to the challenges in data acquisition, this study did not take into account traffic volumes. Existing studies have concluded that high traffic volumes appear to increase crash frequency. Therefore, it needs to be considered to explore local associations between crash frequency and traffic volumes. Secondly, although the GWR and GW-RF models exhibit more robustness in many ZCTAs, their accuracy becomes somewhat compromised in the western and mid-western regions due to limited crash samples. This suggests that additional crash data should be included to further enhance the performance of the GW model in these regions. Last but not least, the severity level of crashes could be further considered to explore the spatial heterogeneity of crash frequency across different severity levels. By conducting joint investigations into crash occurrences and severity, more effective and informative results can be attained to provide countermeasures of reducing crashes for researchers and policy-makers.

CRediT authorship contribution statement

Shuli Wang: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Kun Gao:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Lanfang Zhang:** Validation, Methodology, Conceptualization. **Bo Yu:** Writing – original draft, Methodology, Data curation. **Said M. Easa:** Writing – original draft, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The research was supported by Chinese Scholarship Council Scholarship, VINNOVA, Sweden (2019-03418) and Area of Advance Transport at Chalmers University of Technology, Sweden. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

Appendix A

The matchup between road types in this study and road types from OSM is shown in Table A.1 and the matching between the land use categories and the POI categories from OSM is presented in Table A.2.

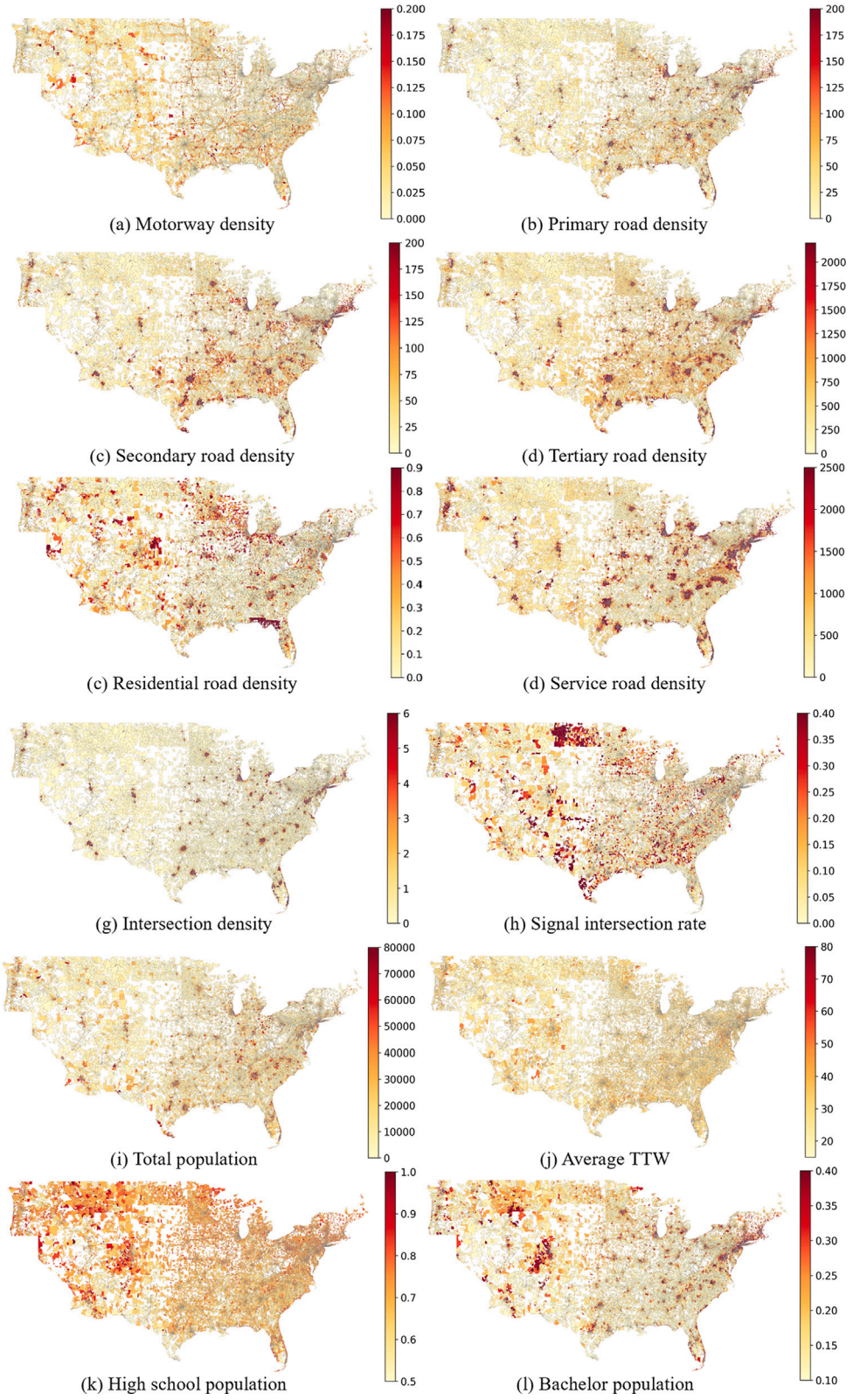


Fig. B.1. ZCTA-level values of all influencing factors investigated.

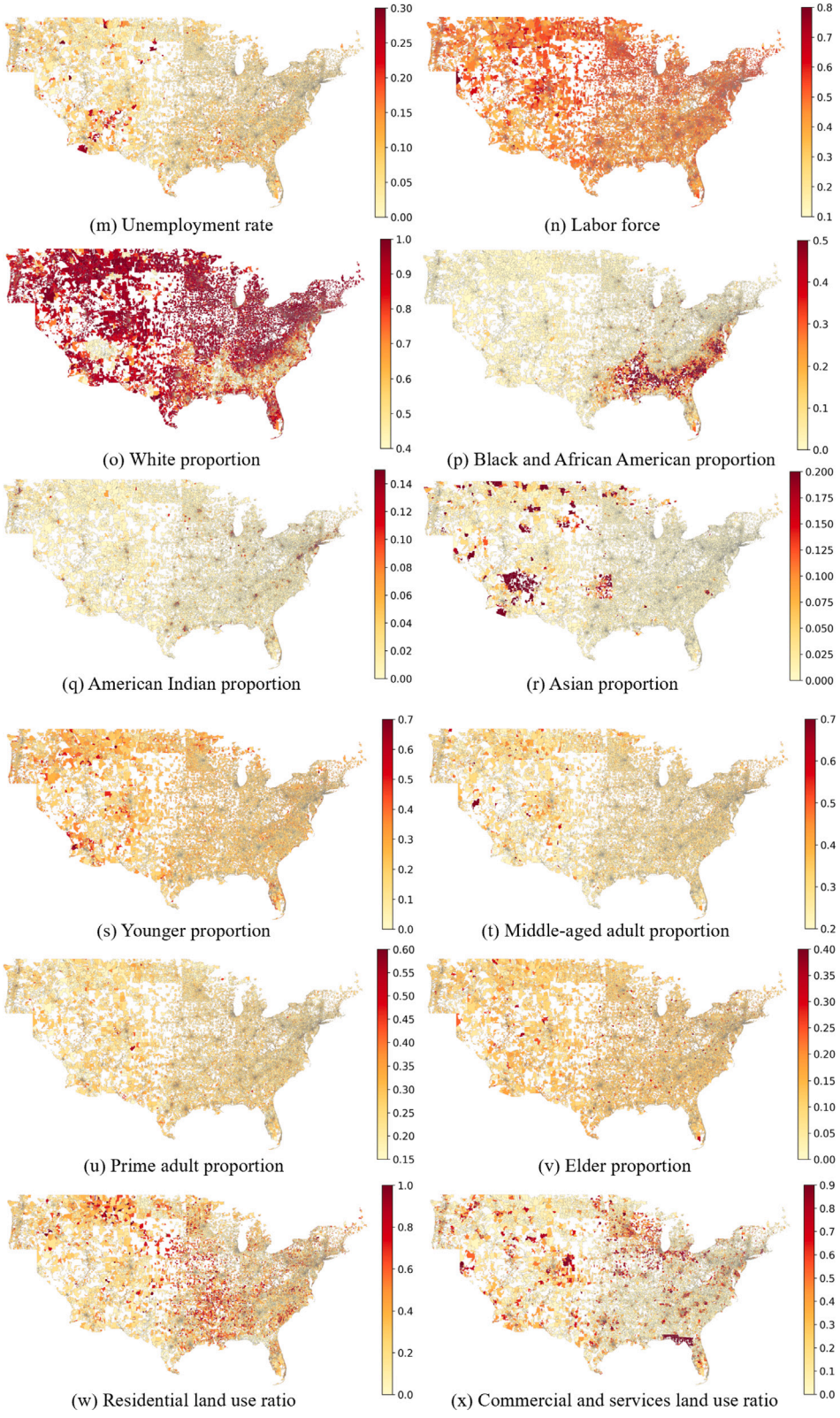
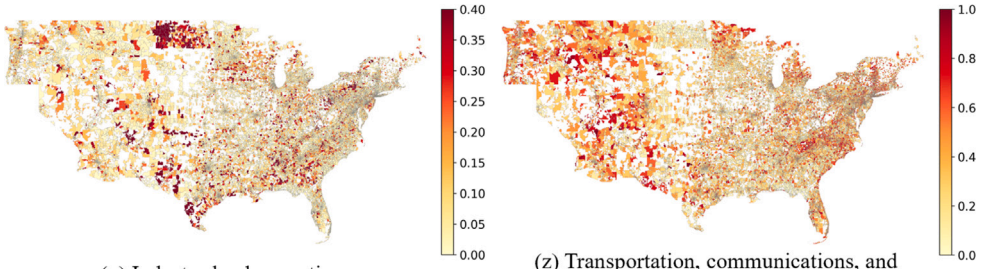


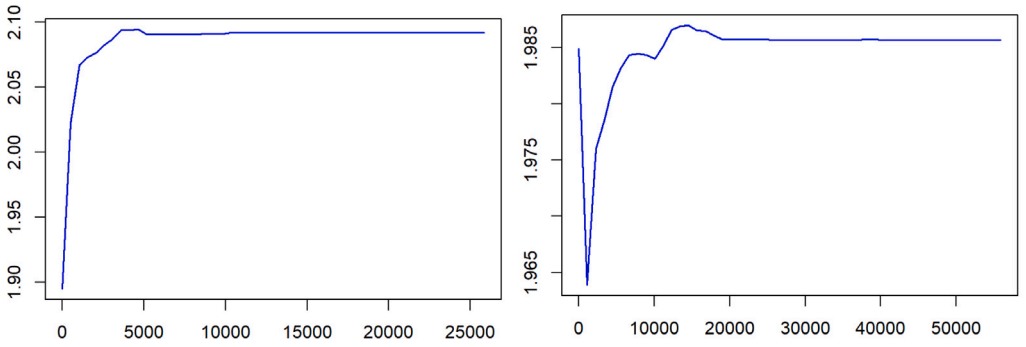
Fig. B.1. (*continued*).



(y) Industry land use ratio

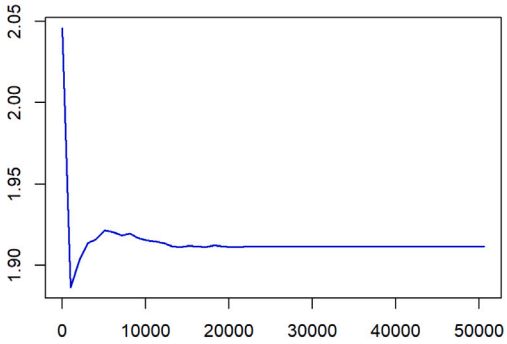
(z) Transportation, communications, and utilities land use ratio

Fig. B.1. (*continued*).

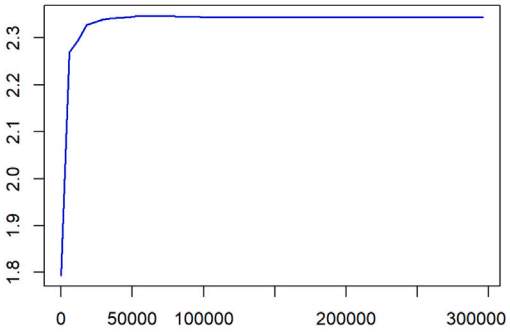


(a) Primary road density

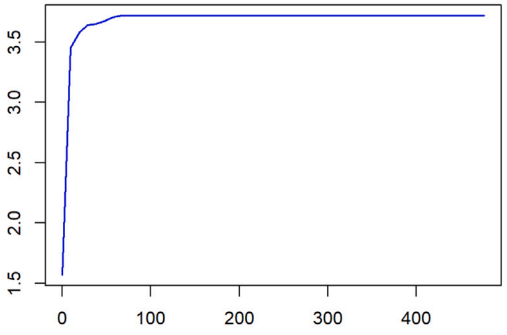
(b) Secondary road density



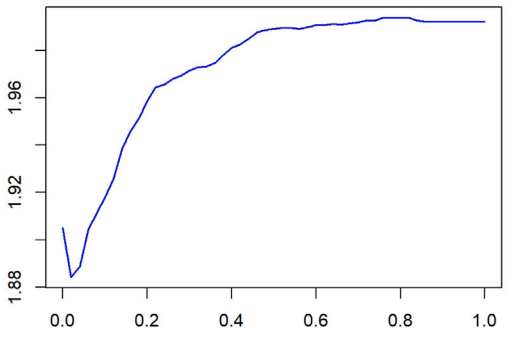
(c) Residential road density



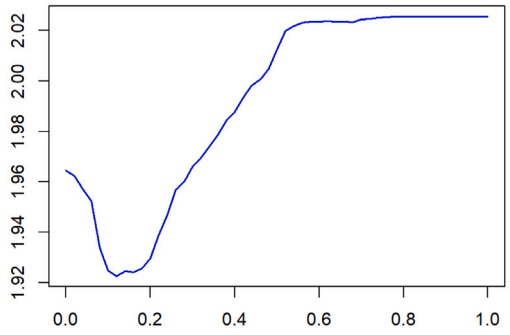
(d) Service road density



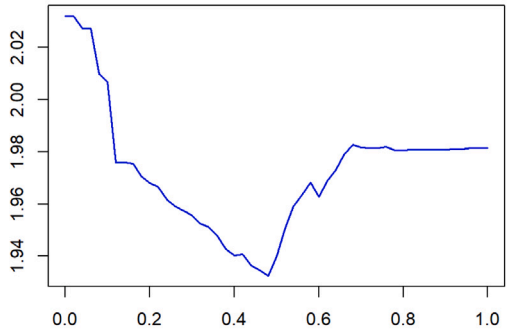
(e) Intersection density



(f) Signal intersection rate

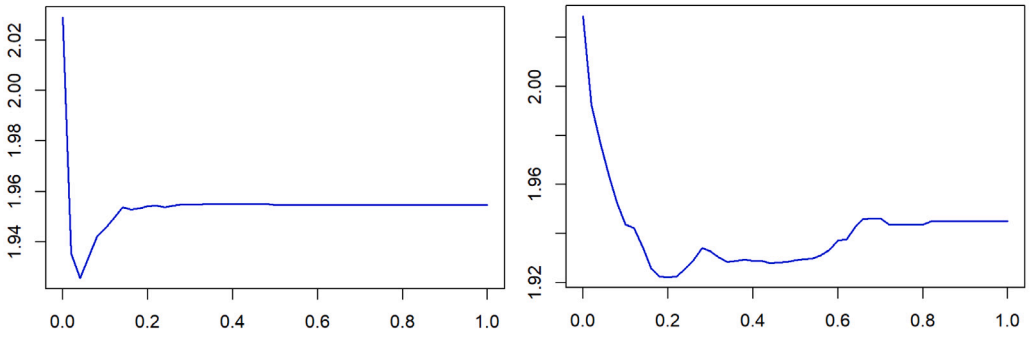


(g) Bachelor proportion



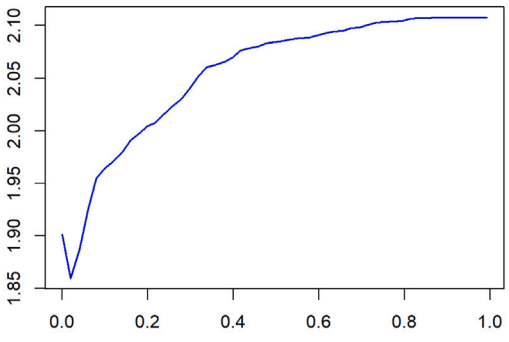
(h) Labor force proportion

Fig. B.2. Partial dependency profiles of the remaining variables of the global random forest model. Noting that the y-coordinate is the average prediction.

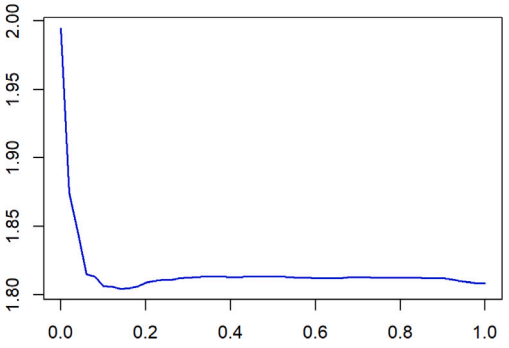


(i) Unemployment rate

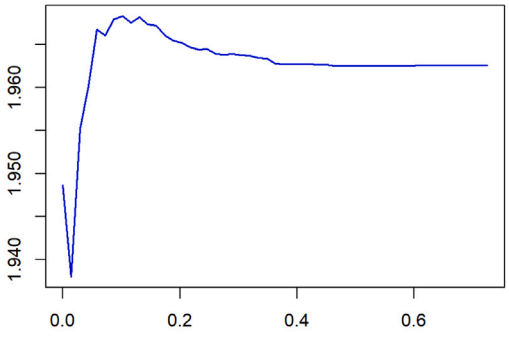
(j) Average travel time to work



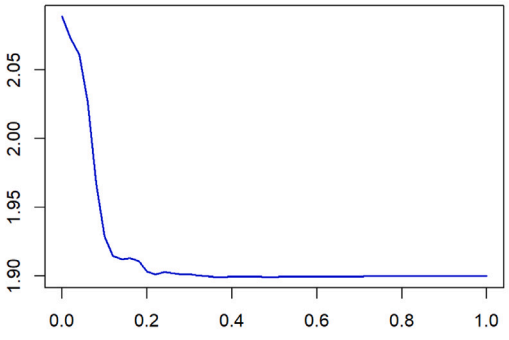
(k) Black and African American proportion



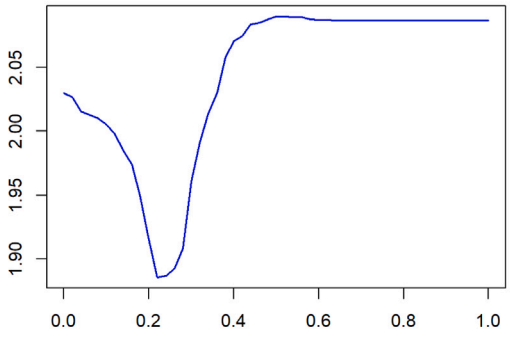
(l) American Indian proportion



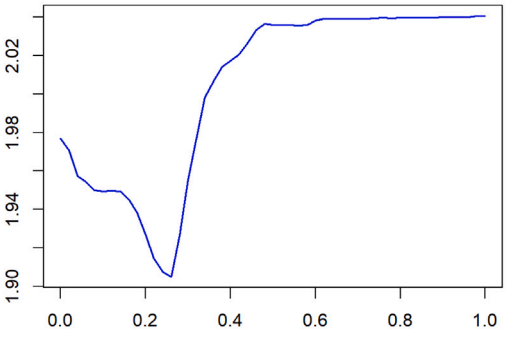
(m) Asian proportion



(n) Young proportion

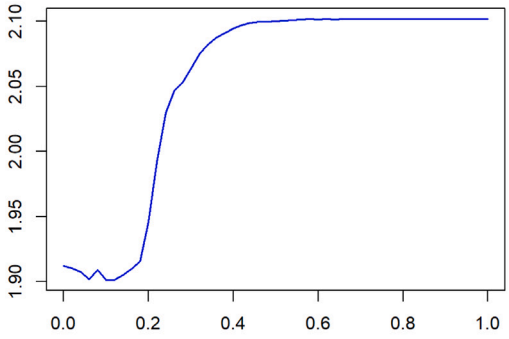


(o) Prime adult proportion

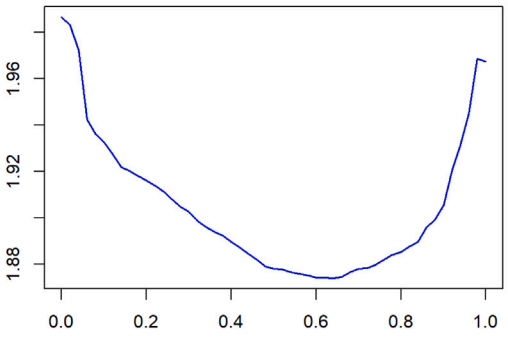


(p) Middle-aged adult proportion

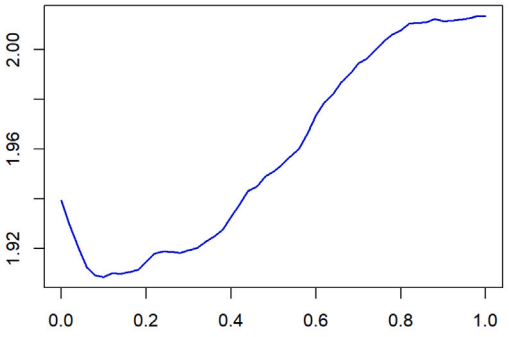
Fig. B.2. (*continued*).



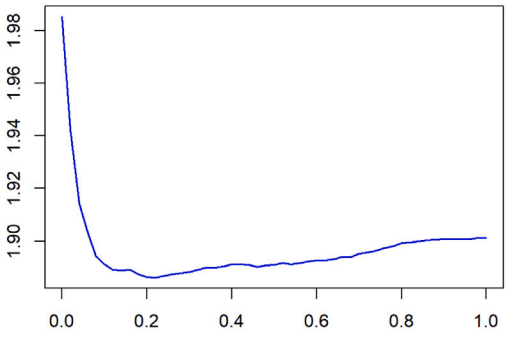
(q) Elder proportion



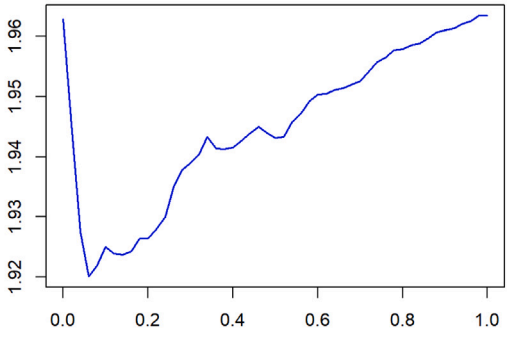
(r) Residential land use ratio



(s) Commercial and services land use ratio

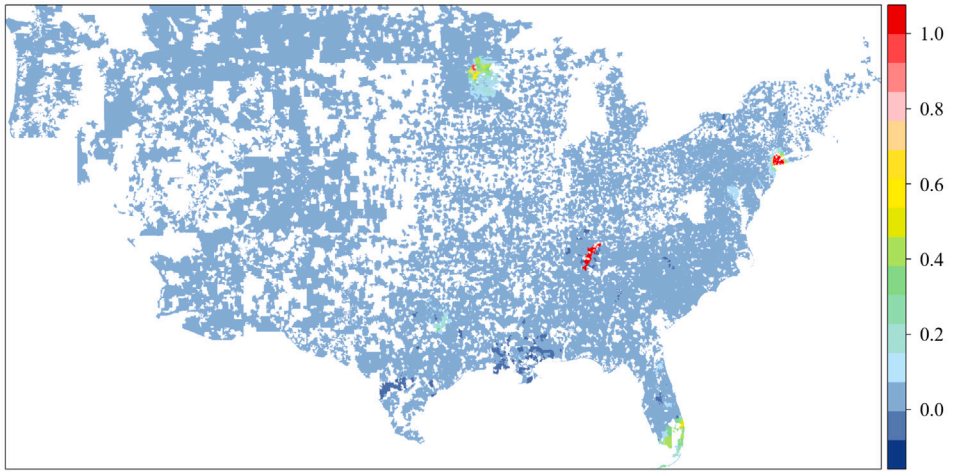


(t) Industry land use ratio

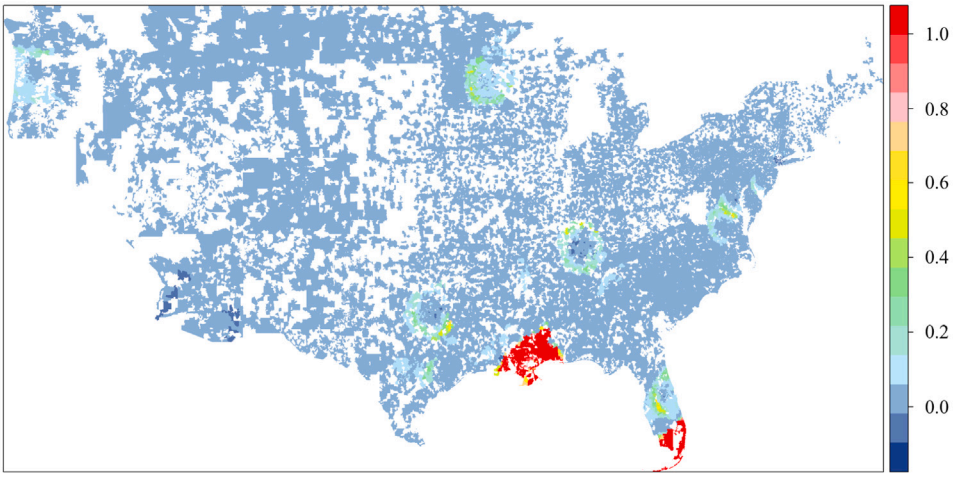


(u) Transportation, communications, and utilities land use ratio

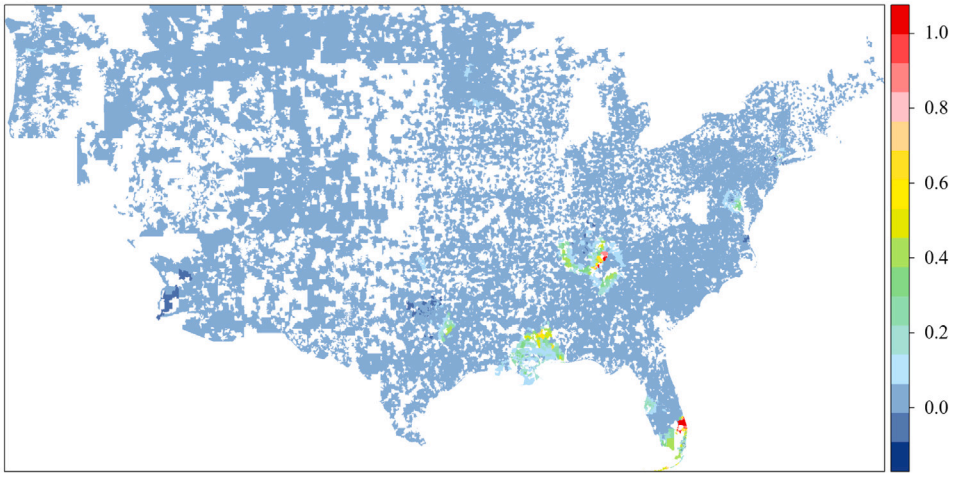
Fig. B.2. (*continued*).



(a) Motorway density

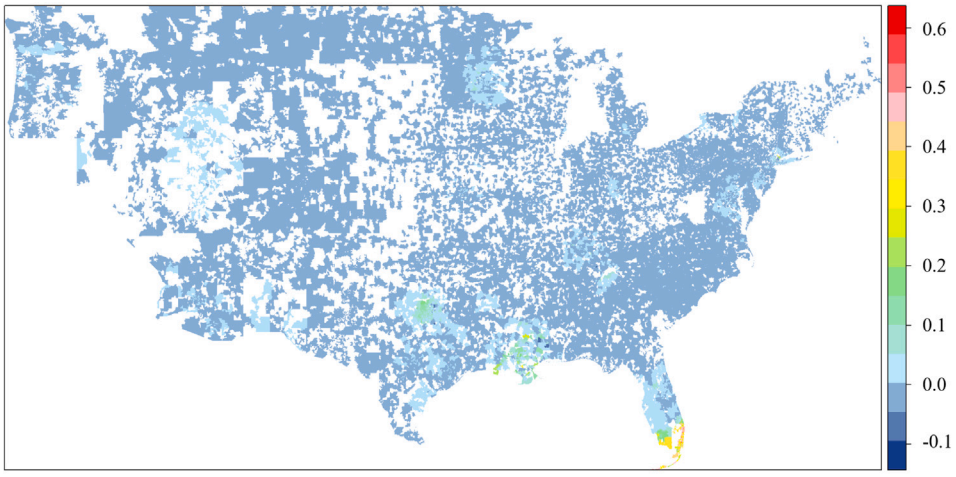


(b) Tertiary road density

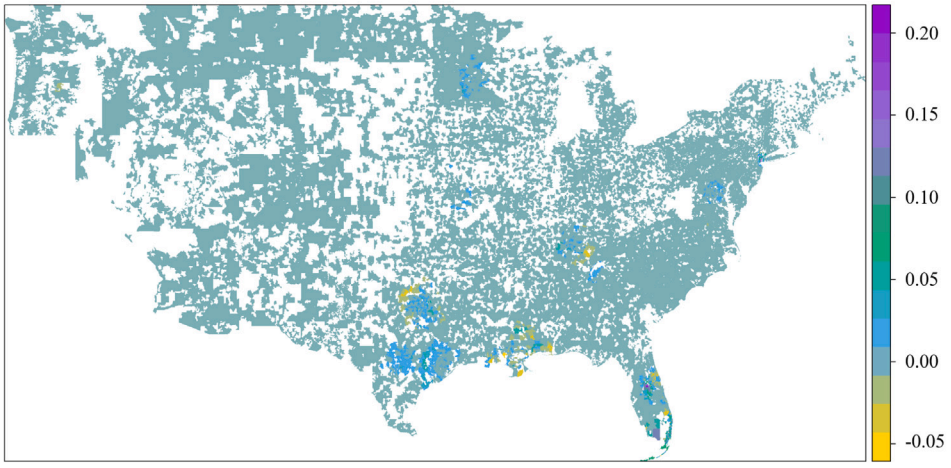


(c) Residential road density

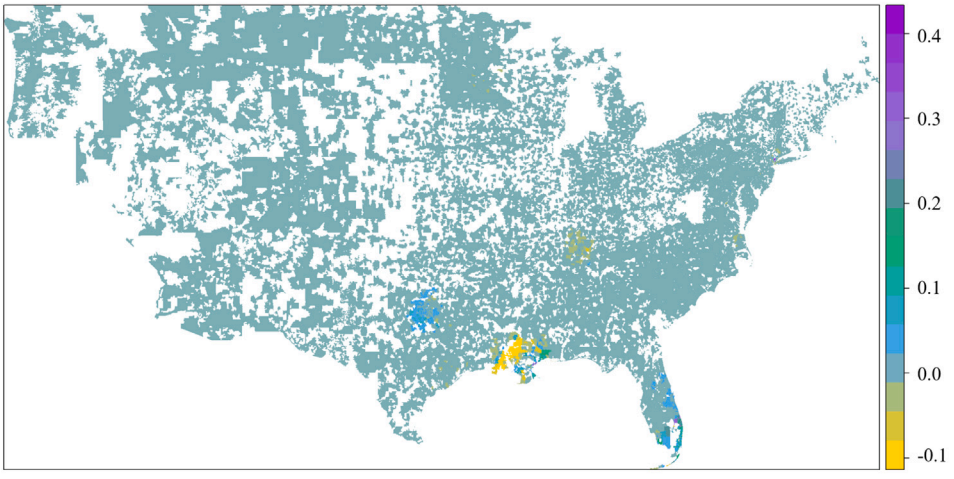
Fig. B.3. Spatial heterogeneity distribution of local effects of the remaining factors on crash frequency.



(d) Signal intersection rate

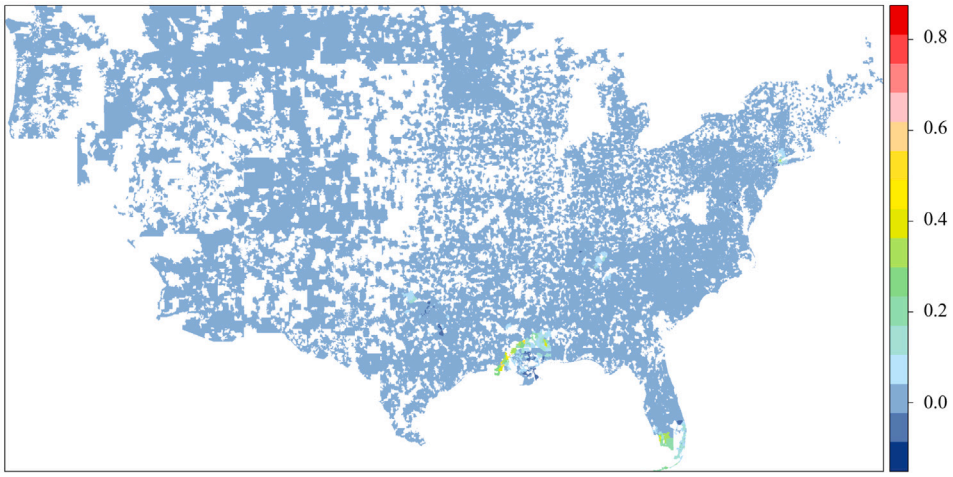


(e) Industry land use ratio

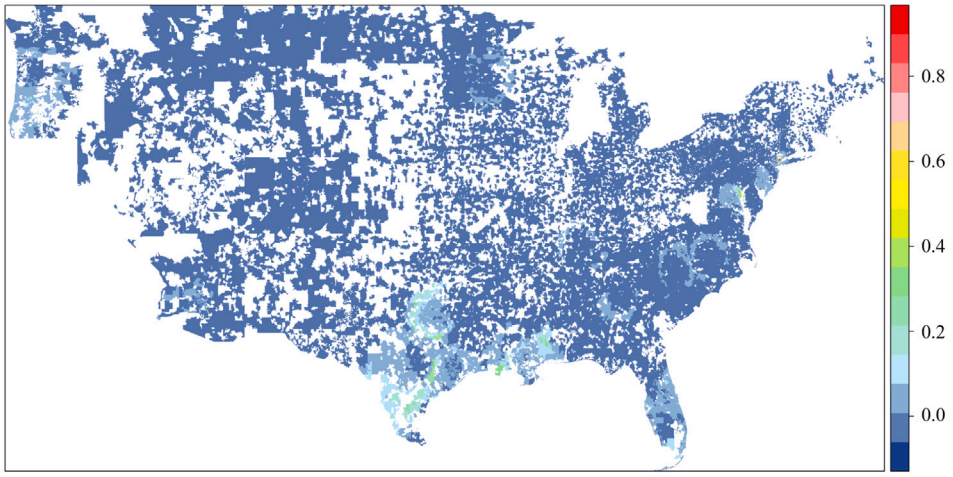


(f) Transportation, communications, and utilities land use ratio

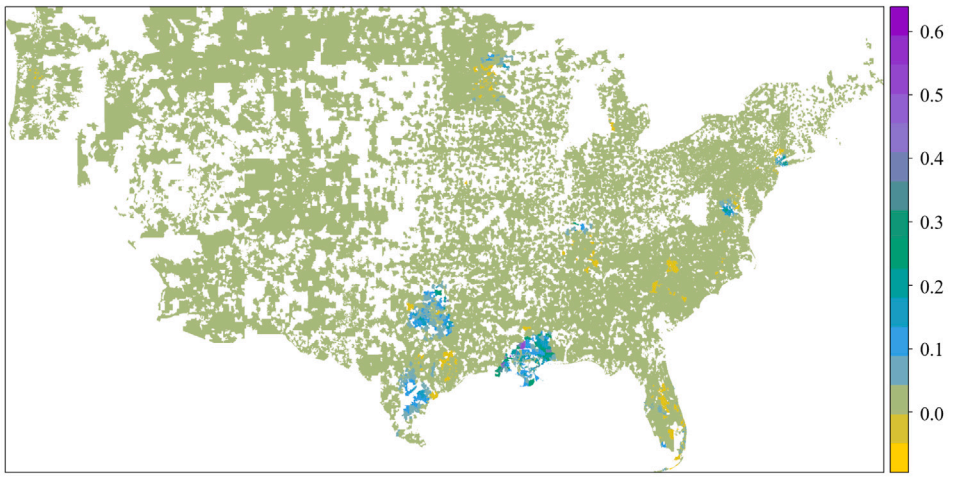
Fig. B.3. (*continued*).



(g) Population density

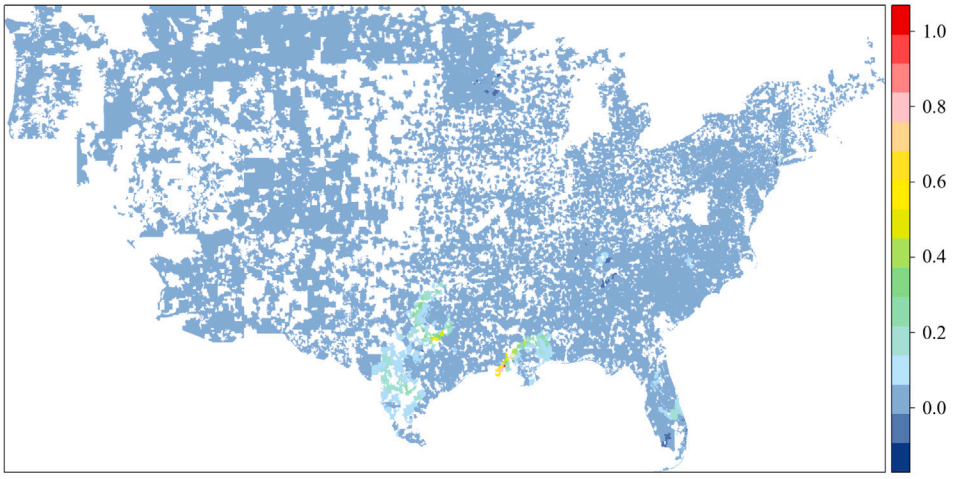


(h) Bachelor proportion

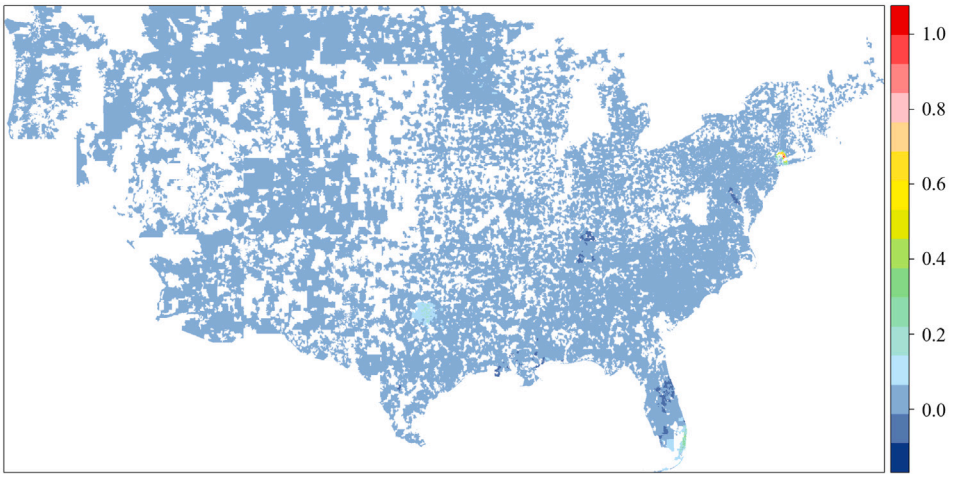


(i) High school proportion

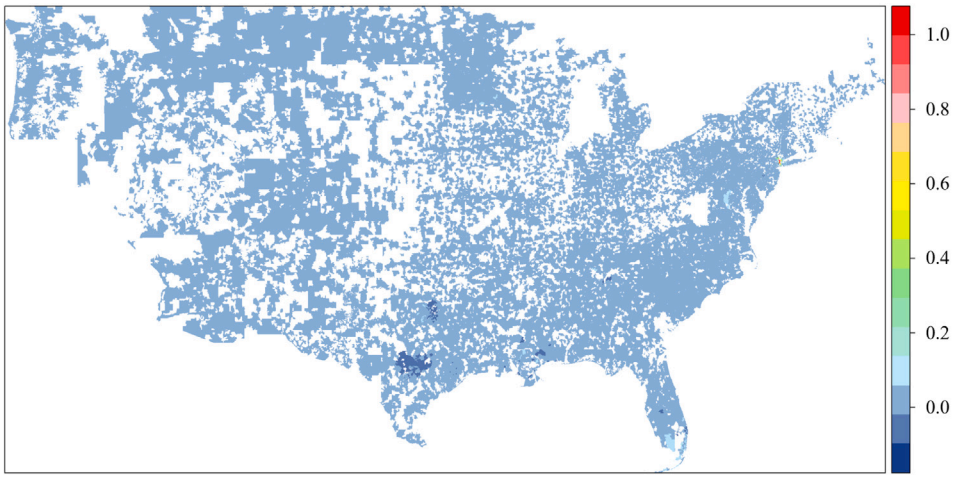
Fig. B.3. (*continued*).



(j) Labor force proportion

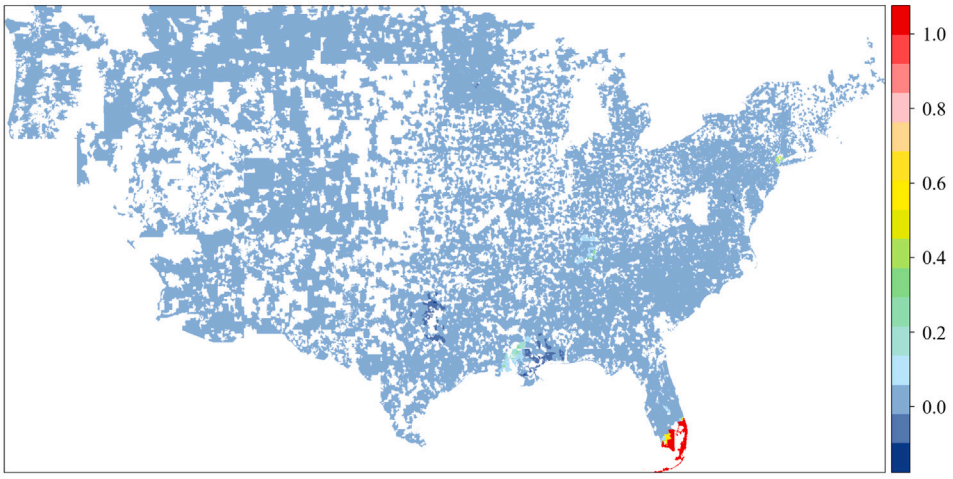


(k) Poverty rate

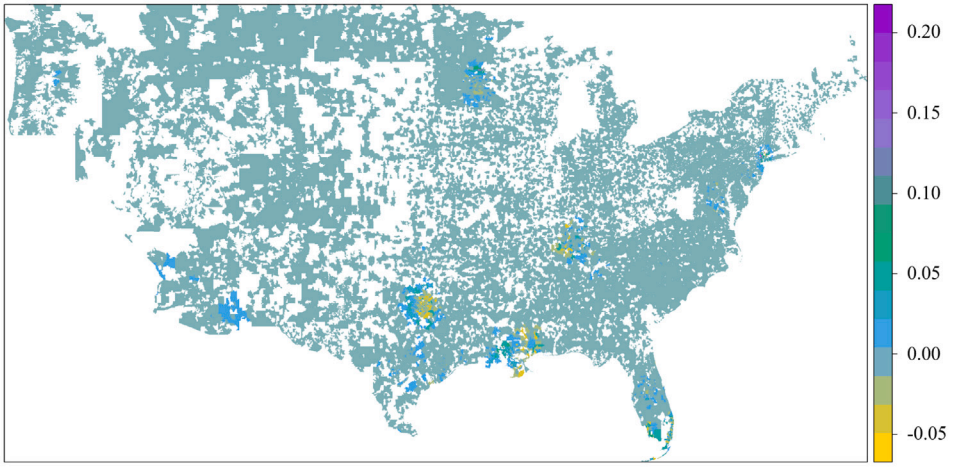


(l) Unemployment rate

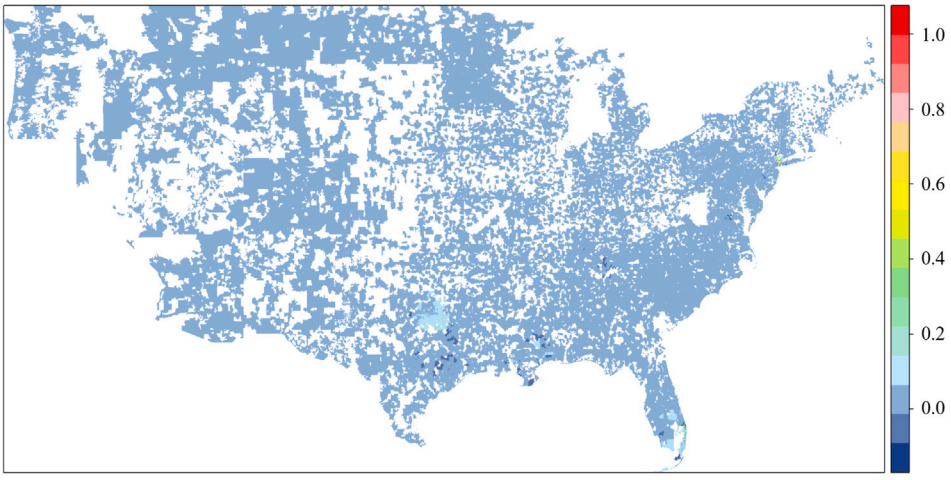
Fig. B.3. (*continued*).



(m) Average Travel Time to Work

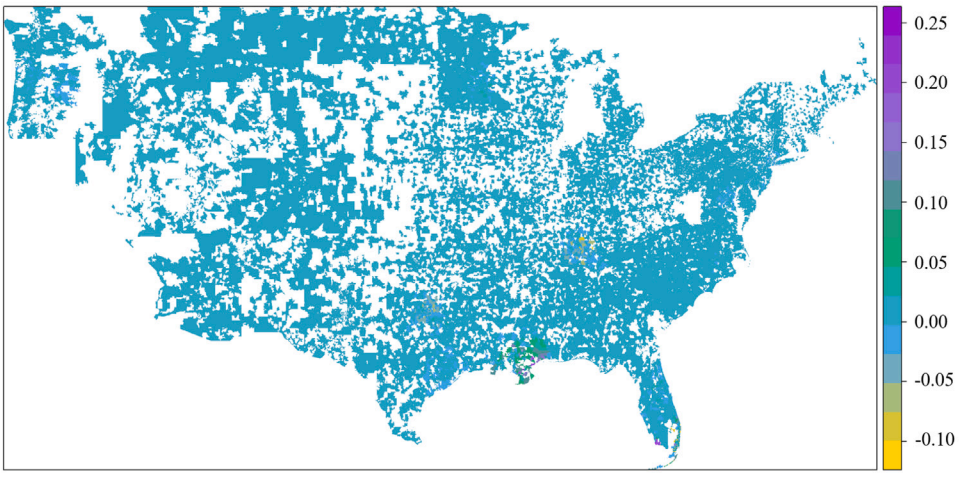


(n) American Indian proportion

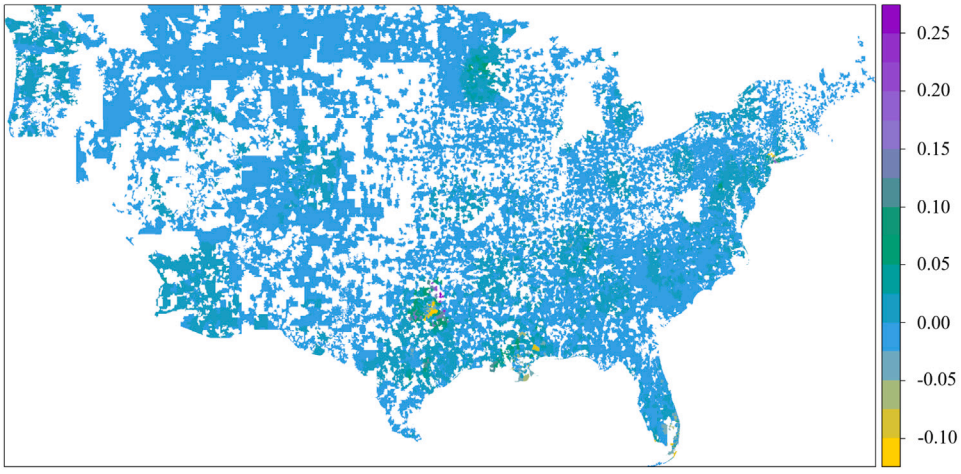


(o) Asian proportion

Fig. B.3. (*continued*).



(p) Prime adult proportion



(q) Middle-aged adult proportion

Fig. B.3. (*continued*).

Appendix B

The spatial distributions of influencing factors of crash frequency at the ZCTA-level are checked and depicted in Fig. B.1. Besides, partial dependency profiles of the remaining variables of the global random forest model are shown in Fig. B.2 (see Fig. B.3).

References

Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. Transp. Res. A 49, 62–75. http://dx.doi.org/10.1016/j.tra.2013.01.030.

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2017. The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. Anal. Methods Accid. Res. 13, 28–51. http://dx.doi.org/10.1016/j.amar.2016.11.001.

Azimian, A., Pyrialakou, V.D., Lavrenz, S., Wen, S., 2021. Exploring the effects of area-level factors on traffic crash frequency by severity using multivariate space-time models. Anal. Methods Accid. Res. 31, 100163.

Bao, J., Liu, P., Qin, X., Zhou, H., 2018. Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data. Accid. Anal. Prev. 120, 281–294. http://dx.doi.org/10.1016/j.aap.2018.08.014.

Bao, J., Liu, P., Yu, H., Xu, C., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accid. Anal. Prev. 106, 358–369. http://dx.doi.org/10.1016/j.aap.2017.06.012.

Bao, J., Yang, Z., Zeng, W., Shi, X., 2021. Exploring the spatial impacts of human activities on urban traffic crashes using multi-source big data. J. Transp. Geogr. 94, 103118.

Bi, H., Ye, Z., Zhu, H., 2022. Examining the nonlinear impacts of built environment on ridesourcing usage: Focus on the critical urban sub-regions. J. Clean. Prod. 350, 131314. http://dx.doi.org/10.1016/j.jclepro.2022.131314.

Cai, Q., Abdel-Aty, M., Lee, J., 2017. Macro-level vulnerable road users crash analysis: A Bayesian joint modeling approach of frequency and proportion. Accid. Anal. Prev. 107, 11–19. http://dx.doi.org/10.1016/j.aap.2017.07.020.

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. Accid. Anal. Prev. 93, 14–22. http://dx.doi.org/10.1016/j.aap.2016.04.018.

Chiou, Y.-C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accid. Anal. Prev. 50, 73–82. http://dx.doi.org/10.1016/j.aap.2012.03.030.

Chiou, Y.-C., Fu, C., Chih-Wei, H., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Anal. Methods Accid. Res. 2, 1–11. http://dx.doi.org/10.1016/j.amar.2013.12.001.

El-Basyouny, K., Sayed, T., 2009. Urban arterial accident prediction models with spatial effects. Transp. Res. Rec. 2102 (1), 27–33.

Gao, K., Yang, Y., Gil, J., Qu, X., 2023. Data-driven interpretation on interactive and nonlinear effects of the correlated built environment on shared mobility. J. Transp. Geogr. 110, 103604. http://dx.doi.org/10.1016/j.jtrangeo.2023.103604.

Gao, K., Yang, Y., Li, A., Qu, X., 2021. Spatial heterogeneity in distance decay of using bike sharing: An empirical large-scale analysis in Shanghai. Transp. Res. D 94, 102814. http://dx.doi.org/10.1016/j.trd.2021.102814.

Gomes, M.J.T.L., Cunto, F., da Silva, A.R., 2017. Geographically weighted negative binomial regression applied to zonal level safety performance models. Accid. Anal. Prev. 106, 254–261.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42 (6), 1556–1565. http://dx.doi.org/10.1016/j.aap.2010.03.013.

Jonathan, A.-V., Wu, K.-F.K., Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accid. Anal. Prev. 87, 8–16. http://dx.doi.org/10.1016/j.aap.2015.11.006.

Lee, J., Abdel-Aty, M., 2018. Macro-level analysis of bicycle safety: Focusing on the characteristics of both crash location and residence. Int. J. Sustain. Transp. 12 (8), 553–560.

Li, X., Chen, W., Zhang, Q., Wu, L., 2020. Building auto-encoder intrusion detection system based on random forest feature selection. Comput. Secur. 95, 101851.

Li, A., Gao, K., Zhao, P., Axhausen, K.W., 2024. Integrating shared e-scooters as the feeder to public transit: A comparative analysis of 124 European cities. Transp. Res. C 160, 104496. http://dx.doi.org/10.1016/j.trc.2024.104496.

Li, S., Zhao, Z., Miaomiao, X., Wang, Y., 2010. Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression. Environ. Model. Softw. 25 (12), 1789–1800. http://dx.doi.org/10.1016/j.envsoft.2010.06.011.

Liu, J., Khattak, A.J., Wali, B., 2017. Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity. Accid. Anal. Prev. 109, 132–142. http://dx.doi.org/10.1016/j.aap.2017.10.012.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transp. Res. A 44 (5), 291–305. http://dx.doi.org/10.1016/j.tra.2010.02.001.

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees. Adv. Neural Inf. Process. Syst. 26.

Luo, Q., Forscher, T., Shaheen, S., Deakin, E., Walker, J.L., 2023. Impact of the COVID-19 pandemic and generational heterogeneity on ecommerce shopping styles – A case study of Sacramento, California. Commun. Transp. Res. 3, 100091. http://dx.doi.org/10.1016/j.commtr.2023.100091.

Luo, Y., Yan, J., McClure, S., 2021. Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. Environ. Sci. Pollut. Res. 28, 6587–6599.

Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. Anal. Methods Accid. Res. 15, 29–40. http://dx.doi.org/10.1016/j.amar.2017.06.001.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22. http://dx.doi.org/10.1016/j.amar.2013.09.001.

Ouyang, Y., Bejleri, I., 2014. Geographic information system–based community-level method to evaluate the influence of built environment on traffic crashes. Transp. Res. Rec. 2432 (1), 124–132.

Pervaz, S., Bhowmik, T., Eluru, N., 2022. Integrating macro and micro level crash frequency models considering spatial heterogeneity and random effects. Anal. Methods Accid. Res. 36, 100238. http://dx.doi.org/10.1016/j.amar.2022.100238.

Pirdavani, A., Bellemans, T., Brijs, T., Kochan, B., Wets, G., 2014. Assessing the road safety impacts of a teleworking policy by means of geographically weighted regression method. J. Transp. Geogr. 39, 96–110. http://dx.doi.org/10.1016/j.jtrangeo.2014.06.021.

Qian, X., Ukkusuri, S.V., 2015. Spatial variation of the urban taxi ridership using GPS data. Appl. Geogr. 59, 31–42.

Qu, X., Lin, H., Liu, Y., 2023. Envisioning the future of transportation: Inspiration of ChatGPT and large models. Commun. Transp. Res. 3, 100103. http://dx.doi.org/10.1016/j.commtr.2023.100103.

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. Accid. Anal. Prev. 40 (4), 1486–1497. http://dx.doi.org/10.1016/j.aap.2008.03.009.

Quiñones, S., Goyal, A., Ahmed, Z.U., 2021. Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. Sci. Rep. 11 (1), 6955.

Rhee, K.-A., Kim, J.-K., ihn Lee, Y., Ulfarsson, G.F., 2016. Spatial regression analysis of traffic crashes in Seoul. Accid. Anal. Prev. 91, 190–199. http://dx.doi.org/10.1016/j.aap.2016.02.023.

Santos, F., Graw, V., Bonilla, S., 2019. A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. PLoS One 14 (12), e0226224.

Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. Accid. Anal. Prev. 90, 152–158. http://dx.doi.org/10.1016/j.aap.2016.02.018.

Wang, Y., Kockelman, K.M., 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accid. Anal. Prev. 60, 71–84. http://dx.doi.org/10.1016/j.aap.2013.07.030.

Wang, X., Yang, J., Lee, C., Ji, Z., You, S., 2016. Macro-level safety analysis of pedestrian crashes in Shanghai, China. Accid. Anal. Prev. 96, 12–21. http://dx.doi.org/10.1016/j.aap.2016.07.028.

Wang, W., Yuan, Z., Yang, Y., Yang, X., Liu, Y., 2019. Factors influencing traffic accident frequencies on urban roads: a spatial panel time-fixed effects error model. PLoS One 14 (4), e0214539. http://dx.doi.org/10.1371/journal.pone.0214539.

Wang, X., Zhang, Q., Yang, X., Pei, Y., Yuan, J., 2022. Traffic safety analysis and model updating for freeways using Bayesian method. J. Transp. Saf. Secur. 15 (7), 1–23.

Wen, H., Zhang, X., Zeng, Q., Lee, J., Yuan, Q., 2019. Investigating spatial autocorrelation and spillover effects in freeway crash-frequency data. Int. J. Environ. Res. Public Health 16 (2), 219.

Wu, D., Zhang, Y., Xiang, Q., 2024. Geographically weighted random forests for macro-level crash frequency prediction. Accid. Anal. Prev. 194, 107370. http://dx.doi.org/10.1016/j.aap.2023.107370.

Xie, K., Ozbay, K., Yang, H., 2019. A multivariate spatial approach to model crash counts by injury severity. Accid. Anal. Prev. 122, 189–198.

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16–25. http://dx.doi.org/10.1016/j.aap.2014.10.020.

Xu, P., Huang, H., Dong, N., Wong, S., 2017. Revisiting crash spatial heterogeneity: A Bayesian spatially varying coefficients approach. Accid. Anal. Prev. 98, 330–337. http://dx.doi.org/10.1016/j.aap.2016.10.015.

Yuan, Y., Cave, M., Zhang, C., 2018. Using Local Moran's I to identify contamination hotspots of rare earth elements in urban soils of London. Appl. Geochem. 88, 167–178. http://dx.doi.org/10.1016/j.apgeochem.2017.07.011, SI: ISEG 2016.

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S., 2017. A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity. Accid. Anal. Prev. 99, 184–191. http://dx.doi.org/10.1016/j.aap.2016.11.018.

Zhang, H., Bao, J., Hong, Q., Chang, L., Yin, W., 2023. Zone-level traffic crash analysis with incorporated multi-sourced traffic exposure variables using Bayesian spatial model. J. Transp. Saf. Secur. 31, 1–24.

Ziakopoulos, A., Yannis, G., 2020. A review of spatial approaches in road safety. Accid. Anal. Prev. 135, 105323. http://dx.doi.org/10.1016/j.aap.2019.105323.