Transactions in GIS

WILEY

Transactions in GIS

RESEARCH ARTICLE

gdverse: An R Package for Spatial Stratified Heterogeneity Family

Wenbo Lv[1,2] | Yangyang Lei[3] | Fangmei Liu[1] | Jianwu Yan[1] | Yongze Song[4] | Wufan Zhao[2]

[1]School of Geography and Tourism, Shaanxi Normal University, Xi'an, China | [2]Urban Governance and Design Thrust, Society Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China | [3]School of Architecture and Design, Harbin Institute of Technology, Harbin, China | [4]School of Design and the Built Environment, Curtin University, Perth, Australia

Correspondence: Wufan Zhao (wufanzhao@hkust-gz.edu.cn)

Received: 20 November 2024 | **Revised:** 12 March 2025 | **Accepted:** 14 March 2025

Funding: This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (42401567) and Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau (2024312159).

Keywords: geographical detector | R package *gdverse* | spatial associations detection | spatial factors exploration | spatial stratified heterogeneity

ABSTRACT

Spatial stratified heterogeneity (SSH) is a prevalent characteristic of geospatial data, which can be effectively modeled and analyzed using the geographical detector and other models from the SSH family. Various related models have been developed, focusing primarily on spatial data discretization, addressing spatial dependence, and capturing complex spatial interactions. Moreover, models incorporating discrete dependent variables have also emerged. These diverse models significantly enhance the ability to analyze and model SSH. However, the lack of a comprehensive and user-friendly software tool has greatly limited their broader application in geospatial analysis and environmental modeling. To address this gap, an R package *gdverse* has been developed to integrate various SSH models, leveraging R's rich statistical and spatial data processing capabilities while natively supporting multicore parallel computing in the widely used R environment. A case study on the determinants of trace element Zn demonstrates the application of the *gdverse* package, showcasing its effectiveness and convenience.

1 | Introduction

Spatial stratified heterogeneity (SSH) refers to the geographical phenomenon in which attributes within the same stratum exhibit greater similarity to each other than to those in other strata. Widely observed in the natural world, this phenomenon provides valuable insights into the underlying drivers of natural processes, highlighting that the natural world is not entirely random (Wang et al. 2016, 2024; Guo et al. 2022). Building upon this concept, Wang et al. (2010) employed the sum of squares and introduced the q value to assess the similarity within and between strata, a key innovation that led to the development of the GeoDetector model (for detailed information, see Section 2.1). As shown in Figure 1, by grounding the model in the q value, GeoDetector became the inaugural model in the SSH family, establishing a solid foundation for the continued development and expansion of the SSH family to address a variety of modeling scenarios.

The SSH family can be systematically categorized into five groups based on the modeling challenges they can address. First, in response to the challenges of discretizing continuous independent variables, the Optimal Parameters-based Geographical Detector (OPGD), the Geographically Optimal Zones-based Heterogeneity Model (GOZH), and the Robust Geographical Detector (RGD) were developed (Song et al. 2020; Luo et al. 2022; Zhang et al. 2022). Discretization methods tailored for specific data distributions, such as heavy-tailed distributions (Hu et al. 2024), have also been integrated with GeoDetector (this aspect will not be elaborated upon in the present study). Second, to address the difficulties in effectively characterizing relationships within and between strata in spatial data using the sum of squares, the

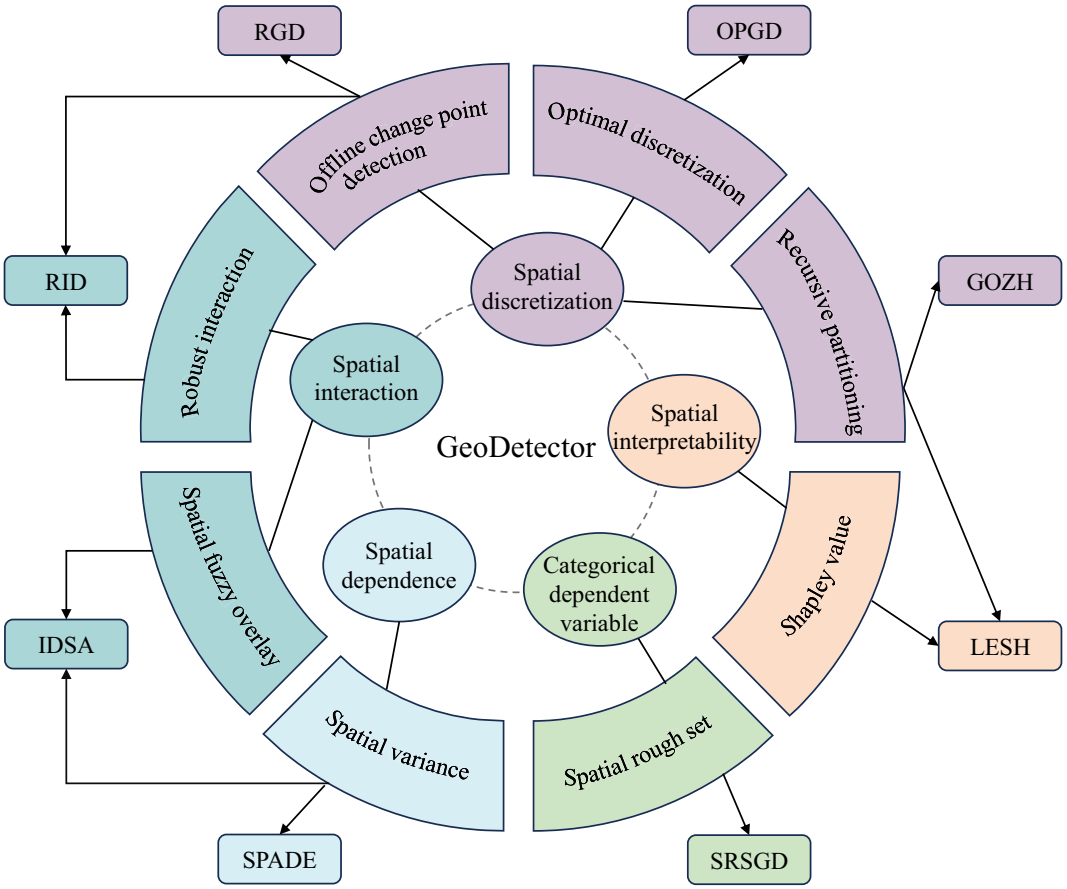

FIGURE 1 | The development of the geodetector model [21] into the spatial stratified heterogeneity (SSH) family.

Spatial Association Detector (SPADE) was introduced (Cang and Luo 2018). Third, to tackle the challenge of stratifying interaction variables in interaction detection, the Interactive Detector for Spatial Associations (IDSA) and the Robust Interaction Detector (RID), based on the RGD model, were developed (Song and Wu 2021; Zhang et al. 2022, 2024). Fourth, to enhance the interpretability of interaction detection results, the Locally Explained Stratified Heterogeneity Model (LESH) was proposed, building upon the GOZH model (Li et al. 2023; Luo et al. 2022). Finally, for scenarios involving discrete dependent variables, Bai et al. (2022) introduced the SRSGD model. In aggregate, these models have strengthened SSH modeling across a variety of contexts, forming a comprehensive SSH family (refer to Table 1 for a thorough exploration of the theoretical foundations underpinning the members of the SSH family).

As previously outlined, the q value provided both the theoretical and computational foundation, while various models from the SSH family served as adaptable variants tailored to different scenarios. Throughout the evolution of these models, issues such as the need for prior knowledge of discretization, the neglect of inherent autocorrelation characteristics in spatial data, and the potential for excessive stratification have been addressed, resulting in reduced bias, enhanced reliability, interpretability, and robustness of the models. This progress has maximized the potential of the SSH family for SSH analysis. However, despite the development of tools such as the GeoDetector Excel add-in, the GeoDetector ArcGIS plugin, the GeoDetector QGIS plugin, the *geodetector* R package, the *GD* R package, and the *idsa* R package (Wang et al. 2024, 2010, 2016; Song et al. 2020; Song and Wu 2021), they still face challenges related to limited usability and slow computational performance. For example, the GeoDetector toolkit requires users to manually discretize variables using external tools; the *GD* package, written entirely in R without parallel processing support, suffers from slow computation when a large number of discretization methods and categories are selected; similarly, the *idsa* package, also implemented in pure R, is hindered by the high computational complexity of the SPADE and IDSA models.

In this instance, the effective application of models in the SSH family in geospatial analysis and environmental modeling has been prevented. To address these challenges, we have developed an R package, *gdverse*, which integrates all the aforementioned models, ensuring fast and stable performance to enable wider application. Three primary advantages are presented under the development of *gdverse*: First, it supports all aforementioned models in the SSH family, providing a unified and streamlined API for users. With preprocessing completed, model analyses can be executed via a single function call. Second, *gdverse* integrates comprehensively with the R ecosystem (R Core Team 2024), allowing for efficient tabular data manipulation and cleaning through the *tidyverse* package (Wickham et al. 2019), spatial vector data handling with the *sf* package (Pebesma 2018), and spatial raster data processing via the *terra* package (Hijmans 2024). Third, the *gdverse* package provides robust parallel computing support, user-friendly output formatting, and a one-run function for effective result visualization by the *ggplot2* package (Wickham 2016).

TABLE 1 | Models in spatial stratified heterogeneity (SSH) family.

Model	Description
GeoDetector (Wang et al. 2010, 2016)	GeoDetetor is the primary model using the equation of the q value to calculate the spatial stratified heterogeneity
OPGD (Song et al. 2020)	OPGD (Optimal Parameters-based Geographical Detector) can identify the most suitable discretization approach and the optimal number of discrete categories by maximizing the q value for each variable
GOZH (Luo et al. 2022)	GOZH (Geographically Optimal Zones-based Heterogeneity Model) employs recursive partitioning trees to discretize continuous variables
RGD (Zhang et al. 2022)	RGD (Robust Geographical Detector) can utilize a variance-based change point detection optimization algorithm after converting variables into sorted rank series to get stable results
SPADE (Cang and Luo 2018)	The concept of spatial variance, derived from spatial dependence as expressed in spatial autocorrelation analysis, is introduced to SPADE (Spatial Association Detector) to incorporate information loss across different levels of discretization
IDSA (Song and Wu 2021)	IDSA (Interactive Detector for Spatial Associations) develops the concept of spatial fuzzy overlay, combining it with the spatial variance calculation from the SPADE model
RID (Zhang et al. 2024)	RID (Robust Interaction Detector) is to maximize the q value for bivariable interactions based on the RGD model
LESH (Li et al. 2023)	LESH (Locally Explained Stratified Heterogeneity Model) integrates Shapley values from game theory to allocate the contribution of each individual variable within an interaction more precisely, building on the GOZH model
SRSGD (Bai et al. 2022)	SRSGD (Spatial Rough Set-based Geographical Detector) employs the concept of spatial rough sets to model nominal target variables
HtGD (Hu et al. 2024)	HtGD (Head/tail Breaks-based Geographical Detector) utilizes the head/tail breaks method to efficiently calibrate spatial stratified heterogeneity for heavy-tailed distributed data

We organized the remainder of this paper in the following manner: Section 2 elaborates on the theoretical foundations of the models in the SSH family integrated into the *gdverse* package. Section 3 offers a comprehensive overview of the methods and functions available within *gdverse*. Section 4 illustrates a practical application of *gdverse*, and Section 5 concludes the study.

2 | Model Rationale Description

As discussed in Section 1, the OPGD, GOZH, and RGD models discretize independent variables using different methods: maximizing the q value of the variables, employing recursive partitioning trees, and detecting change points based on the variance of the independent variable's rank series, respectively. The RID model, on the other hand, focuses on maximizing the bivariate interaction q value. As none of these four models has an explicit computational formula and they are all highly reliant on the native geographical detector (GeoDetector) model, we will provide a detailed description of the computational methods of the other models in this section, without reiterating the principles of these four models.

2.1 | (Native) Geographical Detector (GeoDetector)

2.1.1 | Factor Detector

In the GeoDetector model, the formula for calculating the q value is presented below (Wang et al. 2010):

$$q = 1 - \frac{\sum_{l=1}^{H} N_l \sigma_l^2}{N\sigma^2} \tag{1}$$

and the significance of the q values calculated by the factor detector can be tested using a noncentral F-distribution (Wang et al. 2016):

$$F = \frac{N-H}{H-1}\frac{q}{1-q} \sim F(H-1, N-H, \varepsilon) \tag{2}$$

$$\varepsilon = \frac{1}{\sigma^2}\left[\sum_{l=1}^{H} \overline{Y}_l^2 - \frac{1}{N}\left(\sum_{l=1}^{H} \sqrt{N_l}\overline{Y}_l^2\right)^2\right] \tag{3}$$

where l represents the strata of the independent variable; N_l is the sample size within stratum l; N represents the total sample size for all strata; σ_l^2 and σ^2 are the variances of the dependent variable in stratum l and all strata, respectively; ε is the noncentral parameter; $\overline{Y}_l^2$ denotes the mean of the dependent variable within stratum l.

2.1.2 | Interaction Detector

To identify interactions between independent variables, we assess whether the combined influence of X_c and X_d on the dependent variable Y enhances, weakens, or remains independent in its explanatory power, calculating each variable's individual q value for its effect on Y: $q(X_c)$ and $q(X_d)$ involved. Subsequently, we calculate the q value for their combined spatial stratification, $q(X_c \cap X_d)$, and contrast it with $q(X_c)$ and $q(X_d)$, revealing the

TABLE 2 | Interaction categories of interaction detector in the GeoDetector model.

Interaction condition	Interaction type
$q(X_c \cap X_d) < \min\left(q(X_c), q(X_d)\right)$	Nonlinear-weaken
$\min\left(q(X_c), q(X_d)\right) < q(X_c \cap X_d) < \max\left(q(X_c), q(X_d)\right)$	Univariable-weaken
$q(X_c \cap X_d) > \max\left(q(X_c), q(X_d)\right)$	Bivariable-enhance
$q(X_c \cap X_d) = q(X_c) + q(X_d)$	Independent
$q(X_c \cap X_d) > q(X_c) + q(X_d)$	Nonlinear-enhance

nature of their interaction effect on Y (Wang et al. 2010, 2016). Results of all interaction detection are provided in Table 2.

2.1.3 | Risk Detector

Use the t-statistic to test whether there is a significant variation in the mean attributes of a certain independent variable across two sub-strata (Wang et al. 2010, 2016):

$$t_{\bar{y}_{l_1}-\bar{y}_{l_2}} = \frac{\overline{Y}_{l_1} - \overline{Y}_{l_2}}{\left[\frac{\mathrm{Var}\left(\overline{Y}_{l_1}\right)}{n_{l_1}} + \frac{\mathrm{Var}\left(\overline{Y}_{l_2}\right)}{n_{l_2}}\right]^{1/2}} \tag{4}$$

The test statistic t approximately follows a Student's t distribution, while the degrees of freedom is computed as

$$df = \frac{\frac{\mathrm{Var}\left(\overline{Y}_{l_1}\right)}{n_{l_1}} + \frac{\mathrm{Var}\left(\overline{Y}_{l_2}\right)}{n_{l_2}}}{\frac{1}{n_{l_1}-1}\left[\frac{\mathrm{Var}\left(\overline{Y}_{l_1}\right)}{n_{l_1}}\right]^2 + \frac{1}{n_{l_2}-1}\left[\frac{\mathrm{Var}\left(\overline{Y}_{l_2}\right)}{n_{l_2}}\right]^2} \tag{5}$$

2.1.4 | Ecological Detector

To evaluate the diverse effects of two independent variables, X_c and X_d, on the spatial distribution of the dependent variable Y while assessing the presence of significant disparities, the F-statistic is used as a measure (Wang et al. 2010, 2016):

$$F = \frac{N_{X_c}\left(N_{X_d}-1\right)SSW_{X_c}}{N_{X_d}\left(N_{X_c}-1\right)SSW_{X_d}} \tag{6}$$

where N_{X_c} and N_{X_d} denote the sample sizes of the two independent variables X_c and X_d, respectively; SSW_{X_c} and SSW_{X_d} represent the sums of the within-layer variances formed by X_c and X_d.

In practical calculations, N_{X_c} and N_{X_d} are generally the same within the same dataset. Therefore, combined with Equation (1), Equation (6) can be simplified to

$$F = \frac{1 - q_{X_c}}{1 - q_{X_d}} \tag{7}$$

2.2 | Locally Explained Heterogeneity (LESH) Model

2.2.1 | The SHAP Power of Determinants (SPD)

The SHAP power of determinants (SPD) can be calculated by the following formula:

$$\theta_{x_i}(S) = \sum_{s \in U \setminus \{x_i\}} \frac{|S|!(|U|-|S|-1)!}{|U|!}\left(g\left(S \cup \{x_i\}\right) - g(S)\right) \tag{8}$$

where $\theta_{x_i}(S)$ refers to the SPD of variable x_i within the set U. Excluding the variable x_i, the subset S is derived from U, which can be signified in the notation $s \in U \setminus \{x_i\}$; $g(S)$ denotes the function employed to compute the Q statistic in relation to the interaction of $|S|$ variables, and $|S|$ indicates the quantity of variables within the set S (Li et al. 2023). In LESH model, the $g(S)$ is utilized by the GOZH model.

2.2.2 | Calculation Process of the LESH Model

Two main steps are involved in the computational process of the LESH model: First, calculating the optimal power of determinants (OPD) using the GOZH model to evaluate the spatial associations between the response variable and each explanatory variable. Second, determining the SHAP power of determinants (SPD) using Equation (8) to assess the individual contributions of each variable and their interactions (Li et al. 2023).

2.3 | Spatial Association Detector (SPADE)

2.3.1 | Spatial Variance

Compared to the GeoDetector model, the SPADE model has developed spatial variance (Cang and Luo 2018):

$$\Gamma = \frac{\sum_u \sum_{v \neq u} \omega_{uv} \frac{(y_u - y_v)^2}{2}}{\sum_u \sum_{v \neq u} \omega_{uv}} \tag{9}$$

where ω_{uv} is the weight between uth location and vth location; y_u and y_v are the dependent variable values at the uth and vth locations, respectively.

2.3.2 | The Power of Spatial and Multilevel Discretization Determinant (PSMD)

As the proportion of the local spatial sum of squares and the global one, the power of spatial determinant (PSD) is calculated as below (Cang and Luo 2018):

$$q_s = 1 - \frac{\sum_{l=1}^{H} N_l \Gamma_l}{N\Gamma} \tag{10}$$

where N_l is a count of samples in l th stratum; Γ_l is the spatial variance of dependent variable within stratum l; H is the total strata number; N is the total sample number, and Γ is the total spatial variance of the dependent variable.

As the ratio of the PSD of the dependent variable and the PSD of the undiscretized independent variable, the compensated power of spatial discretization determinant (CPSD) is calculated according to the following formula (Cang and Luo 2018):

$$Q_s = \frac{q_s}{q_{s_{inforkep}}} = \frac{1 - \frac{\sum_{l=1}^{H} N_l \Gamma_{l_{dep}}}{N\Gamma_{total_{dep}}}}{1 - \frac{\sum_{l=1}^{H} N_l \Gamma_{l_{ind}}}{N\Gamma_{total_{ind}}}} \tag{11}$$

The power of spatial and multilevel discretization determinant (PSMD) is the mean of CPSD values, as defined by Equation (11), across all discretization levels:

$$PSMD_{Q_s} = MEAN(Q_s) \tag{12}$$

2.3.3 | Calculation Process of the SPADE Model

Three steps make up the execution process of the SPADE model: First, for each independent variable, a specific number of discretization categories and various discretization methods are selected, along with the construction of a spatial weight matrix. Second, Equations (9–12) are applied to calculate the PSMD values for each independent variable. Finally, the original data is randomly shuffled, and the second step is repeated 99,999,9999, ... times to obtain the corresponding p values (Cang and Luo 2018).

2.4 | Interactive Detector for Spatial Associations (IDSA)

2.4.1 | Spatial Fuzzy Overlay

Spatial fuzzy overlay is grounded in the theory of fuzzy sets, which articulates the fuzzy relationship between geographical variables via fuzzy membership functions. Standardizing the observed values of these variables into fuzzy numbers, these functions thereby capture the spatial distribution of the geographical elements. In the IDSA model, fuzzy membership functions correspond to the normalized mean risk values obtained from the GeoDetector model's risk detection component (Song and Wu 2021):

$$\left[f_n(X_1)\ \ f_n(X_2)\ \ \ldots\ \ f_n(X_m)\right] = z\left(\left[\mu(X_1)\ \ \mu(X_2)\ \ \ldots\ \mu(X_m)\right]\right) \tag{13}$$

where X is an explanatory variable, f_n is the fuzzy membership function, μ is the mean risk score obtained by the risk detector, with z representing a normalization function.

The process of integrating fuzzy numbers, representing the fuzzy relations among variables, is employed to calculate the combined fuzzy number resulting from the interaction of these variables (Song and Wu 2021):

$$f_n(X_1 \cap X_2 \cap \ \ldots\ \cap X_m) = \mathrm{F}\left(f_n(X_1), f_n(X_2), \ \ldots\ , f_n(X_m)\right) \tag{14}$$

where F denotes a fuzzy operator. When F is defined as "and," the minimization of fuzzy numbers is employed; conversely, when F is defined as "or," the maximization of fuzzy numbers is utilized.

2.4.2 | The Power of Spatial Determinant of an Interaction of Explanatory Variables (PID-IEV)

As the proportion of the sum of local spatial variance in overlay zones for individual variables with interaction to the total global spatial variance of these variables, the value of PSD-IEV is calculated as below (Song and Wu 2021):

$$\phi_{pid} = 1 - \frac{\sum_{i=1}^{m} \sum_{p=1}^{n_i} N_{i,p} \tau_{i,p}}{\sum_{i=1}^{m} N_i \tau_i} \tag{15}$$

where $N_{i,p}$ is the sample size at stratum p for an individual variable x_i; $\tau_{i,p}$ is the local spatial variance; N_i is the total of samples; and τ_i is global spatial variance of variable x_i.

The PID-IEV is a proportion between the PSD value of dependent variables using Equation (10) and the PSD-IEV value for interaction independent variables using Equation (15):

$$Q_{IDSA} = \frac{\theta_{dep}}{\phi_{pid}} \tag{16}$$

Equations (1), (10), and (15) essentially differ only in the type of variance utilized in their calculations. In particular, Equation (1) employs conventional statistical variance, whereas Equations (10) and (15) make use of the spatial variance as defined in Equation (9). Equation (10), compared to Equation (1), incorporates the variance that accounts for the spatially weighted cross-product, introducing the consideration of spatial dependence (Cang and Luo 2018). Moreover, Equations (10) and (15) provide respective descriptions tailored to the SPADE and IDSA model.

2.4.3 | Calculation Process of the IDSA Model

The following steps are involved in the IDSA model: First, determine the optimal parameters for dividing spatial variables into distinct zones, guided by data quantity and practical requirements. Second, calculate the q values for each variable based on the SPADE model. Third, determine the optimal variable interactions through spatial fuzzy overlay (Equation 14) and calculate the PID (Equation 16), which quantifies the interaction effect of variables on the response. Finally, evaluate the model's effectiveness using indicators such as the quantity of variables, a count of overlay zones, and the percentage of significantly differentiated zone pairs.

2.5 | Spatial Rough Set-Based Geographical Detector (SRSGD)

2.5.1 | x-Local Approximation Quality (x-LAQ)

As the proportion of the correctly classified observations to the total observations in that region, the x-local approximation quality (x-LAQ) quantifies the local explanatory power of a

factor within a specific region and is calculated as follows (Bai et al. 2022):

$$\text{x-LAQ}_i = \frac{|\{x \in M_i : f(x) = y(x)\}|}{|M_i|} \tag{17}$$

where M_i denotes the set of instances in region i, $f(x)$ and $y(x)$ are the predicted value and the actual value, respectively.

2.5.2 | Average Local Explanatory Power (ALED)

Representing the mean explanatory power of a factor across all regions (Bai et al. 2022), the average local explanatory power (ALEP) is calculated as the average of the x-LAQ values across all regions:

$$ALEP = \frac{1}{m}\sum_{i=1}^{m} \text{x-LAQ}_i \tag{18}$$

where m represents the total count of regions.

2.5.3 | Spatial Entropy (SE)

Spatial entropy (SE) measures the uncertainty associated with the spatial distribution of a factor (Bai et al. 2022). It is defined as

$$SE = -\sum_{i=1}^{m} ALEP_i \log(ALEP_i) \tag{19}$$

2.5.4 | Three Geographical Detectors Based on Spatial Rough Set

To assess average local explanatory power while identifying spatial heterogeneity, the spatial rough set-based factor detector (SRSF Detector) is developed. Leveraging average local approximation quality and spatial entropy, it evaluates the extent to which conditional features account for the target variable across different locations. Revealing patterns in explanatory power distribution, the SRSF Detector identifies regions where model predictions are more or less reliable (Bai et al. 2022).

Focusing on the comparison of the average local explanatory power among different conditional features, the spatial rough set-based ecological detector (SRSE Detector) highlights which features are more influential in explaining the target variable by examining the differences in explanatory strength between feature pairs, which can aid in feature selection and in understanding the relative importance of factors shaping the geographical phenomenon (Bai et al. 2022).

The interactions among conditional features and any additional explanatory power contributed by new features can be recognized by the spatial rough set-based interaction detector (SRSI Detector). SRSI Detector evaluates the increase in explanatory strength with the addition of new features, helping to determine if they provide complementary insights or redundant information, essential for unraveling complex relationships among multiple factors and understanding their combined effect on the target variable (Bai et al. 2022).

3 | The *gdverse* Package

Implemented in the R statistical computing environment (R Core Team 2024), all computationally intensive functions within the *gdverse* package are using the R package *parallel* to enable multicore parallel computing for fast computation. The majority of functions in the *gdverse* package are implemented via the *tidyverse* package (Wickham et al. 2019), which is designed to facilitate efficient computation with tabular data. In addition, the *sf* package (Pebesma 2018) is employed to provide enhanced capabilities for handling spatial data. Moreover, certain methods for variable discretization and spatial variance calculations are implemented in C++ and wrapped using *Rcpp* (Eddelbuettel and François 2011) within the *sdsfun* package to ensure computational efficiency (see Table 3 for details on the performance advantages of *gdverse*). It is noteworthy that the RGD and RID models use the Python package *ruptures* (Truong et al. 2020) and the R package *reticulate* (Ushey et al. 2024) partly for variable discretization based on an optimization algorithm for variance-based change point detection (Zhang et al. 2022, 2024). Figure 2 shows a summary of the functions in *gdverse*. All function parameters of *gdverse* can be found at https://stscl.github.io/gdverse/reference/.

To demonstrate the performance advantages of the *gdverse* package, we used the *gstat* package to generate datasets of sizes 1k, 5k, 10k, 50k, and 100k through unconditional Gaussian simulation (Pebesma 2004; Gräler et al. 2016). Each dataset includes one dependent variable and six independent variables. On an Intel Core i9-13900K Processor, we conducted 10 benchmark runs for each dataset to compare the performance of the *GD* package and the *gdverse* package (including scenarios using a single-core computation and a 6-core parallel computation) in fitting the OPGD model. Table 3 presents the results, demonstrating that while *gdverse* achieves the same computational outcomes as *GD*, it generally operates faster, with its performance advantage becoming more pronounced as the dataset size increases. When the dataset size is 1k, the execution time of *gdverse* is slightly higher than that of *GD* due to internal data structure conversions, but the performance difference is minimal. For datasets exceeding 10k, *gdverse* still maintains an advantage in single-core computation, with 6-core parallel computation offering even faster speeds. Users can select the appropriate cores for larger datasets to accelerate computation.

For models discussed in Section 1, the *gdverse* package offers functions with identical names but in lowercase (see Table 4). In addition, it provides "sesu_opgd()" and "sesu_gozh()" functions based on the OPGD model and GOZH model, respectively, designed to determine the optimal spatial analysis scale using models from the SSH family. A "cores" parameter is incorporated into all of these functions to specify the number of cores to utilize for parallel computing. By default, parallel computing is disabled, with the default setting of "cores" typically set to 1. When processing large data sets, it is recommended to select an appropriate number of computing cores

TABLE 3 | Comparison of the runtime performance of the OPGD model between the *GD* package and the *gdverse* package (single-core computation and parallel computation using six cores). The metrics—mean, median, minimum, and maximum—represent the average, median, shortest, and longest execution times (s) based on 10 benchmark runs across different data sizes.

		Execution times (s)			
Data size	**Package**	**Mean**	**Median**	**Min**	**Max**
1 k	*GD*	0.71	0.70	0.68	0.80
1 k	*gdverse* (1 core)	1.44	1.44	1.42	1.52
1k	*gdverse* (6 cores)	2.24	2.23	2.19	2.36
5k	*GD*	3.65	3.63	3.58	3.74
5k	*gdverse* (1 core)	1.59	1.58	1.54	1.68
5k	*gdverse* (6 cores)	2.32	2.31	2.27	2.41
10k	*GD*	8.94	8.92	8.77	9.19
10k	*gdverse* (1 core)	1.86	1.89	1.72	1.93
10k	*gdverse* (6 cores)	2.31	2.29	2.26	2.43
50k	*GD*	145.31	145.30	144.48	146.18
50k	*gdverse* (1 core)	7.52	7.51	7.34	7.81
50k	*gdverse* (6 cores)	5.79	5.78	5.72	5.87
100k	*GD*	604.01	604.09	601.26	606.80
100k	*gdverse* (1 core)	22.86	22.87	22.46	23.19
100k	*gdverse* (6 cores)	9.56	9.52	9.46	9.77

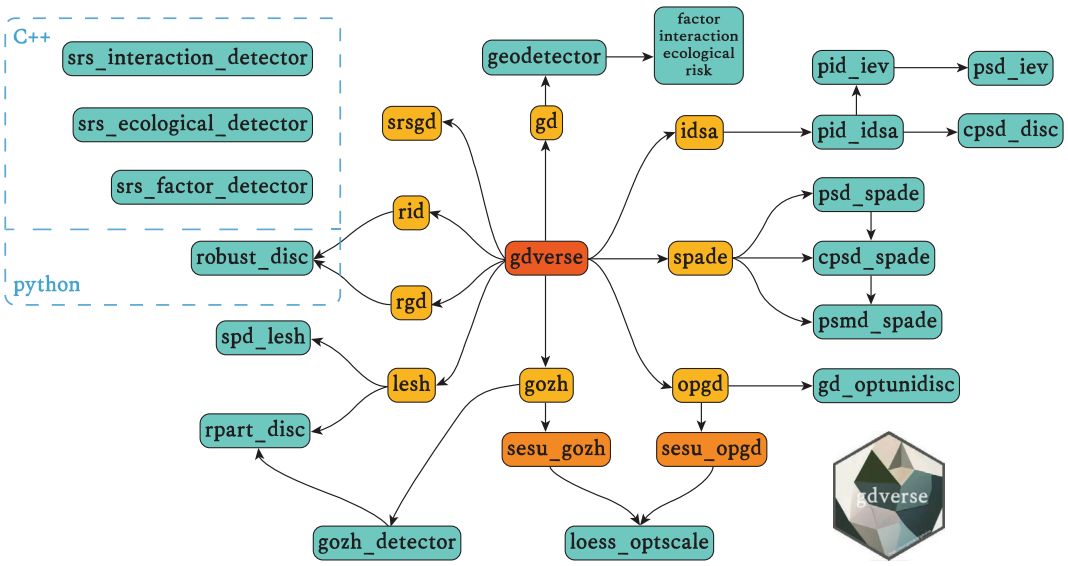

FIGURE 2 | Functions organization in the *gdverse* package.

in accordance with the specifications of your own computer equipment.

In the *gdverse* package, data input can be provided as a *data.frame* or *tibble* in R, including *data.frame* and *tibble* extensions from the *sf* package (R Core Team 2024; Wickham et al. 2019; Pebesma 2018). For other data types, explicit conversion to these formats is required before performing the analysis (the main input parameters required by the key functions in *gdverse* are listed in Table 4). Model outputs in *gdverse* are structured as R *list* objects, where each *list* component is either a *tibble* or a numerical metric, facilitating easy result saving or export. We have developed a comprehensive set of predefined plotting functions with *ggplot2* to support the

TABLE 4 | Spatially stratified heterogeneity (SSH) family supported by *gdverse* package.

Model	Function	Requirements of input
GeoDetector	`geodetector()`	Model formula, data
OPGD	`opgd()`	Model formula, data, discretization numbers and methods
GOZH	`gozh()`	Model formula, data
LESH	`lesh()`	Model formula, data
SPADE	`spade()`	Model formula, data, discretization numbers and methods, spatial weight matrices
IDSA	`idsa()`	Model formula, data, discretization numbers and methods, spatial weight matrices, overlay method
RGD	`rgd()`	Model formula, data, discretization numbers
RID	`rid()`	Model formula, data, discretization numbers
SRSGD	`srsgd()`	Model formula, data, spatial weight matrices

visualization of model results. Users can further customize these visualizations using the *ggplot2* syntax (Wickham 2016). In addition, to help users quickly get started, it is important to highlight the differences in input and output formats between *gdverse* and two key R packages in the SSH: *GD* and *geodetector*. Neither *GD* nor *geodetector* directly supports the widely used spatial vector data objects in the *sf* format in R (Pebesma 2018); both accept only *data.frame* or *tibble* as input. The *geodetector* package focuses on the GeoDetector model (Wang et al. 2010, 2016, 2024) and does not provide support for result visualization. In contrast, the *GD* package uses the base R plotting system to offer basic visualization support (R Core Team 2024), though its customization options are less user-friendly and flexible compared to the *ggplot2*-based implementation in *gdverse* (Wickham 2016). Moreover, the *GD* package supports only the GeoDetector and OPGD models, offering fewer modeling options than *gdverse* (Song et al. 2020). In addition to its highly encapsulated model functions, *gdverse* offers a more granular API for customized model execution. To address computationally intensive algorithms, *gdverse* leverages C++ implementations via *Rcpp* interfaces in the *sdsfun* package, including functions like variable discretization (Eddelbuettel and François 2011; Lv 2024). Overall, the *gdverse* API is scientifically structured and user-friendly, ensuring seamless support for data input and result output. Given its handling of large data requirements, *gdverse* also enables multicore parallel computing, guaranteeing efficient, simple, and reliable operation.

4 | Application on the Determinants of Trace Element zn

Using the zn dataset from the *geosimilarity* package (Song 2022), we demonstrate the application of the *gdverse* package and perform a comparative analysis of the SSH family it encompasses. As discussed in Section 1, SRSGD is designed for scenarios involving a discrete dependent variable; IDSA facilitates the determination of optimal interactions while accounting for spatial autocorrelation; LESH is suited for quantifying the contributions of dual roles in interaction detection processes; and RID employs a robust discrete methodology to identify the nature of interactions. Given the specific contexts and distinct requirements of these methods, an extensive comparison is not essential, as their application domains are well defined. To further illustrate, we conducted a comparative analysis of four models—namely OPGD, GOZH, RGD, and SPADE—that differ in their approaches to discretizing continuous independent variables and calculating the *q* value, using the following code:

```
zn = sf::st_as_sf(geosimilarity::zn,coords =
c("Lon","Lat"), crs=4326)
# detect the number of available physical CPU
cores
cores = parallel::detectCores(logical =
FALSE)
# OPGD
opgd_m = gdverse::opgd(Zn ~., data = zn,
discnum=3:15, cores = cores)
# GOZH
gozh_m = gdverse::gozh(Zn ~., data = zn,
cores = cores)
# RGD
rgd_m = gdverse::rgd(Zn ~., data = zn,
discnum=3:15, cores = cores)
# SPADE
spade_m = gdverse::spade(Zn ~., data = zn,
permutations=999, discnum=3:15, cores =
cores)
```

A concise introduction to the four methods is necessary to highlight their differences in application and offer guidance for users who have not yet made a definitive model choice. OPGD, grounded in the direct observation of samples, is often employed as a baseline method and for evaluating the robustness of alternative approaches, as it does not always achieve the maximum *q* value (Zhang et al. 2022). GOZH, which incorporates recursive partitioning trees in the discretization process, is particularly well-suited for large datasets due to its computational efficiency. RGD, while yielding superior results, is computationally intensive and thus unsuitable for large datasets. In particular, the robust discretization approach used in RGD may produce a high *q* value, occasionally exceeding the true value. SPADE, on the other hand, is advantageous in scenarios where the response variable exhibits moderate to strong spatial autocorrelation (Cang and Luo 2018; Zhang et al. 2023).

The comparison of the results for each method in our application is presented in Table 5, where variables marked with a

superscript "a" failed to pass the significance test. Zn refers to the trace element zinc. "Mine" serves as an indicator of proximity to mining sites. NDVI, the normalized difference vegetation index, reflects the status of vegetation. "SOC" is used to represent the soil organic carbon content. "Slope" indicates terrain steepness, while pH measures the acidity of the soil. "Road" denotes the distance to roads, and "Water" represents the distance to water bodies. "Elevation" corresponds to the height above sea level, and "Aspect" denotes the compass direction that a slope faces (Song 2022). As the Moran's I for Zn in the dataset is only 0.230 (with a p value of 1.898e-32), indicating weak spatial autocorrelation, it does not require specific consideration. On the other hand, compared to the results of the other three models, the SPADE model did not yield all statistically significant results, as the "Elevation" variable failed to pass the significance test, making SPADE unsuitable in this context. The results from the RGD model reveal remarkably high q values for "Elevation" and "Aspect." The inflated q values suggest that the RGD model may not be optimal for this application. Given the consistency in the magnitude of q values calculated by the GOZH and OPGD models, we chose to proceed with GOZH for further analysis.

TABLE 5 | Q values and their corresponding significance for variables in the zn dataset from the *geosimilarity* package, calculated using methods from the *gdverse* package. Columns 2–5 are labeled with the names of the methods and display the q value and significance for each variable.

Variable	OPGD	GOZH	RGD	SPADE
Mine	0.246	0.221	0.227	0.184
NDVI	0.197	0.228	0.279	0.212
SOC	0.143	0.244	0.223	0.240
Slope	0.128	0.171	0.255	0.145
pH	0.118	0.173	0.198	0.155
Road	0.111	0.121	0.131	0.258
Water	0.110	0.128	0.189	0.240
Elevation	0.059	0.110	0.202	0.117[a]
Aspect	0.052	0.071	0.131	0.251

[a]Failed to pass the significance test.

Insights into the determinants of Zn concentration are provided by analyzing the q values of the determinants and variable interactions. The q values of the determinants (factor detector, see Figure 3) reveal that SOC (organic carbon content of the soil), with a value of 0.244, is the most influential factor among the nine variables considered. This is consistent with the understanding that Zn is a component of soil organic matter, and soils with higher SOC tend to exhibit greater heavy metal content. NDVI and Mine (distance to mining sites) follow, with q values of 0.228 and 0.221, respectively. Topographical variables, with the exception of Slope, show lower q values, while variables associated with human activities have a more substantial q value, suggesting that human activities or their indirect effects may have a significant impact on Zn concentration. To capture variable interactions, the LESH model was employed, as it not only identifies interactions between variable pairs but also introduces the SHAP values to quantify the respective contributions of each variable involved in the interaction. The following code demonstrates how to easily implement the LESH model and visualize the resulting output by the *gdverse* package.

```
# LESH
lesh_m = gdverse::lesh(Zn ~., data = zn,
cores = cores)
plot(lesh_m, pie = TRUE, scatter = TRUE)
```

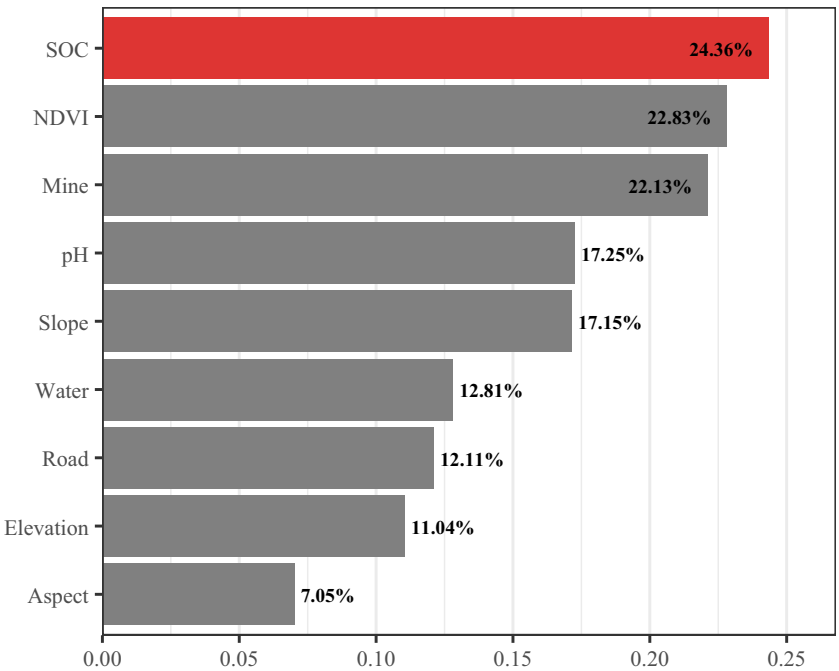

FIGURE 3 | Q values of the determinants for Zn concentration as identified by the GOZH model.

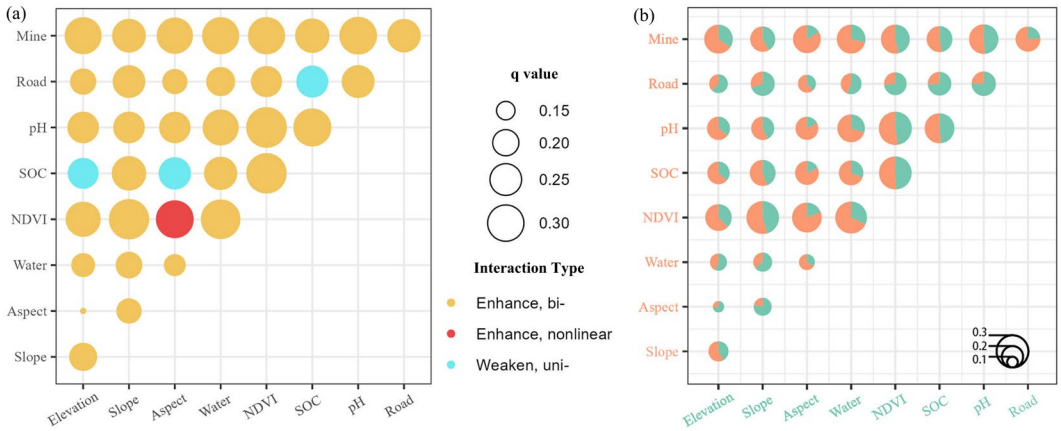

FIGURE 4 | Contribution of variable interactions to Zn concentration. (a) An overview for the identification of interaction types, while (b) offers a clearer visualization of the individual contributions of variables within each interaction.

The results of the variable interactions underscore bivariable relationships among the variable pairs. Mine and NDVI exhibit strong synergistic effects, as indicated by the larger bubble sizes in the interaction detector (see Figure 4). In particular, the interaction between NDVI and Aspect shows a nonlinear enhancement, with the q value of the interaction surpassing the sum of their individual q values. Specifically, the q value for the interaction between Aspect and NDVI is 0.304, with Aspect contributing 0.056 and NDVI contributing 0.248. When considered separately, the q values for Aspect and NDVI are 0.071 and 0.228, respectively. In addition, SOC exhibits a weakening effect when interacting with factors such as Road, Elevation, and Aspect, highlighting the complex and interdependent nature of these environmental variables.

5 | Conclusion

The use of geographical detector models is becoming increasingly prevalent, largely due to the simplicity and elegance of their mathematical principles, as well as their explanatory power. Under this circumstance, we developed an R package that integrates various related models in the SSH family, addressing the gap in application tools for modeling SSH. A brief comparison of the models supported by *gdverse* and some modeling results was provided in Section 4. Further examples of *gdverse* package usage are provided in the vignettes constructed within this package, which can be found at https://stscl.github.io/gdverse/articles/. Drawing upon the features of the models incorporated in *gdverse* and prior studies centered on SSH models, we have identified two typical scenarios for *gdverse* applications:

One is the analysis of spatial influence factors and the identification of spatial interactions based on SSH, also the most typical example of the application of the SSH family, such as the GeoDetector, OPGD, SPADE, and IDSA model. By applying factor detector and interaction detector from the GeoDetector model, along with some derived improved models, it is possible to effectively explore the spatial relationships of complex spatial dependence and nonlinear spatial interactions between variables, while obtaining relatively suitable spatial interpretability (Wang et al. 2010, 2024; Song et al. 2020).

The other involves estimating the scale effects of spatial units and determining the optimal scale for spatial analysis. By calculating the q values across various models from the SSH family at multiple spatial scales, it is possible to effectively assess the impact of scale effects in spatial analysis by comparing changes in q values or their rankings across scales. An optimal spatial analysis scale can then be selected based on the highest q value or through methods such as local scatter plot smoothing. A rigorous framework was provided for understanding and selecting the most appropriate spatial resolution for analyzing spatial heterogeneity (Li et al. 2023; Song and Wu 2021; Jacoby 2000).

To conclude, the *gdverse* package offers a comprehensive suite of tools for analyzing SSH, incorporating models like GeoDetector, OPGD, RGD, SPADE, LESH, and others for spatial association and interaction analysis, providing robust statistical methods to identify and explain SSH by assessing the impact of various factors on spatial distributions. With efficient integration of the *tidyverse*, *sf*, and *Rcpp* frameworks, *gdverse* enables streamlined, reproducible workflows for SSH analysis, making it a powerful tool for geospatial research. The *gdverse* package can be a powerful addition to existing SSH analysis tools, enabling efficient modeling of geospatial data.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available at https://github.com/SpatLyu/gdverse_tgis.

References

Bai, H., D. Li, Y. Ge, J. Wang, and F. Cao. 2022. "Spatial Rough Set-Based Geographical Detectors for Nominal Target Variables." *Information Sciences* 586: 525–539. https://doi.org/10.1016/j.ins.2021.12.019.

Cang, X., and W. Luo. 2018. "Spatial Association Detector (Spade)." *International Journal of Geographical Information Science* 32: 2055–2075. https://doi.org/10.1080/13658816.2018.1476693.

Eddelbuettel, D., and R. François. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40: 1–18. https://doi.org/10.18637/jss.v040.i08.

Gräler, B., E. Pebesma, and G. Heuvelink. 2016. "Spatio-Temporal Interpolation Using Gstat." *R Journal* 8: 204–218. https://journal.r-project.org/archive/2016/RJ-2016-014/index.html.

Guo, J., J. Wang, C. Xu, and Y. Song. 2022. "Modeling of Spatial Stratified Heterogeneity." *GIScience & Remote Sensing* 59: 1660–1677. https://doi.org/10.1080/15481603.2022.2126375.

Hijmans, R. J. 2024. "terra: Spatial Data Analysis." https://doi.org/10.32614/CRAN.package.terra. R package version 1.7-83.

Hu, B., T. Wu, Q. Yin, J. Wang, B. Jiang, and J. Luo. 2024. "Calibrating Spatial Stratified Heterogeneity for Heavy-Tailed Distributed Data." *Annals of the American Association of Geographers* 114: 1568–1586. https://doi.org/10.1080/24694452.2024.2351002.

Jacoby, W. G. 2000. "Loess:: A Nonparametric, Graphical Tool for Depicting Relationships Between Variables." *Electoral Studies* 19: 577–613. https://doi.org/10.1016/s0261-3794(99)00028-1.

Li, Y., P. Luo, Y. Song, L. Zhang, Y. Qu, and Z. Hou. 2023. "A Locally Explained Heterogeneity Model for Examining Wetland Disparity." *International Journal of Digital Earth* 16: 4533–4552. https://doi.org/10.1080/17538947.2023.2271883.

Luo, P., Y. Song, X. Huang, et al. 2022. "Identifying Determinants of Spatio-Temporal Disparities in Soil Moisture of the Northern Hemisphere Using a Geographically Optimal Zones-Based Heterogeneity Model." *ISPRS Journal of Photogrammetry and Remote Sensing* 185: 111–128. https://doi.org/10.1016/j.isprsjprs.2022.01.009.

Lv, W. 2024. "sdsfun: Spatial Data Science Complementary Features." https://doi.org/10.32614/CRAN.package.sdsfun. R package version 0.6.0.

Pebesma, E. 2018. "Simple Features for r: Standardized Support for Spatial Vector Data." *R Journal* 10: 439. https://doi.org/10.32614/RJ-2018-009.

Pebesma, E. J. 2004. "Multivariable Geostatistics in S: The Gstat Package." *Computers & Geosciences* 30: 683–691.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Song, Y. 2022. "Geographically Optimal Similarity." *Mathematical Geosciences* 55: 295–320. https://doi.org/10.1007/s11004-022-10036-8.

Song, Y., J. Wang, Y. Ge, and C. Xu. 2020. "An Optimal Parameters-Based Geographical Detector Model Enhances Geographic Characteristics of Explanatory Variables for Spatial Heterogeneity Analysis: Cases With Different Types of Spatial Data." *GIScience & Remote Sensing* 57: 593–610. https://doi.org/10.1080/15481603.2020.1760434.

Song, Y., and P. Wu. 2021. "An Interactive Detector for Spatial Associations." *International Journal of Geographical Information Science* 35: 1676–1701. https://doi.org/10.1080/13658816.2021.1882680.

Truong, C., L. Oudre, and N. Vayatis. 2020. "Selective Review of Offline Change Point Detection Methods." *Signal Processing* 167: 107299. https://doi.org/10.1016/j.sigpro.2019.107299.

Ushey, K., J. Allaire, and Y. Tang. 2024. "reticulate: Interface to 'Python'." https://doi.org/10.32614/CRAN.package.reticulate. R package version 1.39.0.

Wang, J., R. Haining, T. Zhang, et al. 2024. "Statistical Modeling of Spatially Stratified Heterogeneous Data." *Annals of the American Association of Geographers* 114: 499–519. https://doi.org/10.1080/24694452.2023.2289982.

Wang, J., X. Li, G. Christakos, et al. 2010. "Geographical Detectors-Based Health Risk Assessment and Its Application in the Neural Tube Defects Study of the Heshun Region, China." *International Journal of Geographical Information Science* 24: 107–127. https://doi.org/10.1080/13658810802443457.

Wang, J.-F., T.-L. Zhang, and B.-J. Fu. 2016. "A Measure of Spatial Stratified Heterogeneity." *Ecological Indicators* 67: 250–256. https://doi.org/10.1016/j.ecolind.2016.02.052.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H., M. Averick, J. Bryan, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4: 1686. https://doi.org/10.21105/joss.01686.

Zhang, H., G. Dong, J. Wang, et al. 2023. "Understanding and Extending the Geographical Detector Model Under a Linear Regression Framework." *International Journal of Geographical Information Science* 37: 2437–2453. https://doi.org/10.1080/13658816.2023.2266497.

Zhang, Z., Y. Song, L. Karunaratne, and P. Wu. 2024. "Robust Interaction Detector: A Case of Road Life Expectancy Analysis." *Spatial Statistics* 59: 100814. https://doi.org/10.1016/j.spasta.2024.100814.

Zhang, Z., Y. Song, and P. Wu. 2022. "Robust Geographical Detector." *International Journal of Applied Earth Observation and Geoinformation* 109: 102782. https://doi.org/10.1016/j.jag.2022.102782.