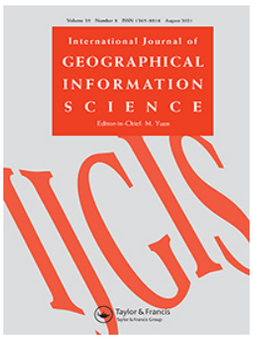

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis20

An interactive detector for spatial associations

Yongze Song & Peng Wu

To cite this article: Yongze Song & Peng Wu (2021) An interactive detector for spatial associations, International Journal of Geographical Information Science, 35:8, 1676-1701, DOI: 10.1080/13658816.2021.1882680

To link to this article: https://doi.org/10.1080/13658816.2021.1882680

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2021, VOL. 35, NO. 8, 1676–1701
https://doi.org/10.1080/13658816.2021.1882680

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

An interactive detector for spatial associations

Yongze Song and Peng Wu

School of Design and the Built Environment, Curtin University, Perth, Australia

ABSTRACT

Geographical variables are usually not independent of each other. Hence, it is necessary to investigate the effect of interactions among explanatory variables on a response variable to characterize spatially enhanced or weakened relationships among all variables. The geographical detector (GD) model identifies zones for each explanatory variable, divides the study area into spatial units by overlapping these zones, and quantifies spatial associations as the power of interactive determinant (PID) between a response variable and explanatory variables. Consequently, the PID values depend upon the distributions of explanatory variables (i.e. spatial characteristics) and the subsequent division of spatial units out of these explanatory variables. This study has therefore proposed an Interactive Detector for Spatial Associations (IDSA) to optimize spatial division and improve PID. IDSA utilizes spatial autocorrelation of each explanatory variable and optimizes spatial units based on spatial fuzzy overlay to compute PID. We test the IDSA on both a simulation study and practical case that analyzes road deterioration in Australia. Results showed that the IDSA model could effectively assess the PID while existing GD overestimated PID. Hence, the IDSA improves the GD with refined spatial units based on explanatory variables to enhance their local spatial associations with a response variable.

ARTICLE HISTORY
Received 24 July 2020
Accepted 25 January 2021

KEYWORDS
Interaction effect; spatial association; geographical detector; spatial fuzzy overlay; spatial heterogeneity

1. Introduction

1.1. Spatial associations of geographical variables

Accurately discovering spatial association, which is the degree of similarity between spatial distribution patterns of a response variable and its potential explanatory variables, is essential for the exploration of geographical factors (Goodchild *et al*. 1992, Fotheringham and Rogerson 2013). Spatial associations are generally constructed based on geo-referencing, which is the process of determining the relationships between geographical locations and the attributes at the given locations (Goodchild *et al*. 2007). Methods for examining spatial associations can be divided into three method categories according to geo-referencing. The first category is based on spatial overlay analysis, which is a fundamental approach in geographical information science in which geographical variables are overlapped in terms of location-based relations. The variables can be applied to both raster and vector data. For example, the structure of overlay maps and error

CONTACT Yongze Song yongze.song@curtin.edu.au

propagation can be investigated for remote sensing and grid data using Geman and Geman's corruption model (Arbia *et al*. 1998). Similarly, overlay analysis of lines and polygons can be performed using uniform spatial grid indexing (Wang *et al*. 2015), and an overlay error propagation process can be used to consider positional errors and the error propagation law (Shi *et al*. 2004).

The second category of methods indicates that the spatial autocorrelation of variables plays a critical role in spatial association. As a unique characteristic of spatial data, spatial autocorrelation measures the presence of spatial pattern variations of a geographical variable observed at a range of locations across space (Haining 2015a, 2015b). During variable exploration, spatial autocorrelation can be presented either as spatial neighbor relations, spatial lag effects, or space-weighted matrices. Examples of such models include spatial lag models, spatial error models (Anselin 2013, Anselin *et al*. 2010), spatial Bayesian hierarchical models (Haining 2003), and geographically weighted regression (GWR) (Fotheringham *et al*. 1998, 2003) and improved GWR models, such as multiscale GWR (Fotheringham *et al*. 2017, Yu *et al*. 2019a), geographically weighted Poisson regression (Nakaya *et al*. 2005, Yu and Peng 2019), as well as geographically and temporally weighted regression (Huang *et al*. 2010).

The third category of methods quantifies spatial associations based on spatial heterogeneity, such as in geostatistical models and the geographical detector (GD) model. For instance, regression kriging and co-kriging explore spatial associations with the support of the spatial heterogeneity of geographical variables that is quantified with spatial variograms and co-variograms under second-order stationarity (Goovaerts 1997, Hengl *et al*. 2004, Garrigues *et al*. 2006). The geo-additive model has been developed to merge the kriging model with nonparametric additive models, which are built using a set of penalized spline regressions to account for the nonlinear relationships between variables (Kammann and Wand 2003, Yu *et al*. 2019b). Kriging models have also been improved by integrating machine learning algorithms during spatial trend modeling (Li *et al*. 2011, Guo *et al*. 2015). In addition, the GD model explores spatial associations based on the theory of spatial stratified heterogeneity, which is a comparison between variance within strata as determined by geographical variables and those between the strata (Wang *et al*. 2010, 2016). A spatial association detector (SPADE) is an improvement on the GD model, as it integrates the spatial characteristics of geographical variables (Cang and Luo 2018).

1.2. Geographical detector (GD) for exploring interaction effects of variables

The GD model explores geographical factors based on the concept of spatial stratified heterogeneity. The model consists of four components: factor, risk, interaction, and ecological detectors. The factor detector, as the core part of the model, quantifies the power of determinant (PD) value of an explanatory variable with a Q value which compares the variance of observations in the stratified variables with that of the whole space (Wang *et al*. 2010, Wang and Chen 2018). The GD Q value is calculated as (Wang *et al*. 2016):

$$Q_{GD} = 1 - \frac{\sum_{k=1}^{n} N_k \sigma_k^2}{N\sigma^2} \tag{1}$$

where $N_k (k = 1, \ldots, n)$ and σ_k^2 are the number and the statistical population variance of observations within k, which is the local zone determined by an explanatory variable; and N and σ^2 are the number and population variance of observations for the entire study area, respectively. Due to its advantages of simpler computation and requiring no statistical assumptions, the GD model has been widely applied in a variety of fields for explanatory variable exploration (Ju *et al*. 2016, Ding *et al*. 2019, Qiao *et al*. 2019, Zhao *et al*. 2020). In reviewing previous applications of the GD model in the last ten years, the model has been primarily applied in spatial heterogeneity research, such as the analysis of patterns and determinants in climate and environment, land use, pollution and human health, and water issues (Song *et al*. 2020a). According to the statistics of the GD website (http://www.geodetector.cn), the model has been cited in more than 1000 papers.

Over the past decade, a series of efforts has been made to improve the GD model. First, spatial data discretization methods were developed as an effective extension of the GD model, which include supervised methods, in which continuous variables are cut with given statistical regulars such as natural or quantile breaks, and unsupervised methods, where breaks were manually determined (Cao *et al*. 2013, 2014, Bai *et al*. 2020). Optimal spatial discretization methods could be determined using maximum Q values (Song *et al*. 2018, 2020a) or by tracking the variation trends of the Q values (Cang and Luo 2018). Second, spatial scale effects were added to the GD model for improving multi-scale analysis, as well as for selecting the optimal spatial scales of explanatory variables (Ju *et al*. 2016, Song *et al*. 2020a). Lastly, the spatial dependence of geographical variables was integrated into the GD model. For example, in the SPADE model, the spatial dependence of the variables was characterized using spatial autocorrelation measures (Cang and Luo 2018).

Geographical variables are not usually independent from each other. Hence, it is desirable to use the interaction of variables that have interaction effects to quantify spatial associations between a given response variable and its associated explanatory variables, a value otherwise known as the power of interactive determinant (PID). Herein, the 'interaction of variables' refers to the spatial overlay of variables, and 'interaction effects' indicate the types of spatial associations or the PID, including the spatially enhanced, weakened, or independent relationships between data in which the joint effects are greater than, less than or equal to the sum of individual effects, respectively (Lavrakas 2008, Ren *et al*. 2014, Ju *et al*. 2016, Ding *et al*. 2019, Bai *et al*. 2019, Zhou *et al*. 2020). In the GD model, the PID can be estimated using an interaction detector (GD-ID) model. The GD-ID model can be used to quantify whether two individual variables have enhanced, weakened, or independent effects on a given response variable (Wang *et al*. 2010). The GD-ID model identifies zones for each explanatory variable, divides the study area into spatial units by overlapping these zones across all explanatory variables, and quantifies spatial associations as the PID between a response variable and explanatory variables within individual spatial units. Consequently, the PID value depends upon how each explanatory variable distributes over space (i.e. spatial characteristics) and the subsequent discretization of spatial units out of these explanatory variables.

Although GD-ID can be used to examine the interaction effects of variables, limitations to its application do exist. For example, the spatial intersection approach is commonly used for spatial overlay, but this application is known to generate numerous finely divided overlay

zones with few or even no observations. Particularly, this phenomenon would be severe in the case of there being more than two explanatory variables and would cause observations with similar geographical features, which should belong to an individual zone, than being divided into multiple zones. This would further lead to a biased estimation of the PID, in which the estimated PID would be higher or lower than the true value due to missing information and limited processing of estimation methods (Calonico *et al*. 2018, Merchant *et al*. 2020). Additionally, the characteristics of spatial dependence of the geographical explanatory variables are ignored in the GD-ID model. For example, the SPADE model applies the spatial dependence of variables in the factor detector that explores the PD of individual variables (Cang and Luo 2018); however, no knowledge is available regarding the integration of the spatial dependence of variables in order to assess interaction effects. Such an absence of the spatial dependence of variables would likely lead to a biased estimation.

To address these gaps, this study has proposed an Interactive Detector for Spatial Associations (IDSA) model based on spatial heterogeneity to effectively explore interaction effects of variables and more accurately quantify the PID. This IDSA model integrates (i) optimal spatial discretization, (ii) spatial autocorrelation characteristics of variables, and (iii) the optimal interaction of variables derived from spatial fuzzy overlay with the GD model. Note that spatial fuzzy overlay is utilized for integrating the uncertainty and vagueness information of geographical attributes in the overlay to achieve a more reasonable understanding of the interaction effects (Phillips *et al*. 2011, Kim *et al*. 2019). Furthermore, a set of effectiveness evaluation indicators for spatial analysis have also been developed in order to assess the model's performance. In this way, the IDSA model was evaluated by comparing with the GD-ID and SPADE-based interaction detector (SPADE-ID) to explore interaction effects of variables and quantifying the PID.

2. The Interactive Detector for Spatial Associations (IDSA) model

The IDSA model, developed for exploring interaction effects variables and the PID, consists of four steps, as shown in Figure 1. The first step is to discretize the spatial variables using the optimal combinations of discretization methods and break numbers. The second step is to calculate the PD of individual variables using the SPADE model. The third step is to identify the optimal interaction of variables with a spatial fuzzy overlay approach and compute the PID with the consideration of spatial autocorrelations of variables. The final step is to evaluate the model's performance in terms of the developed effectiveness indicators. Each of these steps are explained in the following subsections. An R 'IDSA' package is developed for computation.

2.1. Optimal spatial data discretization

The first step of IDSA is to perform spatial data discretization with optimal parameters, which have been derived using appropriate strategies. We recommend two applicable strategies based on the number of observations and practical requirements. Strategy one is applicable for small amounts of data, for which there is no strict data number threshold and practical requirements can be considered. The optional spatial discretization methods include both supervised methods, such as quantile, equal, natural, geometric, and standard deviation breaks, and unsupervised methods. The recommended optional break

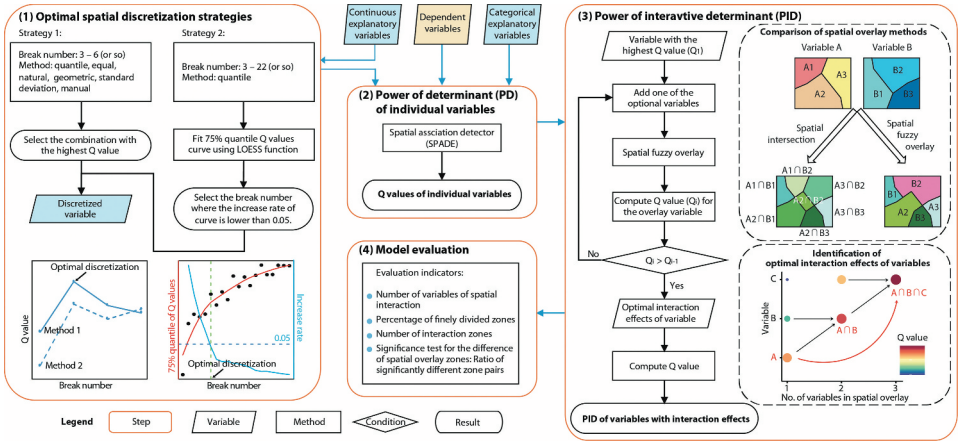

Figure 1. Schematic overview of spatial association analysis using the Interactive Detector for Spatial Associations (IDSA) model.

number is an integer sequence from 3 to 6 (or so). Then, the optimal combination of discretization methods and break numbers is such that the highest Q value is computed using the SPADE model. Strategy two is utilized for a relatively large amount of data. The recommended break number is an integer sequence from 3 to 22 (or so), and the discretization method is the quantile break. Here, the Q values are computed for explanatory variables with all optional parameter combinations. Then, a curve of the 75th quantile Q values is applied to reveal the PD variation trend of the primary explanatory variables using a locally estimated scatterplot smoothing (LOESS) model. The LOESS model is a flexible, nonlinear local regression model suitable for modeling complex processes without theoretical regression forms (Jacoby 2000, Barco *et al*. 2020). The optimal break number is determined to be the point where the increase rate of the curve is lower than 5%, which is a trade-off between a small break number and high Q values.

2.2. Power of determinant (PD) of individual variables

The second step is to compute the PD of individual variables using the SPADE Q value. In the SPADE model, the study area is divided into multiple zones for each individual variable. Then, the power of spatial determinant (PSD) is defined as a ratio between the local spatial variance of the variable-determined overlay zones and the entire study area (Cang and Luo 2018):

$$\theta = 1 - \frac{\sum_{k=1}^{n} N_k \tau_k}{N\tau} \tag{2}$$

where $N_k (k = 1, \ldots, n)$ and τ_k are the number of observations and spatial variability in zone k, respectively, and N and τ are the total number of observations and global spatial variability. The spatial variance is defined as the mean of the spatially weighted cross-product of the spatial autocorrelation measures to the present characteristics of the geographical variables:

$$\tau = \frac{\sum_i \sum_{j \neq i} w_{ij} G_{ij}}{\sum_i \sum_{j \neq i} w_{ij}} \tag{3}$$

where w_{ij} is an element of the spatially weighted matrix between locations i and j, and G_{ij} is the attribute similarity between the two locations. Here, G_{ij} is presented as the semi-squared difference $\frac{1}{2}\left(x_i - x_j\right)^2$, where x is the observation at a given location, and assumptions of spatial autocorrelation should be satisfied. Finally, the SPADE Q value is a ratio of PSD values between a response variable and explanatory variables:

$$Q_{SPADE} = \frac{\theta_r}{\theta_e} \tag{4}$$

where θ_r and θ_e are PSD values of response and explanatory variables, respectively, in which θ_e is used to adjust biases due to the missing information during spatial discretization.

2.3. Power of interactive determinant (PID)

The third step aims to identify the optimal interaction of explanatory variables and compute the PID. This stage consists of three key components: (i) the spatial fuzzy overlay of the overlapping variables, (ii) identification of the optimal interaction of variables via an iteration process, and (iii) computing the PID with the consideration of certain spatial autocorrelation characteristics of the variables.

2.3.1. Spatial fuzzy overlay

The spatial fuzzy overlay is a process of overlapping geographical variables in terms of their fuzzy relations across space (Zadeh 1973, Yu *et al*. 2018, Bustince *et al*. 2010). These fuzzy relations assume that the geographical variables are related to one another to some extent (Dubois and Prade 1983), and are quantified using fuzzy membership functions. Let m be the number of geographical explanatory variables for generating a spatial interaction variable, and fuzzy membership functions, presented as fuzzy numbers, be equal to the normalized mean risk values derived from the risk detector in the GD model:

$$[\, f_n(X_1) \quad f_n(X_2) \quad \ldots \quad f_n(X_m)\,] = g([\, \eta(X_1) \quad \eta(X_2) \quad \ldots \quad \eta(X_m)\,]) \tag{5}$$

where $X_i (i = 1, \ldots, m)$ is an explanatory variable, $f_n(X_i)$ is the fuzzy number, $\eta(X_i)$ are mean risk values, and $g()$ is a normalization function:

$$g(X) = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{6}$$

The mean risk values are the mean values of the response variable in the spatial overlay zones, which are determined by explanatory variables, where continuous variables should be discretized (Wang *et al*. 2016). Therefore, fuzzy numbers can demonstrate fuzzy relations between variables.

The integration of fuzzy numbers which represent the fuzzy relations of the variables is performed to estimate the combined fuzzy number of the interaction of variables. The integration process is computed with the fuzzy AND operator, which can help identify the least common geographical configurations from the fuzzy number variables:

$$f_n(X_1 \cap X_2 \cap_{\ldots} \cap X_m) = \min(f_n(X_1), f_n(X_2), \ldots, f_n(X_m)) \tag{7}$$

where '$\cap$' denotes the spatial overlay process and $X_1 \cap X_2 \cap_{\ldots} \cap X_m$ is the interaction of variables. Thus, the overlay zones of variables with the interaction effects are determined by the fuzzy number of the overlay variable.

2.3.2. Estimation of the PID

The PSD of an interaction of explanatory variables (PSD-IEV) is calculated as a ratio between the sum of the local spatial variance at overlay zones of individual variables with interaction effects and the sum of the global spatial variability of the individual variables:

$$\varphi = 1 - \frac{\sum_{i=1}^{m} \sum_{k=1}^{n_i} N_{i,k} \tau_{i,k}}{\sum_{i=1}^{m} N_i \tau_i} \tag{8}$$

where $N_{i,k} (i = 1, \ldots, m; k = 1, \ldots, n_i)$ is the number of observations at zone k for an individual variable X_i, which has interaction effects with other variables; $\tau_{i,k}$ is the local spatial variance; and N_i and τ_i are the number of observations and global spatial variability of variable X_i, respectively. The spatial variance τ is calculated with equation (3) to capture the characteristics of the geographical variables using spatial autocorrelations. Finally, to adjust biases due to the missing information during spatial discretization, the PID value is calculated as a ratio between the PSD value of a response variable and the PSD-IEV value:

$$Q_{IDSA} = \frac{\theta_r}{\varphi} \tag{9}$$

The range of the PID value is between 0 and 1.

2.3.3. Identification of the optimal interaction of variables

The identification of the optimal interaction of variables is an iteration process, in which one of the optional variables will be added to the spatial fuzzy overlay process at each iteration until the optimal interaction of variables is derived. The initial variable of the iteration is an individual variable with the highest SPADE Q value. Optionally, the initial variables can be variables with the first, second or third highest SPADE Q values. Then, one of the other optional variables is added to perform the spatial fuzzy overlay. The PID of the updated explanatory variables with interaction effects is then computed using equation (9) and the outcome yields the value of Q_i. In this stage, the added variable should help find the maximum Q value of the updated interaction of variables. The next stage is to justify if more variables should be added to generate the optimal interaction of variables. If $Q_i > Q_{i-1}$, the updated interaction of variables is optimal. If $Q_i \leq Q_{i-1}$, a different optional variable should be added which allows for the maximum Q value of the updated interaction of variables. Once the optimal interaction of variables is produced, the PID of the variables with the optimal interaction can be computed using equation (9). In addition to the iteration method used in this study, another approach is to identify the optimal interaction from all possible interactions of variables, but it is time-consuming for large datasets.

2.4. Model performance evaluation

This study also proposes a set of model performance evaluation indicators to assess the effectiveness of the models for exploring factors or explanatory variables with interaction effects. These indicators are used to examine whether the model can identify spatial stratified heterogeneity with a minimum number of variables, as well as to determine significantly different spatial overlay zones. Model performance is evaluated by the following four indicators.

The first indicator is the number of individual explanatory variables used for examining the interaction effects. We assume that a more effective model for exploring interaction effects of variables should use fewer spatial explanatory variables.

The second indicator is the percentage of finely divided zones that are determined by the interaction of variables. If multiple spatial variables overlap to compute the PID, numerous overlay zones with only minimal or no observations are likely to be produced. This phenomenon would lead to a critically biased estimation of the PID, as the sum of the local spatial variance would be affected by zones with uncertain variance derived from having only minimal observations. Furthermore, the spatial variance cannot be computed for zones with only one observation. Thus, a powerful model could essentially reduce the amount of finely divided zones. In this study, the finely divided zones were defined as zones with only one observation.

The third indicator is the number of overlay zones. Fewer overlay zones indicate that the model can more effectively present spatial heterogeneity through the interaction of variables.

Finally, a t-test is performed to assess if values of the response variable between a pair of overlay zones are significantly different. The difference between a pair of zones, u and v, is tested using a t-test approach:

$$t = \frac{\bar{Y}_u - \bar{Y}_v}{\sqrt{\sigma_u^2/N_u + \sigma_v^2/N_v}} \tag{10}$$

where $\bar{Y}$, σ^2, and N are the mean values, variance, and number of observations in each of the two zones, respectively. Then, the percentage of significantly different zone pairs among all pairs of overlay zones is computed to indicate if a model can effectively identify the interaction effects of variables.

3. IDSA simulation study

A simulation study was designed to present the general steps of the IDSA model and evaluate the model's performance. The simulation study consists of a spatial data simulation, IDSA-based interaction effects exploration of variables, and evaluation of the model's performance. Three geographical variables with 10×8 grids were generated, including a response variable Y and two explanatory variables X_a and X_b, with the assumptions that the variables have spatial autocorrelations and that the response and explanatory variables are correlated. The spatial distributions of the simulation data are mapped in Figure 2.

The IDSA-based interaction effects exploration of variables was performed according to the steps presented in the methods section. First, the optimal spatial discretization

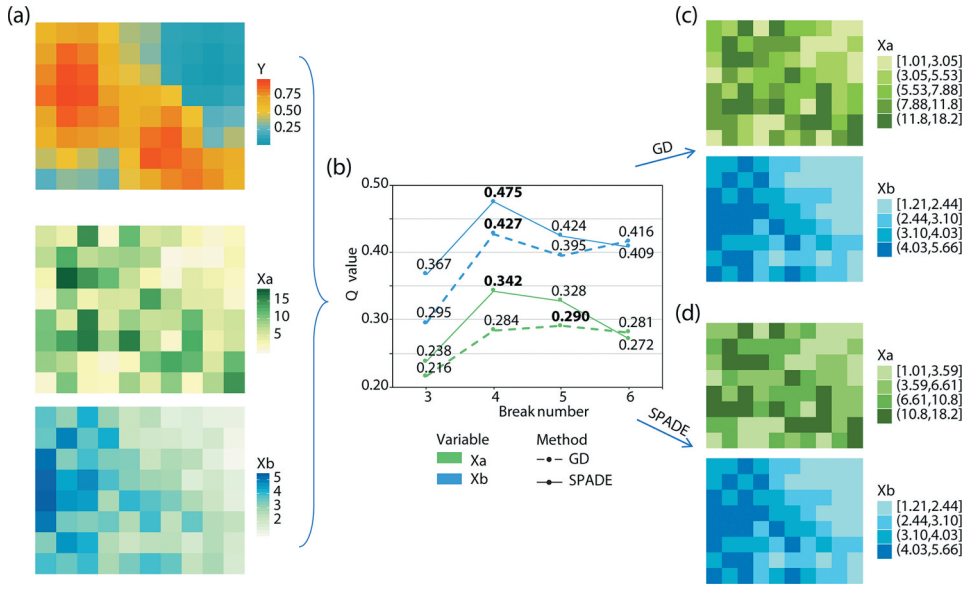

Figure 2. Simulation data distributions (a) and optimal spatial discretization (b) using the geographical detector (GD) (c) and the spatial association detector (SPADE) (d) approaches, respectively.

parameter was identified, and second, the Q values of individual variables were computed using the SPADE model. As only 80 observations were present, strategy one of the optimal spatial discretization was used in this step, in which the optional parameter combinations were formed using the quantile break method and with an integer sequence from 3 to 6. The PDs of variables derived by the factor detector of the GD and SPADE models are compared in Figure 2. Results show that when compared with the SPADE-based results, the PDs of variables were relatively low in the GD model. In the SPADE results, the optimal break numbers of X_a and X_b were 4 and 4, respectively; however, numbers computed using the GD model were 5 and 4.

Figure 3 shows the process and results of identifying the interaction effects of the variables. Figure 3(a) shows the mean risk values of Y due to explanatory variables X_a and X_b, and Figure 3(b) illustrates the distributions of the respective fuzzy numbers. Figure 3(c) shows the distributions of the fuzzy overlay, spatial zones determined by the interaction of variables $X_a \cap X_b$, and regions of predominant variables that were used in the interaction. The study area was divided into seven zones by the interaction of variables. The last map in Figure 3(c) demonstrates that variable X_a had more contributions than X_b in the interaction in the southwest region and had fewer contributions in the northeast region. The results of the IDSA model are compared with those of the GD-ID and the SPADE-ID. The spatial overlay of both GD-ID and SPADE-ID were performed with the intersection approach. It should be noted that SPADE-ID is a direct combination of the spatial intersection of variables and the SPADE model for computing PID. Table 1 lists the PID of the interaction ($Q_{X_a \cap X_b}$) as computed by the GD-ID, SPADE-ID, and IDSA models, and indicates that the PID of variables with interaction effects were overestimated by the former two models.

The last step was to evaluate model performance using a set of evaluation indicators (Table 1 and Figure 4). Table 1 lists the number of spatial overlay zones, number of overlay

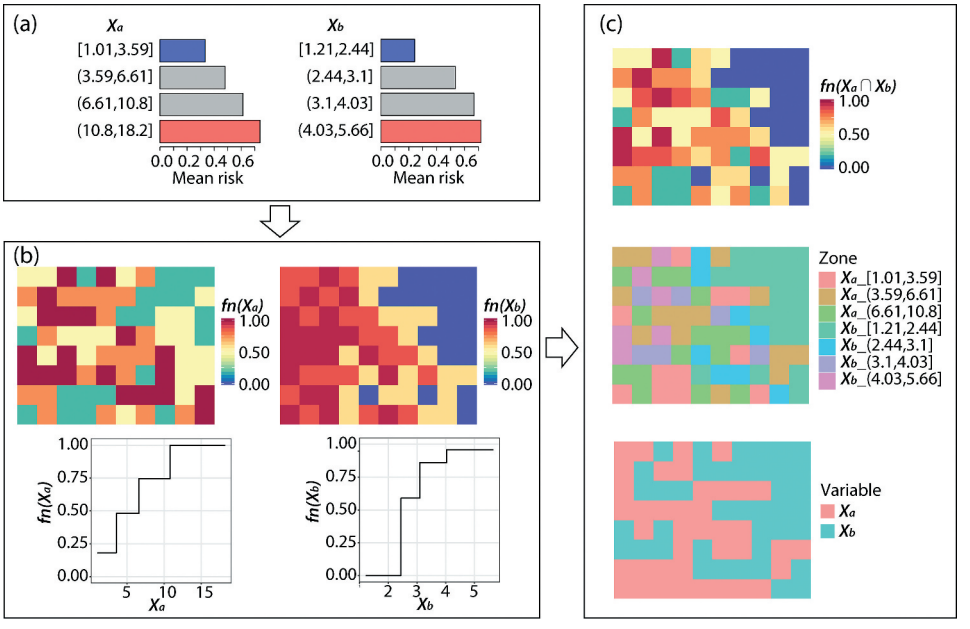

Figure 3. Process and results of the IDSA simulation study: mean risks determined by explanatory variables (a), fuzzy number distributions (b), and fuzzy overlay results (c).

Table 1. Summary of the power of interactive determinant (PID) for the simulation study and zonal difference test.

Model	$Q_{X_a \cap X_b}$	Number of overlay zones	Number of overlay zone pairs	Ratio of significantly different zone pairs
GD-ID	0.653	20	190	22.11%
SPADE-ID	0.623	16	120	24.17%
IDSA	0.597	7	21	42.86%

zones pairs, and ratio of significantly different zone pairs as estimated by the t-test. Additionally, three finely divided overlay zones (3/20) occurred in the GD-ID model and one (1/16) occurred in the SPADE-ID model; however, no finely divided zones were found in the IDSA model. Figure 4 shows a comparison of difference tests for the zone pairs in the three models, together with the corresponding cumulative significance distributions. Thus, when compared with GD-ID and SPADE-ID, the IDSA model could identify an optimal interaction of variables and more effectively capture geographical information from explanatory variables using spatial heterogeneity analysis.

4. Application in exploring road infrastructure performance factors

4.1. Study area and data

Road network deterioration in Mid-West Gascoyne, Western Australia (WA) is an essential example of smart and sustainable road infrastructure system management for Australia, as well as globally. The primary reasons for its significance are its critical location and

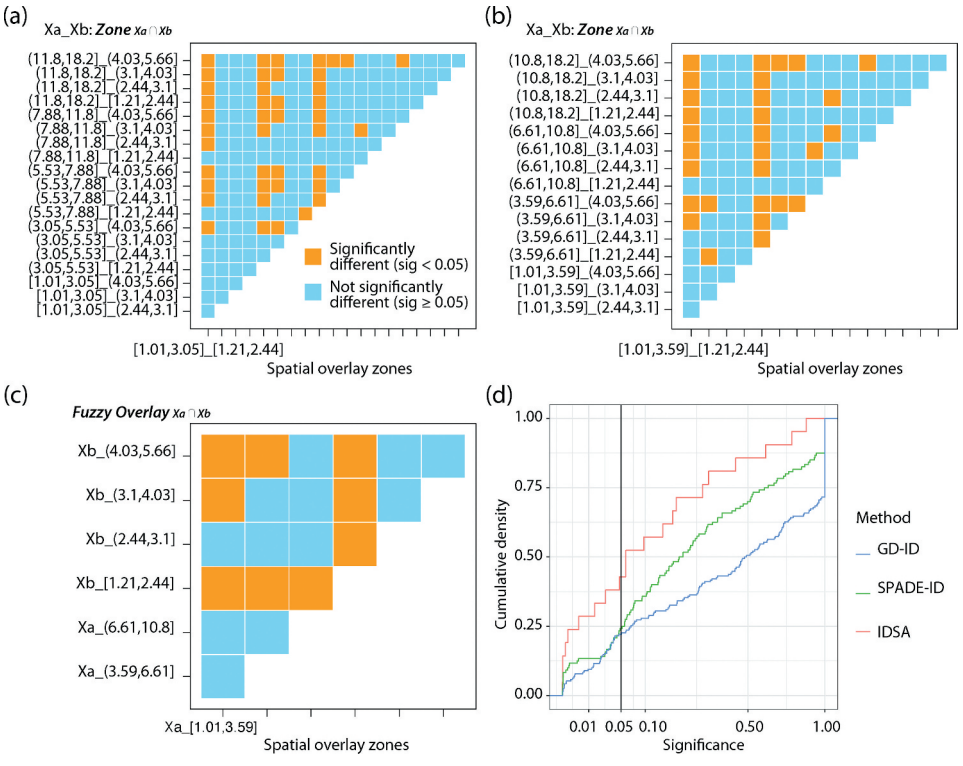

Figure 4. Difference test for the zone pairs determined by the interaction of variables in the GD-ID (a), SPADE-ID (b), and IDSA (c) models, and corresponding cumulative significance distributions (d).

multiple network functions (Song *et al*. 2020b). The road network, comprised of 3649 km, links four of the most important regions in WA, including the northern region which is one of the largest mineral, oil, and gas production regions in the world, as well as the southern region which is one of the most globally productive agricultural regions. WA also features most of the primary Australian ports are located along the Indian Ocean, as well as Perth, the capital city, which hosts 70% of the WA's total population (Figure 5). According to geographical location, the road network can be divided into three parts: the northern, southern, and eastern roads. This road network can also be used for a performance evaluation of typical road deterioration, which can be assessed using updated Traffic Speed Deflectometer (TSD) data for Australia (The Government of Western Australia, Main Roads Western Australia and ARRB Australia Road Research Board (ARRB) 2018). These high-resolution TSD data are collected using a heavy-vehicle-based laser scanner system to accurately monitor the structure, functions, and surface conditions of the roads. Laser scanner cloud points are collected by the system for every 10-m interval along the road.

This study explored the explanatory variables with interaction effects of road deterioration performance in the Mid-West Gascoyne region. In the study, approximately 0.365 million observations of road deflection were collected across the entire main roads network, where deflection is one of the major indicators of road performance (Song *et al*. 2018). The deflection data were preprocessed using a spatial heterogeneity-based segmentation (SHS) model for homogeneous segmentation and for effective road

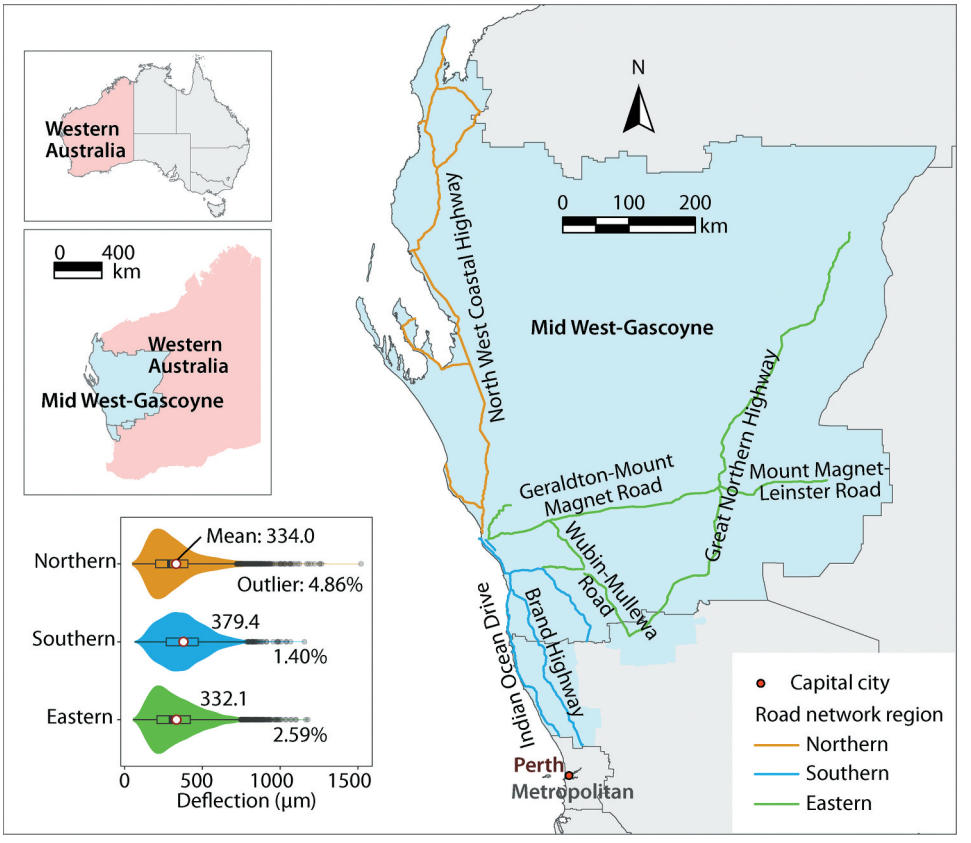

Figure 5. Road network in the Mid-West Gascoyne region in Western Australia.

infrastructure management (Song *et al*. 2020b). The SHS model uses relatively few segments to present the road performance data, where homogeneity of the data within the segments and heterogeneity between the segments are high. As a result, 12,594 road segment-based deflection data were generated; and the statistical summaries of deflection in the three network regions are presented in Figure 5. The mean deflection in the southern region was higher than that in the northern or eastern regions. However, the northern region had more high-value outliers, indicating high deterioration risks.

Corresponding to the road segment-based deflection data, three categories of potential explanatory variables were collected, including traffic vehicle burdens, as well as meteorological and soil moisture variables, which were converted to road segment data (Table 2). The masses of the traffic vehicles and heavy vehicles were estimated using a segment-based regression kriging method with traffic volume monitoring data and estimated individual vehicle masses from the restricted access vehicles network standard (Song *et al*. 2019). The temperature was sourced from a 1-km resolution Land Surface Temperature (LST) MOD11A2 system using a Moderate Resolution Imaging Spectroradiometer (MODIS) (Wan *et al*. 2015). Precipitation and soil moisture variables, including surface soil moisture, runoff, actual evapotranspiration, and deep drainage, were sourced from the 1-km resolution Australian Soil

Table 2. Potential explanatory variables of road infrastructure performance.

Category	Code	Variable	Unit
Traffic vehicle burden	tmv	Annual daily mean total mass of traffic vehicles	ton
	phm	Percentage of heavy vehicle masses	%
Meteorological	temp	Annual mean temperature	℃
	prec	Annual mean precipitation rate	mm
Soil moisture	sm	Surface soil moisture	%
	ro	Runoff	mm
	ae	Actual evapotranspiration	mm
	dd	Deep drainage	mm

Resources Information System dataset (Johnston *et al*. 2003, Bureau of Meteorology Australian Government 2019). Both meteorological and soil moisture data were then converted to road segment data.

4.2. Optimal spatial discretization and PD of individual variables

Due to the numerous observations in this case, the optimal spatial data discretization process (which was strategy two mentioned in the methods section) was used to explore the optimal discretization parameters. The optional parameters consisted of a combination of the quantile break method and optional break numbers of an integer sequence, from 3 to 22. Figure 6 shows the processes of optimal break number selection for the IDSA model for three regions in the road network. For instance, Figure 6(a) and (b) show the process in the northern road network. Figure 6(a) shows the SPADE Q values of the explanatory variables when the discretization break number ranged from 3 to 22, while Figure 6(b) shows the curves for the LOESS function, which was fitted to the 75th quantile Q values and the rate of increase. In general, with the increase in break numbers, Q values gradually increased, meaning that the PD of the variables was enhanced when more spatial zones were given by the discretized variables. The 75th quantile curve from the Q values showed an increasing trend. Although this trend increased with the break numbers, the increasing rate gradually reduced. When the break number was greater than 6, the increasing rate was lower than 5%; thus, 6 was the optimal break number for the variables in the northern network. Similarly, optimal spatial data discretization was performed for the explanatory variables in the southern and eastern networks; the results are listed in Table 3.

To compare the PD estimation for the SPADE and GD models for individual variables, GD-based optimal spatial data discretization and PD computations were performed. The results are presented in Figure 7 and Table 3. Similar to the results shown in the simulation study, the Q values computed by the SPADE model were generally higher than those in the GD model. This indicates that the PD of the individual variables may have been underestimated by the GD model.

The PD results for the individual explanatory variables explored using the SPADE model are presented in Figure 8. For all three regions, traffic vehicle burden variables, especially the total mass of the vehicles, generally had higher contributions to road performance than the meteorological and soil moisture variables; however, regional disparities remained. In the northern region, the total mass of vehicles and percentage of heavy vehicles contributed 42.2% and 31.2%, respectively, followed by runoff, which contributed 29.9%. In the

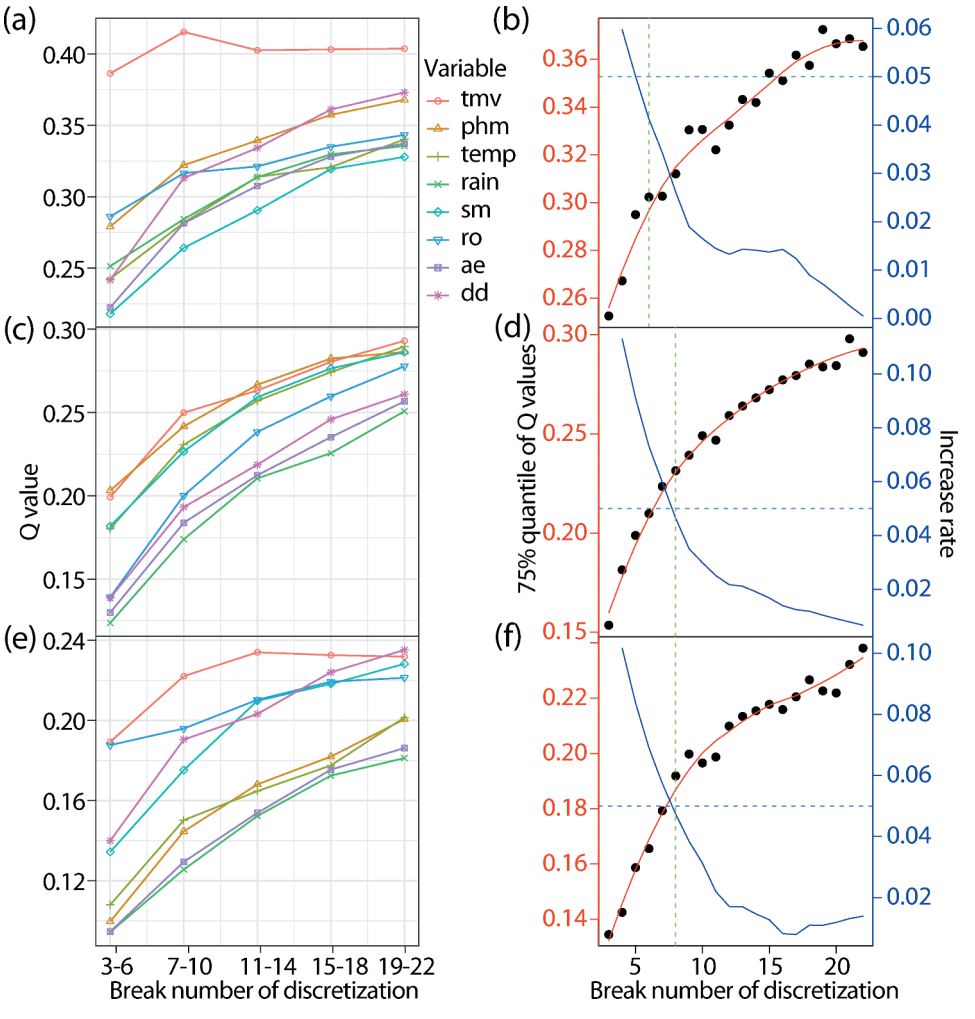

Figure 6. SPADE-based PD and optimal spatial data discretization in the northern (a and b), southern (c and d), and eastern (e and f) road networks.

Table 3. Breaks numbers and PD of optimal spatial data discretization for explanatory variables.

Region	Model	Break number	$Q_{0.75}$
Northern	GD	8	0.229
	SPADE	6	0.302
Southern	GD	8	0.153
	SPADE	8	0.231
Eastern	GD	16	0.052
	SPADE	8	0.192

southern region, the total mass of vehicles and percentage of heavy vehicles were still major variables, contributing 26.4% and 24.7%, respectively. This was followed by temperature, which contributed 22.6%. In the eastern region, the total mass of vehicles contributed

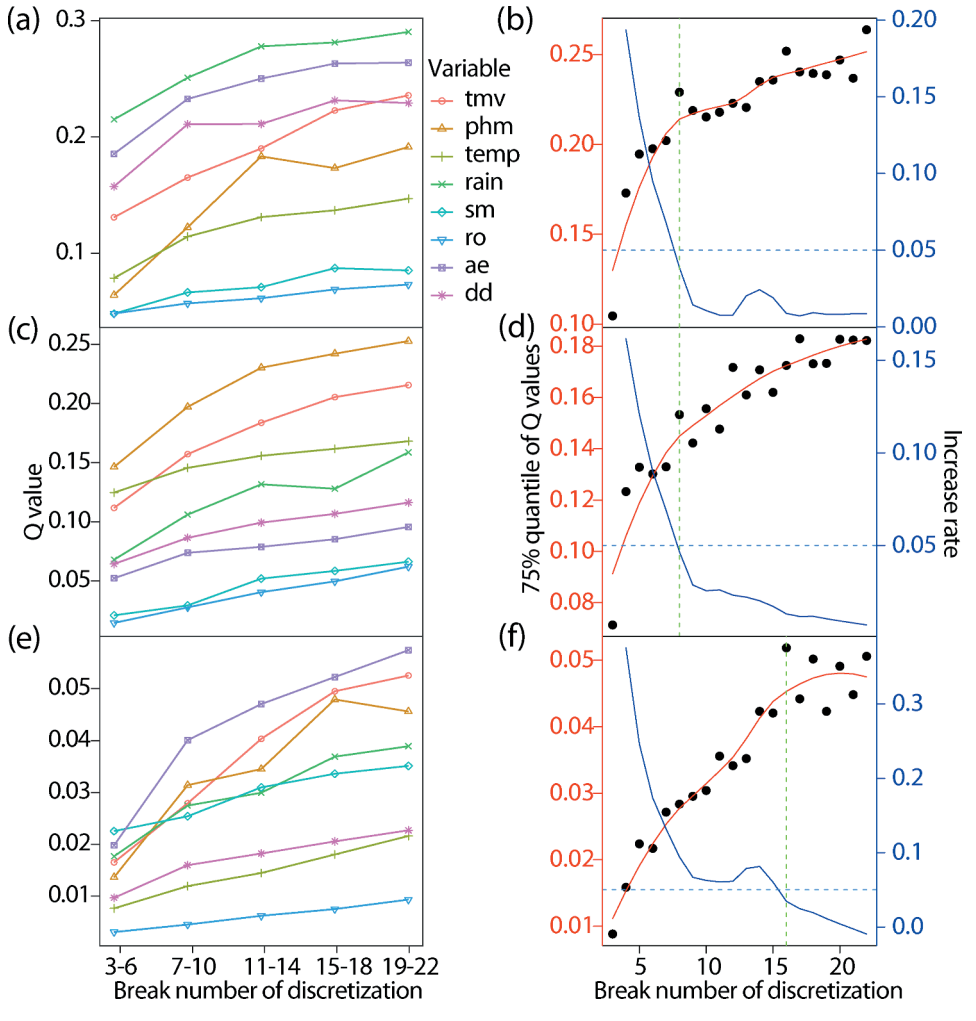

Figure 7. GD-based PD and optimal spatial data discretization in the northern (a and b), southern (c and d), and eastern (e and f) road networks.

22.9%, followed by soil moisture variables, and runoff and deep drainage, which contributed 19.8% and 19.0%, respectively.

4.3. PID and model performance evaluation

The optimal interaction of variables was identified using an iteration process supported by spatial fuzzy overlay and characterization of the geographical variables using spatial autocorrelations. The process and results of the optimal interaction effects of variables for the northern, southern, and eastern road networks are illustrated in Figure 9. For instance, in the northern region, the iteration process started with the total mass of vehicles, which was the variable with the highest PD, as explored by the SPADE model. Next, one of the other variables was added to

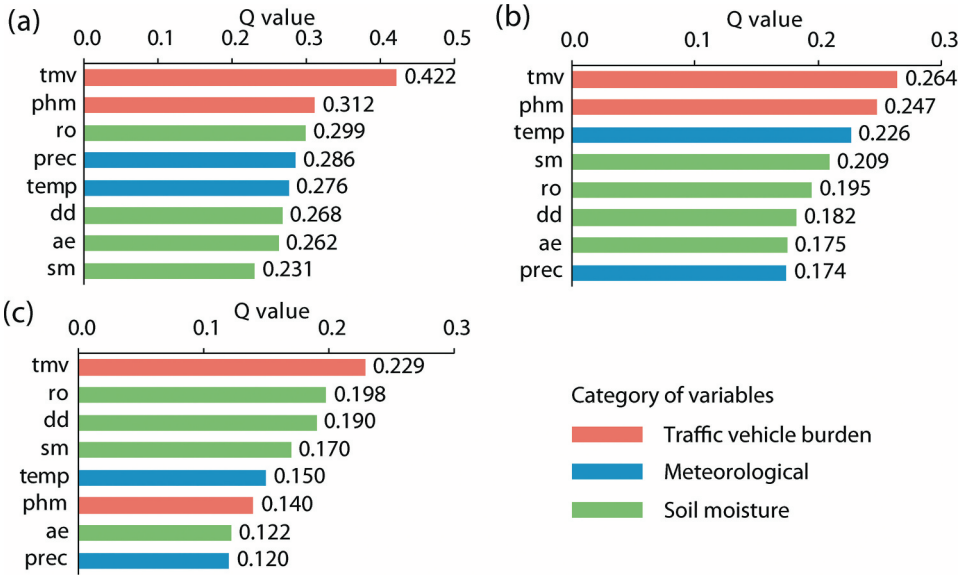

Figure 8. PD for the individual explanatory variables explored using the SPADE model in northern (a), southern (b) and eastern (c) road networks.

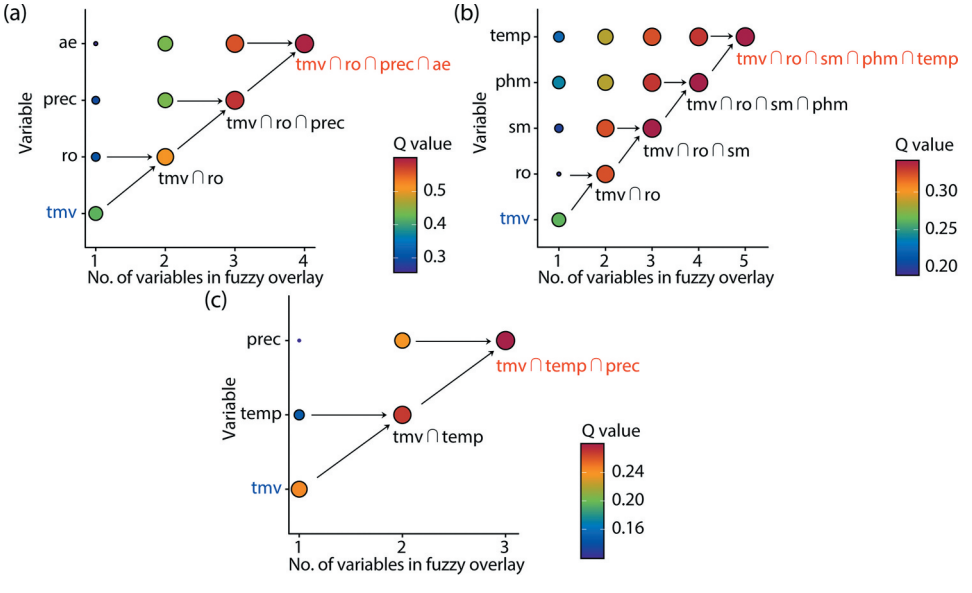

Figure 9. Determination of the optimal interaction of variables for the northern (a), southern (b), and eastern (c) road networks.

assess the interaction of variables using a spatial fuzzy overlay approach. As a result, the interaction between the total mass of the vehicles and runoff ($tmv \cap ro$) was the optimal interaction, as it had the highest PID and $Q_{tmv \cap ro} > Q_{tmv}$. Then, one more variable was added to the interaction $tmv \cap ro$,

which explored the interaction with the highest PID, and was also higher than the previous PID. This process was repeated in order to determine the overall optimal interaction of variables until no interaction had a higher PID. Finally, the optimal interaction of road performance variables in the northern road network was $tmv \cap ro \cap \prec \cap ae$, and its contribution to road performance was computed using the IDSA model to be 59.1%. This process was also performed for the southern and eastern road networks; the results showed that the optimal interaction of variables for the southern and eastern road networks were $tmv \cap ro \cap sm \cap phm \cap temp$, which contributed 33.9%, and $tmv \cap temp \cap \prec$, which contributed 27.4%. Compared with evaluating every possible interaction of variables ($2^8 = 256$), only 18, 22, and 13 combinations were required in this process of studies in the northern, southern and eastern areas, accounting for 7.03%, 8.59% and 5.08% of all possible combinations, respectively. Thus, the results demonstrate the effectiveness of the process for identifying optimal interactions of variables in the IDSA model. Figure 10 shows the distributions of the selected variables with the interaction effects during spatial fuzzy overlay, as well as the percentages of the variables that have primary contributions in the fuzzy overlay. This map presents the relative contributions of the variables that comprised the interaction for each regional road

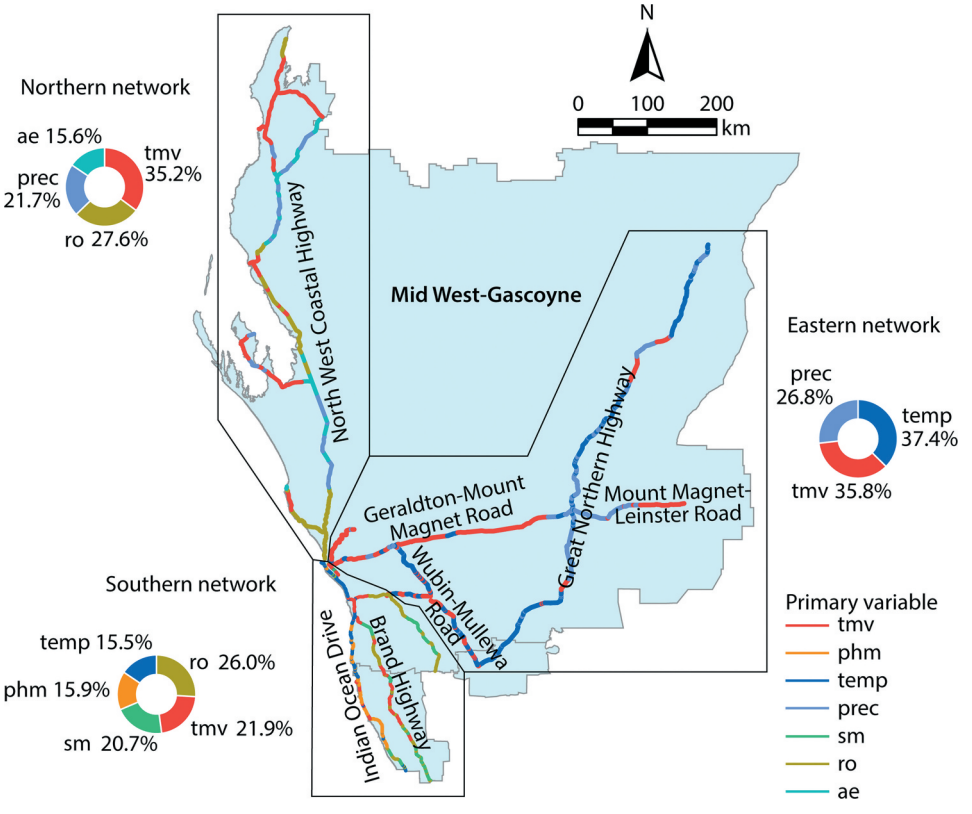

Figure 10. Distributions of selected variables with interaction effects and percentages of variables that have primary contributions in the fuzzy overlay.

network. These are essential for regional deterioration analysis and strategic road maintenance decision-making.

In this study, the performance of the IDSA model was compared with the GD-ID and the SPADE-ID models (Table 4). First, the analysis of all three road network regions demonstrated that the IDSA model could use fewer explanatory variables with interaction effects to explain the road performance. Among the eight potential explanatory variables, the IDSA model used only three to five variables for spatial analysis of the three road networks. This is lower than that required for the other two models.

Second, the IDSA model could effectively adjust the PID overestimation of the interaction of variables, as derived from the GD-ID model. The overestimation was reduced between 6.3% and 52.7% in the three regions.

Third, the results also show that in the IDSA model, the interaction of variables had substantially less spatial overlay zones compared with other models. The overlay zones determined from the interaction of variables by the IDSA model were only 0.9–3.6% (in three regions) of those examined by the GD-ID model, and only 4.5–10.7% of those examined by the SPADE-ID model. It should be noted that fewer overlay zones can effectively reduce PID estimation errors and critically decrease the amount and probability of finely divided zones. Figure 11 shows a statistical summary of the observation numbers in the overlay zones with density plots. In the analysis of both the GD-ID and the SPADE-ID models, most overlay zones only contained a few observations, resulting in severely biased PID estimations. Contrastingly, the density curves for the observation numbers in the overlay zones, as derived from the IDSA model, tended to be bell-shaped, meaning that most of the values were clustered in the middle sections. Thus, the IDSA model could effectively reduce estimation errors caused by numerous finely divided zones of the interaction of variables.

Fourth, due to the significant reduction in the number of overlay zones, the number of overlay zone pairs simultaneously decreased. The numbers of overlay zone pairs in the northern, southern, and eastern network analyses were 136, 231, and 190 in the IDSA-based analysis, respectively. However, for the GD-ID- and SPADE-ID-based analyses, the numbers of overlay zone pairs were 0.28 million, 0.12 million, and 2.55 million, and 12.6 thousand, 119.8 thousand, and 54.6 thousand, respectively.

Fifth, no finely divided zones were present in the IDSA-based spatial analysis for any of the three regions. Finely divided zones in the GD-ID- and SPADE-ID-based analyses were 21.5–43.5% and 3.8–16.5%, respectively, for the three computed regions.

Finally, the ratios of significantly different ($p < 0.05$) zone pairs for the IDSA-based analysis were 49.0–82.4% for the three regions, which was substantially higher than the ratios in the GD-ID- (14.4–42.7%) and SPADE-ID- (34.8–64.1%) based analyses. Significantly different probability distributions between the overlay zone pairs are shown in Figure 12, revealing that the IDSA-based analysis had higher probabilities of significantly different overlay zone pairs than the other two models. This implies that greater heterogeneous patterns of the response variables could be identified by the IDSA-based interaction of explanatory variables. Therefore, all model performance evaluation indicators demonstrated that the IDSA model is a powerful method for examining interaction effects of explanatory variables.

Table 4. Optimal interaction effects of variables and model performance evaluation.

Region (NO[1])	Model	Interaction (NV[2])	PID[3]	NOZ[4]	NOZP[5]	NOZ without FDZs[6]	NOZP without FDZs	Percentage of FDZs	RSDZP[7]	RSDZP compared with GD-ID
Northern (4342)	GD-ID	Interaction of all variables (8)	0.631	752	282,376	590	173,755	21.54%	42.70%	/
	SPADE-ID	tmv ∩ ro ∩ ae ∩ prec (4)	0.527	159	12,561	153	11,628	3.77%	64.12%	1.502
	IDSA	**tmv∩ro∩prec∩ae (4)**	**0.591**	**17**	**136**	**17**	**136**	**0.00%**	**82.35%**	**1.929**
Southern (2853)	GD-ID	Interaction of all variables (8)	0.569	610	185,745	478	114,003	21.64%	32.64%	/
	SPADE-ID	tmv ∩ ro ∩ sm ∩ ae ∩ temp ∩ phm (6)	0.381	490	119,805	409	83,436	16.53%	34.87%	1.068
	IDSA	**tmv∩ro∩sm∩phm∩temp (5)**	**0.339**	**22**	**231**	**22**	**231**	**0.00%**	**65.80%**	**2.016**
Eastern (5399)	GD-ID	Interaction of all variables (8)	0.579	2257	2,545,896	1276	813,450	43.46%	14.39%	/
	SPADE-ID	tmv ∩ sm ∩ temp ∩ prec ∩ ae ∩ ro ∩ dd (7)	0.313	331	54,615	293	42,778	11.48%	35.20%	2.446
	IDSA	**tmv∩temp∩prec (3)**	**0.274**	**20**	**190**	**20**	**190**	**0.00%**	**48.95%**	**3.402**

[1]NO: Number of observations; [2]NV: Number of variables; [3]PID: Power of interaction determinant; [4]NOZ: Number of overlay zones; [5]NOZP: Number of overlay zone pairs; [6]FDZ: Finely divided zones, which are overlapped zones with only one observation; [7]RSDZP: Ratio of significantly different (sig. < 0.05) zone pairs.

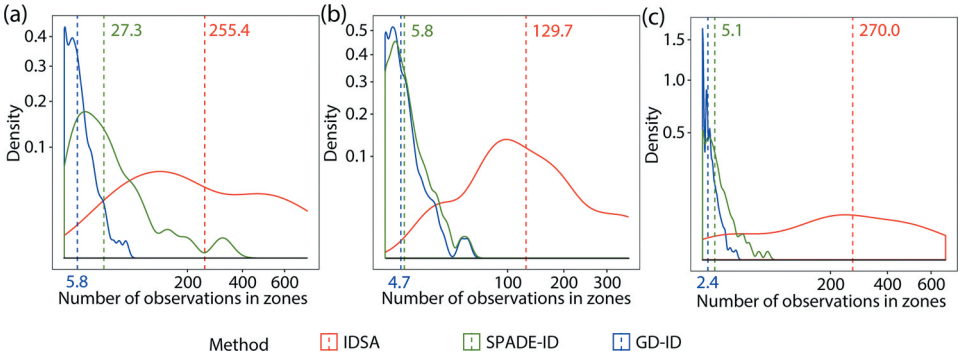

Figure 11. Density of observation numbers in the overlay zones in the northern (a), southern (b), and eastern (c) road networks.

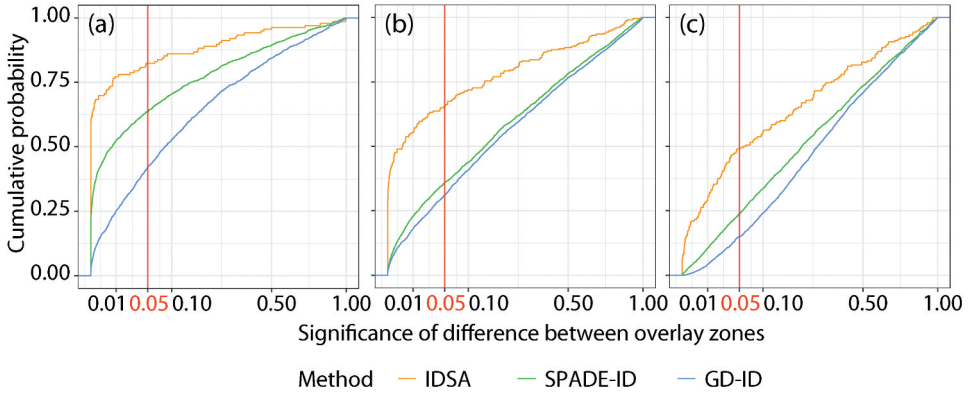

Figure 12. Significantly different probability distributions between the overlay zone pairs in the northern (a), southern (b) and eastern (c) road networks.

5. Discussion

This study proposes an IDSA model to examine spatial associations between a response variable and explanatory variables with interaction effects and presents a comprehensive investigative analysis of interaction effects of the variables based on the spatially stratified heterogeneity theory. The IDSA model had the following advantages in exploring factors or explanatory variables with interaction effects. First, the spatial fuzzy overlay process allowed for a more reasonable understanding of the physical interaction effects of geographical objects compared with that of spatial intersection process used in the GD-ID model. This is because the spatial fuzzy overlay could identify common geographical configurations with fuzzy membership functions and integrate the related uncertainty and vagueness information into the overlay, thereby achieving a more reasonable understanding of the interaction effects of the objects. Local geographical primary variables were identified

within the interaction of the variables, which was helpful for learning about the trade-offs of the geographical variables during spatial overlay. This concept is significantly important for understanding the nature and interaction of geographical variables. In recent research, diverse multi-variable regression models and machine learning algorithms have been developed to address geographical issues; however, the investigation of interaction of geographical variables has been limited. Therefore, the IDSA model provides a reasonable solution for characterizing the interaction of geographical variables for examining explanatory variables. It also provides an innovative concept of the interaction effects of variables that can be integrated into other models, such as machine learning algorithms and regression, Bayesian, and geostatistical models.

Additionally, IDSA is effective in dealing with the interaction effects of variables. For example, the IDSA-based optimal interaction of variables was able to use the least numbers of explanatory variables, spatial overlay zones, and finely divided zones to explain a response variable. Further, the overlay zones determined by the interaction of variables could identify greater heterogeneous patterns of the response variables, compared with the spatial intersection-based interaction detector models. These greater heterogeneous patterns were presented as indicators of a low percentage of finely divided zones and a high ratio of significantly different overlay zone pairs. This means that the interaction of variables derived from the IDSA model revealed more heterogeneity of the geographical variables than did those derived from the GD-ID or the SPADE-ID models.

This research did feature certain limitations. Firstly, the fuzzy AND operator was used for computing the spatial fuzzy overlay of the variables in terms of the physical processes of the interactions with the geographical variables. However, multiple other innovative fuzzy operators exist which could also be tested to improve accuracy. In this study, we tested a few other operators that have similar fuzzy AND characteristics, such as fuzzy sum, fuzzy product, and fuzzy gamma operators. However, the results were not reasonable from the perspective of physical interactions with the geographical attributes, and more estimation bias was present. Thus, we recommend that further innovative fuzzy operators should be developed to optimize the spatial fuzzy overlay process. Moreover, it is recommended to develop more effective methods for identifying the optimal interaction effects of variables based on the strategy in the IDSA model. Finally, we highly recommend that researchers attempt to integrate the IDSA model, or its concepts, with other models to examine spatial associations and explore explanatory variables in order to improve overall modeling accuracy and effectiveness. This includes machine learning algorithms and regression, Bayesian, and geostatistical models. It is also recommended to compare the results of variable exploration using the IDSA model with that of other models, such as machine learning algorithms and models developed with the considerations of statistical inference instead of geo-referencing, such as multivariable regression, nonlinear regression, and the maximum entropy model. Comparison studies should be performed for various types of spatial data, such as point-based, line segment-based, areal and raster data.

6. Conclusion

This study developed an IDSA model based on spatial heterogeneity in order to quantify the PID more accurately and improved the understanding of the interaction effects of geographical variables. Both a simulation and practical case study were utilized to showcase the calculation processes, results, and performance of the IDSA model. Results demonstrate that the IDSA model is a powerful tool for examining the spatial associations between a response variable and its associated explanatory variables with interaction effects. Compared with the GD model, which ignores characteristics of spatial dependence, the IDSA model characterizes the spatial autocorrelations of variables to more accurately assess the PD of individual variables. Further, the IDSA model uses spatial fuzzy overlay and the process of identifying optimal interaction effects of variables to then estimate the PID. Compared with GD-ID and SPADE-ID, the IDSA model could improve the accuracy and effectiveness of estimating the PID while requiring fewer variables of spatial interaction, a significantly reduced number of finely divided zones, substantially less interaction zones, and a higher ratio of significantly different zone pairs. We therefore highly recommended that the IDSA model, and its concepts, be applied and integrated into studies of spatial analysis.

Acknowledgments

We would like to thank Qindong Li, Brett Belstead and Tom McHugh from Main Roads Western Australia, Government of Western Australia, for providing their practical knowledge in road surface performance data application, practical decision making and policies development for road infrastructure asset management. We would like to thank Prof. May Yuan and anonymous reviewers for their constructive suggestion and comments for improving this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the Australian Government through the Australian Research Council's Discovery Early Career Researcher Award funding scheme (Project No. DE170101502) and Discovery Project (Project No. DP180104026).

Notes on contributors

Yongze Song is a Lecturer at Curtin University, Australia, and a Fellow of the Royal Geographical Society (with IBG), United Kingdom. His current research interests include geospatial analysis methods, spatial statistics, sustainable development, and infrastructure management.

Peng Wu is a Professor at Curtin University, Australia, and an Australian Research Council (ARC) DECRA Fellow. His research interests include sustainable construction, lean production and construction, production and operations management, and life cycle assessment.

ORCID

Yongze Song http://orcid.org/0000-0003-3420-9622
Peng Wu http://orcid.org/0000-0002-3793-0653

Data and codes availability statement

The simulation data and codes that support the findings of this study are available with a DOI at http://doi.org/10.6084/m9.figshare.13636157. The model code can be run in R using '*model.R*' with the simulation dataset '*simdata.rda*'. Data and codes used to create each figure are presented in the file '*Data and codes.pdf*', where data of Figures 2, 3 and 4 are computed with the simulation dataset '*simdata.rda*', data of Figures 6, 7, 9, 11 and 12 are presented in '*data.figure6.rda*', '*data.figure7.rda*', '*data.figure9.rda*', '*data.figure11.rda*' and '*data.figure12.rda*', respectively, and data of other figures are listed in tables in the file '*Data and codes.pdf*'. The raw road deterioration performance data that support the practical experiment of this study were provided by Main Roads Western Australia and cannot be made publicly due to data use restrictions.

References

Anselin, L., 2013. *Spatial econometrics: methods and models*. Netherlands: Springer Science & Business Media.

Anselin, L., Syabri, I., and Kho, Y., 2010. *GeoDa: an introduction to spatial data analysis. Handbook of applied spatial analysis*. Berlin, Heidelberg: Springer, 73–89.

Arbia, G., Griffith, D., and Haining, R., 1998. Error propagation modelling in raster GIS: overlay operations. *International Journal of Geographical Information Science*, 12 (2), 145–167. doi:10.1080/136588198241932.

Bai, H., *et al*. 2020. Incorporating spatial association into statistical classifiers: local pattern-based prior tuning. *International Journal of Geographical Information Science*, 34 (10), 2077–2114. doi:10.1080/13658816.2020.1737702.

Bai, L., *et al*., 2019. Quantifying the spatial heterogeneity influences of natural and socioeconomic factors and their interactions on air pollution using the geographical detector method: a case study of the Yangtze River Economic Belt, China. *Journal of Cleaner Production*, 232, 692–704. doi:10.1016/j.jclepro.2019.05.342.

Barco, S., *et al*. 2020. Trends in mortality related to pulmonary embolism in the European Region, 2000–15: analysis of vital registration data from the WHO Mortality Database. *The Lancet Respiratory Medicine*, 8 (3), 277–287. doi:10.1016/S2213-2600(19)30354-6.

Bureau of Meteorology Australian Government, 2019. Australian landscape water balance [online]. Available from: http://www.bom.gov.au/water/landscape [Accessed May 2019].

Bustince, H., *et al*. 2010. Overlap functions. *Nonlinear Analysis, Theory, Methods & Applications*, 72 (3–4), 1488–1499. doi:10.1016/j.na.2009.08.033.

Calonico, S., Cattaneo, M.D., and Farrell, M.H., 2018. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113 (522), 767–779. doi:10.1080/01621459.2017.1285776.

Cang, X. and Luo, W., 2018. Spatial association detector (SPADE). *International Journal of Geographical Information Science*, 32 (10), 2055–2075. doi:10.1080/13658816.2018.1476693.

Cao, F., Ge, Y., and Wang, J., 2013. Optimal discretization for geographical detectors-based risk assessment. *GIScience & Remote Sensing*, 50 (1), 78–92. doi:10.1080/15481603.2013.778562.

Cao, F., Ge, Y., and Wang, J., 2014. Spatial data discretization methods for geocomputation. *International Journal of Applied Earth Observation and Geoinformation*, 26, 432–440. doi:10.1016/j.jag.2013.09.005

Ding, Y., *et al*., 2019. Using the geographical detector technique to explore the impact of socio-economic factors on PM2.5 concentrations in China. *Journal of Cleaner Production*, 211, 1480–1490. doi:10.1016/j.jclepro.2018.11.159.

Dubois, D. and Prade, H., 1983. Ranking fuzzy numbers in the setting of possibility theory. *Information Sciences*, 30 (3), 183–224. doi:10.1016/0020-0255(83)90025-7.

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2003. *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester, UK: John Wiley & Sons.

Fotheringham, A.S., Charlton, M.E., and Brunsdon, C., 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment & Planning A*, 30 (11), 1905–1927. doi:10.1068/a301905.

Fotheringham, A.S. and Rogerson, P., 2013. *Spatial analysis and GIS*. Boca Raton, FL: CRC Press.

Fotheringham, A.S., Yang, W., and Kang, W., 2017. Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, 107 (6), 1247–1265. doi:10.1080/24694452.2017.1352480.

Garrigues, S., *et al*. 2006. Quantifying spatial heterogeneity at the landscape scale using variogram models. *Remote Sensing of Environment*, 103 (1), 81–96. doi:10.1016/j.rse.2006.03.013.

Goodchild, M., Haining, R., and Wise, S., 1992. Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems*, 6 (5), 407–423. doi:10.1080/02693799208901923.

Goodchild, M.F., Yuan, M., and Cova, T.J., 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21 (3), 239–260. doi:10.1080/13658810600965271.

Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. New York, NY: Oxford University Press on Demand.

Guo, P., *et al*., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. *Geoderma*, 237-238, 49–59. doi:10.1016/j.geoderma.2014.08.009.

Haining, R., 2003. *Spatial data analysis: theory and practice*. London, UK: Cambridge University Press.

Haining, R., 2015a. Spatial Autocorrelation. *In*: J.D. Wright, ed. *International Encyclopedia of the Social & Behavioral Sciences*. 2nd Edn ed. Oxford: Elsevier, 105–110.

Haining, R., 2015b. Spatial Sampling. *In*: J.D. Wright, ed. *International encyclopedia of the social & behavioral sciences*. 2nd Edn ed. Oxford: Elsevier, 185–190.

Hengl, T., Heuvelink, G.B., and Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120 (1–2), 75–93. doi:10.1016/j.geoderma.2003.08.018.

Huang, B., Wu, B., and Barry, M., 2010. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24 (3), 383–401. doi:10.1080/13658810802672469.

Jacoby, W.G., 2000. Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19 (4), 577–613. doi:10.1016/S0261-3794(99)00028-1.

Johnston, R., *et al*. 2003. ASRIS: the database. *Australian Journal of Soil Research*, 41 (6), 1021–1036. doi:10.1071/SR02033.

Ju, H., *et al*. 2016. Driving forces and their interactions of built-up land expansion based on the geographical detector – a case study of Beijing, China. *International Journal of Geographical Information Science*, 30 (11), 2188–2207. doi:10.1080/13658816.2016.1165228.

Kammann, E.E. and Wand, M.P., 2003. Geoadditive models. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 52 (1), 1–18.

Kim, Y.-H., Choe, K.-U., and Ri, R.-K., 2019. Application of fuzzy logic and geometric average: a Cu sulfide deposits potential mapping case study from Kapsan Basin, DPR Korea. *Ore Geology Reviews*, 107, 239–247. doi:10.1016/j.oregeorev.2019.02.026

Lavrakas, P.J., 2008. *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage Publications.

Li, J., *et al*. 2011. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling and Software*, 26 (12), 1647–1659. doi:10.1016/j.envsoft.2011.07.004.

Merchant, C.J., Saux-Picart, S., and Waller, J., 2020. Bias correction and covariance parameters for optimal estimation by exploiting matched in-situ references. *Remote Sensing of Environment*, 237, 111590. doi:10.1016/j.rse.2019.111590

Nakaya, T., *et al*. 2005. Geographically weighted poisson regression for disease association mapping. *Statistics in Medicine*, 24 (17), 2695–2717. doi:10.1002/sim.2129.

Phillips, T., *et al*. 2011. Modeling moulin distribution on Sermeq Avannarleq glacier using ASTER and WorldView imagery and fuzzy set theory. *Remote Sensing of Environment*, 115 (9), 2292–2301. doi:10.1016/j.rse.2011.04.029.

Qiao, P., *et al*., 2019. Quantitative analysis of the factors influencing spatial distribution of soil heavy metals based on geographical detector. *Science of the Total Environment*, 664, 392–413. doi:10.1016/j.scitotenv.2019.01.310.

Ren, Y., *et al*. 2014. Geographical modeling of spatial interaction between human activity and forest connectivity in an urban landscape of southeast China. *Landscape Ecology*, 29 (10), 1741–1758. doi:10.1007/s10980-014-0094-z.

Shi, W., Cheung, C.-K., and Tong, X., 2004. Modelling error propagation in vector-based overlay analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59 (1–2), 47–59. doi:10.1016/j.isprsjprs.2004.05.003.

Song, Y., *et al*. 2018. Segment-based spatial analysis for assessing road infrastructure performance using monitoring observations and remote sensing data. *Remote Sensing*, 10 (11), 1696. doi:10.3390/rs10111696.

Song, Y., *et al*. 2019. Traffic volume prediction with segment-based regression kriging and its implementation in assessing the impact of heavy vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 20 (1), 232–243. doi:10.1109/TITS.2018.2805817.

Song, Y., *et al*. 2020a. 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 (5), 593–610. doi:10.1080/15481603.2020.1760434.

Song, Y., *et al*. 2020b. A spatial heterogeneity-based segmentation model for analyzing road deterioration network data in multi-scale infrastructure systems. *IEEE Transactions on Intelligent Transportation Systems*, 1–11, Early Access. doi:10.1109/TITS.2020.3001193

The Government of Western Australia, Main Roads Western Australia and ARRB Australia Road Research Board (ARRB), 2018. *An evaluation of the Traffic Speed Deflectometer (TSD) for Main Roads Western Australia*. Leederville: ARRB.

Wan, Z., Hook, S., and Hulley, G., 2015. MOD11A2 MODIS/Terra land surface temperature/emissivity 8-day L3 global 1km SIN Grid V006. *NASA EOSDIS Land Processes DAAC*, 10. https://doi.org/10.5067/MODIS/MOD11A2.006

Wang, J., *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 (1), 107–127. doi:10.1080/13658810802443457.

Wang, J., Zhang, T., and Fu, B., 2016. A measure of spatial stratified heterogeneity. *Ecological Indicators*, 67, 250–256. doi:10.1016/j.ecolind.2016.02.052

Wang, L. and Chen, L., 2018. The impact of new transportation modes on population distribution in Jing-Jin-Ji region of China. *Scientific Data*, 5 (1), 170204. doi:10.1038/sdata.2017.204.

Wang, S., *et al*., 2015. An effective algorithm for lines and polygons overlay analysis using uniform spatial grid indexing. In: 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), Fuzhou, China. IEEE, 175–179.

Yu, H., *et al*. 2019a. Inference in multiscale geographically weighted regression. *Geographical Analysis*, 52 (1), 87–106. doi:10.1111/gean.12189.

Yu, H. and Peng, Z.-R., 2019. Exploring the spatial variation of ridesourcing demand and its relationship to built environment and socioeconomic factors with the geographically weighted poisson regression. *Journal of Transport Geography*, 75, 147–163. doi:10.1016/j.jtrangeo.2019.01.004

Yu, Q., Liu, D., and Wang, S., 2018. A fuzzy overlay analysis model for raster map layers. *Journal of Image and Graphics*, 9 (7), 832–836.

Yu, S., *et al*., 2019b. Estimation and inference for generalized geoadditive models. *Journal of the American Statistical Association*, 15 (530), 761–774.

Zadeh, L.A., 1973. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3 (1), 28–44. doi:10.1109/TSMC.1973.5408575.

Zhao, Y., *et al*., 2020. Cadmium source identification in soils and high-risk regions predicted by geographical detector method. *Environmental Pollution*, 263, 114338. doi:10.1016/j.envpol.2020.114338.

Zhou, C., *et al*. 2020. The contribution rate of driving factors and their interactions to temperature in the Yangtze River Delta region. *Atmosphere*, 11 (1), 32. doi:10.3390/atmos11010032.