International Journal of Geo-Information

MDPI

Article

An Analysis of the Spatial Variations in the Relationship Between Built Environment and Severe Crashes

Onur Alisan * and Eren Erman Ozguven

Department of Civil and Environmental Engineering, FAMU-FSU College of Engineering, Tallahassee, FL 32310, USA; eozguven@eng.famu.fsu.edu
* Correspondence: oalisan@fsu.edu

Abstract: Traffic crashes significantly contribute to global fatalities, particularly in urban areas, highlighting the need to evaluate the relationship between urban environments and traffic safety. This study extends former spatial modeling frameworks by drawing paths between global models, including spatial lag (SLM), and spatial error (SEM), and local models, including geographically weighted regression (GWR), multi-scale geographically weighted regression (MGWR), and multi-scale geographically weighted regression with spatially lagged dependent variable (MGWRL). Utilizing the proposed framework, this study analyzes severe traffic crashes in relation to urban built environments using various spatial regression models within Leon County, Florida. According to the results, SLM outperforms OLS, SEM, and GWR models. Local models with lagged dependent variables outperform both the global and generic versions of the local models in all performance measures, whereas MGWR and MGWRL outperform GWR and GWRL. Local models performed better than global models, showing spatial non-stationarity; so, the relationship between the dependent and independent variables varies over space. The better performance of models with lagged dependent variables signifies that the spatial distribution of severe crashes is correlated. Finally, the better performance of multi-scale local models than classical local models indicates varying influences of independent variables with different bandwidths. According to the MGWRL model, census block groups close to the urban area with higher population, higher education level, and lower car ownership rates have lower crash rates. On the contrary, motor vehicle percentage for commuting is found to have a negative association with severe crash rate, which suggests the locality of the mentioned associations.

Keywords: built environment; severe crashes; multi-scale geographically weighted regression; multi-scale geographically weighted regression with lagged dependent variable

Citation: Alisan, O.; Ozguven, E.E. An Analysis of the Spatial Variations in the Relationship Between Built Environment and Severe Crashes. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 465. https://doi.org/10.3390/ijgi13120465

Academic Editor: Hartwig H. Hochmair and Wolfgang Kainz

Received: 9 October 2024
Revised: 18 December 2024
Accepted: 20 December 2024
Published: 22 December 2024

1. Introduction

In 2021, roadway traffic crashes resulted in 1,354,840 fatalities globally, translating to an average of 3711 deaths per day [1]. The economic burden of these crashes is substantial; for example, in 2019, the total financial cost of roadway crashes in the United States was approximately USD 339.8 billion, compared to USD 242 billion in 2010 [2,3]. Additionally, during this period, the annual average number of traffic-related fatalities was 34,885 [4].

Recent trends indicate a significant shift in traffic fatalities from rural to urban areas [5]. In 2019, urban roadways accounted for 56.2% of traffic fatalities, while rural roadways accounted for 43.8% [3]. This shift corresponds to a 15% increase in urban vehicle miles traveled and a 0.3% reduction in rural vehicle miles traveled since 2009 [6]. This changing pattern highlights the need for a deeper understanding of the factors contributing to urban traffic crashes.

The substantial negative impacts of traffic crashes, both human and material, have led to an extensive body of research on traffic safety. A significant portion of this research focuses on the relationship between the built environment and traffic crashes, particularly

in urban areas where many fatalities occur. Studies in this field can be categorized primarily by the spatial unit of analysis: individual spots (e.g., roadway intersections or segments), zonal areas (e.g., neighborhoods, traffic analysis zones, or census tracts), and regional areas (e.g., urban clusters, counties, or metropolitan areas) [7–9]. Additionally, these studies vary based on the statistical methodologies employed, such as Poisson regression or Bayesian modeling [7,10].

In terms of the level of analysis, research at the zonal level is predominant compared to spot-level and regional-level studies [7,9]. This preference is largely due to the availability of built environment data at the zonal level, which facilitates a better understanding of the impacts of built environment factors on traffic crashes [11,12]. Zonal-level analysis is particularly useful for examining aggregated crash data, addressing randomness, and investigating crash types such as pedestrian or severe crashes, which occur less frequently than property damage-only crashes [9]. In contrast, spot-level studies typically focus on "black spots" and are more concerned with traffic volumes and roadway design, rather than built environment characteristics. Regional-level studies, while valuable for policy-oriented and long-term analyses, are less suited for detailed insights into local crash dynamics.

In terms of methodology, many studies have utilized global statistical models to explore associations between crashes and built environment factors. Global models assume uniform relationships between explanatory variables and crash outcomes across different locations and often use aggregated data for analysis [13]. Common global statistical approaches include Poisson regression [14], negative binomial/Poisson-gamma models [11,14,15], zero-inflated models [14,16,17], generalized additive models [17,18], and multivariate models [19,20].

While global models are effective at identifying overall patterns between crashes and their causes, a key drawback of this approach is that these models do not account for spatial characteristics. They overlook factors like spatial autocorrelation and non-stationarity. Most global statistical models assume that observations are independent, but spatial occurrences, including crash events, tend to be spatially correlated. Moreover, in global models, the relationship between the dependent and independent variables is assumed to be the same over space. However, the relationships would vary from one location to another [13]. The first drawback could be handled by introducing spatial components in global models, but those models also cannot capture spatial non-stationarity. To address this limitation, local models like geographically weighted regression (GWR) have been employed. GWR allows for varying relationships between variables across different locations, capturing spatial variations that global models miss.

Among the studies using GWR in crash analysis, Gomes et al. applied Geographically Weighted Negative Binomial Regression (GWNBR) to examine the associations between injury crashes and exposure, network characteristics, socioeconomic factors, and land use [21]. Huang et al. used GWR to analyze spatial relationships between crashes and built environment [13], Tang et al. used a geographically weighted Poisson quantile regression (GWPQR) model to investigate the spatial effects on crash frequency [22], Aljoufie et al. examined the relationship between pedestrian fatalities and other built environmental variables by geographically weighted Poisson regression (GWPR) [23], and Zafri and Khan used geographically weighted logistic regression (GWLR) for examining the associations between pedestrian crash severity and built environment factors [24].

A fundamental limitation of traditional GWR is its use of a single bandwidth, which may not capture scale differences in relationships. This limitation has been addressed by multi-scale geographically weighted regression (MGWR), which incorporates multiple bandwidths to better capture spatial heterogeneity. Among the recent studies on traffic safety that employ MGWR, Qu et al. examined the influences of point-of-interest on traffic crashes [25], Li et al. explored the association between collision risk and the social vulnerability index [26], Zafri and Khan investigated the relationships between the built environment and pedestrian crash occurrences [27], Tang et al. investigated the spatial variations in moving-vehicle crashes and fixed-object crashes in relation to the road net-

work, geographic, demographic and socioeconomic, and land-use variables [28], Liu et al. analyzed the relationship between social disparities and traffic crash frequency [29], and Kuo et al. examined the spatial heterogeneity between e-bike and motorcycle crashes and environmental variables [30].

Though MGWR is the latest and most advanced version among the aforementioned spatial models, it is computationally intensive [31]. As every spatial model has pros and cons, it is important to draw a spatial framework to guide model selection. There are formerly introduced frameworks [27,32], and the main aim of our study is to propose a more comprehensive framework. In addition, considering the emerging use of MGWR, research investigating local and spatially varying relationships between severe traffic crashes and urban built environment remains sparse, where severe traffic crashes are assumed to consist of incapacitating injury and fatal injury crashes. Thus, the second aim of this study is to fill this gap by examining the relationship between severe crashes and built environment factors within the proposed spatial framework, comparing modeling results, and providing a detailed analysis that could guide future research and policy development. It focuses on variables related to urban density, design, and diversity, and integrates extensive demographic and socioeconomic data provided at the block level [33–35]. Furthermore, this study employs a combination of global and local spatial regression models to capture spatial variations in severe crash risk. To optimize the explanatory power of the models, Genetic Algorithm and Tabu-Search metaheuristic methods are utilized to identify the most significant variables. The analysis is applied to data obtained for Leon County, Florida.

The contributions of this research are twofold. First, it extends existing studies by integrating spatial regression models within a spatial modeling framework. Second, it offers insights into the spatial dynamics of traffic safety considering local spatial variations by thoroughly examining how built environment factors relate to severe traffic crashes.

This paper is organized as follows: Section 2 introduces the study area, the dataset, and the methodology. Section 3 presents the results, and Section 4 discusses the results. Section 5 concludes this paper.

2. Materials and Methods

2.1. Study Area and Data

The proposed methodology was applied to a case study of Leon County located in the northwestern part of Florida, U.S. Tallahassee, the capital city of the State of Florida, is located in Leon County. According to the US Census 2020, the total population of Leon County is 292,198, and 196,169 people are settled at the urban core that includes Tallahassee [35]. The location of the study area (with county borders of Florida), the urban areas in Leon County, and the census block group borders are shown in Figure 1.

The dataset mainly consisted of the crash records for Leon County for the 2017–2019 period, gathered from the Florida Department of Transportation [36]. For this study, severe crashes, which include fatal injury (within 30 days) and incapacitating injury crashes, are considered according to the State of Florida injury classification. The respective classes are K (fatal injury) and A (suspected serious injury), respectively, according to the KABCO injury classification scale [37,38]. Several steps were processed to analyze crashes at the population block group level. Crash data consist of geocoded points in space, but population block groups are polygonal units. One way to determine the crash frequency of each population block group is to count the number of crashes that occurred in the respective census block group. However, as crashes usually happen on roadway segments, and roadway segments are also the borders of geographic units, there is a conflict about assigning a crash to a single geographic unit. As the geographic units become smaller (e.g., from county to census tract), the number of conflicting assignments would increase since the divisions and roadway intersections increase.

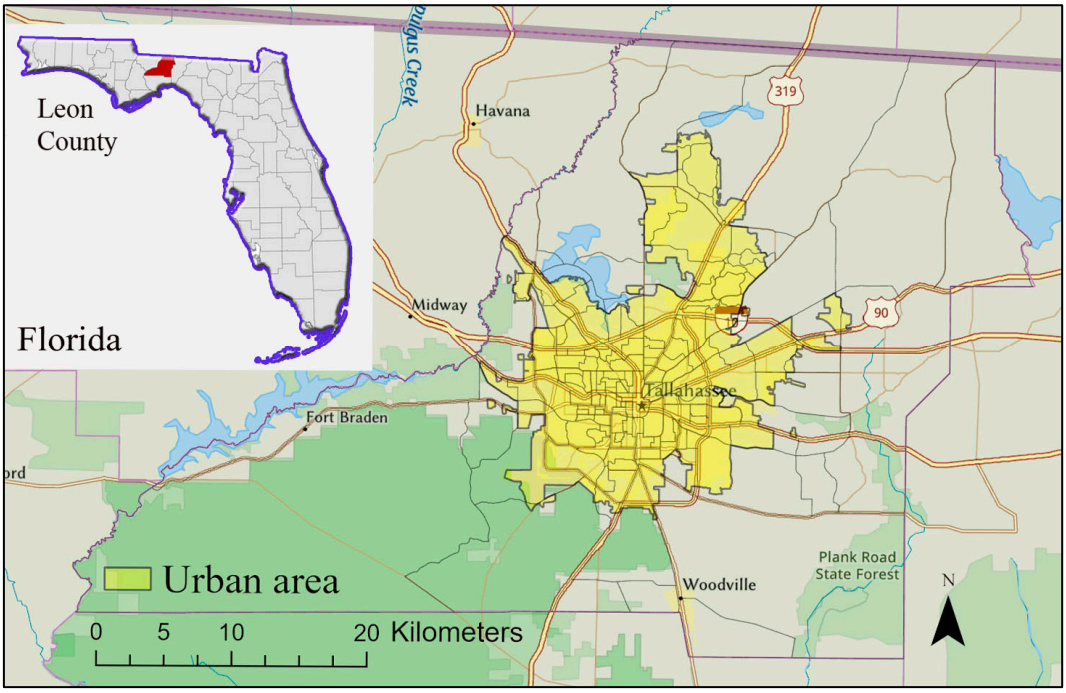

Figure 1. Study area and census block groups.

Moreover, it could be assumed that a crash happening at an intersection is affected by all the bounding geographic units [11]. In this study, the crash data were assigned to the population block groups using buffers. The buffer length was assumed to be 125 m (410 feet) for two reasons: (a) to consider high-width roadways, and (b) to capture the higher number of units impacting crashes at the city centers.

For crash rate calculation, all crash instances were spatially joined to census block groups. Then, the crash rate of each block group was calculated by Equation (1):

$$R_i = \frac{C_i \times 100,000,000}{VMT_i \times 3} \quad (1)$$

where R_i is the crash rate, C_i is the total number of severe crashes during the study period (2017–2019), and VMT_i is the annual vehicle miles traveled for census block group i [13].

Several other datasets were utilized in this study, such as the 2019 American Community Survey data, which had demographic, socioeconomic, and land use information. Traffic volume (annual average daily traffic) and street network data were gathered from the Florida Department of Transportation. A similar assignment procedure with crashes was applied to traffic volume assignments. The averages of traffic volumes in 2017, 2018, and 2019 were assigned to block groups. Moreover, the total roadway length for each block group was calculated by summing the respective roadway assignments. Built environment data were mainly gathered from the Smart Location Database (SLD) [34]. We created variables using other variables structured under the five main built environment characteristics: density, diversity, design, transit accessibility, destination accessibility, and travel. Moreover, demographics and economic structure were used as another main characteristic dimension for which the data were gathered from US Census Bureau [35]. Selecting, merging, and joining operations to preprocess the built environment dataset were carried out in ArcGIS Pro [39].

In total, there were more than 300 independent variables. Despite this availability of substantial number of variables, not all could be used for the spatial regression analyses. Therefore, correlation analysis was used to select the candidates and eliminate severe collinearity problems. In this analysis, the threshold values were chosen as -0.7 and 0.7. Moreover, the decision on the variables to be selected was based on the work of Alisan

et al. [40]. The variables were sorted according to the main characteristics of the built environment. Table 1 presents the final list of variables and their descriptive statistics.

Table 1. Descriptive statistics of dependent and independent variables.

Variable [1]	Explanation	Mean	SD	Max	Min
Crash rate	Number of severe crashes per 100 million VMT	26.61	45.74	526.48	0.00
LN_crash_rate	Natural logarithm of cash rate	2.73	1.13	6.27	0.00
D0A	Urban indicator (1 Urban, 0 Rural)	0.88	0.33	1.00	0.00
D0B	Total population (×1000 persons)	1.64	0.86	5.62	0.15
D0B_B_P	% of Black population	32.08	25.82	99.20	0.00
D0B_H_P	% of Hispanic population	6.05	5.48	34.61	0.00
D0B_U5_P	% of population under the age of 5	5.02	3.80	15.87	0.00
D0B_O65_P	% of population above the age of 65	14.46	10.30	53.38	0.00
D0B_M_P	Male % of population	47.66	7.16	78.87	29.44
D0C_HHS	Average household size (persons)	2.42	0.44	4.07	1.11
D0C_OC_P	% of occupied housing units	87.17	11.14	100.00	29.34
D0C_B90_P	% of housing units built before 1990	60.78	23.32	100.00	3.45
D0C_A0_P	% of households with zero automobiles	7.35	8.98	47.40	0.00
D0C_A1_P	% of households with one automobile	38.30	13.55	79.24	8.30
D0D_L9	% of population 25 years and more with no schooling completed	1.53	2.76	25.17	0.00
D0D_CE	% of population 25 years and more with at least bachelor's degree	27.47	16.33	65.33	0.00
D0E	Total number of workers (×1000 workers)	0.69	0.38	2.43	0.17
D0E_HI	Median household income (×USD 1000)	56.97	32.17	158.19	0.00
D0E_LWH	% of low wage (less than USD 1250/month) workers (home location)	27.01	8.56	50.69	13.85
D1A_LA	Land area (×1000 acres)	2.59	8.09	81.64	0.07
D1B	Gross population density on unprotected land (total population/land area)	4.40	4.12	22.10	0.03
D1C_5_ENT	Gross entertainment employment density	0.38	0.90	6.76	0.00
D1C_5_IND	Gross industrial employment density (number of industrial jobs/land area)	0.13	0.27	2.26	0.00
D1C_5_OFF	Gross office employment density (number of office jobs/land area)	0.64	3.34	42.21	0.00
D1C_5_RET	Gross retail employment density (number of retail jobs/land area)	0.23	0.57	5.55	0.00
D2A_JP_HH	Jobs per household (total employment/number of households)	1.50	3.66	33.12	0.00
D2A_WRK_EM	Workers per job (number of workers/total employment)	11.00	38.31	443.00	0.03
D2B_E5_MX	Employment entropy (five-tier employment entropy)	0.66	0.24	0.98	0.00
D2C_TRIP_E	Trip productions and trip attractions equilibrium index (0–1)	0.37	0.30	0.99	0.00
D2C_WRK_MX	Workers per job equilibrium index (0–1)	0.31	0.32	0.98	0.00
D2D_NDX	Employment mix index (0–100)	5.34	2.21	9.00	0.40
D3A_AO	Network density of auto-oriented links per square mile (per square mile)	0.89	1.20	7.41	0.00
D3A_MM	Network density of multi-modal links per square mile (miles per square mile)	1.77	1.95	10.73	0.00
D3B_MM3	Intersection density of multi-modal intersections having three legs per square mile (number of 3-leg intersections/area)	8.13	8.88	40.09	0.00
D3B_MM4	Intersection density of multi-modal intersections having four or more legs per square mile (number of 4-leg intersections/area)	3.29	6.04	46.69	0.00
D3D_R1_P	% of primary roads	3.43	8.75	40.75	0.00
D3D_R2_P	% of secondary roads	16.02	10.94	48.03	0.00
D3D_R2_D	Density of secondary roads (miles per square mile)	2.28	2.21	13.76	0.00
D3E	Total sidewalk length (miles)	9.28	9.95	53.05	0.00
D4C_W_P	Transit ridership % of workers	2.21	1.70	7.00	0.00
D4D	Number of bus stops (stops)	18.14	18.54	117.00	0.00
D5B	Walking index (1–20)	9.85	4.09	18.00	2.00
D6A_M_P	% of motorized modes for commuting (excluding transit)	88.67	10.51	100.00	40.20
D6B_AA	Annual average daily traffic (average of 3 years) for all modes (×1000 vehicles)	129.91	104.88	592.40	3.37
Total number of census block groups (n): 174					

[1] Variable classification codes: D0: Demographic and socioeconomic, D1: Density, D2: Diversity, D3: Design, D4: Transit Accessibility, D5: Destination Accessibility, D6: Travel.

2.2. Spatial Analysis

The selected methodology is responsive to the two phenomena of spatial data: autocorrelation and non-stationarity. Autocorrelation is the dependence of observations on neighboring observations, which violates the assumption of independence of the observations. Non-stationarity is related to having different observations due to the location. The existence of these two phenomena should be detected to gain better insight into the spatial data, and the statistical analysis would better respond to deal with them. In general, spatial autocorrelation could be handled with global regression models that have spatial components, and spatial non-stationarity is commonly conceptualized with local regression models. The details of those models are given in the following section.

2.2.1. Global Regression Models

In global regression models, the relationship is assumed to be stationary over the study area. The ordinary least squares (OLS) model has the assumption of independent observations, which is usually violated in spatial cases due to the relationship or dependence of the variables. The spatial dependence would occur in two ways: (a) a variable of interest at one location is jointly impacted by the same variable at neighboring locations, and (b) the dependence is solely observed on the error terms [41]. Within the global modeling framework, these spatial dependencies could be configured in two major ways.

The spatial lag model (SLM) considers the spatial relationship or dependence of the variables. In SLM, a spatial lag parameter configures the spatial relationships through a spatial weight matrix and the dependent variable as follows:

$$y_i = a_0 + \sum_{k=1,m} a_k x_{ik} + \rho \sum_{j \in J_i} w_{ij} y_j + \varepsilon_i \quad (2)$$

where ρ is the estimated lag parameter [41–43]. W is an $n \times n$ spatial weights matrix and w_{ij} is the spatial weight between the spatial unit i and j, where J_i is the set of neighbors of the block group i. The weights are determined either by distance decay or contiguity, and the elements of each row are standardized as the row sums are equal to one [42]. Thus, the spatially weighted sum of the neighboring crash rates is treated as an explanatory variable.

The spatial error model (SEM) assumes the relationship or dependence at the error terms among different spatial units rather than the direct impacts between variables. The spatial dependence or relationship is structured through the spatial weight matrix W, and the error terms are as follows [41–43]:

$$\begin{aligned} y_i &= a_0 + \textstyle\sum_{k=1,m} a_k x_{ik} + \varepsilon_i, \\ \varepsilon_i &= \lambda \textstyle\sum_{j \in J_i} w_{ij} \varepsilon_j + u_i \end{aligned} \quad (3)$$

where λ is the estimated spatial error coefficient, ε_i is the spatially correlated error term and u_i is the error term of the ith block group.

The global models would handle autocorrelation, though they assume spatial stationarity of the association between the response and predictor variables. Thus, local models are developed to explain any non-stationarity the global models could not.

2.2.2. Local Regression Models

To overcome the drawback of global models, geographically weighted regression (GWR) was introduced [44,45]. GWR recognizes that relationships between variables across the entire study area can vary spatially. By allowing regression parameters to differ locally, GWR provides a more nuanced understanding of spatially varying relationships [45,46]. All relationships are defined at spatial units, and each spatial unit has its own coefficients for the independent variables in the model:

$$y_i = a_{0i} + \sum_{k=1,m} a_{ki} x_{ik} + \varepsilon_i \quad (4)$$

where a_{0i} is the intercept parameter, and a_{ki} is the kth built environment factor coefficient of the ith census block group centroid. A weighting scheme is used in coefficient estimation [45,46]:

$$\hat{\alpha}_i = \left(X^T W_i X\right)^{-1} X^T W_i y \tag{5}$$

where $\hat{\alpha}_i = (\alpha_{i0}, \ldots, \alpha_{im})^T$ is a vector of location-specific regression coefficients, X is the matrix of independent variables, y is a vector of independent variables, and W_i is a diagonal matrix with the weights of each observed data for census block group i. W_i is calculated by a kernel function. The Gaussian and bi-square kernel functions are the most used techniques for implementing distance weights. In Gaussian kernel functions, even the farthest neighbors have an impact on a given census block group. In contrast, in the bi-square function, the neighboring relations are assumed to be valid up to a threshold distance (the bandwidth), which is more reasonable considering the crash rates, e.g., a block group 20 miles away from another block group would not be related. Thus, in this study, a bi-square kernel function is employed as follows:

$$W_{ij} = \begin{cases} \left[1 - \left(d_{ij}/b\right)^2\right]^2, & if\ d_{ij} < b \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where W_{ij} is the weight between census block groups i and j, d_{ij} is the distance among them, and b is the bandwidth that is determined based on the distance or the number of nearest neighbors [47]. The kernels could be fixed or adaptive. The fixed kernel function sets the same bandwidth parameter for each location, which may not be responsive to heterogeneous or sparsely distributed locations. The adaptive one ensures that a nearest-neighbor scheme uses the same number of observations for each local regression model [48]. This study uses an adaptive kernel function to respond to a non-homogenous spatial configuration consisting of urban and rural census block groups. Moreover, the optimal bandwidth selection could be carried out by several methods, such as cross-validation (CV), Akaike Information Criterion (AIC), and corrected Akaike Information Criterion (AIC_c) [49]. The AIC_c method is employed in this study as it prevents more complex models that use higher degrees of freedom by penalizing smaller bandwidths [48].

As GWR works with a fixed bandwidth (i.e., distance or number of neighbors), the spatial variability in some variables could not be captured. The bandwidth represents the scale of variability in variables; as the bandwidth increases, the model converges to global models, while as the bandwidth decreases, the model has the highest locality. The scale of variability may not be the same for all processes, variables, or locations. For those cases where the relationship between the dependent and independent variables has its own scale, flexible bandwidths are proposed in the multi-scale geographically weighted regression (MGWR) model [49] as follows:

$$y_i = a_{bw_0 i} + \sum_{k=1,m} a_{bw_k i} x_{ik} + \varepsilon_i \tag{7}$$

where $a_{bw_k i}$ indicates that varying bandwidths (bw_k) are allowable for each local parameter estimate associated with a census block group i. Thus, not only do the regression coefficients vary at different locations, but they also vary at different spatial scales for each independent variable.

MGWR extends GWR by allowing for the simultaneous examination of relationships at multiple spatial scales. This approach recognizes that spatial relationships may vary across space and different scales or levels of spatial aggregation. An MGWR model can be fitted through a backfitting algorithm. Each relation set for the kth explanatory variable is formulated as an additive term f_k in the following model [49,50]:

$$y = \sum_{k=0,m} f_k + \varepsilon \quad (8)$$

Calibration is a stepwise method; at each step, the solution of Equation (8) is followed by GWR solutions for each variable k that yield an optimum bandwidth bw_k and the estimates of that iteration are carried to the following variable until all the variables are visited. The overall process continues until convergence criteria are met [49,50].

According to Oshan et al. (2019), removing the restriction on the variability in the relationships being at the same spatial scale can minimize over-fitting, reduce bias in the parameter estimates, and mitigate collinearity [48]. Thus, MGWR could be preferred over GWR to investigate spatial heterogeneity and scale.

2.2.3. GWR and MGWR with Spatially Lagged Dependent Variable (GWRL and MGWRL)

GWR and MGWR models might respond to autocorrelation problems. For the cases in which autocorrelation persists, introducing a lagged form of the dependent variable as an independent variable is a common method [51]. The spatially lagged term for the dependent variable is calculated as follows:

$$lag(Y_i) = \sum_{j=1,n,\ j\neq i} \frac{Y_j}{d_{ij}} \quad (9)$$

where $lag(Y_i)$ is the spatially lagged value of Y_i for census block group i, j is one of the other census block groups, n is the number of block groups, and d_{ik} is the distance between block group i and j [51].

2.3. Modeling Framework

We propose a spatial modeling framework to analyze the impact of the built environment on severe crash rates, incorporating considerations of spatial autocorrelation and non-stationarity. This framework is adopted from [27,32]. As an initial step, all the variables are scaled for consistency since it is advised for local regression models [48].

Starting with a non-spatial global model OLS, the framework first determines the best subset of variables by minimizing the AIC_c value. As the best OLS model is determined and solved (Figure 2, box 1), the next step is to analyze the residuals for autocorrelation using Moran's I statistic (Figure 2, box 2). If the Moran's I value for OLS residuals is insignificant (i.e., no autocorrelation), the Breusch–Pagan test is used to identify spatial non-stationarity in the OLS model residuals (Figure 2, box 3). A significant result indicates spatial non-stationarity in the relationships between the dependent and independent variables, highlighting the need for local spatial regression models such as geographically weighted regression (GWR) and multi-scale geographically weighted regression (MGWR) [27,32]. Otherwise, the OLS model is enough to explain the relationship between the dependent and independent variables (Figure 2, box 1).

For significant Moran's I statistic (Figure 2, box 4), the framework proceeds with global spatial regression models, i.e., spatial lag model (SLM) and spatial error model (SEM) [27,32,52]. To determine whether the SLM or SEM model is more suitable, we performed Lagrange Multiplier (LM) tests [53] (Figure 2, box 4).

The SLM model is appropriate if only the LM-lag test is significant (Figure 2, box 5). The SEM model is preferred if only the LM-error test is significant (Figure 2, box 6). When both tests are significant, we examine the Robust LM-lag and Robust LM-error tests (Figure 2, box 7). The SEM model is chosen if only the Robust LM-error test is significant (Figure 2, box 6), while the SLM model is selected if only the Robust LM-lag test is significant (Figure 2, box 5). If both tests are significant, the model with the lower p-value (higher significance) is selected [27,53]. Different from the previous studies, we also applied the Breusch–Pagan test (Figure 2, box 8) for the SLM and SEM settings, switching to local models for responding to non-stationarity. If the system is stationary according to

the Breusch–Pagan Test, either SLM (Figure 2, box 5) or SEM (Figure 2, box 6) is selected, whichever directs the path to the Breusch–Pagan Test.

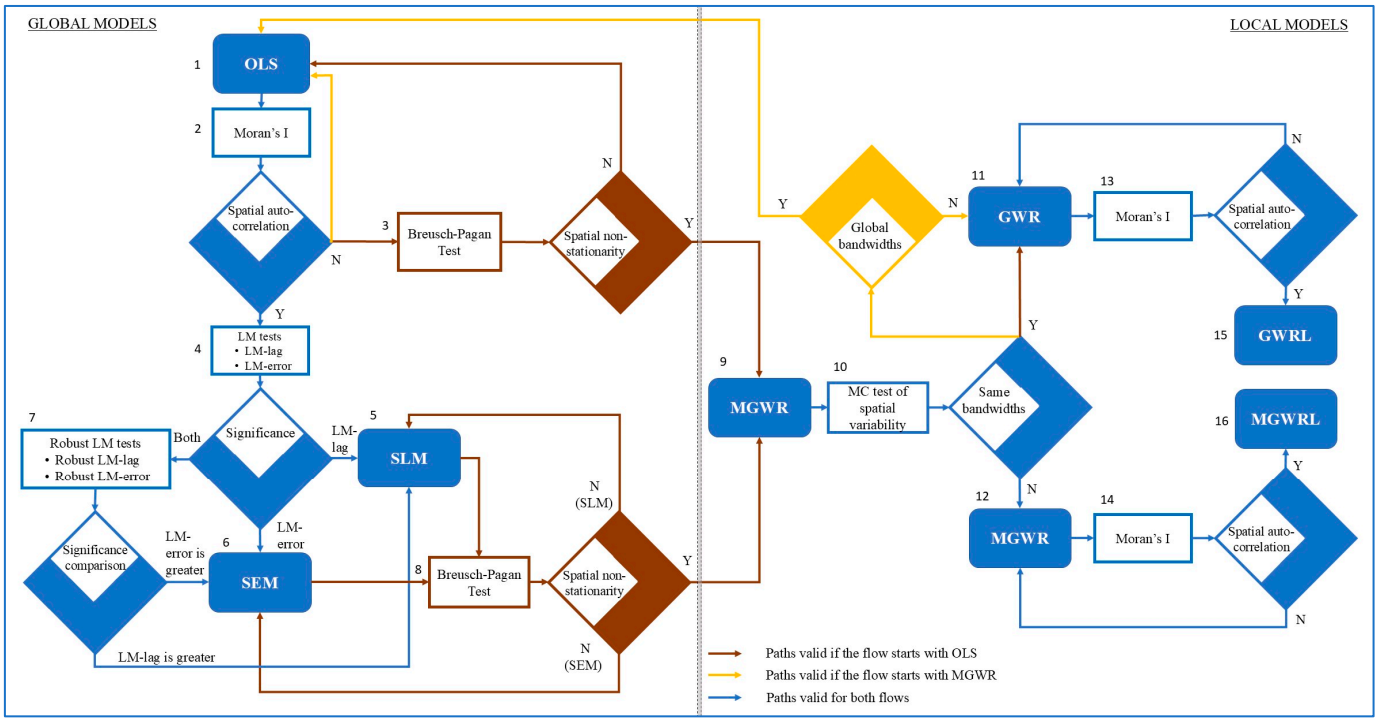

Figure 2. Spatial modeling framework.

If non-stationarity exists, the framework continues with local models. To determine the most suitable model between GWR and MGWR, we analyzed the bandwidths in the MGWR model (Figure 2, box 9) by Monte Carlo test of spatial variability (Figure 2, box 10). The GWR model is preferred if the bandwidths for all independent variables are similar (Figure 2, box 11); however, if all the bandwidths are global, the OLS model is the resulting model (Figure 2, box 1).

Conversely, if one or more variables exhibited different trends in bandwidths, the MGWR model has been deemed more appropriate (Figure 2, box 12) [27,32]. One more step is taken for the GWR and MGWR models for the consideration of autocorrelation. If Moran's I statistic is significant (Figure 2, box 13 for GWR; box 14 for MGWR), the spatially lagged GWR and MGWR models, GWRL (Figure 2, box 15) and MGWRL (Figure 2, box 16), respectively, are proposed [51]. Proposing lagged dependent variables for the local models is another extension for the previous crash studies.

This framework could be started either from global (OLS) (Figure 2, box 1) or local (MGWR) modeling (Figure 2, box 9). In this study, we started from the global side (OLS) and solved all the models to compare their performances.

We applied an adaptive kernel for bandwidth selection, using the "Golden Section" method for bandwidth optimization and AIC_c as the criterion for local spatial models. Note that, at each step and model, multicollinearity is checked by the variance inflation factor (VIF) and condition number (CN), such that VIF should be less than 10 and CN should be less than 30 [32].

The global models are solved by *spatialreg* package in R [54], and local models are solved by MGWR 2.2.1 software [48]. After developing all seven models, we evaluated their performance based on R^2, adjusted R^2, Akaike Information Criterion (AIC), corrected

Akaike Information Criterion (AIC_c) mean absolute deviation (MAD), and root mean square error (RMSE) as given in Equations (10) and (11), respectively.

$$MAD = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{10}$$

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

2.4. Subset Selection

Variable subset selection is one of the most critical steps in statistical modeling. Successful variable selection can remove redundant variables, simplify models, and improve model performance [26]. There are 42 variables in the dataset and using them all would not yield the best regression models. It is not computationally possible to check for all the variables whether to be used in models or not (as there are 2^42 alternatives). Thus, combinatorial optimization algorithms are viable alternatives to decide on the variable subsets. To choose the best set of variables, this study used Genetic Algorithm (GA) and Tabu-Search (TS) metaheuristics [40] by employing R packages *GA* [55] and *tabuSearch* [56], respectively. Both algorithms are calibrated with test sets. The objective function to be minimized for both algorithms is set to be the AIC_c value. The results of these metaheuristic algorithms are then compared, and the best of the two are presented.

3. Results

The spatial distribution of the severe (i.e., severe injury and fatal) crashes and VMT-adjusted crash rates of Leon County for 2017–2019 are given in Figure 3. The crashes are clustered around the city center, as hot spot analysis and Local Moran's I analysis depict. Also, in the northern part, there are cold spots with low–low clustering. Moran's I analysis for the crash rate returns an index value of 0.4, with a z-score of 9.27 and *p*-value of zero, which are clear indications of clustering in the dependent variable, severe crash rate.

In the second step, using the subset optimization algorithms (Genetic Algorithm and Tabu-Search yielded the same result), the best indicator variable subset is determined as an OLS model (the resulting set of variables yielded the lowest AIC_c value). The correlation matrix of the selected 12 variables by this algorithm is given in Figure 4 (details of the variables will be discussed later). As seen in Figure 4, there are no warning signals regarding multicollinearity.

As the variables and the optimum OLS model are determined, the next step is to check the OLS model and residuals in terms of spatial features, i.e., checking for autocorrelation and non-stationarity (note that all the tests applied for the global models are given in Table 2). The significant Jarque–Bera Test for OLS (see Table 2) indicates the failure of the normality assumption in OLS. The residuals of the OLS model have a significant Moran's I value (see Table 3, *p*-value) representing significant spatial autocorrelation, indicating the requirement of either a spatial global model (SLM or SEM) or local models. In Table 3, the condition number (CN) for the diagnosis of multicollinearity is given for all models. As seen from Table 3, there are no collinearity issues, as all CN values are lower than 30. Moreover, for autocorrelation detection, Moran's I values, respective z-values, and *p*-values for Moran's I values are given for all models.

For the determination of which model to select between global spatial models, SLM and SEM, Lagrange Multiplier (LM) tests are applied. According to Anselin, if one model is significant, the significant model should be selected. If they are both significant, robust LM tests are applied, and, in this round, the model with the highest significance should be selected [53]. LM tests are significant for both SLM and SEM; thus, robust tests are applied (see Table 2). In the robust tests, SLM was found to be more significant than SEM. So, the results of SLM are investigated. Rho (the lag parameter) in SLM is 0.39 and is significant with a *p*-value of 0 (see Table 4, in which the regression results of global models are given;

Table 5 has the results of the local models); moreover, Moran's I value is not significant with a *p*-value of 0.75 (Table 3); so, the model confirms the spatial correlation and eliminates it well. At that point, it should be checked whether non-stationarity exists. As seen in Table 2, the Breusch–Pagan test results for the global models are presented, and accordingly, the null hypothesis of having the same variations is rejected; thus, spatial non-stationarity should be investigated. This variability in variances leads the way to local regression models.

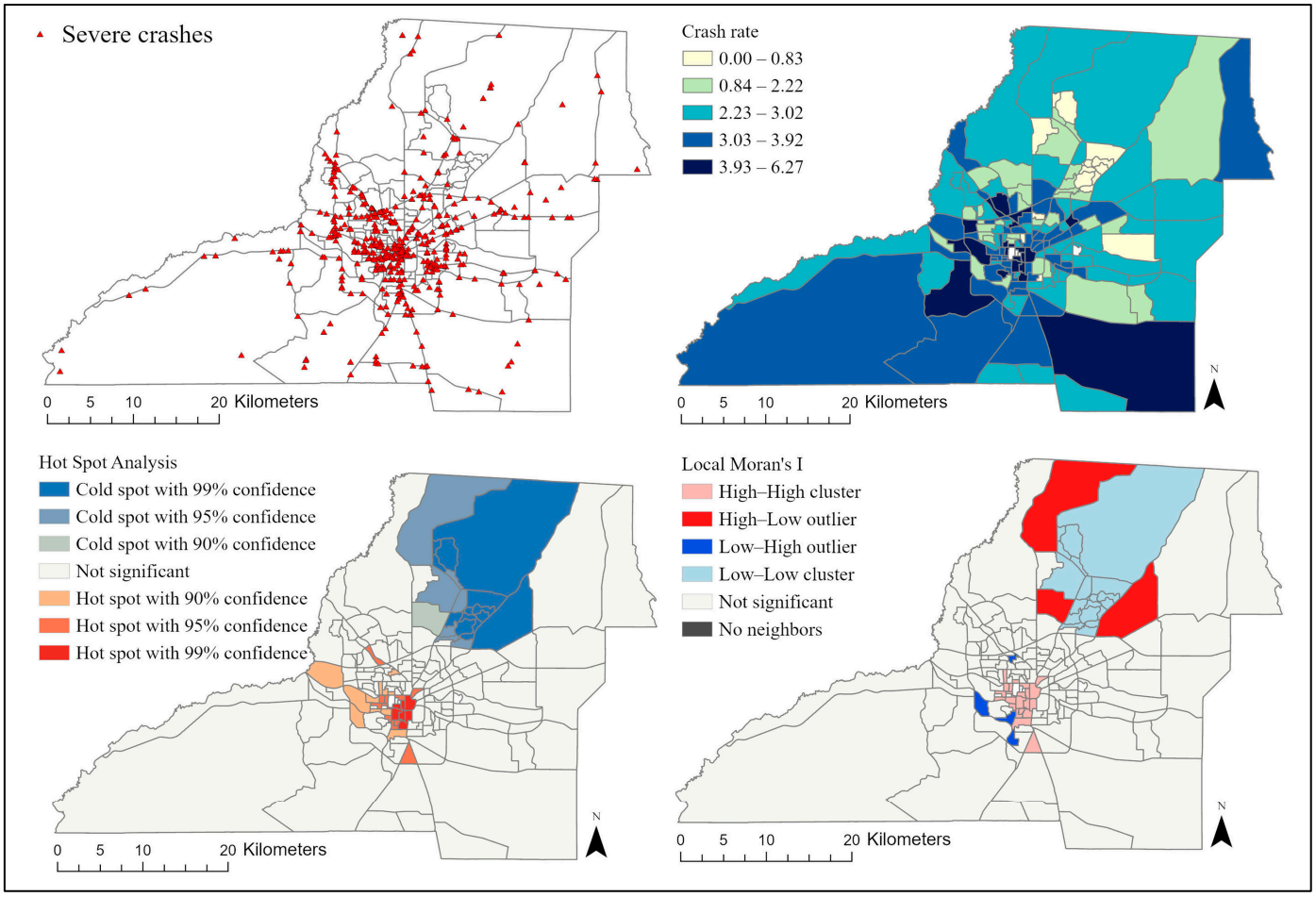

Figure 3. Spatial indicators of crashes.

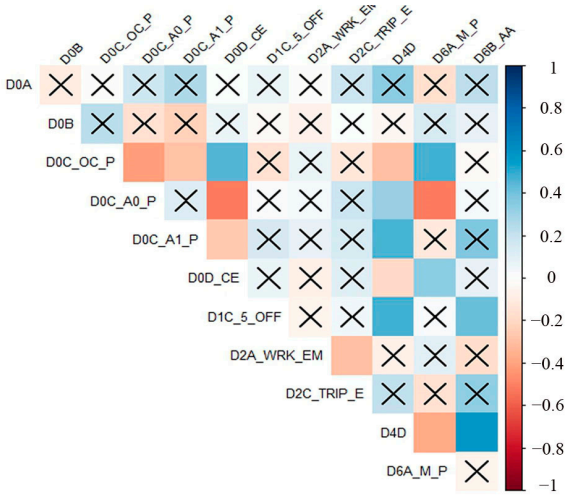

Figure 4. Correlation matrix of independent variables.

Table 2. Test results for global regression models.

	Jarque–Bera Test	
	Test-Stat	*p*-Value
OLS	14.35	0.00
	Breusch–Pagan Test	
	Test-stat	*p*-Value
OLS	31.84	0.00
SLM	25.56	0.01
SEM	27.96	0.01
	Lagrange Multiplier (LM) Test	
	Test stat	*p*-Value
SLM	26.22	0.00
SEM	15.87	0.00
SLM (robust)	10.67	0.00
SEM (robust)	0.32	0.57

Table 3. Multicollinearity and autocorrelation diagnoses.

	OLS	**SLM**	**SEM**	**GWR**	**GWRL**	**MGWR**	**MGWRL**
Condition number (CN)	3.45	3.45	3.45	4.09	4.15	4.19	3.82
Moran's I (two-tailed)	0.18	0.01	0.00	0.14	0.00	0.06	−0.02
z-score	4.19	0.32	0.13	3.31	0.24	1.39	−0.35
p-value	0.00	0.75	0.90	0.00	0.81	0.16	0.72

Table 4. Results of global models.

Global Models															
Variable	**Estimate**			**Std. Error**			**t-Value**	**z-Value**	**z-Value**	***p*-Value**					
	OLS	**SLM**	**SEM**	**OLS**	**SLM**	**SEM**	**OLS**	**SLM**	**SEM**	**OLS**		**SLM**		**SEM**	
Intercept	0.00	−0.04	−0.03	0.04	0.04	0.08	0.00	−0.97	−0.37	1.00		0.33		0.71	
D0A	−0.18	−0.17	−0.17	0.05	0.04	0.05	−3.70	−3.87	−3.46	0.00	***	0.00	***	0.00	***
D0B	−0.31	−0.29	−0.31	0.05	0.04	0.04	−6.62	−7.04	−7.27	0.00	***	0.00	***	0.00	***
D0C_OC_P	−0.15	−0.11	−0.14	0.06	0.05	0.05	−2.70	−2.10	−2.63	0.01	**	0.04	*	0.01	**
D0C_A0_P	−0.11	−0.12	−0.08	0.06	0.05	0.06	−1.86	−2.33	−1.48	0.07	.	0.02	*	0.14	
D0C_A1_P	−0.12	−0.14	−0.13	0.06	0.05	0.05	−2.07	−2.78	−2.46	0.04	*	0.01	**	0.01	**
D0D_CE	−0.44	−0.36	−0.38	0.06	0.05	0.06	−7.63	−6.68	−6.38	0.00	***	0.00	***	0.00	***
D1C_5_OFF	−0.08	−0.06	−0.09	0.05	0.05	0.05	−1.52	−1.21	−1.87	0.13		0.22		0.06	.
D2A_WRK_EM	−0.21	−0.20	−0.18	0.05	0.04	0.04	−4.47	−4.79	−4.40	0.00	***	0.00	***	0.00	***
D2C_TRIP_E	0.17	0.17	0.15	0.05	0.04	0.05	3.32	3.69	3.23	0.00	**	0.00	***	0.00	***
D4D	0.15	0.07	0.16	0.07	0.06	0.08	2.12	1.10	2.08	0.04	*	0.27		0.04	*
D6A_M_P	−0.11	−0.12	−0.13	0.06	0.05	0.05	−1.97	−2.42	−2.40	0.05	.	0.02	*	0.02	*
D6B_AA	0.43	0.41	0.40	0.06	0.05	0.06	7.11	7.63	6.68	0.00	***	0.00	***	0.00	***
Rho		0.39			0.07			5.22				0.00	***		
Lambda			0.49			0.09			5.20					0.00	***

Significance codes: . *p*-Value < 0.1; * *p*-Value < 0.05; ** *p*-Value < 0.01; *** *p*-Value < 0.001.

Table 5. Results of local models.

Local Models												
Variable	**Mean**				**Min**				**Max**			
	GWR	**GWRL**	**MGWR**	**MGWRL**	**GWR**	**GWRL**	**MGWR**	**MGWRL**	**GWR**	**GWRL**	**MGWR**	**MGWRL**
Intercept	0.01	0.00	0.01	0.02	−0.07	−0.01	−0.02	0.02	0.08	0.02	0.03	0.03
D0A	−0.17	−0.16	−0.16	−0.16	−0.23	−0.23	−0.20	−0.21	−0.12	−0.12	−0.12	−0.11
D0B	−0.33	−0.28	−0.32	−0.31	−0.40	−0.32	−0.43	−0.46	−0.24	−0.24	−0.14	−0.12
D0C_OC_P	−0.14	−0.10	−0.16	−0.09	−0.21	−0.11	−0.44	−0.10	−0.11	−0.08	0.00	−0.09
D0C_A0_P	−0.12	−0.15	−0.15	−0.15	−0.20	−0.19	−0.16	−0.18	−0.07	−0.11	−0.14	−0.13

Table 5. *Cont.*

Local Models												
Variable	**Mean**				**Min**				**Max**			
	GWR	**GWRL**	**MGWR**	**MGWRL**	**GWR**	**GWRL**	**MGWR**	**MGWRL**	**GWR**	**GWRL**	**MGWR**	**MGWRL**
D0C_A1_P	−0.14	−0.15	−0.11	−0.13	−0.19	−0.17	−0.12	−0.14	−0.07	−0.12	−0.09	−0.12
D0D_CE	−0.43	−0.35	−0.41	−0.37	−0.48	−0.39	−0.57	−0.38	−0.36	−0.33	−0.25	−0.36
D1C_5_OFF	−0.08	−0.06	−0.05	−0.05	−0.18	−0.07	−0.05	−0.05	−0.03	−0.05	−0.05	−0.05
D2A_WRK_EM	−0.22	−0.21	−0.31	−0.30	−0.24	−0.22	−0.58	−0.74	−0.18	−0.19	−0.14	−0.06
D2C_TRIP_E	0.17	0.15	0.15	0.12	0.14	0.13	0.13	0.12	0.21	0.18	0.20	0.14
D4D	0.16	0.06	0.18	0.09	0.06	0.02	0.13	0.08	0.30	0.08	0.30	0.09
D6A_M_P	−0.12	−0.15	−0.11	−0.13	−0.18	−0.16	−0.28	−0.37	−0.04	−0.13	0.23	0.18
D6B_AA	0.40	0.41	0.37	0.39	0.36	0.40	0.33	0.38	0.50	0.43	0.49	0.39
lagY		0.30		0.20		0.22		−0.10		0.36		0.42

Firstly, MGWR is solved to examine varying bandwidths, as Comber et al. [32] suggested. As seen in Table 6, the fitted model has different bandwidths (number of block groups) for several variables. For example, the percentage of motorized mode (D6A_M_P) has a bandwidth of 61 census block groups, meaning that it is not a global variable. The Monte Carlo simulation test confirms its spatial variability result with a significance value of 0.02, whereas the urban variable (D0A) is a regional variable with a bandwidth of 172 census block groups. If the bandwidths were similar for all variables and lower than the total number of block groups, GWR would be the better alternative, presenting different associations but with the same local scale for all variables. Since the bandwidths vary, MGWR is more appropriate for this spatial dataset. For the final step of this framework, the autocorrelation of the residuals is checked. As seen from Table 3, Moran's I significance value is 0.16, which is not significant with 95% confidence. So, it could be assumed that MGWR responded to the autocorrelation problem in the global models.

Table 6. Spatial variability in MGWR model.

Variable	Bandwidth	Monte Carlo Test
D0A	172	0.02
D0B	132	0.16
D0C_OC_P	87	0.13
D0C_A0_P	172	0.90
D0C_A1_P	170	0.75
D0D_CE	80	0.17
D1C_5_OFF	172	1.00
D2A_WRK_EM	75	0.21
D2C_TRIP_E	166	0.35
D4D	147	0.07
D6A_M_P	61	0.02
D6B_AA	142	0.27

One important contribution of this paper is the introduction of the lagged variable in the spatial analysis framework. If the autocorrelation problem persists, the MGWR with a lagged dependent variable should be solved. As seen in Table 3, MGWRL has a Moran's I significance value of 0.72, which is almost as insignificant as SLM and SEM. It could be observed from the same table that Moran's I value of the GWR model is as significant as the one in OLS; so, GWR could not respond to the autocorrelation problem. However, GWRL has an insignificant autocorrelation. Thus, introducing the lagged dependent variable helps reduce autocorrelation in the local regression models.

We solved all the models to compare their performances better. The results in terms of variables for the global and local models are presented in Tables 4 and 5, respectively. The global models yielded similar results regarding variable significance, and Rho in the lag model and lambda in the error model capture the dependence.

The goodness-of-fit measures are given in Table 7. When comparing performance measures, SLM outperformed the OLS, SEM, and GWR models. SLM also had a lower AICc value and less significant autocorrelation than MGWR; however, probably due to spatial non-heterogeneity, SLM had lower R^2 and higher error measures (MAD and RMSE) than MGWR. Moreover, the lagged versions of local models outperformed both the global and generic versions of the local models in all performance measures. Finally, MGWR and MGWRL outperformed GWR and GWRL.

Table 7. Goodness-of-fit measures.

Model	AIC	AICc	R^2	MAD	RMSE
OLS	308.57	311.22	0.70	0.40	0.54
SLM	287.09	290.15	0.74	0.38	0.50
SEM	293.79	296.85	0.73	0.38	0.51
GWR	302.16	309.47	0.74	0.37	0.51
GWRL	283.30	289.13	0.76	0.36	0.49
MGWR	276.41	294.60	0.81	0.33	0.44
MGWRL	262.13	278.42	0.82	0.32	0.43

4. Discussion

This exploratory study is conducted to fill the knowledge void related to the relationship between urban environments and severe crashes. A diverse suite of methods is applied in order to develop a spatial modeling framework by drawing paths between these global and local models. There are several important findings that can assist planners and policymakers for better management of traffic operations with a focus on the occurrence of severe traffic crashes in relation to urban built environments:

From the optimization algorithm, with the OLS model, the following variables constituted the best subset: urban indicator (D0A), total population (D0B), percentage of occupied housing units (D0C_OC_P), percentage households with no automobile (D0C_A0_P), percentage households with one automobile (D0C_A1_P), percentage of population with at least bachelor's degree (D0D_CE), office employment density (D1C_5_OFF), workers per job (D2A_WRK_EM), trip productions and trip attractions equilibrium index (D2C_TRIP_E), number of bus stops (D4D), percentage of motorized modes for commuting (D6A_M_P), and annual average daily traffic (D6B_AA). We solved all the spatial models for this set of variables. Note that the spatial framework ended up with the MGWR model, and the results of MGWR are very similar to MGWRL (see Table 5). For discussion, the results of the best model (according to the goodness of fit measures presented in Table 7), MGWRL, are presented.

For the local models, the variable coefficients and their significance levels would vary for each census block group. Thus, the coefficients and their significance levels are depicted in Figure 5. According to the results, the percentage of occupied housing units (D0C_OC_P), office employment density (D1C_5_OFF), and number of bus stops (D4D) were not significant for any census block group. First, Leon County and the City of Tallahassee exhibit a strong reliance on personal vehicles, with limited transit system usage, which may be the reason for D4D being insignificant. In addition, Tallahassee's downtown, while present, is not as densely populated or vibrant as those found in larger cities, which can justify the reasoning for D0C_OC_P and D1C_5_OFF. The spatial variations in the rest of the variables are depicted in Figure 5 (note that the significance level was selected as 95%).

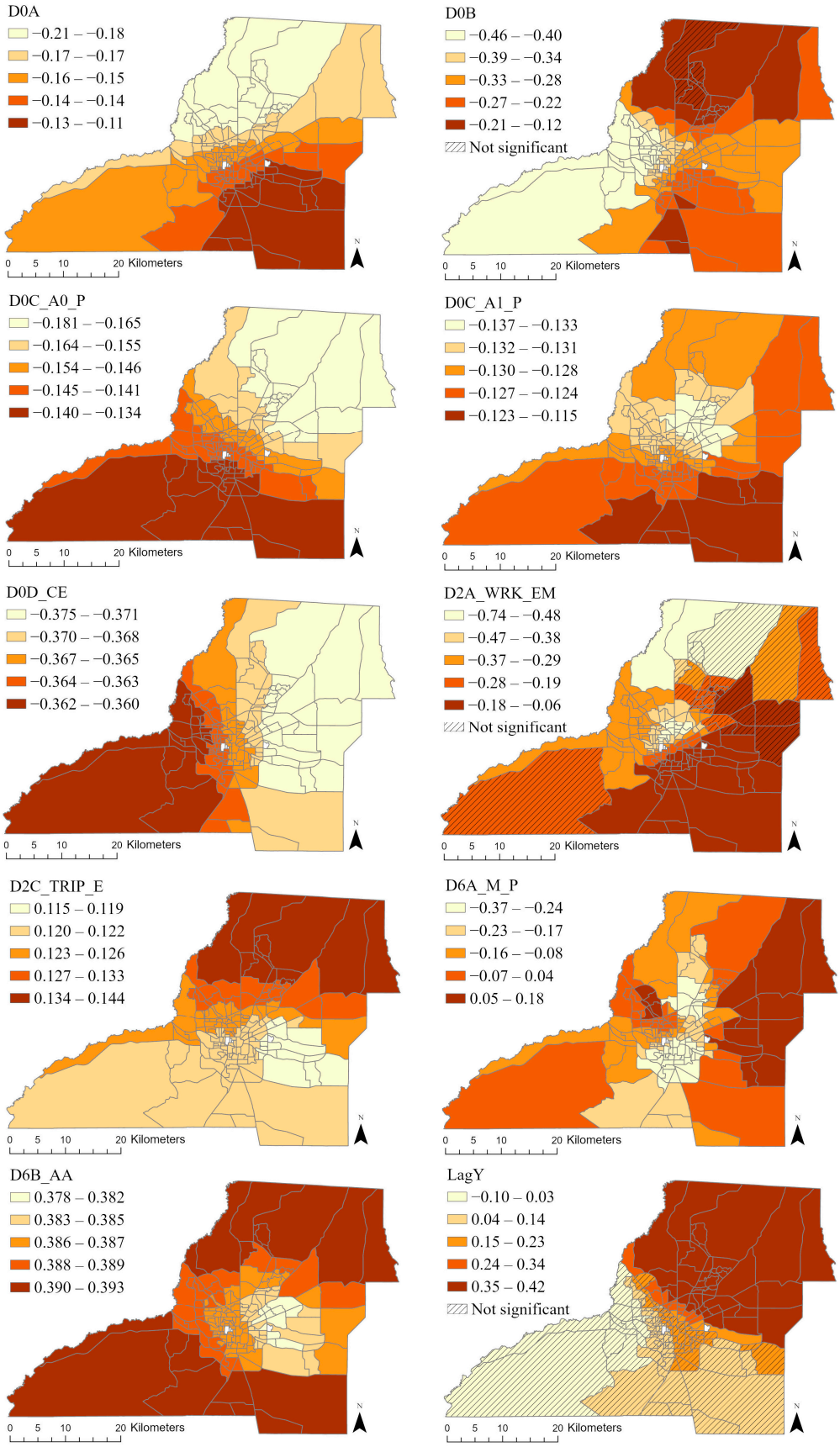

Figure 5. Spatial variation in coefficients.

Variables other than trip equilibrium index (D2C_TRIP_E), percentage of motorized modes for commuting (D6A_M_P), and annual average daily traffic (D6B_AA) have negative associations with crash rate at varying levels (this negative association is also similar in global models). Accordingly, urban areas with a higher population tend to have fewer severe crashes. Note that Leon County hosts the City of Tallahassee, a mid-sized city relatively smaller than cities like Miami or Orlando. The urban core of the city, although it possesses higher traffic than other locations of the county, has lower speeds in general and does not have high congestion levels, which may be the reason for this finding. In the literature, there are also contradictory findings about these variables. For example, distance to the nearest urban area is found to be negatively associated with crash counts for the macro-level crash counts [57], while the severity of the crashes was found to be higher around rural areas, which could be attributed to the developed and better-provided roadway network in cities and towns compared to rural areas [58,59]. On the other hand, the total population usually has positive associations with crashes [60–62] since more pedestrian and vehicle activities are expected with a higher population, and the population is generally higher in urban areas. However, the population could also negatively impact severe crashes as higher population densities could also be associated with public transit use, walkable neighborhoods, or slower driving speeds due to congestion [29]. In addition, there are crash hot spots around the city center, which could be due to the fatal crashes happening around the fringe of the city center (rather than in the core downtown area), where speeds are higher. As speed increases, the likelihood of severe crashes increases.

Moreover, a lower number of automobiles (zero or one) compared to the average car ownership rate is found to be negatively associated with the severe crash rate, which both supports [29,63] and contradicts some previous research [60,64]. The negative associations could be attributed to walkability and public transportation usage around the two universities hosted in the city, and the positive associations with crash rates could be related to underserved infrastructure and lower maintenance levels, especially on the rural county roadways of the county.

Educational level and workers per job are also found to have negative relationships with the severe crash rate. This finding conforms to the literature as education level is negatively associated with crashes [54,61,65]. In the studied area, highly educated people living in the northeast and southeast parts of the City of Tallahassee would generally seek to avoid risky activities and obey traffic rules, as also confirmed by other studies [65]. This is since the level of education is usually positively associated with income, and higher income levels also lead to car ownership with better safety performances [66]. Another variable negatively associated with the crash rate is workers per job (D2A_WRK_EM). As this variable indicates residential areas for larger values, the crash rate is expected to be lower compared to working zones, conforming to former studies [66,67]. In general, the central region of the county hosts more job opportunities compared to other portions of the county with residential areas.

Traffic volume (D6B_AA) is found to have positive associations, which is the most common exposure factor in the literature [57,68,69]. As the volume of traffic increases, the likelihood of crashes increases, as clearly observed in the core downtown area of the City of Tallahassee, in the southeast section of the city (which is developing rapidly), and around the interstate highway I-10 to the north of the city. Another positive association is with trip equilibrium index (D2C_TRIP_E), which provides insights into the balance of jobs and residential areas. The more balanced block groups are around the city center of Tallahassee. So, they would indicate more exposure and the likelihood of crashes as traffic volume and travel demand increases [70]. Transportation officials, especially those who maintain and operate the roadways close to downtown Tallahassee, should be aware of the consequences of this high crash risk. However, there are also contradictory findings stating that more balanced neighborhoods would reduce crashes [71].

Local variations other than the coefficient variations with all negative and all positive associations are observed for the percentage of motorized modes for commuting

(D6A_M_P). This variable is expected to be positively associated with crashes as automobile usage is an exposure variable that increases the likelihood of crashes. However, only the negative associations around the city center and the southern part of Leon County are significant. Transportation officials can use the proposed framework to analyze these characteristics of associations, which can help with identifying the possible reasons behind the hotspot and coldspot locations. Walking mode is very common around the city center where the highest trip attractors, the universities (Florida A&M University and Florida State University), are located. This is critical since a younger population is attributed with inexperience, greater inattentiveness, and riskier behavior while driving, and needs to be carefully studied by transportation officials. Also, there are lower-income neighborhoods with higher walking and transit usage (although limited in the city, possibly due to low vehicle ownership), which would impact that negative association, as former studies suggest [72–74], though that is not confirmed in all studies [27].

For the lag variable, the positive associations to the north of the county are significant. Those block groups have positive relations with the crash rates of their neighboring block groups. Some affluent communities in the north have a common negative association between income and crash rates. Those cold spots were also depicted previously (see Figure 3). These northern communities are especially critical since there is a substantial number of older adults living in those areas. This part of the city also has multiple intersections near I-10, and complex signalizations and design features. As redesigning roadways or intersections would be very costly, city officials should try to find smarter ways to alleviate these problems, especially in regions like northern Tallahassee that have high elderly populations.

5. Conclusions

This study's main contribution is the implementation of an extended spatial regression framework in which global and local spatial regression models are embedded to analyze the impacts of the built environment on traffic safety, where Leon County, Florida, is selected as the study area. Optimization algorithms are employed to select the best subset of variables among an extensive set of built environment variables. Then, global and local model performances are evaluated depending on the modeling outcomes and performance concerning two spatial phenomena: autocorrelation and non-stationarity.

The findings indicate that SLM outperforms OLS, SEM, and GWR models. The lagged versions of local models outperformed both the global and generic versions of the local models in all performance measures. Moreover, MGWR and MGWRL outperformed GWR and GWRL. The better performance of local models compared to global models proves the existence of spatial non-stationarity; so, the relation between the dependent and independent variables varies over space. The better performance of models with lagged dependent variables than those without dependent variables shows spatial autocorrelation, meaning that the spatial distribution of severe crashes is correlated. Finally, the better performance of multi-scale local models (MGWR and MGWRL) compared to classical local models (GWR and GWRL) is due to the varying influences of independent variables with different bandwidths.

According to the MGWRL results, some variables are similar, and some have mixed relationships with the severe crash rate over space. For example, AADT is positively related to the severe crash rate for all block groups, but having a bachelor's degree is negatively associated with the crash rate. Other than that, motor vehicle percentage for commuting has mixed relationships. For some block groups, the relationship is positive; for others, it is negative. Moreover, the significance of this relationship varies spatially.

There are some limitations to this study. The first limitation is the generalizability of the case study findings. The major contribution of geographically weighted regression models is the formulation of local relationships. Therefore, the relations have local characteristics and are not averaged under global relationships. This property bears generalizability concerns. The model should be applied to cities with similar characteristics and those with

different structures, such as ones with multiple cores, so that patterns could be observed with a higher number of samples. The second limitation is related to the case study itself, which is the conclusiveness of the results. There should be a fine-tuned investigation of the results, especially regarding the local relations, as those relations would be site-specific. The third one is that the dataset did not have any identifiers for the vehicle type; so, all severe crashes are considered irrespective of vehicle type. Moreover, only severe crashes are investigated due to concerns about crash reports being inaccurate for non-incapacitating, possible injury and no injury crashes. The final one is related to methodology and data preprocessing. Crash events are singular events in space, and for a macro-level analysis, they should be aggregated and assigned to zonal units. Several other boundary assignment methods are available; however, no standard method has been defined yet. The robustness of the results should be tested using different boundary assignment methods.

For future direction, micro-level relations could be used to further investigate and better understand those macro-level outcomes. Also, the analysis could be repeated for other locations to have better insight into local built environment characteristics that would impact severe crash occurrences. One other direction could be the application of optimization algorithms for variable subset selection to MGWR. Moreover, the single timeframe could be expanded to multiple periods to analyze not only the spatial but also the temporal variations. Finally, other modeling practices, such as Bayesian spatial models, could be employed to compare the results and performances.

Author Contributions: Onur Alisan: conceptualization, methodology, data curation, formal analysis, visualization, writing—original draft preparation, Eren Erman Ozguven, writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Science Foundation (NSF) grant CMMI-2101091, the US Department of Energy grant DE-EE0010427, and Rural, Equitable and Accessible Transportation (REAT) Center, a Tier-1 University Transportation Center (UTC) funded by the United States Department of Transportation (USDOT), through the agreement number 69A3552348321. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The National Science Foundation and the US Government assume no liability for the contents or use thereof.

Data Availability Statement: The necessary crash data were acquired from the Florida Department of Transportation (FDOT) Safety Office through the Department of Highway Safety and Motor Vehicles (DHSMV) Crash Analysis Reporting (CAR) system. Access to this source is limited to authorized users, including FDOT staff, consultants, governmental agencies, and universities, pending approval by the FDOT.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. WHO. Death on Roads. Available online: https://extranet.who.int/roadsafety/death-on-the-roads/#ticker (accessed on 10 June 2024).
2. Blincoe, L.; Miller, T.R.; Zaloshnja, E.; Lawrence, B.A. The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised). 2015. Available online: https://trid.trb.org/view/1311862 (accessed on 10 June 2024).
3. Blincoe, L.; Miller, T.R.; Wang, J.S.; Swedler, D.; Coughlin, T.; Lawrence, B.; Guo, F.; Klauer, S.; Dingus, T. The Economic and Societal Impact of Motor Vehicle Crashes. 2023. Available online: https://rosap.ntl.bts.gov (accessed on 15 June 2024).
4. NHTSA. Traffic Crashes Cost America Billions in 2019. 2023. Available online: https://www.nhtsa.gov/press-releases/traffic-crashes-cost-america-billions-2019 (accessed on 29 July 2024).
5. ITF. Road Safety Annual Report 2018. 2018. Available online: https://www.itf-oecd.org/road-safety-annual-report-2018 (accessed on 9 June 2024).
6. National Center for Statistics and Analysis. Rural/Urban Comparison of Traffic Fatalities: 2018 Data (Traffic Safety Facts. Report No. DOT HS 812 957). 2020. Available online: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812957 (accessed on 9 June 2024).
7. Ziakopoulos, A.; Yannis, G. A review of spatial approaches in road safety. *Accid. Anal. Prev.* **2020**, *135*, 105323. [CrossRef] [PubMed]
8. Ewing, R.; Dumbaugh, E. The Built Environment and Traffic Safety. *J. Plan. Lit.* **2009**, *23*, 347–367. [CrossRef]

9. Merlin, L.A.; Guerra, E.; Dumbaugh, E. Crash risk, crash exposure, and the built environment: A conceptual review. *Accid. Anal. Prev.* **2020**, *134*, 105244. [CrossRef]
10. Lord, D.; Mannering, F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transp. Res. Part. A Policy Pract.* **2010**, *44*, 291–305. [CrossRef]
11. Ouyang, Y.; Bejleri, I. Geographic Information System–Based Community-Level Method to Evaluate the Influence of Built Environment on Traffic Crashes. *Transp. Res. Rec. J. Transp. Res. Board.* **2014**, *2432*, 124–132. [CrossRef]
12. Ukkusuri, S.; Hasan, S.; Aziz, H.M.A.A. Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency. *Transp. Res. Rec. J. Transp. Res. Board.* **2011**, *2237*, 98–106. [CrossRef]
13. Huang, Y.; Wang, X.; Patton, D. Examining spatial relationships between crashes and the built environment: A geographically weighted regression approach. *J. Transp. Geogr.* **2018**, *69*, 221–233. [CrossRef]
14. Khattak, A.J.; Wang, X.; Zhang, H. Spatial Analysis and Modeling of Traffic Incidents for Proactive Incident Management and Strategic Planning. *Transp. Res. Rec. J. Transp. Res. Board.* **2010**, *2178*, 128–137. [CrossRef]
15. Ukkusuri, S.; Miranda-Moreno, L.F.; Ramadurai, G.; Isa-Tavarez, J. The role of built environment on pedestrian crash frequency. *Saf. Sci.* **2012**, *50*, 1141–1151. [CrossRef]
16. Chen, P.; Shen, Q. Identifying high-risk built environments for severe bicycling injuries. *J. Saf. Res.* **2019**, *68*, 1–7. [CrossRef] [PubMed]
17. Chen, P.; Shen, Q. Built environment effects on cyclist injury severity in automobile-involved bicycle crashes. *Accid. Anal. Prev.* **2016**, *86*, 239–246. [CrossRef] [PubMed]
18. Tasic, I.; Elvik, R.; Brewer, S. Exploring the safety in numbers effect for vulnerable road users on a macroscopic scale. *Accid. Anal. Prev.* **2017**, *109*, 36–46. [CrossRef]
19. Aguero-Valverde, J.; Jovanis, P.P. Spatial Correlation in Multilevel Crash Frequency Models. *Transp. Res. Rec. J. Transp. Res. Board.* **2010**, *2165*, 21–32. [CrossRef]
20. Abdel-Aty, M.; Lee, J.; Siddiqui, C.; Choi, K. Geographical unit based analysis in the context of transportation safety planning. *Transp. Res. Part. A Policy Pract.* **2013**, *49*, 62–75. [CrossRef]
21. Gomes, M.J.T.L.; Cunto, F.; Silva, A.R. Geographically weighted negative binomial regression applied to zonal level safety performance models. *Accid. Anal. Prev.* **2017**, *106*, 254–261. [CrossRef] [PubMed]
22. Tang, J.; Gao, F.; Liu, F.; Han, C.; Lee, J. Spatial heterogeneity analysis of macro-level crashes using geographically weighted Poisson quantile regression. *Accid. Anal. Prev.* **2020**, *148*, 105833. [CrossRef]
23. Aljoufie, M.; Tiwari, A. Modeling road safety in car-dependent cities: Case of Jeddah city, Saudi Arabia. *Sustainability* **2021**, *13*, 1816. [CrossRef]
24. Zafri, N.M.; Khan, A. Using geographically weighted logistic regression (GWLR) for pedestrian crash severity modeling: Exploring spatially varying relationships with natural and built environment factors. *IATSS Res.* **2023**, *47*, 325–334. [CrossRef]
25. Qu, X.; Zhu, X.; Xiao, X.; Wu, H.; Guo, B.; Li, D. Exploring the Influences of Point-of-Interest on Traffic Crashes during Weekdays and Weekends via Multi-Scale Geographically Weighted Regression. *ISPRS Int. J. Geoinf.* **2021**, *10*, 791. [CrossRef]
26. Li, X.; Yu, S.; Huang, X.; Dadashova, B.; Cui, W.; Zhang, Z. Do underserved and socially vulnerable communities observe more crashes? A spatial examination of social vulnerability and crash risks in Texas. *Accid. Anal. Prev.* **2022**, *173*, 106721. [CrossRef] [PubMed]
27. Zafri, N.M.; Khan, A. A spatial regression modeling framework for examining relationships between the built environment and pedestrian crash occurrences at macroscopic level: A study in a developing country context. *Geogr. Sustain.* **2022**, *3*, 312–324. [CrossRef]
28. Tang, X.; Bi, R.; Wang, Z. Spatial analysis of moving-vehicle crashes and fixed-object crashes based on multi-scale geographically weighted regression. *Accid. Anal. Prev.* **2023**, *189*, 107123. [CrossRef] [PubMed]
29. Liu, J.; Das, S.; Khan, M.N. Decoding the impacts of contributory factors and addressing social disparities in crash frequency analysis. *Accid. Anal. Prev.* **2024**, *194*, 107375. [CrossRef] [PubMed]
30. Kuo, P.F.; Sulistyah, U.D.; Putra, I.G.B.; Lord, D. Exploring the spatial relationship of e-bike and motorcycle crashes: Implications for risk reduction. *J. Saf. Res.* **2024**, *88*, 199–216. [CrossRef] [PubMed]
31. Lin, P.; Hong, Y.; He, Y.; Pei, M. Advancing and lagging effects of weather conditions on intercity traffic volume: A geographically weighted regression analysis in the Guangdong-Hong Kong-Macao Greater Bay Area. *Int. J. Transp. Sci. Technol.* **2024**, *13*, 58–76. [CrossRef]
32. Comber, A.; Brunsdon, C.; Charlton, M.; Dong, G.; Harris, R.; Lu, B.; Lü, Y.; Murakami, D.; Nakaya, T.; Wang, Y.; et al. A Route Map for Successful Applications of Geographically Weighted Regression. *Geogr. Anal.* **2023**, *55*, 155–178. [CrossRef]
33. Center for Neighborhood Technology. Housing + Transportation Affordability Index. Available online: https://www.cnt.org/tools/housing-and-transportation-affordability-index (accessed on 10 November 2020).
34. Smart Growth EPA. Smart Location Mapping. Available online: https://www.epa.gov/smartgrowth/smart-location-mapping#SLD (accessed on 10 November 2020).
35. Census Bureau Data. Available online: https://data.census.gov (accessed on 9 September 2024).
36. FDOT Safety Office, Department of Highway Safety and Motor Vehicles (DHSMV). Crash Analysis Reporting (CAR) System – Reports and Database. Available online: https://www.fdot.gov/safety/safetyengineering/crash-data-systems-and-mapping (accessed on 16 July 2023).

37. FHWA. KABCO Injury Classification Scale and Definitions by State. Available online: https://safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf (accessed on 25 November 2024).
38. NHTSA. Model Minimum Uniform Crash Criteria. Available online: https://www.nhtsa.gov/traffic-records/model-minimum-uniform-crash-criteria (accessed on 25 November 2024).
39. ESRI. *ArcGIS Pro, Release 3.1*; Environmental Systems Research Institute: Redlands, CA, USA, 2023.
40. Alisan, O.; Tuydes-Yaman, H.; Ozguven, E.E. Tabu-Search-Based Combinatorial Subset Selection Approach to Support Investigation of Built Environment and Traffic Safety Relationship. *Transp. Res. Rec. J. Transp. Res. Board.* **2022**, *2677*, 588–609. [CrossRef]
41. Anselin, L.; Rey, S. Properties of Tests for Spatial Dependence in Linear Regression Models. *Geogr. Anal.* **1991**, *23*, 112–131. [CrossRef]
42. Won Kim, C.; Phipps, T.T.; Anselin, L. Measuring the benefits of air quality improvement: A spatial hedonic approach. *J. Environ. Econ. Manag.* **2003**, *45*, 24–39. [CrossRef]
43. Ward, M.; Gleditsch, K. *Spatial Regression Models*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2008.
44. Brunsdon, C.; Fotheringham, A.S.; Charlton, M.E. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geogr. Anal.* **1996**, *28*, 281–298. [CrossRef]
45. Lu, B.; Charlton, M.; Harris, P.; Fotheringham, A.S. Geographically weighted regression with a non-Euclidean distance metric: A case study using hedonic house price data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 660–681. [CrossRef]
46. Fotheringham, A.S.; Charlton, M.E.; Brunsdon, C. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environ. Plan. A* **1998**, *30*, 1905–1927. [CrossRef]
47. Mollalo, A.; Vahedi, B.; Rivera, K.M. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Sci. Total Environ.* **2020**, *728*, 138884. [CrossRef]
48. Oshan, T.M.; Li, Z.; Kang, W.; Wolf, L.J.; Fotheringham, A.S. MGWR: A *Python* implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS Int. J. Geoinf.* **2019**, *8*, 269. [CrossRef]
49. Yang, W. An Extension of Geographically Weighted Regression with Flexible Bandwidths. Ph.D. Thesis, University of St. Andrews, Fife, Scotland, UK, 2014. Available online: http://hdl.handle.net/10023/7052 (accessed on 15 September 2024).
50. Fotheringham, A.S.; Yang, W.; Kang, W. Multiscale Geographically Weighted Regression (MGWR). *Ann. Am. Assoc. Geogr.* **2017**, *107*, 1247–1265. [CrossRef]
51. Fotheringham, A.S.; Yue, H.; Li, Z. Examining the influences of air quality in China's cities using multi-scale geographically weighted regression. *Trans. GIS.* **2019**, *23*, 1444–1464. [CrossRef]
52. Sisman, S.; Aydinoglu, A.C. A modelling approach with geographically weighted regression methods for determining geographic variation and influencing factors in housing price: A case in Istanbul. *Land Use Policy.* **2022**, *119*, 106183. [CrossRef]
53. Anselin, L. Exploring Spatial Data with GeoDa: A Workbook. 2005. Available online: https://www.geos.ed.ac.uk/~gisteac/fspat/geodaworkbook.pdf (accessed on 10 September 2024).
54. Bivand, R.; Millo, L.; Piras, G. A Review of Software for Spatial Econometrics in R. *Mathematics* **2021**, *9*, 1276. [CrossRef]
55. Scrucca, L. GA: A Package for Genetic Algorithms in R. *J. Stat. Softw.* **2013**, *53*, 1–37. Available online: http://www.jstatsoft.org/v53/i04/ (accessed on 11 July 2024). [CrossRef]
56. Domijan, K. tabuSearch: Tabu-search Algorithm for Binary Configurations. R. 2018. Available online: https://cran.r-project.org/web/packages/tabuSearch/index.html (accessed on 11 November 2020).
57. Cai, Q.; Abdel-Aty, M.; Lee, J.; Huang, H. Integrating macro- and micro-level safety analyses: A Bayesian approach incorporating spatial interaction. *Transp. A Transp. Sci.* **2018**, *15*, 285–306. [CrossRef]
58. Wang, J.; Ji, L.; Ma, S.; Sun, X.; Wang, M. Analysis of Factors Influencing the Severity of Vehicle-to-Vehicle Accidents Considering the Built Environment: An Interpretable Machine Learning Model. *Sustainability* **2023**, *15*, 12904. [CrossRef]
59. Wang, J.; Ma, S.; Jiao, P.; Ji, L.; Sun, X.; Lu, H. Analyzing the Risk Factors of Traffic Accident Severity Using a Combination of Random Forest and Association Rules. *Appl. Sci.* **2023**, *13*, 8559. [CrossRef]
60. Lee, J.; Yasmin, S.; Eluru, N.; Abdel-Aty, M.; Cai, Q. Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. *Accid. Anal. Prev.* **2018**, *111*, 12–22. [CrossRef]
61. LaScala, E.A.; Gerber, D.; Gruenewald, P.J. Demographic and environmental correlates of pedestrian injury collisions: A spatial analysis. *Accid. Anal. Prev.* **2000**, *32*, 651–658. [CrossRef]
62. Yu, C.Y.Y. Built environmental designs in promoting pedestrian safety. *Sustainability* **2015**, *7*, 9444–9460. [CrossRef]
63. Feyzollahi, M.; Pineau, P.O.; Rafizadeh, N. Drivers of Driving: A Review. *Sustainability* **2024**, *16*, 2479. [CrossRef]
64. Nam, Y.; Hawkins, J.; Butler, D.; Aldridge, N.; Elayan, M.; Yoo, K.I. *Modeling Pedestrian and Bicyclist Crash Exposure with Location-Based Service Data*; SPR-FY23(025); Nebraska Department of Transportation: Lincoln, NE, USA, 2024.
65. Wang, T.; Yu, J.; Chen, Y.; Ma, C.; Ye, X.; Chen, J. Factors Associated with the Severity of Motor Vehicle Crashes Involving Electric Motorcycles and Electric Bicycles: A Random Parameters Logit Approach with Heterogeneity in Means. *Transp. Res. Rec. J. Transp. Res. Board.* **2023**, *2677*, 691–704. [CrossRef]
66. Xu, C.; Li, H.; Zhao, J.; Chen, J.; Wang, W. Investigating the relationship between jobs-housing balance and traffic safety. *Accid. Anal. Prev.* **2017**, *107*, 126–136. [CrossRef] [PubMed]
67. Khondakar, B.; Sayed, T.; Lovegrove, G. Transferability of Community-Based Collision Prediction Models for Use in Road Safety Planning Applications. *J. Transp. Eng.* **2010**, *136*, 871–880. [CrossRef]

68. Guerra, E.; Dong, X.; Kondo, M. Do Denser Neighborhoods Have Safer Streets? Population Density and Traffic Safety in the Philadelphia Region. *J. Plan. Educ. Res.* **2019**, *42*, 654–667. [CrossRef]
69. Yu, C.Y.; Xu, M. Local Variations in the Impacts of Built Environments on Traffic Safety. *J. Plan. Educ. Res.* **2018**, *38*, 314–328. [CrossRef]
70. Ding, H.; Sze, N.N.; Li, H.; Guo, Y. Roles of infrastructure and land use in bicycle crash exposure and frequency: A case study using Greater London bike sharing data. *Accid. Anal. Prev.* **2020**, *144*, 105652. [CrossRef] [PubMed]
71. Xiao, D.; Ding, H.; Sze, N.N.; Zheng, N. Investigating built environment and traffic flow impact on crash frequency in urban road networks. *Accid. Anal. Prev.* **2024**, *201*, 107561. [CrossRef] [PubMed]
72. Cai, Q.; Abdel-Aty, M.; Lee, J.; Eluru, N. Comparative analysis of zonal systems for macro-level crash modeling. *J. Saf. Res.* **2017**, *61*, 157–166. [CrossRef]
73. Lee, J.; Abdel-Aty, M.; Cai, Q. Intersection crash prediction modeling with macro-level data from various geographic units. *Accid. Anal. Prev.* **2017**, *102*, 213–226. [CrossRef]
74. Chen, P.; Zhou, J. Effects of the built environment on automobile-involved pedestrian crash frequency and risk. *J. Transp. Health* **2016**, *3*, 448–456. [CrossRef]