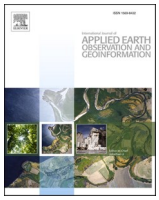
International Journal of Applied Earth Observations and Geoinformation 111 (2022) 102834

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



The second dimension of spatial association

Yongze Song

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

ARTICLE INFO

Keywords:
Spatial association
Spatial statistics
Spatial prediction
Machine learning
Geochemical mapping
Geographical characteristics

ABSTRACT

A reasonable and adequate understanding of spatial association between geographical variables is the basis of spatial statistical inference and geocomputation, such as spatial prediction. Most of the current models for exploring spatial association of variables are constructed using data at sample locations. In this study, approaches for exploring spatial association using observations at sample locations are defined as the first dimension of spatial association (FDA). However, geographical information outside sample locations is usually missing in current models. To address this issue, this study proposes the concept of the second dimension of spatial association (SDA), which is an approach that extracts geographical information at locations outside samples for exploring spatial association. Based on the concept of SDA, three SDA models, including SDA-based multivariate linear regression, machine learning (i.e., random forest), and geostatistical models (i.e., random forest kriging), are developed for examining spatial association and predicting spatial distributions of trace elements, including Cr and Cu, in a mining region in Western Australia. Model accuracy is evaluated by comparing with corresponding FDA models. A new R package "SecDim" is developed to conduct SDA models. Results show that SDA models have a series of advantages in examining spatial association compared with FDA models. First, the accuracy of spatial prediction can be critically improved by SDA compared with the FDA, although identical explanatory variables and models are used for the modeling and prediction. Second, SDA can effectively indicate the multi-scale effects and diverse information within explanatory variables of the geographical environment at local ranges using the second dimension variables. Third, SDA can avoid underestimating high values and overestimating low values in the general FDA-based statistical models, machine learning, and geostatistical models. Finally, SDA models provide more smooth spatial predictions across space than that predicted by FDA models and avoid massive fluctuations at local ranges. The concept of SDA provides new insight into geographical information-based spatial association. SDA and multiple types of SDA models have great potential for more accurate and effective spatial statistical inference and geocomputation in diverse fields.

1. Introduction

Spatial statistical inference is an essential approach for addressing issues of the natural and social environment using geospatial data (Haining, 1993; Wang et al., 2020; Gao et al., 2022). Spatial statistical inference is the process of statistical modeling for spatial data based on geographical characteristics (Jacquez, 1999; Cressie and Moores, 2021). It has been widely implemented in spatial analysis and geocomputation, such as assessing spatial patterns, identifying spatial clustering regions, exploring factors, spatial predictions, and geographical decision making (Hoef et al., 2018; Song et al., 2021b). A reasonable understanding of spatial association between geographical variables is a basis of effective spatial statistical inference.

In most current models, spatial association is explored using data at sample locations or spatial units. In this study, approaches for exploring spatial association using observations at sample locations or spatial units are defined as the first dimension of spatial association (FDA). FDA methods have been widely applied in research and practice due to the advantages of easy access to data from identical samples and the availability of multiple types of models. The current FDA models for exploring spatial association can be classified into the following categories. The first category is aspatial models, where the most commonly used approaches include statistical models and machine learning algorithms. Compared with statistical models, machine learning algorithms can deal with relatively complex data and have high modeling accuracy (Chen et al., 2017; Hengl et al., 2018). The primary characteristic of aspatial models for exploring spatial association is that the relationships between response and explanatory variables are directly constructed using observation data at the exact corresponding locations. The geographical characteristics of data and location information are usually

E-mail address: yongze.song@curtin.edu.au.

https://doi.org/10.1016/j.jag.2022.102834
Received 27 March 2022; Received in revised form 28 April 2022; Accepted 19 May 2022
Available online 28 May 2022

ignored in the models.

The second category is models developed based on spatial dependence of data, such as geographically weighted regression and spatial Bayesian hierarchical models (Haining and Haining, 2003). In the models, spatial association is constructed between data of a response variable at sample locations and data of explanatory variables at surrounding sample locations using various types of space-weighted matrices. The models can effectively depict the spatial non-stationarity of geographical variables using locally varied relationships between response and explanatory variables (Brunsdon et al., 1996) The third category includes models of spatial heterogeneity characteristics. The examples consist of geostatistical models (Krige, 1951; Goovaerts et al., 1997), i.e., kriging family models, the geo-additive model (Kammann and Wand, 2003), and spatial stratified heterogeneity models (Wang et al., 2010; Song et al., 2020). For instance, kriging models quantify the spatial variations of response variable data using variogram functions. Spatial stratified heterogeneity models, such as the interactive detector of spatial association (IDSA) (Song and Wu, 2021) and geographically optimal zones-based heterogeneity (GOZH) (Luo et al., 2022), assess the spatial disparity of response variables using the spatial zones determined by explanatory variables. However, geographical information outside sample locations or spatial units of observations is generally not included in the above models.

Geographical data across the space of a study area contains essential information on geographical characteristics (Walford, 2002; Song et al., 2020). Therefore, it is necessary to develop models to extract diverse geographical information outside sample locations and spatial units of observations. Only a few models have been developed to demonstrate the importance of geographical information outside sample locations in modeling spatial association, such as land use regression and geographical similarity-based approaches. In land-use regression, percentages of different types of land use within buffers of certain distances from sample locations are generally used as explanatory variables for spatial modeling (Hoek et al., 2008; Xu et al., 2019). In geographical similarity-based models, spatial association is investigated using data of explanatory variables at both observation and unknown locations (Zhu et al., 1997; Zhu et al., 2018). However, models for extracting diverse geographical information outside sample locations are still limited, and more effective extraction of geographical information across the whole space is still increasingly required.

To address the above issues, this study proposes the concept of the second dimension of spatial association (SDA), which is an approach that extracts in-depth information about the geographical environments from locations outside sample points or spatial units of observations for exploring spatial association. Based on the concept of SDA, the second dimension models (SDMs), including SDA-based multivariate linear regression (MLR), random forest (RF), and random forest kriging (RFK), i.e., an integration between RF and regression kriging, are developed for examining spatial association by extracting more information about the geographical environment outside sampling locations. The SDA-MLR, SDA-RF, and SDA-RFK represent typical SDA-based statistical models, machine learning algorithms, and spatial machine learning or machine learning-based geostatistics (Hengl et al., 2018), respectively. The SDMs are implemented in spatial predictions of trace elements, including Cr and Cu, in a mining region in Western Australia. The model accuracy, uncertainty, and effectiveness of SDA models in spatial predictions are evaluated by comparing them with corresponding FDA models.

The remainder of this article is arranged as follows. Section 2 presents the concepts of FDA and SDA and the detailed computation and analysis steps of SDMs. Section 3 presents the practical application, including study area and data, and experiment design, of predicting spatial distributions of trace elements in a mining region in Western Australia. Section 4 shows SDA-based results of spatial predictions of trace elements and model validation. Section 5 discusses the findings from the study and the advantages and contributions of SDA for examining spatial association, and Section 6 concludes this study.

2. Method

2.1. Concept of the second dimension of spatial association (SDA)

This study proposes the concept of the second dimension of spatial association (SDA), which corresponds to the idea of the first dimension of spatial association (FDA). Fig. 1 shows a comparison of a few typical scenarios of FDA and SDA. As explained in Section 1, in this study, FDA is defined as an approach that explores spatial association using observations at sample locations or spatial units. Most of the current models for exploring spatial association are FDA models. For instance, MLR and machine learning construct relationships between a response variable and explanatory variables at the exact sample locations, and geographically weighted regression explores relationships between a response variable at a given location and explanatory variables at its surrounding sample locations.

In this study, SDA is an approach that extracts in-depth information about the geographical environment from locations outside sample points or spatial units of observations to explore spatial association. Fig. 1 shows two typical scenarios of SDA. One scenario is that all data of explanatory variables within a specific range from sample locations will be used for modeling. The other scenario is that the data of explanatory variables outside sample locations will be selected using reasonable strategies, and a part of the data will be used for modeling. The new variables computed using the data of explanatory variables outside sample locations are named the second dimension variables. Different strategies can be developed to generate the second dimension variables to extract the in-depth geographical information from data. An explanatory variable can create a series of second dimension variables. The above characteristics of SDA demonstrate SDA's flexibility and great potential for dealing with multiple types of spatial data. In this study, the second dimension models (SDMs) are developed based on searching range and probability parameters, which are explained in the following section, to generate the second dimension variables and explore spatial association.

2.2. SDA models

This study developed SDMs to explore spatial association based on the concept of SDA. Fig. 2 shows a schematic overview of SDMs, which includes the following four stages. The first stage characterizes the geographical environment using relatively high-resolution data of explanatory variables. For instance, the objective of a study is spatial prediction at 1-km resolution using ground samples, and the collected explanatory variables across space have different spatial resolutions, such as 30 m for elevation data and 250 m for vegetation coverage data. In FDA models, explanatory variables with different spatial resolutions are generally re-sampled to a 1-km solution for consistent modeling and prediction. However, massive geographical information is lost during re-sampling, and diverse geographical characteristics within the 1-km grids and surrounding sample locations are ignored. In SDMs, explanatory variables with different spatial resolutions can be directly used in modeling. In practical solutions, a part of the high-resolution data can be re-sampled to course resolutions to reduce computation time and resources. Still, it is recommended that the resolutions of explanatory variables should be higher than the resolution of spatial predictions. In the above example, the resolutions of explanatory variables should be higher than 1 km, such as 100 m, 250 m, or 500 m.

The second stage is generating the second dimension variables using proper strategies. In SDMs, a searching range and probability-based approach is developed to generate the second dimension variables. For an explanatory variable X_i, data within a set of distances from sample locations are selected using searching range parameters b values, such as $b = 0.1$, 0.2, 0.5, 1, 2, and 5 km. Next, a quantile operator is used to evaluate the statistical distribution characteristic of data within a searching range from a sample location, where a series of probability

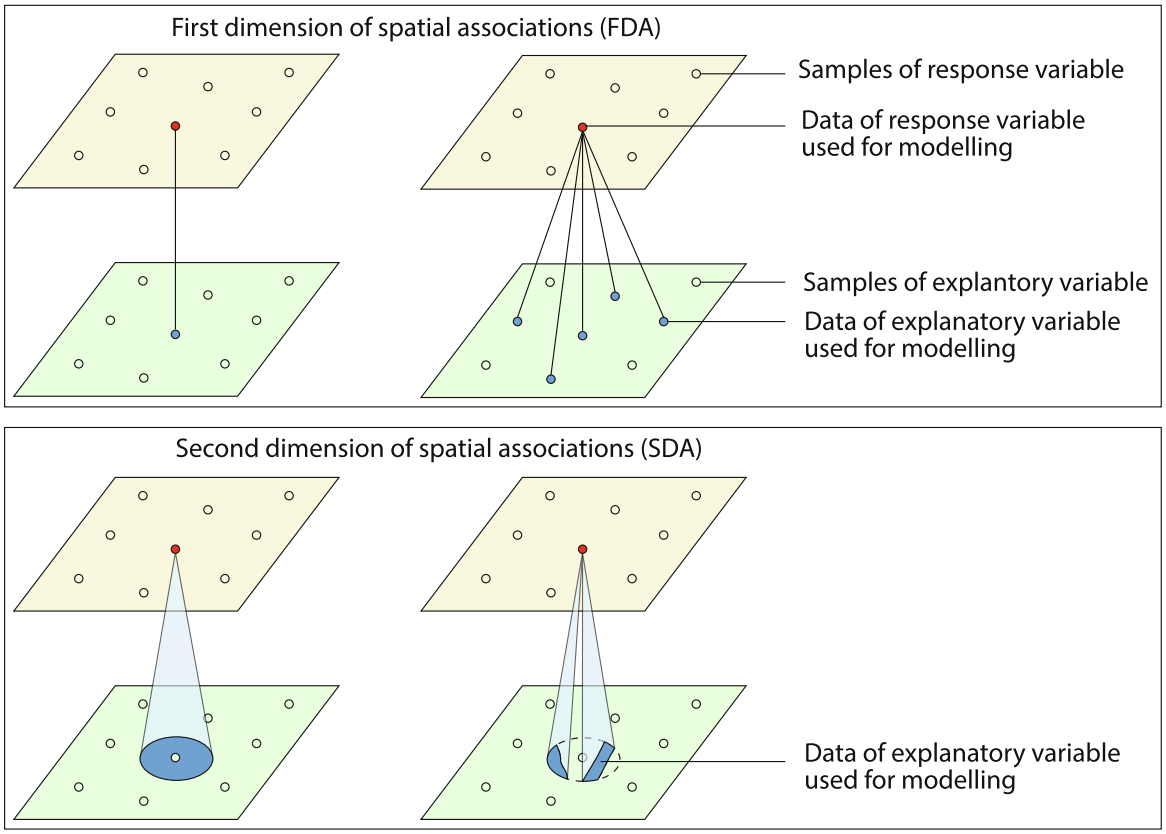


Fig. 1. Comparison of concepts of the first and the second dimensions of spatial association (FDA and SDA).

parameters τ are used in the quantile operator, such as $\tau = 0, 0.1, \ldots, 0.9$, and 1. If $\tau = 0$, 0.5, and 1, the second dimension variables are the minimum, median, and maximum data values within a searching range, respectively. A series of the second dimension variables are generated using the combinations of searching range and probability parameters. For an explanatory variable X_i, $X_i \in \mathbf{X}$, the corresponding second dimension variables can be recorded as $X_i(\mathbf{b}, \boldsymbol{\tau})$. For instance, if the approach uses 10 b values and 10 τ values, as a result, $X_i(\mathbf{b}, \boldsymbol{\tau})$ includes 100 variables.

The third stage is selecting the second dimension variables. In statistical models, variable selection can be performed using various methods, such as step-wise regression, correlation analysis, and principal component analysis. In SDMs, a correlation analysis and multicollinearity analysis-based approach is used for the variable selection due to the potential high multicollinearity among $X_i(\mathbf{b}, \boldsymbol{\tau})$, which consists of the following three steps. First, correlation analysis is performed for the second dimension variables of individual explanatory variables. Variables with significant correlations are selected and ordered according to the absolute values of correlation coefficients. The selected and ordered the second dimension variables are recorded as $X_i'(\mathbf{b}, \boldsymbol{\tau})$. Second, multicollinearity analysis with a measure of variance inflation factor (VIF) is used to remove multicollinearity within $X_i'(\mathbf{b}, \boldsymbol{\tau})$. The highest absolute correlation coefficient variable is the first variable in regression modeling. Then, variables from the ordered list are sequentially added to the first variable, and a linear regression is performed. If the VIF values of variables are lower than 10, the variable is selected, but if some of the VIF values are higher than 10, variables with the highest VIF are removed (Song et al., 2021a). This process is performed sequentially until all variables are tested. The VIF values of the selected variables $X_i'(\mathbf{b_j}, \boldsymbol{\tau}_k)$ are all lower than 10, meaning that they are uncorrelated. Finally, the second dimension variables $X_i'(\mathbf{b_j}, \boldsymbol{\tau}_k)$ from multiple explanatory variables are further selected using the multicollinearity analysis. The process of multicollinearity analysis is similar to the above step. The final selected variables are $\mathbf{X}'(\mathbf{b_p}, \boldsymbol{\tau}_q)$, where $\mathbf{X}' \subset \mathbf{X}, \mathbf{b_p} \subset \mathbf{b}, \boldsymbol{\tau}_q \subset \boldsymbol{\tau}$, variables are significantly correlated with the response variable, and $\mathbf{X}'(\mathbf{b_p}, \boldsymbol{\tau}_q)$ are uncorrelated.

The last stage of SDMs is modeling and spatial prediction. In this study, the SDMs include SDA-MLR, SDA-RF, and SDA-RFK, representing SDA-based statistical models, machine learning algorithms, and spatial machine learning or machine learning-based geostatistics (Hengl et al., 2018), respectively. The SDMs can be presented as:

$$Y(\mathbf{u}) = f(\mathbf{X}'(\mathbf{v}, \mathbf{b_p}, \boldsymbol{\tau}_q)) \quad (1)$$

where $Y(\mathbf{u})$ is the observations at sample locations $\mathbf{u}$, $\mathbf{X}'(\mathbf{v}, \mathbf{b_p}, \boldsymbol{\tau}_q)$ is the selected second dimension variables at locations $\mathbf{v}$, which are outside sample locations $\mathbf{u}$, and determined by the searching range parameter $\mathbf{b_p}$ and probability parameter $\boldsymbol{\tau}_q$, and f is the relationship function. A new R package "SecDim" (The Second Dimension for Spatial Association) is developed to performe SDA analysis (https://cran.r-project.org/web/packages/SecDim/vignettes/SecDim.html). In the geostatistical model, i.e., SDA-RFK, prediction uncertainty is computed as:

$$\delta(\mathbf{u}) = \frac{\sigma_{\widehat{Y}}(\mathbf{u})}{\widehat{Y}(\mathbf{u})} \quad (2)$$

where $\delta(\mathbf{u})$ is the prediction error uncertainty, and $\sigma_{\widehat{Y}}(\mathbf{u})$ is the square root of the variance of the kriging prediction.

3. Application: spatial prediction of trace elements

3.1. Study area and data

In this study, SDMs are implemented in predicting spatial distributions of trace elements, including Cr and Cu, in a mining region in the

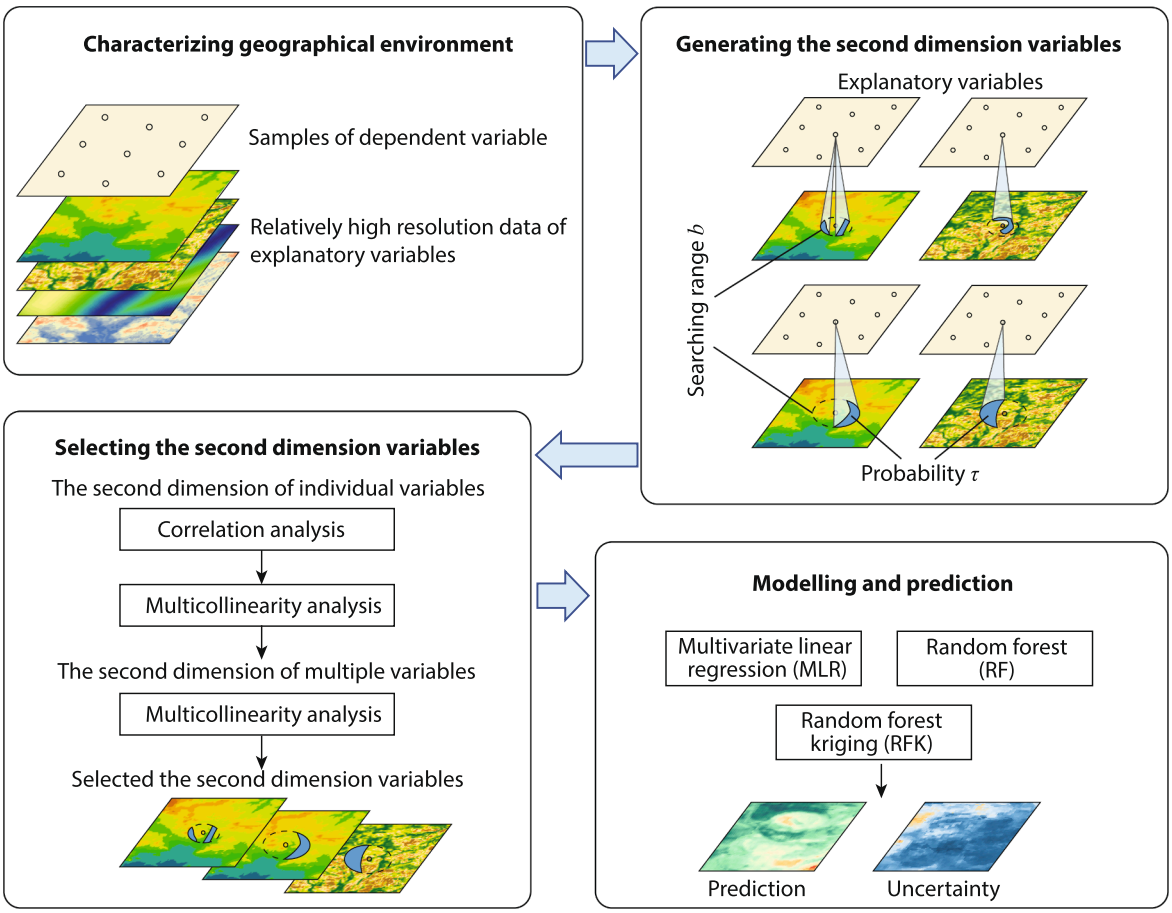


Fig. 2. Schematic overview of the second dimension model (SDM) for assessing spatial association.

western Meekatharra Local Government Area (LGA) in Western Australia. Fig. 3 a shows the location and basic geographical environments, such as elevation, rivers, lakes, and roads, of the study area. The total area of this region is about 10.215 km^2. Fig. 3 b and c show the spatial distributions of Cr and Cu samples, respectively. The trace element samples are collected from the Geological Survey of Western Australia (GSWA) Geochemistry data (Department of Mines, 2022). More details of the GSWA data set can be found in the studies about the descriptions and applications of the data set (Morris et al., 1998; Wells et al., 2016; Morin-Ka et al., 2019). This study averages repeated observations of trace element data of Cr and Cu in the study area at the same sample sites. As a result, trace elements at 614 sample locations are collected. The maps of trace elements show that the high-concentration Cr samples are primarily distributed in the south-eastern regions, and low-concentration samples are in the northern and western areas. The high-concentration Cu samples are located in the northern and central regions, and the low-concentration samples are in the other areas. In addition, Table 1 shows a statistical summary of trace element observations. The mean concentrations of Cr and Cu are 190.8 ppm and 50.8 ppm, respectively. Statistical summaries also indicate that the observations of Cr and Cu both show right-skewed distributions.

3.2. Experiment design

The application of SDMs in predicting spatial distributions of trace elements is designed as follows. The expected spatial resolution of trace element predictions is 1 km across the study area. First, explanatory variables are collected from remote sensing data and geospatial analysis to characterize the geographical environment. Second, SDMs are implemented in modeling relationships between trace elements and data of explanatory variables outside sample locations. Third, the modeling accuracy is evaluated using cross-validation and compared with FDA models. Finally, the relationships between trace elements and data of explanatory variables outside sample locations explored by SDMs are used for spatial prediction, and the prediction uncertainty and effectiveness are evaluated. The above steps are explained in the following paragraphs.

First, nine explanatory variables are collected for characterizing the geographical environment in this study, including elevation, slope, aspect, distance to water, vegetation coverage, soil organic carbon (SOC), pH, and distance to roads (see Table 2). Among the variables, data of the terrain-related variables, including elevation, slope, and aspect, are derived from the Australian Smoothed Digital Elevation Model (Geosciences Australia, 2015) and processed using Google Earth Engine (GEE) (Gorelick et al., 2017). The spatial resolution of the terrain-related variables is about 30 m. The distance to water is computed as the distances to rivers and lakes in the study area, where rivers and lakes data are derived from the GEODATA TOPO 250 K Series 3 data set (Geoscience Australia, 2006). The annual mean normalized difference vegetation index (NDVI) in 2020 is used to describe the vegetation coverage, and it is sourced from the 250-m resolution product of the Moderate Resolution Imaging Spectroradiometer (MODIS) Terra Vegetation Indices (MOD13Q1.006) using GEE (Didan, 2015). The soil property is characterized using SOC and pH data of the Soil and Landscape Grid of Australia product with about 93 m resolution using GEE (Rossel et al., 2014). The distance to roads is calculated using the road network data in the open data portal of Main Roads Western Australia (Main Roads Western Australia, 2020). In SDMs, raster data with original resolutions can be directly used for modeling. In this study, to reduce the computation time and resources, data of all explanatory

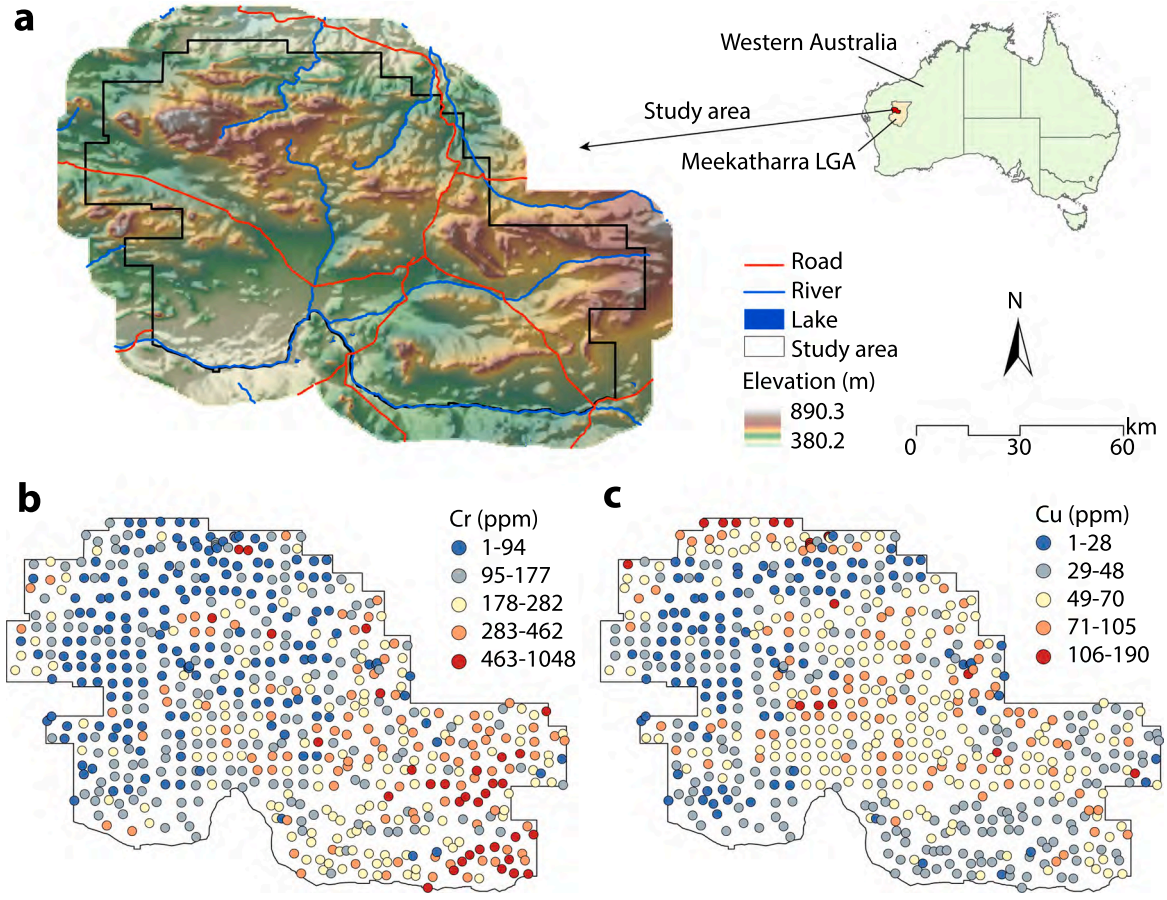


Fig. 3. Study area (a) and spatial distributions of trace element data, Cr (b) and Cu (c).

Table 1
Statistical summary of trace element sampling observations.

Trace element	Mean	Min	1st Qu.	Median	3rd Qu.	Max	SD	CV
Cr (ppm)	190.8	1	85	155	262	1048	143.0	0.750
Cu (ppm)	50.8	1	33	49	64	190	25.3	0.497

SD: Standard deviation; CV: Coefficient of variation; 1st Qu.: The first quartile; 3rd Qu.: The third quartile.

Table 2
Statistical summary of explanatory variables.

Variable	Code	Mean	Min	Median	Max	SD	CV
Elevation (m)	Elevation	518.8	380.2	514.4	890.3	58.9	0.114
Slope	Slope	0.847	0.009	0.515	14.460	1.050	1.241
Aspect (°)	Aspect	192.4	0.9	197.8	359.1	88.0	0.457
Distance to water (km)	Water	8.361	0.000	7.363	28.680	5.885	0.704
Vegetation coverage (NDVI)	NDVI	1.834	0.503	1.765	4.510	0.406	0.221
Soil organic carbon (SOC)	SOC	0.780	0.569	0.779	1.066	0.058	0.074
Soil pH	pH	5.515	5.040	5.496	6.178	0.177	0.032
Distance to roads (km)	Roads	11.811	0.000	10.583	38.269	8.278	0.701

SD: Standard deviation; CV: Coefficient of variation.

variables are re-sampled to 500-m resolution, higher than the 1-km resolution of the expected spatial predictions. Data of explanatory variables at locations within the 10-km range of regions outside the study area are also collected for the second dimension modeling.

Next, SDMs are implemented in modeling relationships between trace elements and data of explanatory variables outside sample locations. The details of SDMs are presented in Section 2. The trace elements data are transformed using a logarithm function due to the skewed statistical distributions, and outliers are removed for reliable modeling using the threshold of mean plus or minus 2.5 times of the standard deviation of data (Song et al., 2021a; Wu and Song, 2022). To generate the second dimension variables, the optional searching ranges b values are 1, 3, 5, 7, and 9 km, and the optional probability τ values are a sequential number from 0 to 1 with an interval of 0.1. This means that 55 the second dimension variables are generated for each variable. Then, correlation and multicollinearity-based variable selection strategies are used for variable selection. The selected variables are used for constructing their relationships with trace elements.

Third, the modeling accuracy is evaluated using cross-validation compared with FDA models. In the cross-validation, the observations are randomly divided into two parts, where 70% of the data are used as the training data, and the remained 30% of the data are used as the validation data. The observations and predictions of the validation data are compared to evaluate the modeling accuracy. The modeling accuracy of SDA-MLR, SDA-RF, and SDA-RFK, are compared with the corresponding accuracy of FDA-MLR, FDA-RF, and FDA-RFK, respectively. The cross-validation indicators include the coefficient of determination R^2, the root mean square error (RMSE), and mean absolute error (MAE). The computation equations are:

$$R^2 = 1 - \frac{\sum (Y_m - \widehat{Y}_m)^2}{\sum Y_m - \overline{Y}} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum (\widehat{Y}_m - Y_m)^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum |\widehat{Y}_m - Y_m| \tag{5}$$

where Y_m ($m = 1,\ldots,N$) is an observation of the response variable, $\widehat{Y}_m$ is the corresponding prediction, and $\overline{Y}$ is the mean value of all observations.

Finally, SDMs are used for predicting spatial distributions of trace elements, and the prediction uncertainties are evaluated. The spatial pattern characteristics of prediction derived by SDMs are compared with that derived using FDA models. To evaluate the difference in local details in the SDA and FDA-based predictions, prediction values along a cross-section in the study area are compared. For RFK models, in addition to predictions, uncertainties of kriging predictions are computed to assess the model performance of FDA-RFK and SDA-RFK. Let the kriging prediction uncertainty of FDA-RFK and SDA-RFK be δ_F and δ_S, respectively. The difference of prediction uncertainty between FDA-RFK and SDA-RFK models is calculated as:

$$\Delta = \delta_F - \delta_S \tag{6}$$

where Δ is the difference of uncertainty, thus, $\Delta > 0$ means that the uncertainty of FDA-RFK-based prediction is higher than that of SDA-RFK-based prediction, and vice versa.

4. Results

4.1. Characterizing geographical environment

Fig. 4 shows the spatial distribution of explanatory variables in the study area and within the 10-km distance range outside the study area. The explanatory variables characterize the geographical environment outside sample locations and derive the second dimension variables.

4.2. Variable selection and SDA modeling

Fig. 5 shows the selected second dimension variables for predicting distributions of Cr and Cu. Ten of the second dimension variables of explanatory variables, distance to water, NDVI, pH, SOC, and distance to roads, are selected for the SDA-based modeling and prediction of Cr. The corresponding SDMs can be presented as:

$$\begin{aligned} Y = f(& X_{pH}^{b=9,\tau=0.9}, X_{Water}^{b=9,\tau=1}, X_{pH}^{b=9,\tau=0.7}, X_{NDVI}^{b=5,\tau=0.1}, X_{NDVI}^{b=7,\tau=0.4}, \\ & X_{NDVI}^{b=9,\tau=0.1}, X_{pH}^{b=3,\tau=0.9}, X_{SOC}^{b=9,\tau=1}, X_{Road}^{b=9,\tau=0.8}, X_{pH}^{b=1,\tau=1}) \end{aligned} \tag{7}$$

In the equation, the second dimension variables are ordered according to the relative importance values in the RF model. The relative importance in RF models is measured by the percentage increase in mean square error (%IncMSE), a commonly used accuracy-based importance indicator, computed with the randomly shuffled values of the out-of-bag

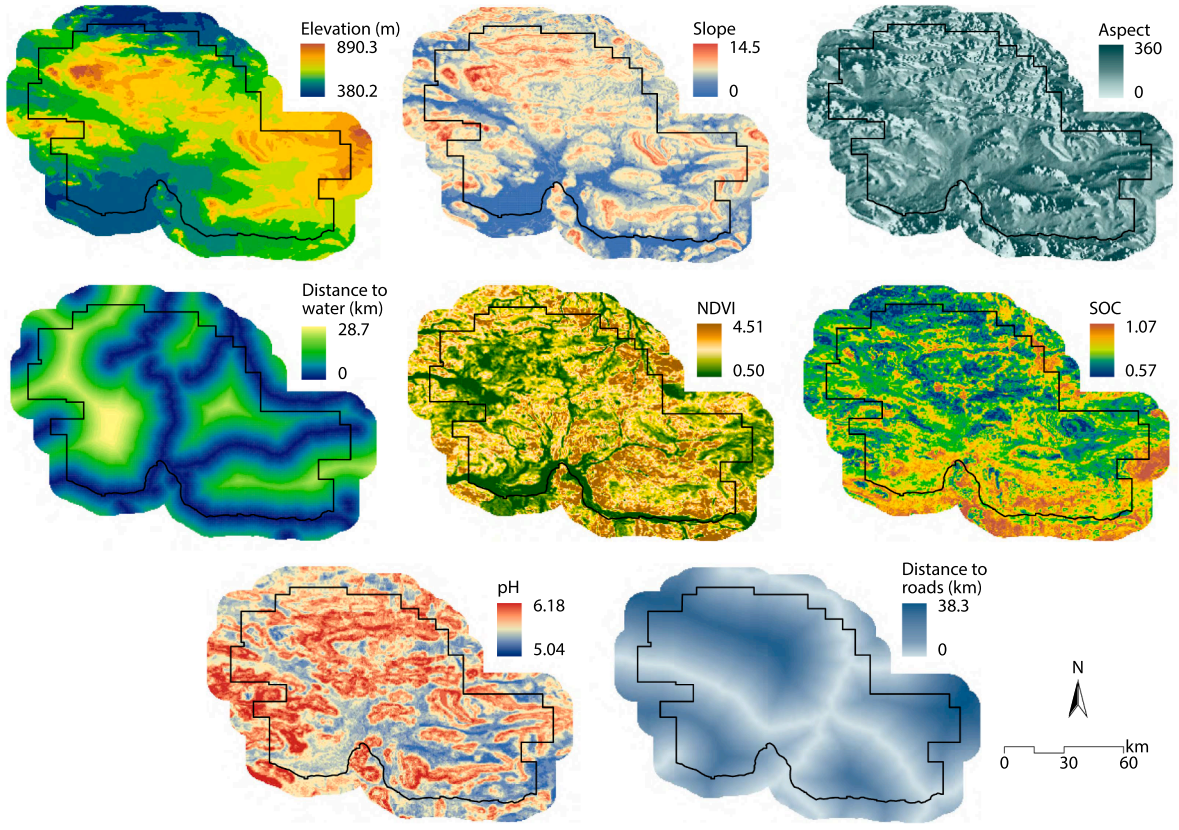


Fig. 4. Spatial distributions of explanatory variables for characterizing geographical environment in the study area.

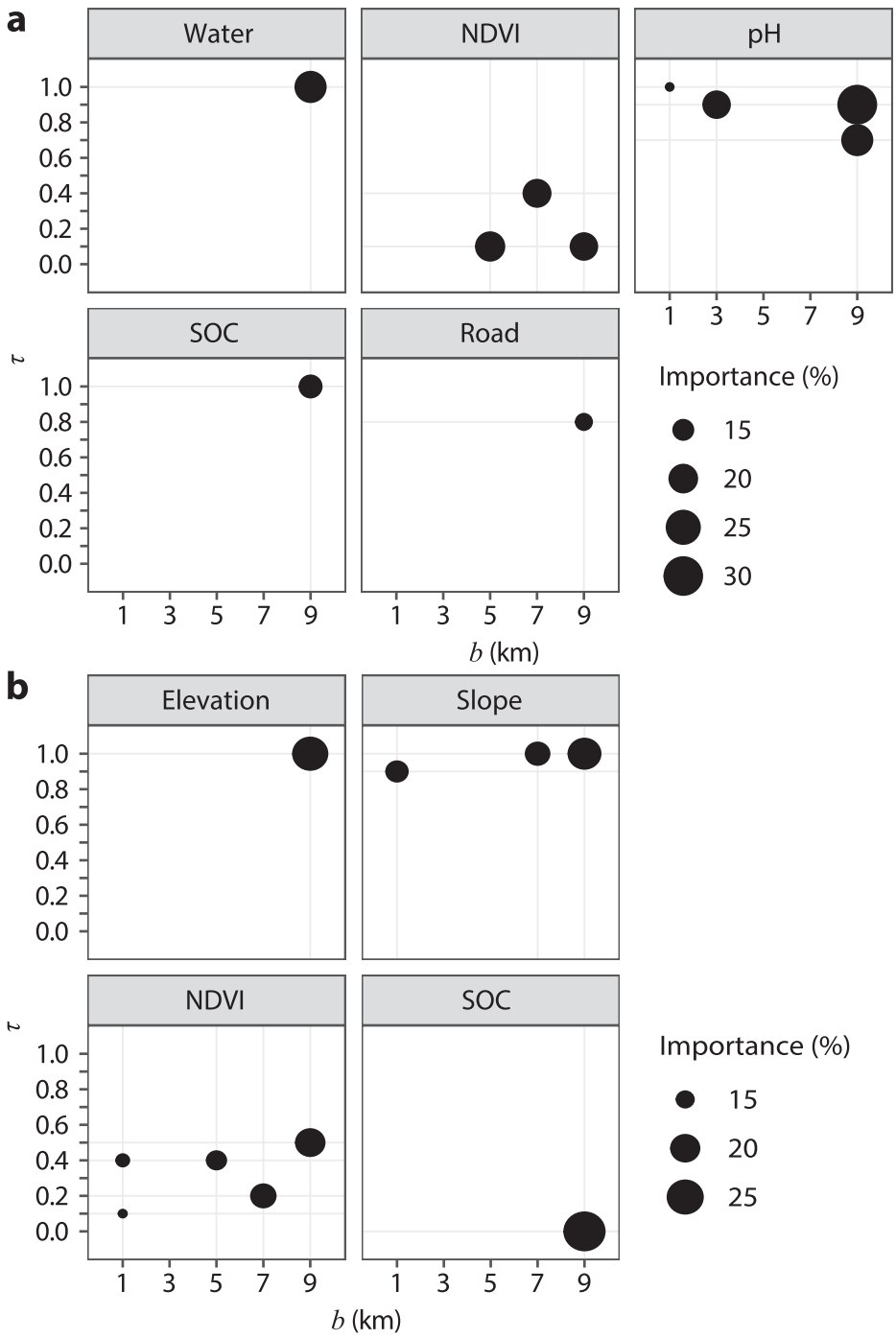


Fig. 5. Selected the second dimension variables for spatial predictions of Cr (a) and Cu (b). b is the searching range parameter, and τ is the probability parameter.

samples (Zhang et al., 2014; Matthew et al., 2011). The variable with a higher value of %IncMSE is more important than other variables in modeling the trace element. Variables with the top three relative importance are $X_{pH}^{b=9,\tau=0.9}$, $X_{Water}^{b=9,\tau=1}$, and $X_{pH}^{b=9,\tau=0.7}$, and their relative importance are 30.2%, 22.2%, and 22.1%, respectively. This means that the soil pH and distance to water are predominant variables of Cr distributions. The distance to water is an essential factor in the spatial distribution of trace elements since water from industrial regions or regions with high concentrations of trace elements may cause the accumulation and transfer of trace elements to the other areas (Liu et al., 2020). Meanwhile, the increase of soil pH can enhance the mobility and sorption of Cr and Cu on clay mineral surface (Kumpiene et al., 2008).

In addition, ten of the second dimension variables of explanatory variables elevation, slope, NDVI, and SOC are selected for the SDA-based modeling and prediction of Cu. The SDMs for modeling Cu can be presented as:

$$Y = f(\ X_{SOC}^{b=9,\tau=0}, X_{Elevation}^{b=9,\tau=1}, X_{Slope}^{b=9,\tau=1}, X_{NDVI}^{b=9,\tau=0.5}, X_{NDVI}^{b=7,\tau=0.2}, X_{Slope}^{b=7,\tau=1}, X_{Slope}^{b=1,\tau=0.9}, X_{NDVI}^{b=5,\tau=0.4}, X_{NDVI}^{b=1,\tau=0.4}, X_{NDVI}^{b=1,\tau=0.1}) \quad (8)$$

The second dimension variables in the equation are also ordered by the relative importance. Variables with the top three relative importance are $X_{SOC}^{b=9,\tau=0}$, $X_{Elevation}^{b=9,\tau=1}$, and $X_{Slope}^{b=9,\tau=1}$, with relative importance of 29.6%, 24.3%, and 22.5%, respectively. The result indicates that the distribution of Cu is closely associated with soil organic carbon and terrain features.

Therefore, the equations of SDMs demonstrate that SDA can effectively indicate the multi-scale effects and diverse information within explanatory variables of the geographical environment at local ranges using the second dimension variables.

4.3. Model validation

Cross-validation is performed to evaluate the accuracy of SDA models. Fig. 6 shows the comparisons between observations and predictions of trace elements in the validation data sets. Results show that compared with FDA models, the correlations between observations and prediction are critically improved by SDA models, including the

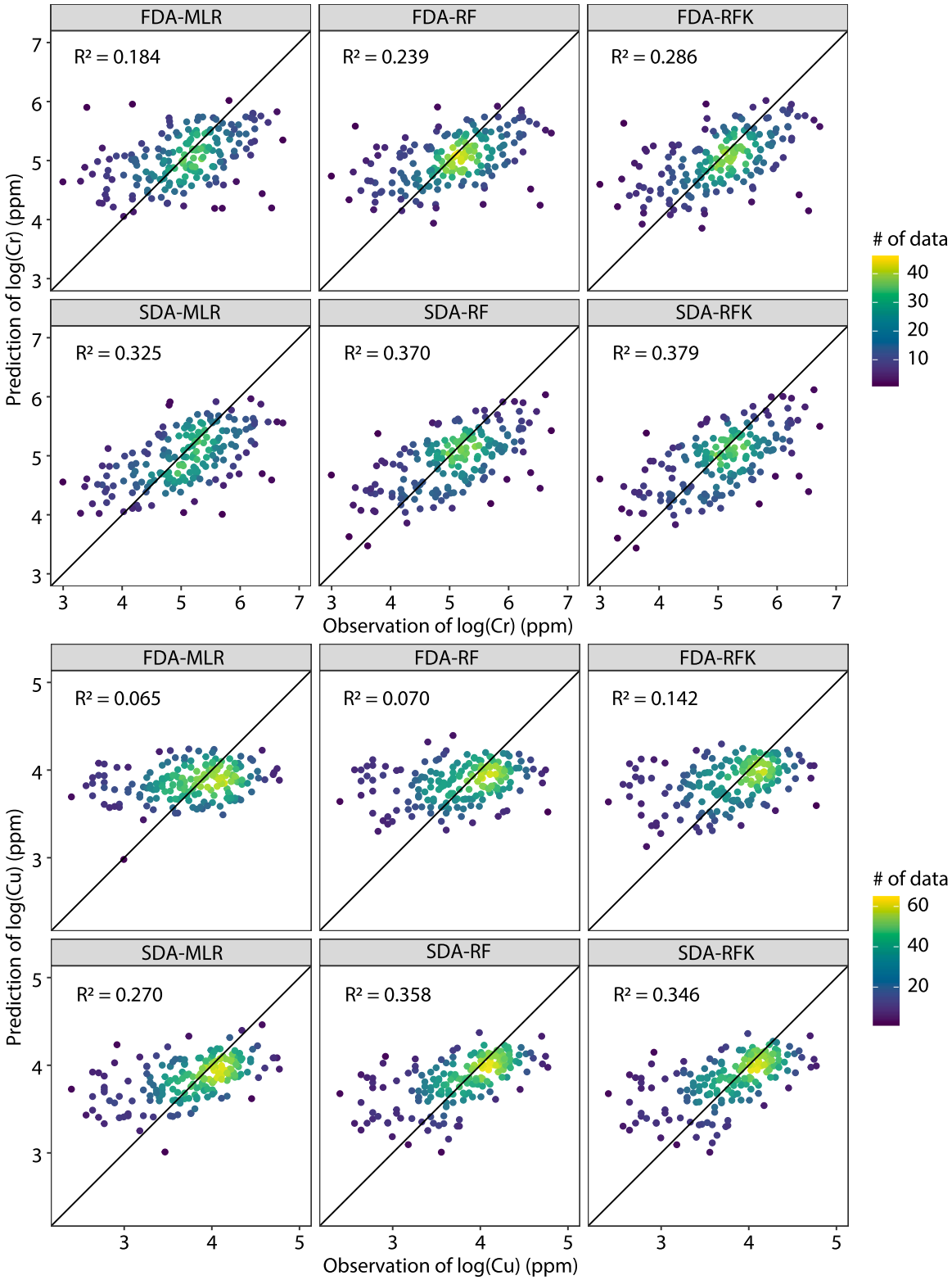


Fig. 6. Comparisons between observations and predictions of trace elements for the validation data in cross validations of the FDA and SDA models.

statistical model MLR, machine learning algorithm RF, and geostatistical model RFK. SDA models can effectively predict very high and low values within data sets. Table 3 shows a summary of cross-validation indicators of FDA and SDA models, and Table 4 presents the accuracy improvements of SDA models compared with FDA models. Results demonstrate that the prediction goodness-of-fit can be critically improved, and SDA models significantly reduce prediction errors. For MLR, RF, and RFK, the R^2 values can be improved by 32.8%–77.0% by SDA-based Cr prediction models, RMSE values are reduced by 5.4%–9.7%, and 2.0%–6.3% reduces MAE values. Among SDA models, the machine learning-based geostatistical SDA-RFK has the highest prediction goodness-of-fit and the lowest prediction errors for predicting Cr. For the prediction models of Cu, the R^2 values are improved by 144.4%–316.6% by SDA models, RMSE values are reduced by 12.1%–18.0%, and 14.5%–19.6% reduces MAE values. SDA-RF has the highest prediction goodness-of-fit and the lowest prediction errors for predicting Cu. In the prediction of Cu, the prediction accuracy of SDA-RF is higher than that of SDA-RFK since RF can depict the spatial variability, and the spatial variance of residuals is not significant.

4.4. Prediction and evaluation

Fig. 7 shows the FDA and SDA-based spatial predictions of trace elements, Cr and Cu. Results indicate the different spatial patterns of predictions, where FDA-based predictions at neighbor locations are critically changed and fluctuated, but SDA-based predictions are smoothly varied across space. Due to the smoothing of the kriging model, FDA-RFK can help smooth predictions. However, predictions derived from all SDA models are smoother than those derived by FDA-RFK.

To further understand the difference between FDA and SDA-based spatial predictions, prediction values are compared along a cross-section from the north-western to the south-eastern regions in the study area (see Fig. 8). The predictions along the cross-section reveal the following advantages of SDA models. First, similar to the above analysis, SDA models provide more smooth spatial predictions across space than that predicted by FDA models. SDA models avoid massive fluctuations at local ranges, where the typical examples include Cr predictions from 118.00 °E to 118.25 °E and from 118.25 °E to 118.50 °E, and Cu predictions from 118.25 °E to 118.50 °E. In addition, the predictions along the cross-section also demonstrate that SDA models can avoid underestimating high values and overestimating low values in the general FDA-based statistical models, machine learning, and geostatistical models.

Fig. 9 shows the comparisons of RFK prediction uncertainties of trace elements Cr and Cu computed using the FDA and SDA models. Results show that SDA models can significantly reduce the prediction uncertainty compared with FDA models. The mean uncertainties of FDA-RFK and SDA-RFK-based Cr predictions are 0.0148 and 0.0137, respectively. This means that the prediction uncertainty can be reduced by 7.4% by SDA-RFK. The mean uncertainties of FDA-RFK and SDA-RFK-based Cu predictions are 0.0101 and 0.0085, respectively, meaning that the prediction uncertainty can be reduced by 16.2% by SDA-RFK. Further, the uncertainty difference is calculated and visualized across the study area. The difference in prediction uncertainty reveals that the SDA-RFK Cr prediction uncertainty at 88.3% of the whole study area is lower than the FDA-RFK uncertainty, and the SDA-RFK Cu prediction uncertainty at 99.0% of the entire study area is lower than the FDA-RFK uncertainty. Therefore, SDA models can effectively reduce the uncertainty of spatial predictions.

Unlike FDA models, SDA models may need more time to select the second dimension variables. In this study, R software programming is used for SDA modeling. The variables and samples are randomly selected from the Cr data set in this study to evaluate the time consumption in SDA-based models. Fig. 10a shows the computation time with different numbers of variables, where the number of samples is 614, the optional searching ranges b values are 1, 3, 5, 7, and 9 km, and the optional probability τ values are a sequential number from 0 to 1 with an interval of 0.1. The results show that the computation time gradually increases with the number of variables. When the number of variables is 1, the computation time is 0.14 s, and when the number of variables is 8, the computation time is 1.99 s. Fig. 10b shows the computation time with different numbers of samples, where the number of variables is 8, b values are 1, 3, 5, 7, and 9 km, and τ values are sequential numbers from 0 to 1 with an interval of 0.1. The results show that the computation time also increases with the number of samples. When the number of samples is lower than 1000, the computation time is lower than 4 s. When the number of samples reaches 10,000, the computation time is 10.18 s. Therefore, in any of the above scenarios, the SDA modeling won't experience high computation burden, even when the sizes of spatial data sets are large.

Table 3
Summary of cross validation of FDA and SDA models for predicting trace elements.

Trace element	Cross validation	FDA-MLR	FDA-RF	FDA-RFK	SDA-MLR	SDA-RF	SDA-RFK
Cr	R^2	0.184	0.239	0.286	0.325	0.370	0.379
	RMSE	0.686	0.655	0.637	0.620	0.606	0.603
	MAE	0.525	0.499	0.486	0.492	0.480	0.476
Cu	R^2	0.065	0.070	0.142	0.270	0.358	0.346
	RMSE	0.501	0.502	0.480	0.440	0.412	0.414
	MAE	0.393	0.384	0.365	0.336	0.309	0.312

FDA: The first dimension of spatial association.
SDA: The second dimension of spatial association.

Table 4
Improvements of model accuracy by SDA models compared with FDA models.

Trace element	Improvement by SDA	MLR	RF	RFK
Cr	R^2 improvement	77.0%	54.9%	32.8%
	RMSE reduction	9.7%	7.6%	5.4%
	MAE reduction	6.3%	3.9%	2.0%
Cu	R^2 improvement	316.6%	411.3%	144.4%
	RMSE reduction	12.1%	18.0%	13.8%
	MAE reduction	14.6%	19.6%	14.5%

5. Discussion

This study proposes the concept of the second dimension of spatial association (SDA) and develops a series of SDA models for modeling spatial association and spatial predictions. The developed SDA models include the SDA-based statistical model MLR, machine learning algorithm RF, and geostatistical model RFK. The SDA models are implemented in predicting spatial distributions of trace elements Cr and Cu in a mining region in Western Australia. Results demonstrate the following advantages of SDA models for exploring spatial association and predictions. First, different from most of the current models, i.e., the first dimension of spatial association (FDA) models, that spatial association is explored using observations at sample locations, SDA models can extract more in-depth geographical information and characteristics than FDA models through building the relationships of a response variable and data of explanatory variables outside sample locations. For instance, in the application of trace element prediction, data of explanatory variables within the 9-km range of sample locations are used for modeling instead of only the data at the sample locations. Second, the SDA models developed in the study can effectively assess the comprehensive multi-scale effects of explanatory variables in modeling since it characterizes the SDA information using a searching range and probability parameters-based approach. Third, due to the in-depth description of geographical information, SDA models can critically improve the accuracy of spatial predictions compared with the FDA models, including

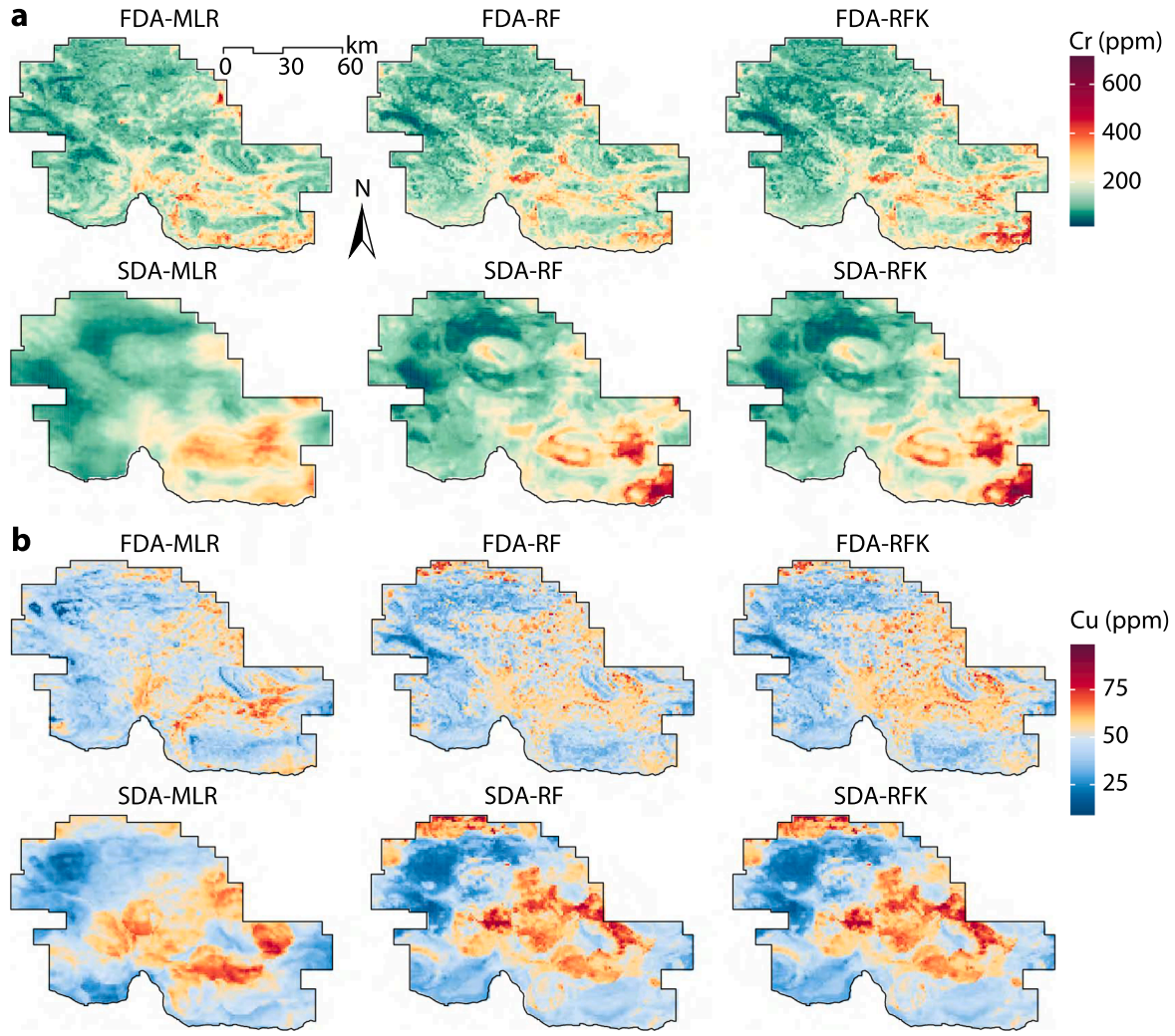


Fig. 7. Spatial predictions of trace elements, Cr (a) and Cu (b), in the study area using the FDA and SDA models.

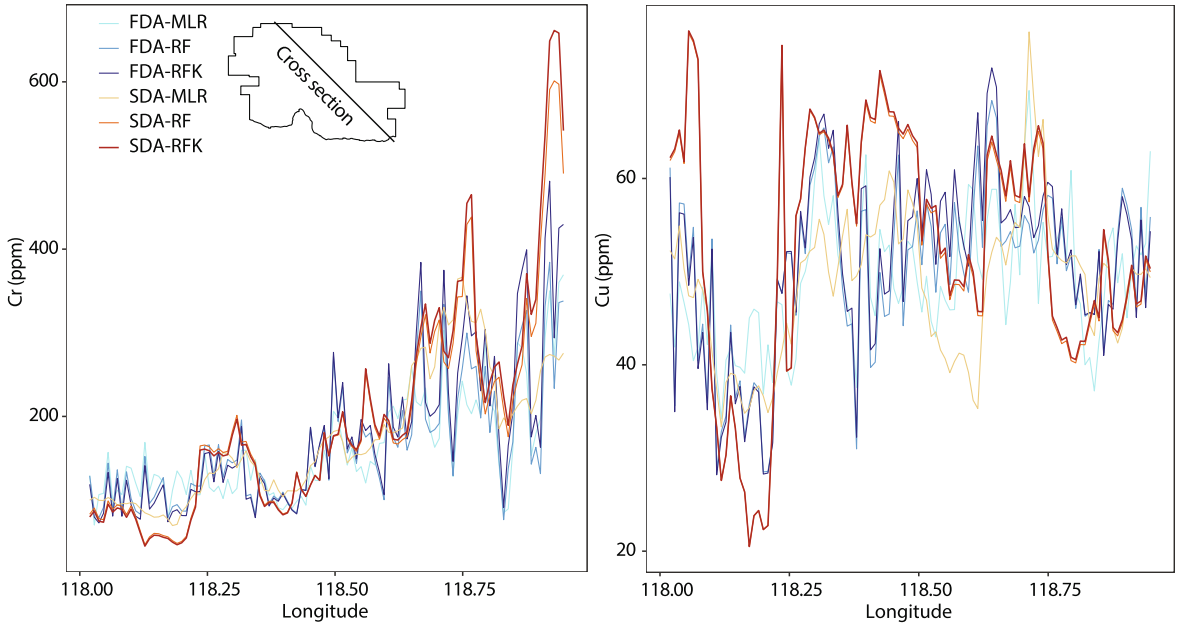


Fig. 8. Comparisons of the FDA and SDA model-based predictions along the cross section from the north-west to the south-east of the study area.

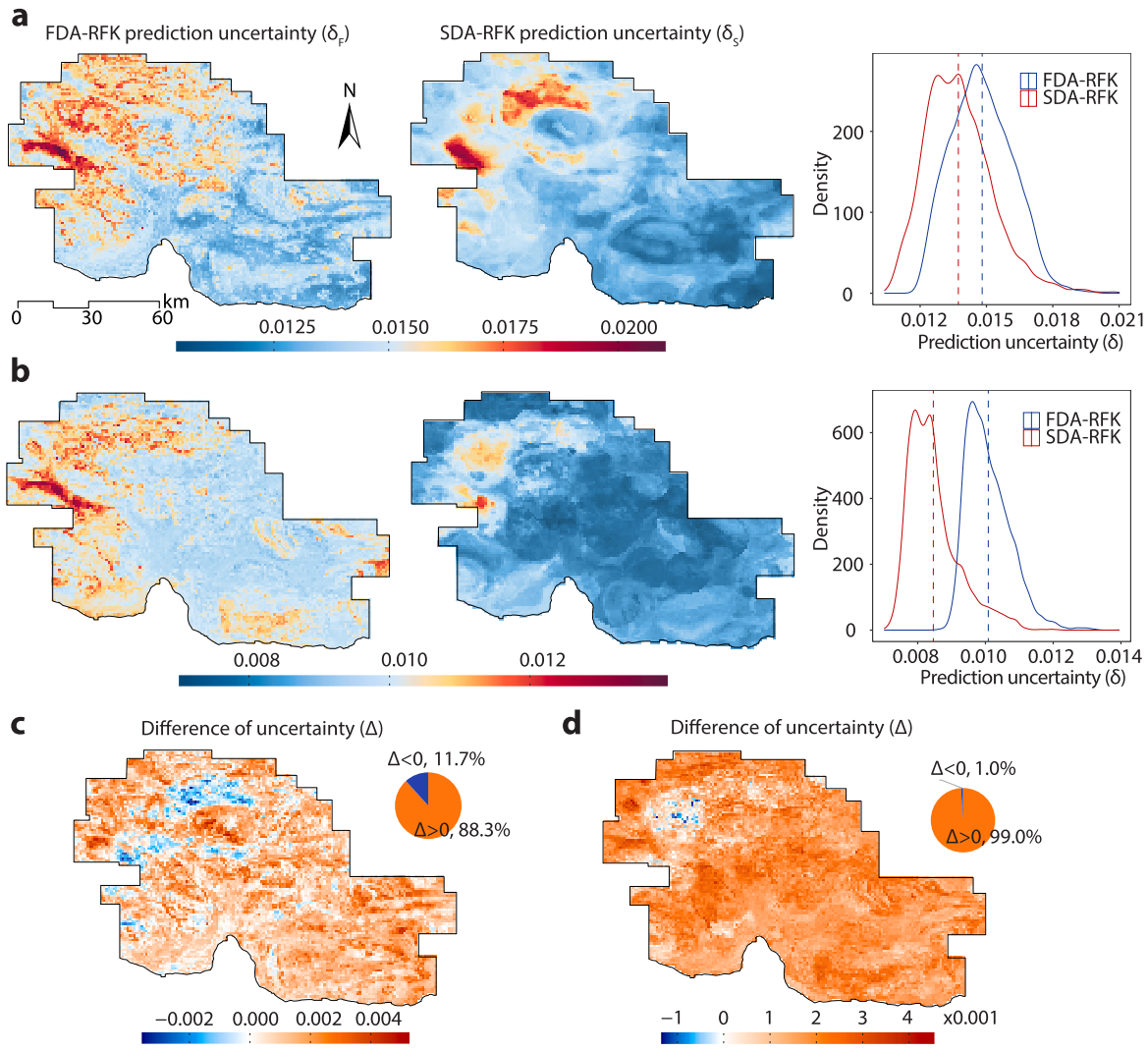


Fig. 9. Uncertainties of spatial predictions of trace element Cr (a) and Cu (b) derived using the FDA and SDA-based random forest kriging (RFK) models, and the corresponding difference of uncertainties (c-d).

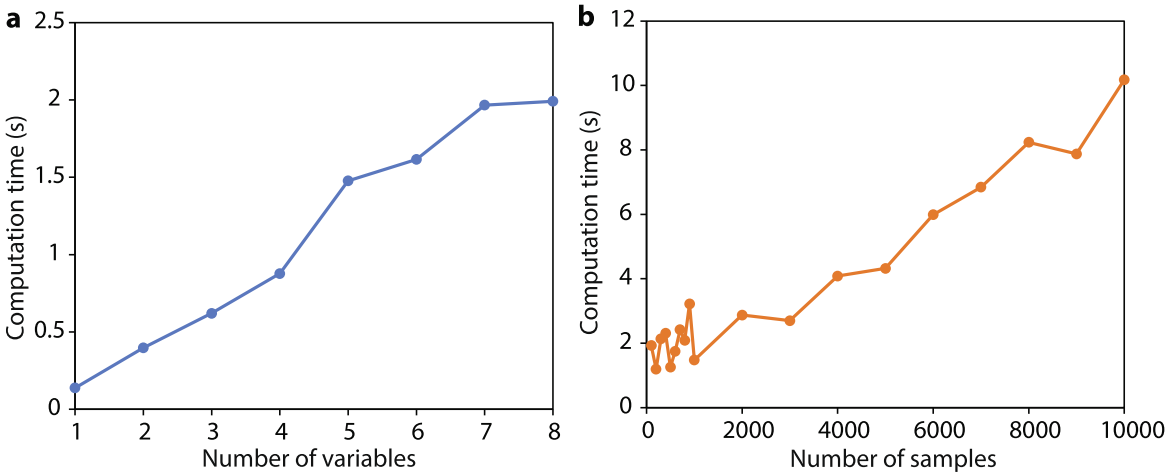


Fig. 10. Comparison of time consumption in the second dimension variable selection stage in SDA-based models. (a) Computation time with different numbers of variables; and (b) computation time with different numbers of samples.

statistical models, machine learning algorithms, and geostatistical models. For instance, the cross-validation R^2 values are improved by 32.8%–77.0% and 144.4%–316.6% by SDA models for Cr and Cu predictions, and the cross-validation RMSE values are reduced by 5.4%–9.7% and 12.1%–18.0% for predicting Cr and Cu, respectively. Fourth, SDA can avoid underestimating high values and overestimating low

values in the FDA models. In FDA models, the characterization of geographical information contained in explanatory variables is insufficient, making it difficult for FDA models to predict the relatively high and low values within a data set. Finally, SDA models provide more smooth spatial predictions across space than that predicted by FDA models and avoid massive fluctuations at local ranges. SDA models also can effectively and significantly reduce the uncertainty of spatial predictions. For instance, the mean prediction uncertainty is reduced by 7.4% and 16.2% by SDA-RFK compared with FDA-RFK in Cr and Cu predictions, respectively, and SDA-RFK-based prediction uncertainty at 88.3% and 99.0% of the whole study area is lower than the FDA-RFK-based uncertainty.

In terms of the analysis in this study, future studies are recommended in the following aspects. First, the concept of SDA can be implemented in more case studies in broader fields. In addition, SDA can also be integrated with different models for exploring spatial association. For instance, SDA can be integrated with geographically weighted regression (GWR) (Brunsdon et al., 1998) and advanced GWR models, such as multiscale GWR (Fotheringham et al., 2017) and geographically and temporally weighted regression (GTWR) (Huang et al., 2010; Fotheringham et al., 2015), for modeling spatial association. In the SDMs presented in this study, including linear regression, machine learning, and machine learning-based geostatistics models, a geospatial variable at a location is affected by the geographical environment from locations outside this location. However, from a geospatial perspective, the integration between SDA and GWR is more complex than the above models. A geospatial variable at a location is affected by the geographical environment from locations outside its surrounding sample locations. Thus, methods for reasonable integration between SDA and GWR should be developed in future studies. Finally, developing strategies for understanding SDA is still a challenge. Therefore, different modeling strategies can be developed to characterize SDA for more accurate, effective, reliable, and robust modeling of spatial association and predictions.

6. Conclusion

An innovative understanding of spatial association is critically essential for addressing complex issues of spatial statistical inference and geocomputation. This study proposed the concept of the second dimension of spatial association (SDA) and developed a series of SDA models for exploring spatial association and predicting distributions of geographical attributes. The concept of SDA provides new insight into geographical information-based spatial association. The SDA-based trace elements prediction study analysis demonstrated that the SDA had a series of advantages in examining spatial association. The advantages of SDA models include extracting more in-depth geographical information and characteristics using data of explanatory variables outside sample locations, effectively assessing multi-scale effects of variables, significantly improving prediction accuracy, avoiding underestimation or overestimation for relatively high or low values, providing more smooth spatial predictions, and significantly reduce prediction uncertainty. SDA and multiple types of SDA models have great potential for more accurate and effective spatial statistical inference and geocomputation in diverse fields.

Declaration of Competing Interest

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

References

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. 28 (4), 281–298.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. Roy. Stat. Soc. Ser. D (Statistician) 47 (3), 431–443.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z., Ma, J., 2017. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 151, 147–160.

Cressie, N., Moores, M.T., 2021. Spatial statistics. arXiv preprint arXiv:2105.07216.

Department of Mines, Industry Regulation and Safety, Government of Western Australia, 2022. Gswa geochemistry. URL https://www.dmp.wa.gov.au/GeoChem-Extract-Geochemistry-1559.aspx (accessed 1 March 2022).

Didan, K., 2015. Mod13q1 modis/terra vegetation indices 16-day l3 global 250m sin grid v006. nasa eosdis land processes daac. https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD13Q1.

Fotheringham, A.S., Crespo, R., Yao, J., 2015. Geographical and temporal weighted regression (gtwr). Geogr. Anal. 47 (4), 431–452.

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale geographically weighted regression (mgwr). Ann. Am. Assoc. Geogr. 107 (6), 1247–1265.

Gao, B., Wang, J., Stein, A., Chen, Z., 2022. Causal inference in spatial statistics. Spatial Stat. 100621.

Geoscience Australia, 2006. Geodata topo 250k series 3. URL https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/metadata/63999.

Geosciences Australia, 2015. Digital elevation model (dem) of australia derived from lidar 5 metre grid. Commonwealth of Australia and Geoscience Australia: Canberra.

Goovaerts, P., et al., 1997. Geostatistics for natural resources evaluation. Oxford University Press on Demand.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Haining, R., 1993. Spatial data analysis in the social and environmental sciences. Cambridge University Press.

Haining, R.P., Haining, R., 2003. Spatial data analysis: theory and practice. Cambridge University Press.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42 (33), 7561–7578.

Huang, B., Wu, B., Barry, M., 2010. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. Int. J. Geogr. Inform. Sci. 24 (3), 383–401.

Jacquez, G.M., 1999. Spatial statistics when locations are uncertain. Geogr. Inform. Sci. 5 (2), 77–87.

Kammann, E., Wand, M.P., 2003. Geoadditive models. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 52 (1), 1–18.

Krige, D.G., 1951. A statistical approach to some basic mine valuation problems on the witwatersrand. J. South Afr. Inst. Min. Metall. 52 (6), 119–139.

Kumpiene, J., Lagerkvist, A., Maurice, C., 2008. Stabilization of as, cr, cu, pb and zn in soil using amendments–a review. Waste Manage. 28 (1), 215–225.

Liu, Y., Fei, X., Zhang, Z., Li, Y., Tang, J., Xiao, R., 2020. Identifying the sources and spatial patterns of potentially toxic trace elements (ptes) in shanghai suburb soils using global and local regression models. Environ. Pollut. 264, 114171.

Luo, P., Song, Y., Huang, X., Ma, H., Liu, J., Yao, Y., Meng, L., 2022. Identifying determinants of spatio-temporal disparities in soil moisture of the northern hemisphere using a geographically optimal zones-based heterogeneity model. ISPRS J. Photogramm. Remote Sens. 185, 111–128.

Main Roads Western Australia, 2020. Road network in western australia. URL https://catalogue.data.wa.gov.au/sv/dataset/mrwa-road-network.

Matthew, W. et al., 2011. Bias of the random forest out-of-bag (oob) error for certain input parameters. Open J. Stat.

Morin-Ka, S., Beardsmore, T., Duuring, P., Guilliamse, J., Burley, L., 2019. The mineral systems atlas—delivering greater value from precompetitive geoscience data. ASEG Extended Abstracts 2019 (1), 1–3.

Morris, P.A., Sanders, A.J., Pirajno, F., Faulkner, J.A., Coker, J., 1998. Regional-scale regolith geochemistry: identification of metalloid anomalies and the extent of bedrock in the archaean and proterozoic of western australia. Taylor, G, Pain, & C. (Eds.). Regolith 98, 101–108.

Viscarra Rossel, R., Chen, C., Grundy, M., Searle, R., Clifford, D., Odgers, N., Holmes, K., Griffin, T., Liddicoat, C., Kidd, D., 2014. Soil and landscape grid national soil attribute maps - soil attribute release 1. v2. URL https://developers.google.com/earth-engine/datasets/catalog/CSIRO_SLGA.

Song, Y., Wu, P., 2021. An interactive detector for spatial associations. Int. J. Geogr. Inform. Sci. 35 (8), 1676–1701.

Song, Y., Wang, J., Ge, Y., Xu, C., 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 Sens. 57 (5), 593–610.

Song, Y., Shen, Z., Wu, P., Viscarra Rossel, R., 2021a. Wavelet geographically weighted regression for spectroscopic modelling of soil properties. Sci. Rep. 11 (1), 1–11.

Song, Y., Thatcher, D., Li, Q., McHugh, T., Wu, P., 2021b. Developing sustainable road infrastructure performance indicators using a model-driven fuzzy spatial multi-criteria decision making method. Renew. Sustain. Energy Rev. 138, 110538.
Ver Hoef, J.M., Peterson, E.E., Hooten, M.B., Hanks, E.M., Fortin, M.-J., 2018. Spatial autoregressive models for statistical inference from ecological data. Ecol. Monogr. 88 (1), 36–59.
Walford, N., 2002. Geographical data: characteristics and sources. John Wiley & Sons.
Wang, J.-F., Li, X.-H., Christakos, G., Liao, Y.-L., Zhang, T., Gu, X., Zheng, X.-Y., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the heshun region, china. Int. J. Geogr. Inform. Sci. 24 (1), 107–127.
Wang, J., Gao, B., Stein, A., 2020. The spatial statistic trinity: A generic framework for spatial sampling and inference. Environ. Model. Softw. 134, 104835.
Wells, M., Laukamp, C., Hancock, E., 2016. Integrated spectral mapping of precious and base metal-related mineral footprints, nanjilgardy fault, western australia. In: GSWA 2016 EXTENDED ABSTRACTS Promoting the prospectivity of Western Australia, p. 26.
Wu, P., Song, Y., 2022. Land use quantile regression modeling of fine particulate matter in australia. Remote Sens. 14 (6), 1370.
Xu, H., Bechle, M.J., Wang, M., Szpiro, A.A., Vedal, S., Bai, Y., Marshall, J.D., 2019. National pm2. 5 and no2 exposure models for china based on land use regression, satellite measurements, and universal kriging. Sci. Total Environ. 655, 423–433.
Zhang, K., Li, Y., Schwartz, J.D., et al., 2014. What weather variables are important in predicting heat-related mortality? a new application of statistical learning methods. Environ. Res. 132, 350–359.
Zhu, A.-X., Band, L., Vertessy, R., Dutton, B., 1997. Derivation of soil properties using a soil land inference model (solim). Soil Sci. Soc. Am. J. 61 (2), 523–533.
Zhu, A.-X., Lu, G., Liu, J., Qin, C.-Z., Zhou, C., 2018. Spatial prediction based on third law of geography. Ann. GIS 24 (4), 225–240.