International Journal of Geo-Information

MDPI

Article

Geospatial Modeling of Autonomous Vehicle Collision Reports: A Traffic Analysis Zone-Based Study of Spatial Heterogeneity in San Francisco

Yao Wu, Haochen Feng, Yichang Shao * and Junzhuo Hu

School of Modern Posts, School of Intelligent Transportation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; wuyao@njupt.edu.cn (Y.W.); 1224097401@njupt.edu.cn (H.F.); hjz001015@foxmail.com (J.H.)

* Correspondence: yshao@njupt.edu.cn

Abstract

Autonomous vehicle collision reports provide valuable geospatial information for understanding how emerging mobility systems interact with urban built environments. However, existing studies have mostly examined autonomous vehicle crashes from a global statistical or traffic safety perspective, with less attention paid to the construction of spatial analysis units, multi-source geospatial data integration, and local spatial non-stationarity. This study develops a Traffic Analysis Zone (TAZ)-based geospatial framework to analyze the spatial distribution of autonomous vehicle collision reports in San Francisco. A total of 427 collision reports from 2015 to 2024 were geocoded and integrated with road infrastructure, land-use, and geographic information data within a unified Geographic Information System (GIS) environment. Several global and spatial regression models, including negative binomial regression, generalized linear models, spatial autoregressive modeling, and geographically weighted regression, were compared to examine the association between built-environment variables and collision report counts. The results show that autonomous vehicle collision reports are spatially clustered and that several road infrastructure variables, including bicycle lanes, traffic signs, and intersections, are associated with their spatial distribution. The geographically weighted regression results further indicate that these associations vary across San Francisco, suggesting clear local spatial non-stationarity. This study contributes a GIS-based workflow for integrating autonomous vehicle collision reports with urban spatial data and provides exploratory evidence for spatially targeted safety diagnosis in smart city and urban mobility contexts.

Keywords: geospatial modeling; spatial heterogeneity; traffic safety; geographically weighted regression; traffic analysis zone

Check for updates

Academic Editors: Wolfgang Kainz, Shivanand Balram, Eric Vaz and Suzana Dragicevic

Received: 23 June 2026
Revised: 7 October 2026
Accepted: 10 October 2026
Published: 13 October 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Autonomous vehicle collision reports are not only traffic safety records but also geospatial event data that reflect the interaction between automated driving systems and urban built environments. Each report contains location information that can be linked with road networks, traffic control facilities, land-use patterns, and administrative or planning units. From the perspective of geographic information science, these reports provide an opportunity to examine how collision events are distributed across urban space and whether their associations with built-environment characteristics vary across locations. Therefore, a Geographic Information System (GIS)-based analytical framework is needed

to transform scattered collision records into spatially structured data and to support spatial statistical modeling at an appropriate geographic scale. This is particularly important when collision records are aggregated into planning units such as the Traffic Analysis Zone (TAZ). Borrowed from transportation research, a TAZ serves as a foundational spatial unit widely utilized in travel demand forecasting and macro-level traffic safety analysis [1]. By adopting the TAZ as the standard spatial division framework, road infrastructure, land use, and traffic control variables can be systematically integrated through spatial overlay and spatial joining.

Recent studies have shown that autonomous vehicles may have different safety patterns from human-driven vehicles under specific road and environmental conditions. For example, Abdel-Aty [2] compared crash data from regions including California and found that vehicles equipped with advanced automated driving systems showed lower crash probability in many scenarios but higher crash rates in low-light conditions and turning maneuvers. These findings suggest that autonomous vehicle safety should be examined not only from a vehicle-technology perspective but also from a spatial and environmental perspective [3–7]. Because collision reports are unevenly distributed across urban areas, GIS-based spatial analysis can help identify where collision reports are concentrated and how their spatial patterns are associated with road infrastructure and land-use characteristics.

Traffic accident research has increasingly shifted from global statistical models [8–10] to spatially explicit frameworks. To capture local non-stationarity and spatial autocorrelation, scholars have widely deployed GIS and spatial econometric models, such as Geographically Weighted Regression (GWR) and Bayesian autoregressive models, to evaluate built-environment impacts on conventional vehicle, pedestrian, and taxi crashes [11–13]. Conversely, current autonomous vehicle (AV) safety research predominantly targets crash severity, scenario classification, and micro-level human-vehicle interactions. These studies heavily rely on machine learning algorithms [14,15] and onboard sensor data to map causal chains in specific maneuvers, such as lane-changing or rear-end conflicts [16,17]. Consequently, a critical geospatial gap remains: while spatial analysis units and regression models are mature in traditional crash studies, existing AV research largely overlooks the geospatial dimension. Specifically, there is limited attention given to constructing standardized spatial analysis units, integrating multi-source GIS data, and exploring the local spatial non-stationarity of AV collision distributions.

To systematically analyze these spatial associations, a robust analytical framework should account for both global baseline patterns and local spatial variations. Traditionally, global regression models, such as Negative Binomial Regression (NBR), Generalized Linear Models (GLM), and Bayesian generalized linear model (Bayesian GLM) specifications, have served as foundational tools for identifying variables statistically associated with collision report counts and resolving statistical issues like data overdispersion [18–22]. Furthermore, to manage unobserved heterogeneity and non-linear associations in both conventional and autonomous vehicle safety evaluations, researchers have increasingly relied on advanced global frameworks, including random-parameters Logit models and hybrid machine learning algorithms [23–27].

Despite their statistical rigor, these global models inherently assume spatial stationarity, estimating a single, uniform average effect across the entire study area. This assumption frequently falls short in urban contexts, where the impacts of transportation infrastructure, land use, and demographic factors exhibit significant geographic variation. To overcome this limitation, spatial econometric tools such as Geographically Weighted Regression (GWR) and Spatial Autoregressive (SAR) models have been introduced to capture spatial dependency and generate localized coefficients [28–30]. Consequently, an integrated approach is indispensable for comprehensive AV safety analysis. Global models are first

required to rigorously screen critical built-environment variables and validate overarching statistical associations, while spatial models are subsequently needed to map local non-stationarity and visualize how these associations vary across diverse urban landscapes.

In summary, existing crash studies have provided useful evidence on the associations between crash outcomes and road, traffic, and environmental variables. However, two gaps remain for autonomous vehicle collision report analysis from a geoinformation perspective. First, many studies rely on global models or non-spatial machine learning methods, which may overlook local spatial non-stationarity in the associations between built-environment variables and collision report counts. Second, there is still limited work on constructing a TAZ-based geospatial database that integrates autonomous vehicle collision reports, road infrastructure, land-use variables, and spatial analysis units in a unified GIS framework. Therefore, a GIS-based spatial modeling framework is needed to examine the spatial clustering of autonomous vehicle collision reports and to map how their built-environment associations vary across urban space.

This study makes three contributions. First, it constructs a TAZ-based geospatial database by integrating autonomous vehicle collision reports, road infrastructure variables, land-use information, and planning-unit boundaries within a unified GIS framework. Second, it combines global count models and spatial models to examine the statistical associations between built-environment variables and autonomous vehicle collision report counts. Third, it applies geographically weighted regression to explore local spatial non-stationarity and to visualize how these associations vary across San Francisco. These contributions link autonomous vehicle safety analysis with geoinformation science and provide exploratory spatial evidence for smart mobility planning and targeted field diagnosis.

2. Materials and Methods

2.1. Data Collection and Processing

The autonomous driving accident data employed in this study are derived from the autonomous vehicle collision reports published by the California Department of Motor Vehicles (CADMV), USA, spanning from 2015 to 2024. To ensure that the accident samples authentically reflect the characteristics and potential risks of autonomous driving systems, this study established strict criteria for sample screening. First, a total of 478 original collision reports published by the CADMV were systematically collected. Second, regarding the spatial scope, the research samples were strictly confined within the administrative boundaries of San Francisco City. In terms of vehicle and driving mode screening, vehicle samples equipped solely with Level 2 advanced driver-assistance systems were explicitly excluded. Furthermore, records in which manual takeover occurred at the time of the accident were filtered out, retaining only accident data where the vehicle was operating in autonomous driving mode both prior to and during the collision. Ultimately, through systematic data cleaning—eliminating invalid records such as coordinate drifts, vague location descriptions, and missing fields—a total of 427 valid accident samples were obtained. These samples were subsequently used to construct the core database for TAZ-level spatial association analysis of autonomous vehicle collision reports. To ensure strict temporal consistency and avoid temporal mismatch, the spatial explanatory variables—including road network characteristics and land-use zoning—were collected and processed on a year-by-year basis corresponding to the 2015–2024 study period. Utilizing historical data archives from OpenStreetMap and the San Francisco Open Data Portal, annual versions of the spatial datasets were extracted. Each AV collision report was spatially joined with the infrastructure and land-use data specific to the year the collision occurred. This dynamic temporal matching process ensures that historical modifications to the built environment are accurately captured and temporally aligned with the collision events.

To ensure the rigorous reproducibility of the GIS data-processing workflow, the exact datasets, coordinate systems, and spatial operations are explicitly documented. The road infrastructure variables were extracted from the vector archives of OpenStreetMap and the San Francisco Municipal Transportation Agency, while land-use zoning and TAZ boundaries were acquired from the San Francisco Open Data Portal. All spatial datasets were initially collected on an annual basis (2015–2024). The spatial explanatory variables were extracted from a comprehensive cross-sectional dataset representing the most recent and complete infrastructure state, and were subsequently joined with the cumulative AV collision reports at the TAZ level.

Spatially, all original data layers were initially aligned using the WGS 84 Geographic Coordinate System. To guarantee the mathematical accuracy of distance and area calculations, all spatial layers were subsequently projected to WGS 1984 UTM Zone 10N (EPSG: 32610) for distance- and area-based calculations.

The core GIS processing procedure in ArcGIS 10.8 involved three systematic steps. (1) Collision locations were translated into spatial point features using coordinate matching and manually verified against municipal basemaps. (2) Road network polylines and land-use polygons were intersected with the TAZ boundary layer to isolate the infrastructure geometries strictly within each zone. (3) The geocoded collision points and the intersected infrastructure features were spatially joined to the TAZ polygons, enabling the precise calculation of the absolute counts and spatially normalized densities for each variable.

As a core spatial unit system in transportation planning, the TAZ plays a crucial role in geographic information acquisition, serving as the standard spatial division framework in the transportation planning domain. This study first obtained the TAZ spatial layer data of San Francisco provided by the California Department of Transportation (Figure 1), which establishes a standardized spatial analysis framework for this research. Within this framework, a multi-dimensional spatial statistical analysis was performed on the autonomous driving accident data. This analysis not only quantified fundamental indicators such as accident frequency but also explored the spatial distribution characteristics and spatial heterogeneity of autonomous driving accidents in depth, based on critical factors including spatial continuity, homogeneity, and compactness [31–34].

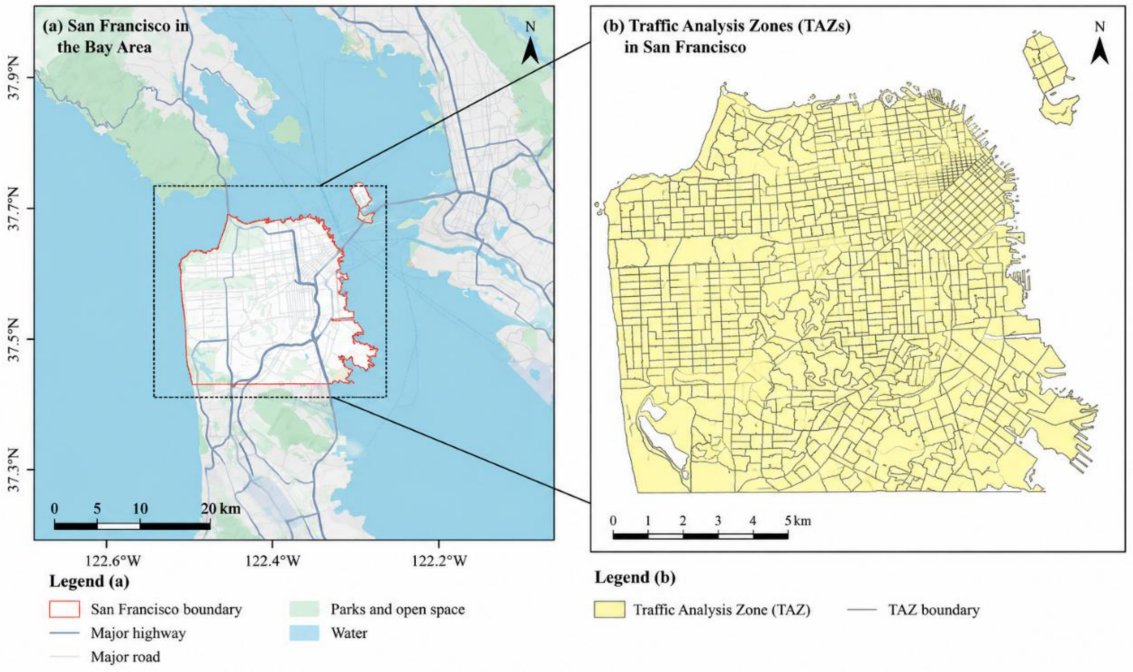


Figure 1. Division of Traffic Analysis Zones in San Francisco City.

Regarding the acquisition of geographic environmental data, this study comprehensively utilized multi-source remote sensing data platforms, including Google Earth, to systematically extract key spatial variables such as land use characteristics, road network layouts, and building distributions within the study area. To ensure the reproducibility of the geospatial workflow, the documentation and correlation mechanisms for key road parameters must be explicitly defined. Specifically, intersection zones were documented as spatial nodes generated by the topological crossing of road centerlines, while speed-limited segments were recorded as continuous polyline vectors carrying specific speed-limit attributes. Because all spatial features were standardized to a unified geographic coordinate system, these infrastructural elements can be seamlessly correlated with any standard street network. To ensure the accuracy and consistency of the spatial data, preprocessing procedures, including image rectification and projection transformation, were executed [35]. Ultimately, the precise correlation between the accident data and multi-dimensional environmental elements was achieved through geospatial overlay analysis and spatial intersection operations. This established a robust, software-agnostic analytical foundation for the subsequent exploration of the associations between autonomous vehicle collision report counts and built-environment variables within a unified geographic framework.

Regarding the extraction of characteristic variables of the road and traffic environment, key road parameters, including the spatial distribution characteristics of intersections and the lengths of speed-limited segments, were systematically collected primarily based on the official road characteristics database of the California Department of Transportation and the Geographic Information System platform of San Francisco. It is important to conceptually distinguish between the 'number' and 'density' variables utilized in this study. 'Number' represents the absolute scale or total prevalence of a specific infrastructure element within a TAZ, capturing the overall capacity or potential exposure. Conversely, 'density' represents the spatial concentration of that infrastructure, calculated by dividing the absolute count by the TAZ area. Both metrics are retained because absolute scale and spatial intensity can independently and differently modulate traffic complexity in heterogeneous urban zones. To achieve precise localization and analysis, the zone codes corresponding to the Traffic Analysis Zone within San Francisco's jurisdiction were further filtered, and the variable information within these specific areas was extracted utilizing automated techniques. Table 1 presents the descriptive analysis of these variables. By utilizing the precise geographic coordinates of the accident locations, the road characteristic data were accurately matched with the Traffic Analysis Zone data layers. The visualization results for several key variables are illustrated in Figures 2 and 3.

The modeling section comprises two distinct components. The first part designates the cumulative number of autonomous vehicle collision reports within each Traffic Analysis Zone during the 2015–2024 study period as the dependent variable. The second part utilizes the annual number of autonomous vehicle collision reports within each TAZ during the same 2015–2024 timeframe as the dependent variable. For both models, the independent variables consistently consist of road infrastructure information (e.g., traffic signals and stop signs) alongside land use information (e.g., activity areas and commercial areas).

Table 1. Descriptive analysis of variables.

Category	Variable	N	Mean	Std. Dev.	Min	Max
Transportation infrastructure environment	Bus stops number	981	5.235	5.602	0	44.000
	Bus stop density	981	0.632	1.189	0	19.919
	Bicycle parking station number	981	5.608	7.688	0	51.000
	Bicycle parking station density	981	0.758	1.084	0	7.093
	Bicycle lane number	981	9.373	8.873	0	66.000
	Bicycle lane density	981	1.019	1.261	0	12.887
	Slow street number	981	0.352	1.181	0	10.000
	Slow street density	981	0.032	0.115	0	1.217
	Painted safety zone number	981	0.251	0.634	0	5.000
	Painted safety zone density	981	0.037	0.134	0	1.745
	Motorcycle parking lot number	981	2.317	9.396	0	127.000
	Motorcycle parking lot density	981	0.585	3.428	0	55.426
	Stop sign number	981	10.881	10.233	0	54.000
	Stop sign density	981	0.904	0.780	0	4.226
	Pedestrian street number	981	26.375	14.565	0	101.000
	Pedestrian street density	981	2.698	1.657	0	15.703
	Speed-restricted segment number	981	27.180	15.567	0	109.000
	Speed-restricted segment density	981	2.758	1.653	0	15.703
	Traffic sign number	981	1.521	1.705	0	11.000
	Traffic sign density	981	0.207	0.306	0	3.591
	Intersection number	981	88.475	109.922	0	956.000
	Intersection density	981	10.020	12.821	0	159.671
	Infrastructure facility number	981	25.812	18.500	0	151.000
	Infrastructure facility density	981	3.122	4.335	0	65.900
Land use	TAZ area	981	1.336	1.497	0	25.116
	Commercial land area	981	0.032	0.336	0	4.090
	Commercial land density	981	0.003	0.026	0	0.443
	Activity land area	981	0.175	0.429	0	3.207
	Activity land density	981	0.020	0.059	0	0.698

Note: 'N' represents the total number of TAZ utilized as the spatial units in this study. The 'Mean' and 'Std. Dev.' indicate the average value and standard deviation of the respective attribute per TAZ. For variables categorized as 'density', the values are spatially normalized, calculated by dividing the absolute count or area of the infrastructural feature by the geographic area of the corresponding TAZ.



Figure 2. Distribution of Bicycle Lane Network in San Francisco City.

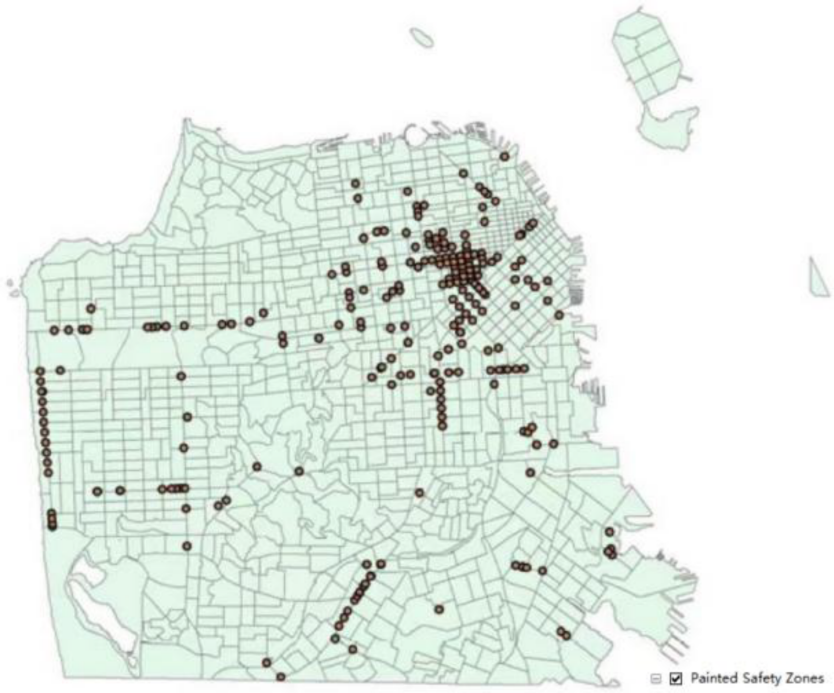


Figure 3. Distribution of Painted Safety Zones in San Francisco City.

2.2. Methodology

The flowchart of the methodology is illustrated in Figure 4.

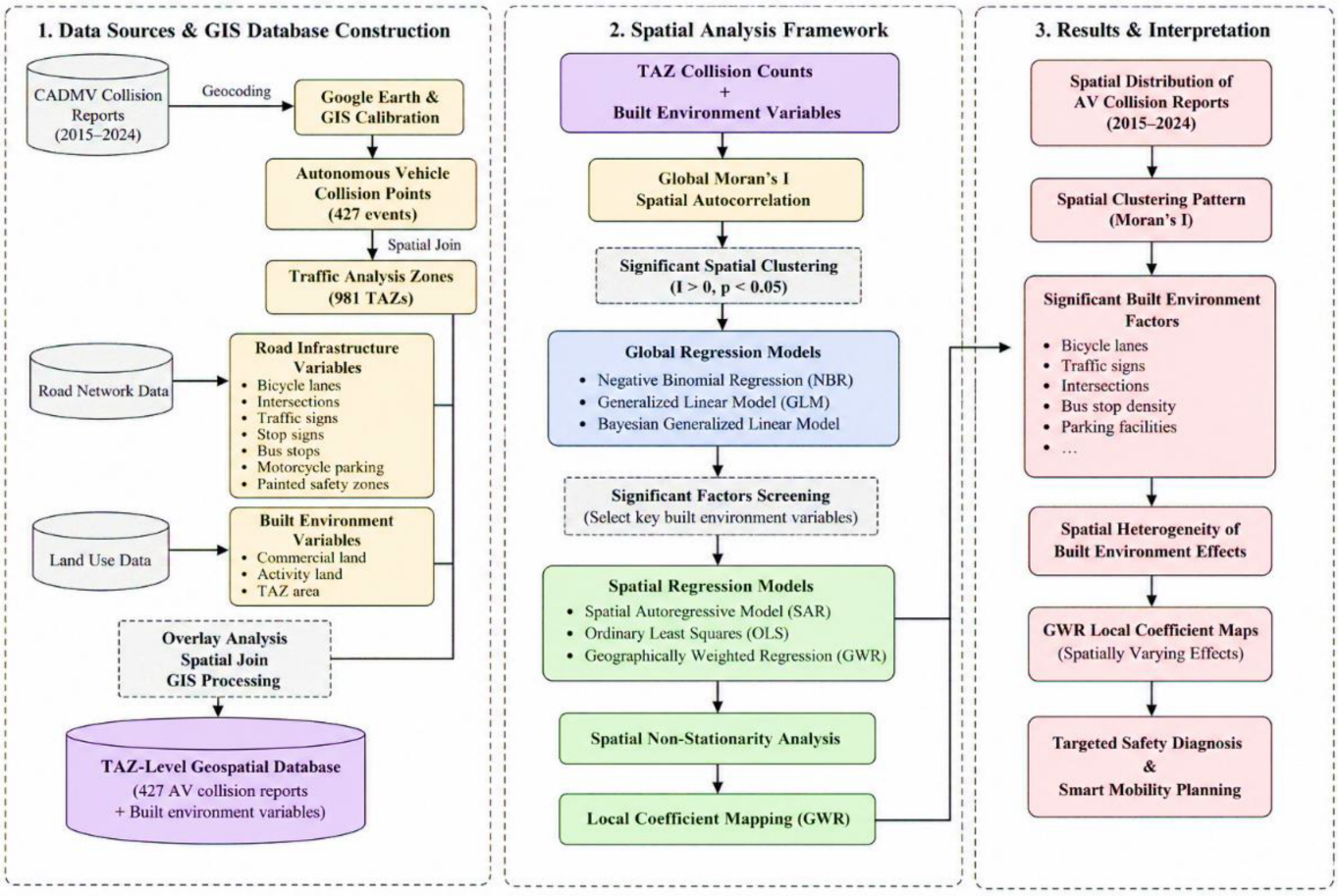


Figure 4. TAZ-Based GIS Workflow for Autonomous Vehicle Collision Report Analysis.

To rigorously evaluate the associations between built-environment variables and collision report counts, this study constructs a comprehensive, multi-perspective analytical framework comprising six distinct models, each serving a specific methodological purpose. The Negative Binomial Regression (NBR) [36,37] serves as the primary global baseline to accommodate the overdispersed count data. To validate this baseline, the Generalized Linear Model (GLM) [38] and Bayesian GLM are incorporated as robust sensitivity checks—the former compares alternative distributional assumptions, while the latter assesses parameter stability and uncertainty from a probabilistic perspective. Furthermore, to address the geographic nature of the data, Ordinary Least Squares (OLS) is utilized strictly as a non-spatial diagnostic baseline to verify residual spatial autocorrelation. Upon confirming spatial dependence, the Spatial Autoregressive (SAR) [39] model is deployed to capture global spatial spillover effects, and finally, Geographically Weighted Regression (GWR) [40] is employed to explicitly explore and visualize local spatial non-stationarity across the study area.

Given the inherent structure and distribution patterns of the autonomous vehicle collision-report dataset constructed herein, the aforementioned models are highly applicable for conducting a comprehensive analysis of the data. Regarding the analysis of the spatial heterogeneity of autonomous driving accidents, the Spatial Autoregressive Model and Geographically Weighted Regression are primarily utilized to address global spatial dependence and local spatial variation, respectively. The characteristics of these six models are summarized in Table 2.

Table 2. Comparison and Introduction of Applicability and Characteristics of Six Models.

Model	Applicable Data	Research Objectives	Result Types
NBR	Overdispersed (variance > mean) count data.	Analyze covariate impacts while resolving Poisson overdispersion.	Regression coefficients, *p*-values, dispersion parameter.
GLM	Continuous, count, or binary variables.	Model non-linear relationships across diverse distributions via link functions.	Regression coefficients, standard errors, goodness-of-fit (e.g., AIC).
Bayesian GLM	Small samples, data with priors, or parameter uncertainty.	Derive posterior distributions and integrate priors to enhance robustness.	Posterior means, credible intervals, MCMC diagnostics.
SAR	Spatially autocorrelated data (e.g., significant Moran's I).	Capture spatial spillover effects from neighboring regions.	Spatial lag coefficients, covariate coefficients, spatial error terms.
OLS	Continuous data satisfying linearity, homoscedasticity, and non-autocorrelation.	Establish linear baselines and evaluate global average effects.	Regression coefficients, R^2, residual spatial autocorrelation tests.
GWR	Spatially non-stationary data with geographically varying relationships.	Reveal localized variations and capture spatial non-stationarity.	Local regression coefficients, spatial variation maps (e.g., heatmaps).

2.2.1. Global Moran's I

Global Moran's I was employed to evaluate the spatial autocorrelation of TAZ-level autonomous vehicle collision report counts. A significantly positive value indicates that collision reports are spatially clustered rather than randomly distributed across the study area. Confirming this global spatial dependency is an essential exploratory step, providing the robust methodological justification required for deploying subsequent spatial econometric frameworks, specifically the SAR and GWR models.

The calculation of Global Moran's I [41] can be expressed as:

$$Moran's\ I = \frac{n}{SumW} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad (1)$$

where n denotes the total number of TAZ within the study area; x_i and x_j represent the observed autonomous driving accident frequencies in spatial units i and j, respectively; $\overline{x}$ is the mean accident count across all spatial units; and w_{ij} represents the elements of the spatial weight matrix W, which defines the geographic proximity or connectivity relationship between spatial units i and j.

The sum of all spatial weights across all sample points can be derived using the equation:

$$\begin{gathered} S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \\ z_i = x_i - \overline{x} \end{gathered} \quad (2)$$

where S_0 represents the aggregate sum of spatial weights across all sample points; w_{ij} denotes the spatial weight between sample points i and j; x_i is the observed variable value of the i sample point; and z_i signifies the deviation of the attribute value of a given sample point from its mean.

2.2.2. Negative Binomial Regression Model

While standard Poisson regression is traditionally used for count data, its rigid assumption of equi-dispersion is frequently violated by clustered collision records. Because our TAZ-level collision data exhibits overdispersion (variance > mean), Negative Binomial Regression (NBR) was adopted. By incorporating a dedicated dispersion parameter, the NBR model effectively accommodates this variance inflation, thereby ensuring accurate parameter estimation.

The probability mass function of the NBR model is given by:

$$P(Y = y) = \frac{\Gamma(y+\theta^{-1})}{\Gamma(y+1)\Gamma(\theta^{-1})} \left(\frac{\mu}{\mu+\theta^{-1}}\right)^{y} \left(\frac{\theta^{-1}}{\mu+\theta^{-1}}\right)^{\theta^{-1}} \quad (3)$$

where y is the dependent variable, and the parameters μ and θ are not fixed constants but are predicted through a linear combination of the independent variables, with $\mu, \theta \in (0, +\infty)$.

The NBR model can be expressed as:

$$log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (4)$$

where μ is the expected value of the dependent variable; $x_1, x_2, \ldots, x_n$ denote the independent variables; and $\beta_0, \beta_1, \ldots, \beta_n$ represent the model parameters, with a range of $(-\infty, +\infty)$.

2.2.3. Generalized Linear Model

The Generalized Linear Model (GLM) extends ordinary linear regression to accommodate response variables within the exponential distribution family. By utilizing specific link functions to connect linear predictors with the response, GLM effectively handles discrete, non-normal count data that violate standard linearity and constant variance assumptions.

Assume that the dependent variables $y_1, y_2, \ldots, y_n$ represent n independent observations following a distribution within the exponential family, whose probability density function is given by:

$$f(y_i|\theta_i,\ \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\ \phi)\right) \tag{5}$$

where θ_i and ϕ are parameters, with $\theta_i \in\ (-\infty, +\infty)$ and $\phi \in\ (0, +\infty)$, while $b(\cdot)$ and $c(\cdot)$ denote functions.

Let $x_1, x_2, \ldots, x_n$ be the observations of the p-dimensional independent variable X corresponding to $y_1, y_2, \ldots, y_n$. Define $\eta_i = x_i^T\beta$, where β represents a $p \times 1$ vector of unknown parameters. Assuming $E(Y_i) = \mu_i$ and that μ_i is related to η_i via the relationship $\eta_i = g(\mu_i)$ for $i = 1, 2, \ldots, n$, the model defined in this manner is referred to as a Generalized Linear Model. In this framework, θ_i is termed the natural parameter, ϕ is the dispersion parameter, and $g(\cdot)$ is called the link function.

To determine the optimal GLM structure, this study compares Gaussian-identity, Poisson-log, and Negative Binomial-log specifications for sensitivity analysis. Because the dependent variable is a sparse, non-negative count, the Negative Binomial-log specification was ultimately selected as the primary count-model for interpretation.

2.2.4. Bayesian Generalized Linear Model

As an extension of the Generalized Linear Model, the Bayesian GLM integrates Bayesian statistical methods to allow for the estimation of probability distributions for model parameters, rather than merely relying on point estimates. This approach offers distinct advantages in handling uncertainty and incorporating prior information.

The core concepts of the Bayesian GLM encompass the Bayesian statistical framework, the baseline GLM structure, and the selection of prior distributions. Within the Bayesian statistical paradigm, model parameters are treated as random variables characterized by prior distributions. These prior distributions reflect prior knowledge concerning the parameters. By applying Bayes' theorem, the posterior distribution of the parameters is derived by combining the likelihood function of the data with the specified prior distribution. The Bayesian GLM retains the fundamental architecture of the classical GLM, which includes the systematic component, the random component, and the link function. The key distinction lies in the estimation process, where the Bayesian GLM estimates the posterior distribution of parameters using Markov Chain Monte Carlo methods or alternative Bayesian computational techniques. The choice of prior distributions exerts a significant impact on the results of the Bayesian GLM. For instance, non-informative priors or weakly informative priors based on domain expertise can be implemented. In this study, the Bayesian GLM is used as a supplementary sensitivity specification rather than as the primary model for inference. In this study, the Bayesian GLM was implemented using an approximate Maximum a Posteriori estimation approach strictly as a comparative baseline, rather than full MCMC sampling.

2.2.5. Spatial Autoregressive Model

The SAR model is a statistical framework designed for analyzing data that exhibit spatial dependency. It presupposes that the correlation among observations is determined not merely by conventional temporal dynamics or other exogenous covariates (e.g., weather conditions, socio-economic factors, or macroscopic traffic volumes), but is inherently shaped by geographic location. By incorporating a spatial weight matrix, the SAR model effectively captures the spatial interdependence among spatial units, thereby facilitating a more accurate delineation of spatial data patterns. Its mathematical formulation is expressed as follows:

$$Y = \rho Wy + X\beta + \varepsilon \tag{6}$$

where Y is the vector of dependent variables; ρ is the spatial autoregressive coefficient with a range of $(-1, 1)$; W represents the spatial weight matrix; Wy denotes the spatial lag of

the dependent variable; X is the matrix of explanatory variables; β signifies the parameter vector, where $\beta \in (-\infty, +\infty)$; and ε is the vector of random errors following a multivariate normal distribution, $\varepsilon \sim N(0,\sigma^2 I)$ (with I being the identity matrix), where the mean of the errors is fixed at 0 and the variance satisfies $\sigma^2 \in (0, +\infty)$.

In this study, W was constructed based on first-order Queen contiguity, which defines TAZs as neighbors if they share at least one common edge or vertex. This specification is particularly suitable for the irregular polygon geometries of the TAZ framework. To facilitate interpretation and ensure that the spatial lag terms represent the weighted average of neighboring values, the matrix was row-standardized, meaning the sum of weights for each row equals one ($\sum_j W = 1$). To ensure the reliability of our findings, sensitivity analyses were conducted using alternative spatial-weight specifications, including Rook contiguity and K-nearest neighbors (K = 4 and K = 8). The results demonstrated that both the significance of the spatial autocorrelation and the direction of the SAR parameter estimates remained highly robust to the choice of the spatial weights matrix. Specifically, across the four spatial weight specifications, the Global Moran's I values ranged from 0.032 to 0.041 ($p < 0.05$), and the estimated SAR spatial autoregressive coefficients ranged from 0.264 to 0.305 ($p < 0.01$).

2.2.6. Ordinary Least Squares Model

The OLS Model is an optimization method used to estimate parameters in a linear regression model, aiming to identify the parameter combination that minimizes the sum of squared errors between the predicted and actual values. Its mathematical formulation is expressed as follows:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + \varepsilon_i \tag{7}$$

where y is the dependent variable; β_0 represents the intercept; x_{ij} denotes the j-th explanatory variable corresponding to the i-th TAZ; β_j is the fitting coefficient of the j-th variable; n signifies the total number of explanatory variable types; and ε_i represents the random error following a normal distribution, $\varepsilon_i \sim N(0, \sigma^2)$, where σ^2 is the variance of the error, the mean of the error is fixed at 0, and $\sigma^2 \in (0, +\infty)$.

2.2.7. Geographically Weighted Regression Model

The GWR Model is a local linear regression approach based on modeling spatially varying relationships, which aims to generate a localized regression model for every location within the study area, thereby effectively addressing the challenges inherent in local spatial analysis and spatial heterogeneity research. Its mathematical formulation is expressed as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{p} \beta_{ij}(u_i, v_i) x_{ij} + \varepsilon_i \tag{8}$$

where y is the dependent variable; x_{ij} represents the j-th explanatory variable for the i-th TAZ; (u_i, v_i) denote the geographic coordinates of the i-th sample point; $\beta_{ij}(u_i, v_i)$ is the local regression coefficient for the j-th variable at the i-th sample point; p indicates the number of independent variables; ε_i is the random error term following a normal distribution, $\varepsilon_i \sim N(0,\sigma^2)$, where σ^2 represents the variance of the error, the mean of the error is fixed at 0, and $\sigma^2 \in (0, +\infty)$; and i denotes the sample point index, with $i = 1, 2, \ldots, n$.

To calibrate the GWR model, an adaptive bi-square spatial kernel was employed to account for the heterogeneous distribution of TAZ. The optimal bandwidth was determined by minimizing the corrected Akaike Information Criterion (AICc), which effectively balances model fit and complexity.

3. Results

3.1. Spatial Association Analysis of Autonomous Vehicle Collision Reports

Based on TAZ-level autonomous vehicle collision report counts in San Francisco during the 2015–2024 study period, this section analyzes the spatial distribution of collision reports and their statistical associations with built-environment variables. The following results describe collision report concentration rather than exposure-normalized crash risk, because TAZ-level AV mileage, operating hours, vehicle presence, and traffic exposure data are not available.

To further investigate the spatial distribution characteristics of the accident data, the proportion of zero values was calculated, and an overdispersion test was conducted for the autonomous driving accident counts in each TAZ. Statistical results revealed that 63.6% of the samples contained zero values, indicating that a substantial number of TAZ recorded no autonomous vehicle collisions during the 2015–2024 study period. This high proportion of zeros demonstrates that autonomous vehicle accidents are spatially sparse and unevenly distributed across the study area. Furthermore, the overdispersion test showed a variance of 0.488 and a mean of 0.435 for the accident frequency (Table 3). Since the variance is only slightly greater than the mean, the data exhibits mild overdispersion. To account for this difference and ensure accurate parameter estimation, the Negative Binomial model was selected over the standard Poisson model.

Table 3. Results of Zero-value proportion and Overdispersion test.

Variable	Zero-Value Proportion	Mean	Variance
Autonomous vehicle collision report count	63.6%	0.435	0.488

Furthermore, given the high proportion of zero counts (63.6%), a Zero-Inflated Negative Binomial specification was additionally estimated to rigorously test for excess zeros originating from a separate data-generating process. However, the comparative analysis revealed that the zero-inflated formulation did not provide a material improvement in model fit (ZINB AIC = 1668.50) compared to the conventional NBR model (AIC = 1669.300). Because the added model complexity did not yield superior goodness-of-fit and the substantive associations remained consistent, the conventional NBR model was ultimately retained for parsimony and structural comparability with the subsequent spatial regression models.

To examine whether TAZ-level autonomous vehicle collision report counts exhibit global spatial dependence, Global Moran's I was calculated using the spatial weights matrix defined in Section 3.1. The positive and statistically significant Moran's I statistic indicates global positive spatial autocorrelation in collision report counts across TAZs (Figure 5). In other words, TAZs with similar collision report counts tend to be located near one another more often than would be expected under complete spatial randomness. However, Global Moran's I does not identify the specific locations of high-high or low-low clusters. Therefore, the result is interpreted as evidence of citywide spatial dependence and supports the subsequent use of spatial regression models.

The NBR model reports a Pseudo R^2 of 0.067 (Table 4). Table 5 was condensed to retain the key statistically associated variables while preserving a compact cross-model comparison (including NBR, GLM, Bayesian GLM, and SAR). For full transparency, the complete all-variable coefficient table is provided in Table A1 of Appendix A. Pedestrian street number, infrastructure facility density, and TAZ area show negative associations, whereas the remaining listed variables show positive associations.

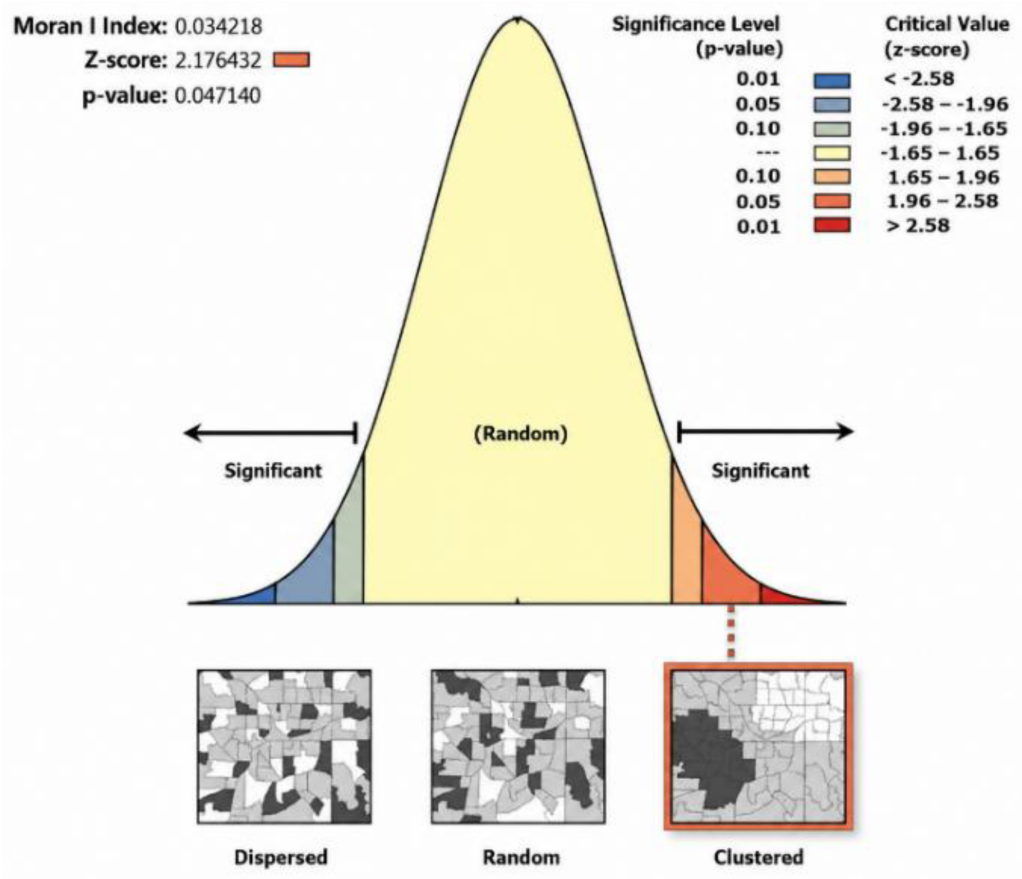


Figure 5. Global Moran's I test for TAZ-level autonomous vehicle collision report counts. Note: The Global Moran's I index is 0.034, accompanied by a Z-score of 2.176 and a permutation-based pseudo *p*-value of 0.047 (derived from 999 spatial permutations). This indicates a weak but statistically significant global positive spatial autocorrelation in the collision reports, rather than a completely random spatial distribution.

Table 4. Results of Negative Binomial Regression, Spatial Autoregressive Model, Ordinary Least Squares Method, and Geographically Weighted Regression Model.

Model	NBR	OLS	SAR	GWR
AIC	1669.300	/	3151.120	/
BIC	1820.844	/	3297.780	/
Log likelihood	−803.650	/	−1545.560	/
AICc	/	3135.500	/	3125.630
Pseudo R^2	0.067	/	0.150	/
R^2	/	0.184	/	/
Adjusted R^2	/	0.176	/	0.228
Local R^2	/	/	/	0.105~0.452
Spatial coefficient (ρ)	/	/	0.285	/
Residual Moran's I	/	0.215	0.015	0.004
Bandwidth	/	/	/	158.000
ENP	/	/	/	84.620

Note: The slash symbol "/" indicates that the statistical metric is not applicable for the respective model.

These patterns may reflect differences in local traffic complexity, roadside activity, mixed motorized and non-motorized traffic interactions, and built-environment configuration. Previous studies have also suggested that bicycle-related facilities and mixed-traffic road environments may involve more complex right-of-way interactions [42]. Rather than

providing a uniform city-wide assessment, the TAZ-based spatial framework quantitatively identifies localized geographic clusters where the estimated associations of specific built-environment variables with AV collision frequencies are stronger in magnitude. By mapping these spatially varying relationships, this method advances the state-of-the-art by offering an exploratory spatial framework. It assists urban planners in identifying localized spatial patterns exhibiting higher collision-report frequencies. Any subsequent safety assessment, such as sign-layout adjustment, intersection safety review, or station and parking-facility planning, should be supported by site-specific safety audits and detailed engineering assessment [43].

Table 5. Coefficient results for NBR, GLM, Bayesian GLM, and SAR models.

Category	Variable/Coef.	NBR	GLM	Bayesian GLM	SAR
Transportation infrastructure environment	Bus stop density	0.384 *	0.427 ***	0.258 *	0.091
	Bicycle parking station number	0.195 *	0.283 ***	0.204 *	0.342 ***
	Bicycle parking station density	0.451 ***	0.518 ***	0.339 **	0.068
	Bicycle lane number	0.276 *	0.362 ***	0.145	0.112
	Painted safety zone density	0.214 *	0.275 **	0.158	0.021
	Motorcycle parking lot density	0.367 ***	0.449 ***	0.312 ***	0.059
	Stop sign density	0.298 **	0.354 ***	0.267 **	0.083
	Pedestrian street number	−0.165 *	−0.231 **	−0.114	−0.058
	Pedestrian street density	−0.102	−0.187 *	−0.095	−0.058
	Traffic sign number	0.532 ***	0.615 ***	0.488 ***	0.573 ***
	Traffic sign density	0.127	0.144	0.109	0.296 **
	Intersection number	0.319 ***	0.388 **	0.305 **	0.461 ***
	Intersection density	0.156	0.263 **	0.197 *	0.124
	Infrastructure facility density	−0.407 ***	−0.472 ***	−0.365 **	−0.092
Land use	TAZ area	−0.589 ***	−0.634 ***	−0.287	−0.375 *

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

For the GLM sensitivity analysis, Gaussian-identity, Poisson-log, and Negative Binomial-log specifications were estimated (Table 6). Because the dependent variable is a sparse non-negative count variable with a high proportion of zero-count TAZs, the Negative Binomial-log specification was retained as the main count-model specification for interpreting the associations between built-environment variables and collision report counts. Gaussian-identity was included only as a linear reference specification.

Table 6. Goodness-of-Fit Comparison of Candidate GLM Specifications for TAZ-Level Collision Report Counts.

Model	AIC	BIC	Log Likelihood
Gaussian-identity	1942.68	2094.22	−940.34
Negative Binomial-log	1669.30	1820.84	−803.65
Poisson-log	1710.80	1857.45	−825.40

Note: Gaussian-identity is included as a linear reference specification. The Negative Binomial-log model achieved the lowest AIC and BIC among the candidate count-model specifications and was retained as the main count-model specification for interpretation.

In the Negative Binomial-log specification, bus stop density, bicycle parking station number, bicycle parking station density, bicycle lane number, painted safety zone density, motorcycle parking lot density, stop sign density, pedestrian street number, pedestrian street density, traffic sign number, intersection number, intersection density, infrastructure facility density, and TAZ area are statistically associated with TAZ-level collision report counts. Pedestrian street number, pedestrian street density, infrastructure facility density,

and TAZ area show negative associations, whereas the remaining listed variables show positive associations.

These results suggest that built-environment characteristics related to transit stops, bicycle-related facilities, roadside activity, traffic control, pedestrian facilities, and spatial unit size are associated with the spatial distribution of collision report counts. The observed associations may reflect differences in local traffic complexity, roadside activity intensity, and the configuration of facilities across TAZs. Furthermore, because a relatively large number of candidate variables were evaluated including a significance threshold of $p < 0.10$, there is an inherent risk of multiplicity. Therefore, the individual marginally significant coefficients must be explicitly regarded as hypothesis-generating spatial associations rather than confirmatory inferences. A primary limitation of this study is the lack of macroscopic traffic exposure data. Without a robust macroscopic exposure denominator, it is difficult to determine whether higher collision counts in certain TAZs are caused by specific built-environment configurations or simply by a higher frequency of AV operations in those zones.

For the Bayesian GLM, a Negative Binomial-log specification was estimated as a supplementary sensitivity model because the dependent variable is a sparse non-negative count variable. The Bayesian GLM results are reported as supplementary evidence and are not used as the primary basis for interpreting built-environment associations (Table 7). The Bayesian GLM sensitivity results provide supplementary evidence that several transit-related, bicycle-related, traffic-control, and intersection-related variables are statistically associated with TAZ-level collision report counts.

Table 7. Fit Statistics for the Bayesian Negative Binomial-log Model for TAZ-Level Collision Report Counts.

Model	AIC	BIC	Log Likelihood
Gamma-log	103.78	240.73	−23.91
Negative Binomial-log	1581.07	1717.93	−762.53

Note: The Bayesian Negative Binomial-log model is reported as a supplementary sensitivity specification.

The OLS model yields an adjusted R^2 of 0.176 (Table 4). Because OLS is inherently a non-spatial linear regression method, it is utilized here strictly as a global baseline to evaluate average associations and to generate residuals for spatial diagnostics. The diagnostic tests reveal significant spatial autocorrelation in the OLS residuals (Residual Moran's I = 0.215, $p < 0.01$). This presence of unmodeled spatial dependence confirms the necessity of deploying dedicated spatial econometric frameworks, such as the SAR and GWR models, to properly account for geographic effects.

To visually corroborate the spatial autocorrelation detected in the global baseline, Figure 6 maps the spatial distribution of the OLS standardized residuals. While standardized residuals typically fall within a normal threshold of −2.5 to 2.5, absolute values exceeding 2.5 signify substantial discrepancies between the model's predictions and the actual observed values at those specific locations. As illustrated, these poorly predicted zones are not randomly scattered but exhibit distinct geographic clustering. Such clustered extreme values suggest that unobserved spatial heterogeneity or localized extreme factors heavily influence collision frequencies at certain observation points, thereby exceeding the predictive capacity of a spatially stationary global model.

Building upon the findings and limitations of the global baseline models, the SAR model was estimated to account for the spatial dependence of the collision data. Compared with the OLS baseline, the SAR model substantially reduces residual spatial autocorrelation,

with Residual Moran's I decreasing from 0.215 to 0.015. This suggests that spatial dependence should be considered when modeling TAZ-level collision report counts. Furthermore, the estimated spatial autoregressive coefficient (ρ = 0.285) and the substantial reduction in residual spatial autocorrelation (Residual Moran's I = 0.015) indicate that the SAR model captures part of the spatial dependence in the data. Under this spatial framework, variables statistically associated with TAZ-level collision report counts in the SAR model include the number of bicycle parking stations, the number of traffic signs, traffic sign density, the number of intersections, and the area of the TAZ. Among these, the area of the Traffic Analysis Zone is negatively correlated with collision report counts, whereas the remaining variables all exhibit a positive correlation.

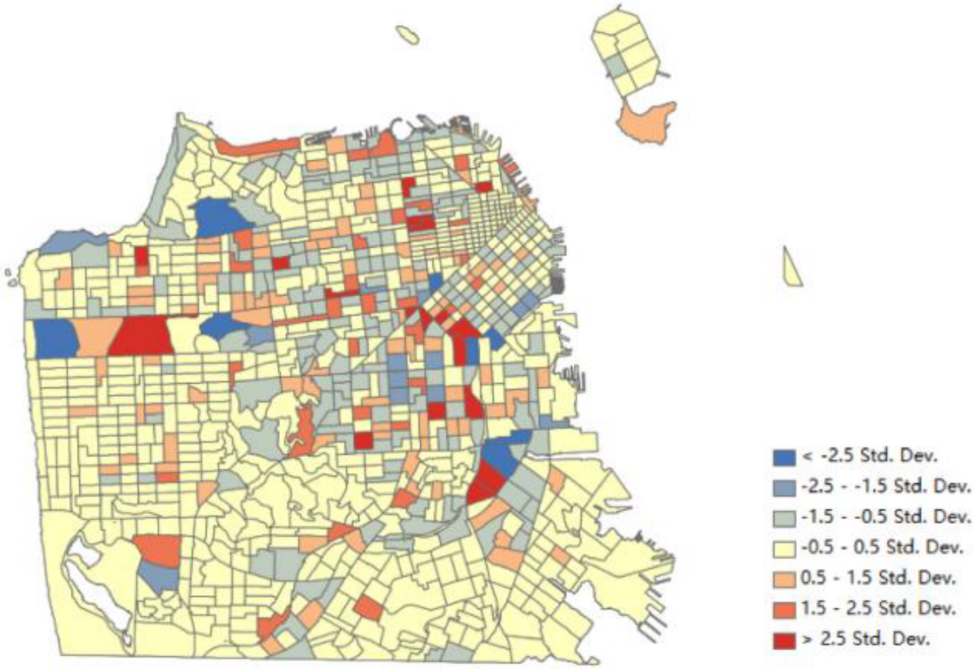


Figure 6. Standardized residual distribution of OLS model for TAZ-level collision report counts during the 2015–2024 study period.

Although the SAR model successfully accounts for global spatial spillovers, it inherently assumes spatial stationarity—meaning the impacts of built-environment variables on collision report counts are presumed to be uniform across the entire city. However, given the complex and heterogeneous urban landscape of San Francisco, these associations are highly likely to exhibit local spatial non-stationarity. To explicitly explore these spatially varying relationships, the GWR model was further employed. Rather than estimating a single global average, GWR computes localized coefficients for each TAZ, offering a granular perspective on how underlying mechanisms shift across different neighborhoods. As reported in Table 4, the application of GWR captures these local variations effectively. The GWR model was calibrated using an adaptive bi-square kernel. Based on the AICc minimization criterion, the optimal bandwidth was determined to be 158 nearest neighbors, resulting in an Effective Number of Parameters (ENP) of 84.62. The model yielded an Adjusted R^2 of 0.228 and an AICc of 3125.63. Because counting models (e.g., NBR) and Gaussian spatial models (e.g., GWR) possess distinct likelihood structures, a direct AIC comparison between them is mathematically inappropriate. However, when compared strictly to the global Gaussian baseline (OLS AICc = 3135.50), the GWR model demonstrates a superior fit. It further reduces residual spatial autocorrelation (Residual Moran's I = 0.004).

As reported in Table 4, the application of GWR captures these local variations effectively (Adjusted R^2 = 0.228; AICc = 3125.63) and further reduces residual spatial autocorrelation (Residual Moran's I = 0.004). Consequently, while global spatial models provide essential overarching diagnostics, GWR serves as a crucial exploratory extension to explore local spatial heterogeneity in the associations between built-environment variables and collision report counts.

Prior to estimating the GWR model, a rigorous multicollinearity diagnosis is essential, as localized regression algorithms are highly vulnerable to linear dependencies among explanatory variables. In this study, a Variance Inflation Factor (VIF) threshold of 10 was applied to detect multicollinearity. Consequently, redundant variables exceeding this critical value were systematically reduced to guarantee the mathematical stability and reliability of the local coefficient estimations (Table 8).

Table 8. Variance inflation factor results for candidate variables.

Variable	VIF	Variable	VIF
Bus stops number	>1000	Pedestrian street density	49.7924
Bus stop density	88.4142	Speed-restricted segment number	20.5342
Bicycle parking station number	>1000	Speed-restricted segment density	50.4211
Bicycle parking station density	79.6360	Traffic sign number	>1000
Bicycle lane number *	2.8156	Traffic sign density	74.4582
Bicycle lane density *	3.2876	Intersection number *	2.1648
Slow street number *	7.4803	Intersection density *	2.4794
Slow street density *	7.3110	Infrastructure facility number	>1000
Painted safety zone number	>1000	Infrastructure facility density	>1000
Painted safety zone density *	4.2772	TAZ area *	1.8192
Motorcycle parking lot number	>1000	Commercial land area	13.6006
Motorcycle parking lot density	706.8733	Commercial land density	13.4709
Stop sign number	>1000	Activity land area *	4.2307
Stop sign density	32.3274	Activity land density *	4.1704
Pedestrian street number	18.3977		

Note: * denotes the variables retained for the Geographically Weighted Regression (GWR) model after applying the VIF < 10 threshold.

After removing the redundant variables, ten explanatory variables were retained for the GWR modeling: the number of bicycle lanes, bicycle lane density, the number of slow streets, slow street density, painted safety zone density, the number of intersections, intersection density, activity land area, activity land density, and the area of the TAZ (Table 8). Furthermore, local-collinearity diagnostics were examined for the retained variables. The local condition numbers across all TAZs remained well below the critical threshold of 30, confirming that the localized regressions were not compromised by severe local multicollinearity. Among these factors, the local coefficients of the number of slow streets and activity land density predominantly exhibit negative correlations with collision report counts. Interestingly, while painted safety zone density and intersection density demonstrated positive correlations in the global baselines (Table 5), their localized impacts in the GWR model exhibit significant spatial non-stationarity, shifting to negative associations in certain local neighborhoods. The remaining variables predominantly demonstrate positive correlations. Compared with the OLS residual map in Figure 6, the GWR residual map in Figure 7 shows less visible geographic clustering. This visual pattern is consistent with the lower residual Moran's I value reported for GWR. The result suggests that the local model reduces part of the remaining spatial autocorrelation in the OLS baseline. However, the residual comparison should be interpreted as exploratory diagnostic evidence rather than as proof that GWR fully captures all local mechanisms or removes all spatial dependence.

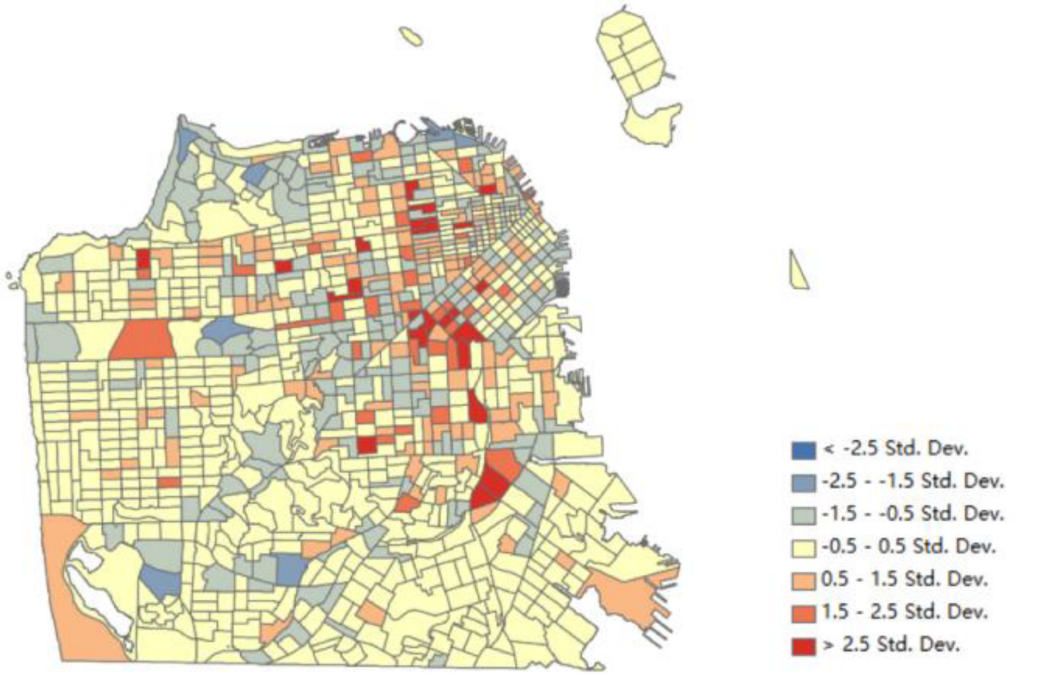


Figure 7. Standardized residual distribution of GWR model for TAZ-level collision report counts during the 2015–2024 study period.

The GWR model was used to explore the local spatial non-stationarity in the associations between built-environment variables and autonomous vehicle collision report counts. Figure 8 presents the spatial distribution of local GWR coefficients. These coefficient maps show where the associations are stronger or weaker across San Francisco.

As illustrated in Figure 8a,b, the local coefficients for bicycle lane number and bicycle lane density show positive associations with collision report counts in most areas, but the strength of these associations varies across space. The coefficient for bicycle lane number tends to be stronger in the eastern part of San Francisco, while the coefficient for bicycle lane density shows a different spatial gradient. These patterns suggest that interactions between autonomous vehicles and bicycle-related infrastructure may differ across local urban environments. As illustrated in Figure 8c,d, the number of slow streets and slow street density show different local coefficient patterns. The number of slow streets is negatively associated with collision report counts in several areas, whereas slow street density shows positive associations in some local areas. This contrast may indicate that the role of slow-street facilities depends on local street layout, surrounding land use, and traffic operation conditions.

As illustrated in Figure 8e, painted safety zone density shows a negative local association with collision report counts in several areas, and the coefficient magnitude varies across space. This pattern may reflect differences in road marking visibility, right-of-way organization, and local traffic management conditions. Therefore, the result should be used to identify areas where marking visibility and right-of-way separation may require field verification.

As illustrated in Figure 8f,g, intersection number and intersection density show spatially varying associations with collision report counts. The local coefficients indicate that intersection-related variables are more strongly associated with collision report counts in some parts of the city than in others. These differences may be related to unobserved variations in intersection design, signal control, pedestrian and bicycle activity, traffic demand, and AV exposure.

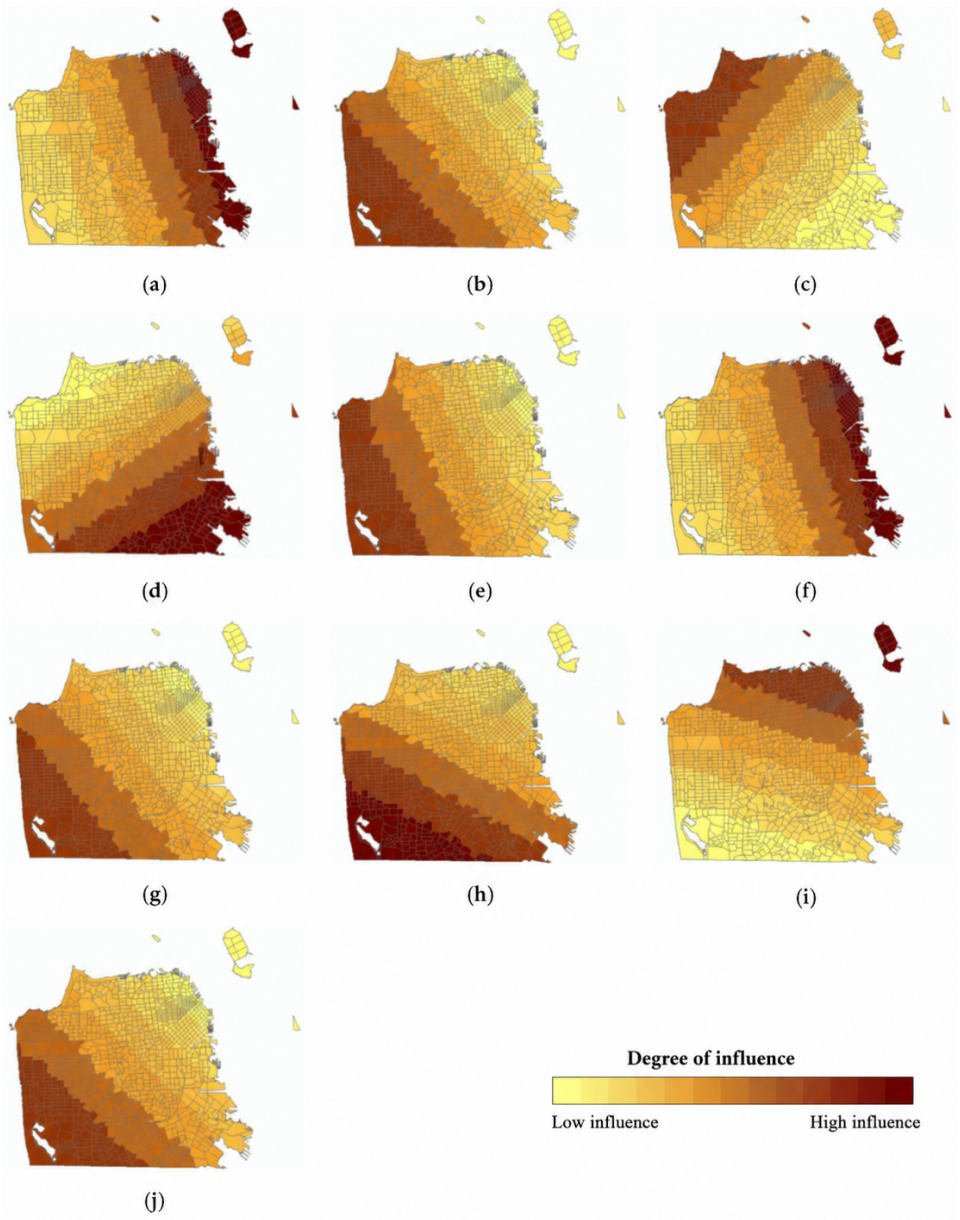


Figure 8. Spatial variation in GWR coefficients for built-environment variables. (**a**) Number of bicycle lanes. (**b**) Density of bicycle lanes. (**c**) Number of slow streets. (**d**) Slow street density. (**e**) Painted safety zone density. (**f**) Number of intersections. (**g**) Intersection density. (**h**) Activity land area. (**i**) Activity land density. (**j**) Area of Traffic Analysis Zone.

As illustrated in Figure 8h–j, activity land area, activity land density, and TAZ area also show spatially varying associations with collision report counts. These patterns suggest that land-use intensity and spatial unit size may affect how collision reports are aggregated and interpreted at the TAZ level. Larger TAZs may include more heterogeneous road and land-use conditions, while activity-related land-use variables may reflect differences in pedestrian activity, trip attraction, and local traffic complexity [44].

Overall, the GWR coefficient maps reveal clear spatial variation in the estimated local coefficients between built-environment variables and AV collision reports. By quantifying these localized variations, this methodology provides a descriptive, data-driven exploratory tool. It enables urban planners and safety engineers to transition from uniform city-wide assumptions to considering localized spatial variations, directly identifying geographic clusters with higher collision frequencies where specific infrastructure elements exhibit stronger estimated associations.

3.2. Descriptive Sensitivity Analysis of TAZ-Level Collision Report Counts

To examine whether the main spatial association patterns remain stable when the dependent variable is measured annually, this section uses the annual number of autonomous vehicle collision reports in each TAZ as a descriptive sensitivity analysis. The purpose of this analysis is not to identify temporal causal mechanisms or policy effects, because the current model does not include year-specific policy variables, AV deployment intensity, or annual exposure measures. Instead, the annual-count analysis is used to assess whether the variables identified in the cumulative-count models show similar statistical associations under a finer temporal aggregation. The descriptive sensitivity analysis is reported as descriptive sensitivity evidence and does not model within-TAZ temporal dependence or year-specific policy effects.

The Negative Binomial Regression model yields a Pseudo R^2 of 0.067. Variables statistically associated with annual TAZ-level collision report counts include bus stop density, the number of bicycle parking stations, bicycle parking station density, the number of bicycle lanes, motorcycle parking lot density, stop sign density, the number of traffic signs, the number of intersections, intersection density, infrastructure facility density, and the area of the TAZ. Among these, infrastructure facility density and the area of the TAZ are negatively associated with annual collision report counts, whereas all other variables exhibit a positive correlation.

For the annual-count GLM analysis, Gaussian-identity, Poisson-log, and Negative Binomial-log specifications were estimated as sensitivity models (Table 9). Because the dependent variable remains a sparse non-negative count variable, the Negative Binomial-log specification was retained as the main count-model specification. Gaussian-identity was included only as a linear reference specification.

Table 9. Goodness-of-Fit Comparison of Candidate GLM Specifications for Annual TAZ-Level Collision Report Counts.

Model	AIC	BIC	Log Likelihood
Gaussian-identity	3836.00	4037.35	−1890.00
Negative Binomial-log	3417.20	3618.35	−1680.50
Poisson-log	3544.40	3738.56	−1745.20

Note: Gaussian-identity is included as a linear reference specification. The Negative Binomial-log specification is retained as the main count-model specification for the descriptive sensitivity analysis.

In the Negative Binomial-log specification, the variables statistically associated with annual TAZ-level collision report counts include bus stop density, bicycle parking station number, bicycle parking station density, bicycle lane number, motorcycle parking lot density, stop sign density, traffic sign number, intersection number, intersection density, infrastructure facility density, and TAZ area. Among these variables, infrastructure facility density and TAZ area show negative associations, whereas the remaining listed variables show positive associations. These results are interpreted as descriptive sensitivity analysis rather than causal effects.

For the annual Bayesian GLM sensitivity analysis, a Negative Binomial-log specification was estimated as a supplementary count-model descriptive sensitivity analysis (Table 10). The annual Bayesian GLM results are reported as supplementary evidence and are not used as the primary basis for causal interpretation. In the Negative Binomial-log specification, the variables statistically associated with annual TAZ-level collision report counts include bus stop density, bicycle parking station number, bicycle parking station density, bicycle lane number, bicycle lane density, painted safety zone density, motorcycle parking lot density, stop sign density, pedestrian street number, pedestrian street density,

traffic sign number, intersection number, intersection density, activity land area, infrastructure facility density, and TAZ area. Among these variables, bicycle lane density, pedestrian street number, infrastructure facility density, and TAZ area show negative associations, whereas the remaining listed variables show positive associations. The annual Bayesian GLM results are used as supplementary evidence rather than as the primary basis for causal interpretation.

Table 10. Goodness-of-Fit Comparison of Candidate Bayesian GLM Specifications for Annual TAZ-Level Collision Report Counts.

Model	AIC	BIC	Log Likelihood
Gamma-log	11,056.40	11,257.75	−5500.20
Negative Binomial-log	3424.37	3625.16	−1684.80

Note: The Bayesian Negative Binomial-log model is reported as a supplementary sensitivity specification.

4. Discussion

This study developed a TAZ-based geospatial framework to examine the spatial distribution of autonomous vehicle collision reports in San Francisco. By integrating 427 collision reports from 2015 to 2024 with road infrastructure, land-use, and planning-unit data across 981 Traffic Analysis Zones, this study constructed a spatially structured database for GIS-based collision report analysis. Several global and spatial regression models were used to examine the associations between built-environment variables and collision report counts and to explore local spatial non-stationarity.

The main findings are as follows. First, autonomous vehicle collision reports showed a weak but statistically significant global positive spatial autocorrelation at the TAZ level (Moran's I = 0.034, permutation-based p-value = 0.047), indicating that these reports were not randomly distributed across urban space [11–13]. The GWR results provided exploratory evidence that the associations between built-environment variables and collision report counts varied across San Francisco [24–26]. This suggests that local spatial non-stationarity should be considered when analyzing autonomous vehicle collision reports within a GIS framework.

Second, the global models suggested that several road infrastructure variables, including bus stop density, bicycle parking station density, bicycle lane number, motorcycle parking lot density, stop sign density, traffic sign number, and intersection number, were statistically associated with collision report counts [18–22]. The local results further showed that the directions and magnitudes of these associations varied across space. These findings indicate that collision report analysis should not rely only on global average effects, but should also consider local spatial differences in urban built environments.

Third, this study provides a geospatial data integration workflow for linking autonomous vehicle collision reports with urban spatial data. The constructed TAZ-level dataset can be updated as new collision reports and infrastructure data become available. This workflow may support spatially targeted safety diagnosis, smart mobility planning, and further research on autonomous vehicle operation in complex urban environments [42]. It also provides a reference for developing similar GIS-based autonomous vehicle collision databases in other regions.

Several limitations should be noted. First and foremost, the central limitation of this study is the absence of detailed AV exposure data, such as AV-kilometers traveled (VKT), operating hours, fleet size, or localized traffic volume. Consequently, the models analyze raw collision-report counts rather than exposure-normalized crash rates or true risk. As a result, positive associations with specific infrastructure or land-use variables (e.g., active commercial zones or intersection density) may simply indicate that these facilities are located in areas where AV activity or general traffic is greater. Due to the lack of TAZ-level autonomous vehicle mileage,

operating hours, vehicle presence, and traffic exposure data, the estimated associations should not be interpreted as causal effects or direct measures of accident risk. Second, the current dataset is limited to San Francisco, and the transferability of the findings to other cities requires further validation. Third, while this study establishes a robust macroscopic spatial framework, it is currently limited by the absence of TAZ-level autonomous vehicle exposure metrics (e.g., VMT) [23]. Future research should integrate comprehensive zonal exposure data to transition from spatial count modeling to exposure-normalized spatial risk analysis [43]. Fourth, there is an inherent methodological limitation in the spatial framework. SAR and GWR models assume normally distributed errors, which is mathematically imperfect for sparse count data. While count-compatible spatial models like Geographically Weighted Negative Binomial Regression offer better theoretical coherence, they frequently face severe local convergence failures when applied to highly sparse datasets (e.g., our 63.6% zero-count rate). Consequently, we adopted a pragmatic approach: the standard NBR model was prioritized for rigorous count estimation, while SAR and GWR served strictly as exploratory tools to visualize macro-level spatial spillovers and local non-stationarity. Future studies should develop robust Bayesian spatial count frameworks to overcome these computational challenges. Fifth, the GWR methodology introduces specific statistical vulnerabilities. Although local multicollinearity was effectively diagnosed and controlled, the multiplicity of localized regressions inherently inflates the risk of false positive rates due to multiple local significance testing. While GWR serves as a powerful exploratory tool, the multiplicity of localized regressions inherently inflates the risk of Type I errors. Therefore, the local coefficients and their significance must be interpreted with appropriate caution, strictly as exploratory spatial associations rather than robust causal effects.

Furthermore, the analytical framework is inherently subject to the Modifiable Areal Unit Problem (MAUP). Because Traffic Analysis Zones were originally designed for macroscopic travel demand forecasting rather than micro-scale road safety analysis, the statistical associations and spatial non-stationarity patterns identified in this study may be sensitive to the choice of the spatial aggregation unit. The specific scale and boundary delineations of TAZs could artificially amplify or mask certain localized infrastructure associations. Therefore, the findings should be interpreted strictly as macro-level spatial trends rather than absolute, scale-independent safety dynamics. Future research should consider multi-scale sensitivity analyses, such as comparing TAZ-based results with uniform hexagon-grid or micro-segment frameworks, to further evaluate the effects of spatial aggregation on autonomous vehicle collision modeling.

5. Conclusions

This study introduces a TAZ-level geospatial framework to explicitly quantify the spatial dependence and local non-stationarity of autonomous vehicle collision distributions, thereby contributing to automated mobility research. Moving beyond traditional global statistical approaches and micro-level human-vehicle interaction studies, this research confirms that AV safety is not solely a uniform technological challenge; rather, it is spatially associated with localized interactions with the urban built environment.

The application of spatial econometric models, particularly GWR, revealed that the estimated associations of road infrastructure—such as bicycle facilities, painted safety zones, and intersection density—vary clearly across different urban landscapes. By standardizing scattered collision records into a structured GIS database, this study equips urban planners and policymakers with an exploratory tool for spatial safety analysis. Rather than relying solely on global assumptions, cities can consider localized environmental contexts in specific geographic clusters exhibiting higher collision frequencies, while AV developers can utilize these spatial associations as preliminary references for their Operational Design

Domains. Although future efforts must incorporate dynamic traffic exposure and temporal variables to derive definitive causal risks, this study establishes a vital geospatial foundation for planning safer and more resilient smart transportation networks.

Author Contributions: Yao Wu: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, and visualization. Haochen Feng: Methodology, software, data curation, visualization and editing. Yichang Shao: Conceptualization, methodology, writing—review and editing, and supervision. Junzhuo Hu: Validation, formal analysis, and writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (52202382) and the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (NY225088).

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors sincerely thank the editors and reviewers for their constructive comments and suggestions, which helped improve this manuscript.

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

Appendix A

Table A1. Coefficient results for NBR, GLM, Bayesian GLM, and SAR model.

Category	Variable/Coef.	NBR	GLM	Bayesian GLM	SAR
Transportation infrastructure environment	Bus stops number	0.042	0.067	0.031	0.015
	Bus stop density	0.384 *	0.427 ***	0.258 *	0.091
	Bicycle parking station number	0.195 *	0.283 ***	0.204 *	0.342 ***
	Bicycle parking station density	0.451 ***	0.518 ***	0.339 **	0.068
	Bicycle lane number	0.276 *	0.362 ***	0.145	0.112
	Bicycle lane density	0.088	0.105	0.063	0.004
	Slow street number	−0.024	−0.011	−0.038	−0.009
	Slow street density	0.053	0.048	0.063	0.032
	Painted safety zone number	0.019	0.026	0.014	0.037
	Painted safety zone density	0.214 *	0.275 **	0.158	0.021
	Motorcycle parking lot number	0.046	0.071	0.028	0.017
	Motorcycle parking lot density	0.367 ***	0.449 ***	0.312 ***	0.059
	Stop sign number	0.008	0.013	0.009	0.044
	Stop sign density	0.298 **	0.354 ***	0.267 **	0.083
	Pedestrian street number	−0.165 *	−0.231 **	−0.114	−0.058
	Pedestrian street density	−0.102	−0.187 *	−0.095	−0.058
	Speed-restricted segment number	0.061	0.084	0.036	0.049
	Speed-restricted segment density	0.035	0.059	0.021	0.028
	Traffic sign number	0.532 ***	0.615 ***	0.488 ***	0.573 ***
	Traffic sign density	0.127	0.144	0.109	0.296 **
	Intersection number	0.319 ***	0.388 **	0.305 **	0.461 ***
	Intersection density	0.156	0.263 **	0.197 *	0.124
	Infrastructure facility number	−0.013	−0.022	−0.017	−0.031
	Infrastructure facility density	−0.407 ***	−0.472 ***	−0.365 **	−0.092
Land use	TAZ area	−0.589 ***	−0.634 ***	−0.287	−0.375 *
	Commercial land area	0.027	0.019	0.033	0.025
	Commercial land density	0.051	0.038	0.046	0.025
	Activity land area	0.084	0.095	0.062	0.073
	Activity land density	0.015	0.028	0.019	0.005

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

References

1. Martínez, L.M.; Viegas, J.M.; Silva, E.A. A traffic analysis zone definition: A new methodology and algorithm. *Transportation* **2009**, *36*, 581–599. [CrossRef]
2. Abdel-Aty, M.; Ding, S. A matched case-control analysis of autonomous vs human-driven vehicle accidents. *Nat. Commun.* **2024**, *15*, 4931. [CrossRef]
3. Zeng, Q.; Wang, X.; Zhang, X.; Wen, H. Analysis of Factors Influencing Seasonal Accident Frequency on Highways Based on Spatiotemporal Interaction Model. *Chin. J. Highw.* **2020**, *33*, 255–263.
4. Fu, H.; Ye, S.; Fu, X.; Chen, T.; Zhao, J. New insights into factors affecting the severity of autonomous vehicle crashes from two sources of AV incident records. *Travel Behav. Soc.* **2024**, *38*, 100934. [CrossRef]
5. Ashraf, M.T.; Dey, K.; Mishra, S.; Rahman, M.T. Extracting rules from autonomous-vehicle-involved crashes by applying decision tree and association rule methods. *Transp. Res. Rec.* **2021**, *2675*, 522–533. [CrossRef]
6. Kutela, B.; Das, S.; Dadashova, B. Mining patterns of autonomous vehicle crashes involving vulnerable road users to understand the associated factors. *Accid. Anal. Prev.* **2022**, *165*, 106473. [CrossRef]
7. Wang, J.; Chen, X.; Jiao, P.; Tan, C.; Wang, Z. Study on the causal interaction effects of traffic accident severity considering the built environment. *Transp. Syst. Eng. Inf.* **2024**, *24*, 272–280.
8. Jiang, N.; Li, N. The Relationship between Meteorological Conditions, Air Quality Index and Vehicle Accidents in Beijing. *J. Beijing Norm. Univ. (Nat. Sci. Ed.)* **2017**, *53*, 609–614.
9. Gao, Y.; Gong, J. Characteristics and Causes Analysis of Road Traffic Accidents in China. *J. Saf. Environ.* **2023**, *23*, 1–12. [CrossRef]
10. Wang, W. Analysis of Factors Influencing Urban Road Traffic Accidents and Traffic Safety Evaluation. Master's Thesis, Chang'an University, Xi'an, China, 2023.
11. Pljakić, M.; Jovanović, D.; Matović, B. The influence of traffic-infrastructure factors on pedestrian accidents at the macro-level: The geographically weighted regression approach. *J. Saf. Res.* **2022**, *83*, 248–259. [CrossRef]
12. Iamtrakul, P.; Chayphong, S.; Zhang, J. Spatially heterogeneous impacts of built environment factors on traffic accidents: A multi-year analysis (2015–2021) in Suburban Bangkok, Thailand. *Transp. Eng.* **2025**, *20*, 100324. [CrossRef]
13. Wang, Y.; Liu, X.; Zhang, H. Analysis of Factors Influencing Taxi Traffic Accidents Considering Spatial Agglomeration Characteristics. *J. Chang. Univ. (Nat. Sci. Ed.)* **2023**, *43*, 99–106.
14. Yan, Y.; Wang, Y.; Zhou, X.; Yuan, H.; Wang, W. Analysis of Factors Influencing the Severity of Autonomous Vehicle Accidents Based on Hybrid Models. *J. Transp. Eng.* **2025**, *25*, 184–196.
15. Zhang, Z.; Yu, X.; Zhu, L.; Sun, Y.; Li, Y. Identification of Key Causes of Traffic Accidents in Autonomous Vehicles Based on CTGAN. *J. South China Univ. Technol. (Nat. Sci. Ed.)* **2025**, *53*, 1–15.
16. Gabaire, M.; Ghomi, H.; Hussein, M. Investigating the contributing factors to autonomous Vehicle-Road user Conflicts: A Data-Driven approach. *Accid. Anal. Prev.* **2025**, *211*, 107898. [CrossRef]
17. Liu, Q.; Wang, X.; Liu, S.; Yu, C.; Glaser, Y. Analysis of pre-crash scenarios and contributing factors for autonomous vehicle crashes at intersections. *Accid. Anal. Prev.* **2024**, *195*, 107383. [CrossRef]
18. Chen, Z.; Xu, W. Analysis of Factors Influencing Highway Traffic Accidents Based on Negative Binomial Distribution. *Traffic Inf. Saf.* **2022**, *40*, 28–35.
19. Ma, C.; Zhang, S.; Ma, Z.; Zhang, Y. Nonlinear negative binomial prediction model for highway traffic accidents. *Chin. J. Highw.* **2018**, *31*, 176–185.
20. Khedher, M.B.B.; Yun, D. Generalized linear models to identify the impact of road geometric design features on crash frequency in rural roads. *KSCE J. Civ. Eng.* **2022**, *26*, 1388–1395. [CrossRef]
21. Zhang, S.; Peng, L.; Huang, J.; Wang, F.; Xu, J. Analysis of Characteristics and Factors Influencing the Severity of Traffic Accidents on Highways in Difficult Mountainous Areas. *Highw. Automob. Transp.* **2024**, *40*, 40–48.
22. Ke, X.; Ding, L.; Zhao, D. Analysis of Factors Influencing the Severity of Electric Bicycle Traffic Accidents Based on Logistic TAN. *J. People's Public Secur. Univ. China (Nat. Sci. Ed.)* **2023**, *29*, 47–54.
23. Wen, H.; Huang, J.; Huang, K.; Zhao, S.; Chen, Z.; Hu, Y. Analysis of Factors Influencing Highway Accidents Integrating Short term Traffic Flow. *J. South China Univ. Technol. (Nat. Sci. Ed.)* **2025**, *53*, 1–13. [CrossRef]
24. He, J. Analysis of Spatial Distribution Characteristics and Influencing Factors of Road Traffic Accidents. Master's Thesis, People's Public Security University of China, Beijing, China, 2024.
25. Hernández, V. A spatial analysis of risk factors associated with road collisions in Ciudad Juarez, Mexico and using a geographically weighted regression approach. *Appl. Geogr.* **2024**, *166*, 103268. [CrossRef]
26. Wang, D.; Chen, C.; Xia, L.; Li, T.; Hao, J.; Xia, Y. Using spatial autoregressive model to analyze the impact of altitude and population density on the tuberculosis epidemic in Sichuan Province. *J. Prev. Med. Inf.* **2022**, *38*, 1535–1540.
27. Guo, R.; Chen, Y.; Zhang, Y.; Yi, F. Analysis of factors contributing to severity of AV crashes at intersections: Insights from the Autonomous Vehicle Operation Incident Dataset Across The Globe (AVOID). *Traffic Inj. Prev.* **2025**, *27*, 420–427. [CrossRef]

28. Wang, S.; Li, Z. Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. *PLoS ONE* **2019**, *14*, e0214550. [CrossRef]
29. Ren, Q.; Xu, M. Heterogeneity in crash patterns of autonomous vehicles: The latent class analysis coupled with multinomial logit model. *Accid. Anal. Prev.* **2025**, *209*, 107827. [CrossRef]
30. Wang, D.; Ding, S.; Abdel-Aty, M.; Barbour, N.; Wang, Z.; Zheng, O. Exploratory analysis of injury severity under different levels of driving automation (SAE Levels 2 and 4) using multi-source data. *Accid. Anal. Prev.* **2024**, *206*, 107692. [CrossRef]
31. Lee, J.; Abdel-Aty, M.; Jiang, X. Development of zone system for macro-level traffic safety analysis. *J. Transp. Geogr.* **2014**, *38*, 13–21. [CrossRef]
32. Pulugurtha, S.S.; Duddu, V.R.; Kotagiri, Y. Traffic analysis zone level crash estimation models based on land use characteristics. *Accid. Anal. Prev.* **2013**, *50*, 678–687. [CrossRef]
33. Zhang, Y.; Wang, X.; Zeng, P.; Chen, X. Centrality characteristics of road network patterns of traffic analysis zones. *Transp. Res. Rec.* **2011**, *2256*, 16–24. [CrossRef]
34. Chen, Y.; Zhang, Z.; Liang, T. Assessing urban travel patterns: An analysis of traffic analysis zone-based mobility patterns. *Sustainability* **2019**, *11*, 5452. [CrossRef]
35. DeMers, M.N. *Fundamentals of Geographic Information Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
36. Zou, Y.; Zhang, Y.; Lord, D. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accid. Anal. Prev.* **2013**, *50*, 1042–1051. [CrossRef]
37. Liu, B.; Li, C.; Hu, J.; Wang, P. A review of the application of regression models in traffic accident and driving risk analysis. *Traffic Transp.* **2024**, *40*, 68–74.
38. He, J.; Fu, Y.; Xiong, J. Research on Small Field Estimation Method Based on Hierarchical Bayesian Generalized Linear Model. *Math. Stat. Manag.* **2019**, *38*, 247–260.
39. Dong, G.; Harris, R. Spatial autoregressive models for geographically hierarchical data structures. *Geogr. Anal.* **2015**, *47*, 173–191. [CrossRef]
40. Brunsdon, C.; Fotheringham, S.; Charlton, M. Geographically weighted regression. *J. R. Stat. Soc. Ser. D (Stat.)* **1998**, *47*, 431–443. [CrossRef]
41. Tiefelsdorf, M.; Boots, B. The exact distribution of Moran's I. *Environ. Plan. A* **1995**, *27*, 985–999. [CrossRef]
42. Zhuang, Y. Multi Source Data-Driven Modeling of Spatiotemporal Distribution Characteristics and Severity of Electric Vehicle Accidents. Ph.D. Thesis, Beijing Jiaotong University, Beijing, China, 2024.
43. Wu, G.; Wen, J.; Pan, J. Research on Safety Evaluation and Improvement of Mountainous Trunk Road Intersections. *Highway* **2021**, *66*, 201–207.
44. Wang, J.; Sun, A.; Sun, X.; Luo, W.; Wu, Y. The spatiotemporal effects of factors affecting the severity of bicycle accidents. *J. Beijing Univ. Technol.* **2026**, *52*, 915–926. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.