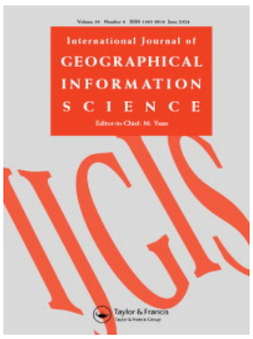

International Journal of Geographical Information Science

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

An ensemble spatial prediction method considering geospatial heterogeneity

Shifen Cheng, Lizeng Wang, Peixiao Wang & Feng Lu

To cite this article: Shifen Cheng, Lizeng Wang, Peixiao Wang & Feng Lu (05 Jun 2024): An ensemble spatial prediction method considering geospatial heterogeneity, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2024.2358052

To link to this article: https://doi.org/10.1080/13658816.2024.2358052

Published online: 05 Jun 2024.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2024.2358052

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

An ensemble spatial prediction method considering geospatial heterogeneity

Shifen Cheng[a,b], Lizeng Wang[a,b], Peixiao Wang[a,b] and Feng Lu[a,b,c,d]

[a]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing, China; [b]University of Chinese Academy of Sciences, Beijing, China; [c]The Academy of Digital China, Fuzhou University, Fuzhou, China; [d]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China

ABSTRACT

Ensemble learning synthesizes the advantages of different models and has been widely applied in the field of spatial prediction. However, the nonlinear constraints of spatial heterogeneity on the model ensemble process make it difficult to adaptively determine the ensemble weights, greatly limiting the predictive ability of the ensemble learning model. This paper therefore proposes a novel geographical spatial heterogeneous ensemble learning method (GSH-EL). Firstly, the geographically weighted regression model, geographically optimal similarity model, and random forest model are used as three base learners to express local spatial heterogeneity, global feature correlation, and nonlinear relationship of geographic elements, respectively. Then, a spatially weighted ensemble neural network module (SWENN) of GSH-EL is proposed to express spatial heterogeneity by exploring the complex nonlinear relationship between the spatial proximity and ensemble weights. Finally, the outputs of the three base learners are combined with the spatial heterogeneous ensemble weights from SWENN to obtain the spatial prediction results. The proposed method is validated on the $PM_{2.5}$ air quality and landslide dataset in China, both of which obtain more accurate prediction results than the existing ensemble learning strategies. The results confirm the need to accurately express spatial heterogeneity in the model ensemble process.

ARTICLE HISTORY
Received 16 August 2023
Accepted 16 May 2024

KEYWORDS
Spatial prediction; spatial inference; spatial heterogeneity; spatial data mining; ensemble learning

1. Introduction

Continuous popularization and development of sensor networks and mobile positioning technologies have made massive spatial data more easily accessible, providing crucial support for the modeling and analysis of complex geographic processes (Karpatne *et al*. 2019, Liu and Biljecki 2022). However, the prevalence of sparse distribution in geospatial data is attributed to the limited capacity of current collection equipment. This limitation severely hampers the accurate portrayal and prediction of natural and

CONTACT Feng Lu luf@lreis.ac.cn

social systems on the land surface (Yao and Huang 2023). To address this issue, spatial prediction techniques utilize limited observed data and geographic features through statistical or machine learning methods to estimate attributes or characteristics of unknown geographic locations (Zhu *et al*. 2018). This technology proves to be crucial in mitigating the impact of spatial data sparsity (Jiang 2019).

Spatial prediction model strategies can be divided into single and multi-model ensemble types. The single modelling strategy typically employs statistical and machine learning methods to learn the spatial dependency structure of the data. However, each model has advantages and disadvantages, and it is often difficult to select the appropriate spatial prediction model for the given study area (Liu *et al*. 2015). Given with the complex changes inherent to geographic processes, the hypothesis space of model learning is usually very large. It thus cannot be effectively explored with limited training samples, and a single model may therefore miss the best fitting function or true distribution of the sample set in the hypothesis space. As a result, optimal spatial prediction results are often difficult to achieve with single models (Hao and Tian 2019). Furthermore, it is often difficult to generalize single models to other study areas (Feng *et al*. 2020).

To further improve the predictive performance and generalization ability of spatial prediction models, existing research has incorporated ensemble learning into geospatial data modelling (Li *et al*. 2017, Requia *et al*. 2020). By constructing and combining multiple heterogeneous base learners, ensemble learning synthesizes the strengths of different base learners to accomplish the learning task, thereby mitigating the problem of model selection in spatial prediction to some extent (Fang *et al*. 2021). According to the different strategies for combining base learners, ensemble learning can be divided into two types: statistics-based and learning-based. The ensemble strategy combining multiple models has been widely proven in the literature to achieve stronger generalization capability and higher accuracy compared to single modelling strategies (Feng *et al*. 2020, Requia *et al*. 2020, Fang *et al*. 2021). However, existing research in designing ensemble strategies often relies on the assumption that the samples are independent and identically distributed, such that the weights of different base learners are globally fixed over the entire spatial range (Li 2019). For example, in the statistics-based average ensemble method, the final ensemble result is obtained by averaging the output results from different base learners. In the learning-based ensemble method, the output results of different base learners are used as input features to train a global model that captures the relationships among the base learners. Both of these ensemble strategies assign the same weight to each base learner at different spatial locations.

Geographic elements are usually heterogeneous across the entire study area, showing obvious spatial heterogeneity (Ge *et al*. 2019). This means global ensemble strategies may perform poorly in some local locations. While some studies have considered spatial heterogeneity by using geographically weighted regression models to obtain a set of spatially varying weights for integrating the predictions from different base learners, improving the robustness of the ensemble model to some extent (Requia *et al*. 2020). However, the geographically weighted regression model (GWR) typically requires the construction of weight kernel functions parameterized by spatial

proximity measures, thereby enabling the modelling of spatially heterogeneous relationships (Wu *et al*. 2021, Hagenauer and Helbich 2022). The simple kernel function structure used in GWR for calculating spatial weights makes it difficult to describe the complex nonlinear effects of spatial proximity on the ensemble weights, leading to difficulties in adequately expressing the spatial heterogeneity of complex geographic relationships (Du *et al*. 2020, 2021).

By focusing on the problem of how to accurately express spatial heterogeneity when designing the ensemble strategy, this paper proposes a novel ensemble spatial prediction method considering spatial heterogeneity. The main contributions are:

1. A geospatial weighted ensemble neural network module (SWENN) is designed. This utilizes the highly abstract expression ability and high-dimensional dynamic learning ability of the neural network to accurately solve the weight kernel function, while fully mining the complex nonlinear relationship between spatial proximity and model ensemble weights. It thus achieves accurate expression of spatial heterogeneity in the ensemble process.
2. A novel geographical spatial heterogeneous ensemble learning method (GSH-EL) is proposed. This embeds the local spatial heterogeneity, global feature correlation, and nonlinear relationship of geographical elements into the SWENN module for ensemble modelling.
3. The proposed method is validated using real geospatial datasets from the perspectives of regression-based spatial prediction tasks and classification-based spatial prediction tasks.

2. Related work

2.1. Spatial prediction with single modelling strategies

The single modelling strategy aims to use a single statistical or machine learning algorithm to describe the spatial correlation between data and estimate the spatial features of unobserved locations. The current methods can be roughly divided into statistics-based methods and machine learning-based methods.

Common statistics-based methods can be classified into the following categories. The first category includes spatial dependence-based models, such as ordinary kriging and spatial bayesian hierarchical model (Saez and Barcelo 2022). The ordinary kriging quantifies the spatial dependency of the target variable through the covariance function on spatial distance. Instead of traditional point support, the variance methods have been developed for data with irregular spatial support, such as area-to-area regression kriging (Ge *et al*. 2014) for estimation of different areas and segment-based regression kriging (Song *et al*. 2019) for integration of the irregular shape of segments. The second category is spatial heterogeneity-based models including sandwich model (Wang *et al*. 2013), P-BSHADE model (Xu *et al*. 2013), geographically weighted regression (Su *et al*. 2022), geo-generalized additive model (Sarker *et al*. 2021), and generalized heterogeneity model (Luo *et al*. 2023). The geographically weighted regression is a classical spatial prediction method of modelling spatially varying relationships

between the target and explanatory variables. The generalized heterogeneity model applies area-to-area kriging to characterize the spatial correlations between different subregions, improving both the overall prediction accuracy across the study area and along strata boundaries. The third category is geographical configuration similarity-based models. For instance, the geographically optimal similarity model (Song 2022a) calculates the similarity of the geographic environments and selects the samples with optimal similarity to infer the target variable for the predicted samples, which provides a new perspective for spatial prediction. In addition, the second-dimension spatial association model (Song 2022b) extracts geographical information at locations outside samples for exploring spatial association, effectively indicating the multi-scale effects and diverse information within explanatory variables of the geographical environment at local ranges using the second-dimension variables.

With the rapid development of artificial intelligence technology in recent years, machine learning models have also been widely used in the field of spatial prediction. Examples include extreme gradient boosting tree (Yi *et al*. 2021), k-nearest neighbor model (Cai *et al*. 2023), Bayesian learning model (Narendra Babu *et al*. 2021), matrix decomposition (Lei *et al*. 2022), tensor decomposition (Chen *et al*. 2022), and deep learning model (Zhu *et al*. 2020, Dai *et al*. 2022, Wang *et al*. 2022, Zeng *et al*. 2022, Yao and Huang 2023). In particular, random forest (Song *et al*. 2021) overcomes problems such as overfitting of a single decision tree by applying the bagging technique to aggregate multiple classification and regression trees. Considering the rich variety of explanatory variables, it can obtain better prediction results than traditional methods. Additionally, the neural network model relies on its powerful expressive and learning capabilities to achieve accurate solutions to complex dynamic and nonlinear problems. The convolutional neural network (Dai *et al*. 2022) is used to improve the kernel function in geographically weighted regression models, capturing the relationship between global spatial proximity information and spatial weights for weighted prediction.

Although there are numerous spatial prediction methods based on a single modelling strategy that effectively enhance the predictive ability and application value of models, these methods are often unable to capture complex geographical processes, with limited predictive accuracy and generalization ability.

2.2. Spatial prediction with ensemble modelling strategies

Spatial prediction methods based on ensemble modelling strategies improve the overall prediction accuracy and generalization performance by integrating the outputs of multiple single models. Current ensemble learning can be divided into homogeneous ensemble learning and heterogeneous ensemble learning according to the type of base learners (Zhou 2012).

Homogeneous ensemble learning refers to the same type of base learner, including ensemble frameworks such as Bagging and Boosting. Bagging and Boosting are used to integrate a range of traditional spatial statistics and machine learning methods to train base learners by sampling or adjusting the distribution of samples to ensure the stability of prediction results (Liu *et al*. 2015, Li *et al*. 2017, Pham *et al*. 2021). These

homogeneous ensemble learning methods improve spatial prediction accuracy to some extent. However, due to their training process involving multiple combinations of the same base learner to construct the ensemble model, if the error of the base learner is large, it may lead to a magnification of errors in the final prediction results.

Compared to homogeneous ensemble learning, heterogeneous ensemble learning can integrate multiple diverse base learners, achieving better coverage of the solution space for the original prediction problem. As a result, it tends to exhibit greater robustness and generalization value (Fang *et al*. 2021). Depending on the ensemble strategy of the base learner, it can be categorized into statistics-based ensemble method and learning-based ensemble method. The statistics-based ensemble method integrates the outputs of multiple base learners using statistical models to obtain the final prediction, such as the weighted average (Deng *et al*. 2016, Shafizadeh-Moghadam *et al*. 2018) and the Bayesian inference model (Murray *et al*. 2019, Yang *et al*. 2019). Learning-based ensembles achieve better prediction results by using another learner to learn how to best integrate outputs from multiple base learners. Common ensemble strategies include neural network model (Cheng and Lu 2017, Cheng *et al*. 2020), ridge regression (Qin *et al*. 2019), logistic regression (Wang *et al*. 2020), multiple linear regression (Li *et al*. 2019, Chen *et al*. 2019, Feng *et al*. 2020) and the deep learning model (Wu *et al*. 2023). However, due to the inherent spatial heterogeneity of geographic elements, the prediction accuracy of the base learner is highly variable across spatial locations. Current research usually assumes that the samples in the entire study area follow an independent and identical distribution when designing ensemble strategies, and uses globally fixed statistical or machine learning models to fit the prediction results of different base learners, ignoring the spatial heterogeneity of ensemble weights.

Considering the existence of spatial heterogeneity in geographic data, relevant research has begun to use geographically weighted correlation models as the ensemble strategy. For example, Li (2019) used GWR integrating three models including the autoencoder-based deep residual network, XGBoost and random forest for high-resolution spatiotemporal estimation of wind speed. Requia *et al*. (2020) integrated multiple types of predictor variables and three machine learning models-neural networks, random forests, and gradient boosting into a geographically weighted generalized additive model, enabling high-resolution estimation of ground-level ozone concentrations. These studies demonstrate that ensemble models are superior to either single model, and considering spatial heterogeneity when designing ensemble strategies can further improve the accuracy of predictive models. However, the calculation of weight kernel functions is the core of expressing spatial heterogeneity in geographic weighted regression models (Wu *et al*. 2021). The current research adopts a simple kernel function structure, which makes it difficult to fully evaluate the complex nonlinear effect of spatial proximity on model ensemble weights, resulting in difficulty in accurately solving the spatial heterogeneity of complex geographical relationships between base learners during the ensemble process. Therefore, there is an urgent need to develop new spatial heterogeneity expression methods to achieve accurate solutions of weight kernel functions.

3. Methodology

3.1. Overall framework

A GSH-EL method is proposed in this paper, as summarized in Figure 1. Firstly, the GWR, geographically optimal similarity model (GOS), and random forest model (RF) are selected as the base learners of the ensemble framework. The observed locations and unobserved locations are input into the three base learners to obtain the output feature vectors. Then, the ordinary least squares method is used to compute the global average ensemble coefficients of the base learner using the feature vectors of the observed locations. Subsequently, the dataset is divided into training, validation, and testing datasets. The SWENN is constructed to solve the weight kernel function using the spatial proximity vector as input to obtain the spatial heterogeneous ensemble weights for the base learners. Finally, the output results of different base learners are

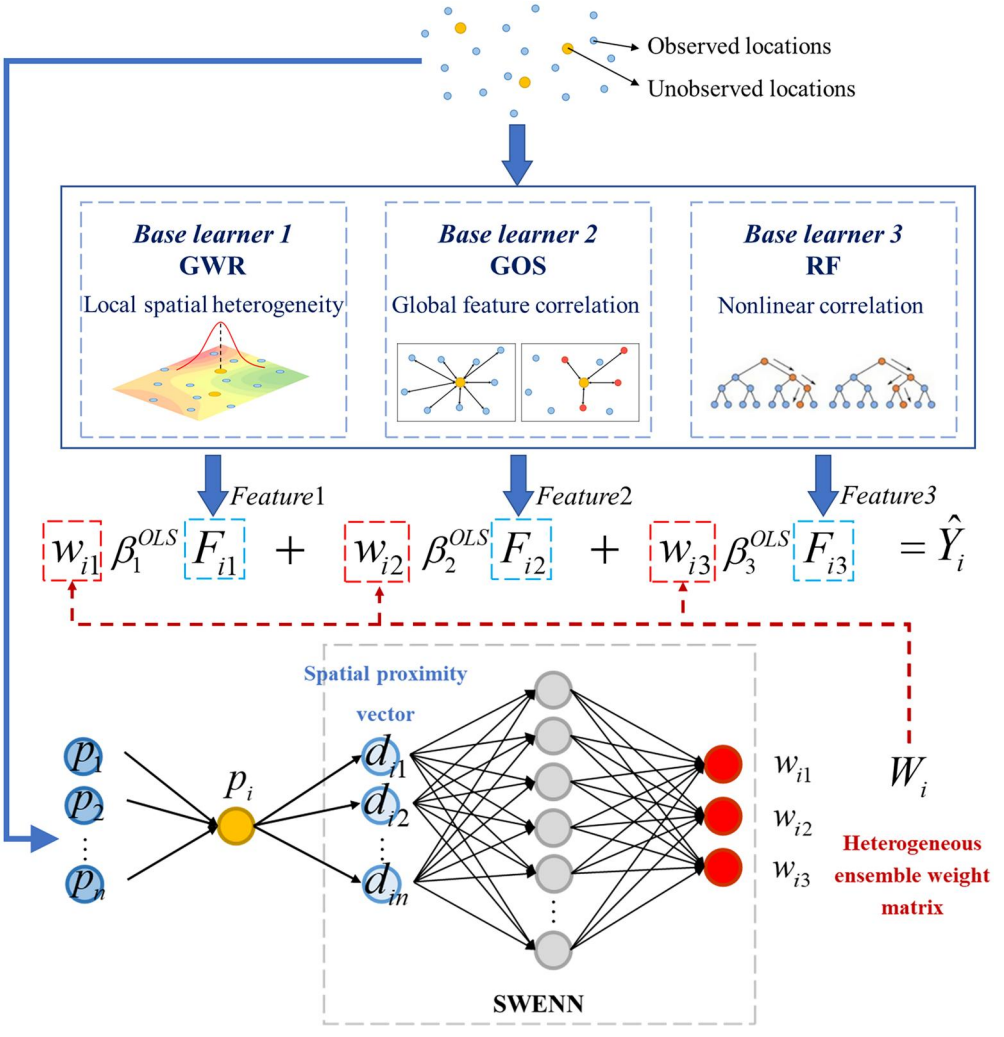

Figure 1. Definition of the geographical spatial heterogeneous ensemble learning (GSH-EL).

integrated by utilizing the spatial heterogeneous ensemble weights and global average coefficients to obtain the final ensemble prediction results.

3.2. Design of the base learner

The GWR, GOS, and RF models are widely employed in the field of spatial prediction. Their diversity allows us to capture differing facets of the underlying data patterns, thereby enhancing the overall predictive capability of the ensemble model. The three models express local spatial heterogeneity, global feature correlation, and nonlinear relationship of geographic elements, respectively.

3.2.1. Geographically weighted regression model

Spatial heterogeneity is prevalent in the modelling of geographical relationships. GWR is one of the main methods for modelling spatial heterogeneity (Brunsdon *et al*. 1996). It performs a local linear weighted regression on the samples by setting local spatial weights, which embeds changes in spatial relationships caused by differences in spatial location in the calculation of the regression coefficients. The predicted value of the target sample y_i is defined as:

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{p} \beta_j(u_i, v_i)x_{ij} + \varepsilon_i \;\; i = 1, 2, \ldots, n \tag{1}$$

where (u_i, v_i) is the spatial coordinates of the sample i, $\beta_j(u_i, v_i)$ is the jth explanatory variable regression coefficient for sample i, which is related to the spatial coordinates (u_i, v_i). ε_i is the random error term for sample i, with $\varepsilon_i \sim N(0, \sigma^2)$ and $Cov(\varepsilon_i, \varepsilon_j) = 0 \;\; (i \neq j)$.

In GWR, the weighted least squares method is commonly used to solve for the regression coefficients of each explanatory variable. Let $\beta(u_i, v_i) = [\beta_0(u_i, v_i), \beta_1(u_i, v_i), \ldots, \beta_p(u_i, v_i)]$ and then the regression coefficient $\hat{\beta}(u_i, v_i)$ for sample i can be expressed as:

$$\hat{\beta}(u_i, v_i) = [X^T W(u_i, v_i)X]^{-1} X^T W(u_i, v_i)Y \tag{2}$$

where $W(u_i, v_i) = diag(w_{i1}, w_{i2}, \ldots, w_{in})$ is the weight matrix at spatial location (u_i, v_i). w_{ij} needs to be solved by setting a specific kernel function, for example Gaussian or Bi-square. The kernel function can be classified into fixed kernel function and adaptive kernel function based on the type of bandwidth. Akaike Information Corrected Criterion (AICc) or cross-validation (CV) is often used to determine the bandwidth size of the kernel function.

3.2.2. Geographically optimal similarity model

The GOS is used as the second base learner of the GSH-EL. It enables the deep coupling of local spatial correlation with global feature correlation. Following the third law of geography (Zhu *et al*. 2018; Song 2022a), GOS demonstrates greater potential for implementing the geographical similarity principle in spatial predictions. By incorporating information from samples with relatively high similarities at any location across space, the GOS model provide a novel perspective for spatial prediction. Instead of directly constructing an explicit relationship between the explanatory

variables and the target variables, the model uses a set of explanatory variables to characterize the geographic configuration of the samples. By calculating the geographic configuration similarity between the unobserved locations and the observed locations, it selects samples with higher similarity for calculation, and the similarity information is used as weights to weight the predictions. Based on p explanatory variables $X = \{X_i\}_i^p \in \mathcal{R}^{p\times 1}$, the similarity $S(k,t)$ between the unobserved locations t and observed location k is computed as:

$$S(k,t) = P\{E(k,t)_i\} \tag{3}$$

$$E(k,t)_i = \exp\left(\frac{(X_i^k - X_i^t)^2}{2(\sigma^2/\delta^t)^2}\right) \tag{4}$$

$$\delta^t = \sqrt{\frac{\sum_{i=1}^n (X_i^k - X_i^t)^2}{n}} \tag{5}$$

where k is the observed location, and t is the unobserved location. P is the minimum operator and $E(k,t)_i$ represents the geographic configuration metric function for the ith explanatory variable. X_i^k is the vector of the ith explanatory variable for the observed location k. X_i^t is the vector of the ith explanatory variable for the unobserved location t. δ_t is the square root of the average deviation from the unobserved location to all observed locations. σ is the standard deviation of the explanatory variable.

The optimal similarity threshold is set using the CV approach. The observed locations above this threshold participate in the final weighting calculation. The predicted value $\hat{Y}_t$ for the unobserved locations is computed as:

$$\hat{Y}_t = \frac{\sum_1^m S(k,t)Y_k}{\sum_1^m S(k,t)} \tag{6}$$

where m is the number of observed locations with similarity above a threshold value, $S(k,t)$ is the similarity of the geographical configuration of the observed location k, and Y_k is the value of the target variable for the observed location k.

3.2.3. Random forest model

In real geographic processes, geographic relationships usually exhibit complex nonlinear characteristics. Compared to the GWR modelling the local linear relationship between explanatory and target variables, the RF generates a tree structure by splitting the features, which provides a good mapping of the nonlinear relationships present in the data. For spatial prediction of discrete and continuous geographic data, the RF is widely used in spatial prediction tasks by integrating multiple classification and regression trees through Bagging to achieve better performance than a single decision tree. Therefore, the RF is used as the third base learner to further consider the nonlinear characteristics of geographical relationships. The process of implementing the RF is as follows:

1. The training set is resampled using a Bootstrap strategy, which is allocated into M sub-training sets S_1, S_2, …, S_M based on the resampling results.
2. A subset of features is selected for constructing each decision tree.

3. Independent decision tree models T_1, T_2, …, T_M are built for the M sub-training datasets.
4. The final output on the testing dataset is determined based on majority voting or averaging for classification and regression tasks, respectively.

3.3. Design of the ensemble strategy

3.3.1. Design of ensemble strategy considering spatial heterogeneity

The commonly used ensemble learning strategies include voting, averaging and meta-learning. Voting is applied to classification tasks, where the final output category is determined by majority rule. The averaging method consists of two strategies: average ensemble (SA) and weighted ensemble (WA). The final prediction result of the SA method is the average of the results of all base learners, while the final prediction result of the WA method is obtained by weighted averaging using the accuracy evaluation metrics of the base learners as weights. The meta-learning method aims to train a meta-model using the output of the base learner as input to get the final prediction result.

The meta-model designed in the current research is usually based on the assumption that the samples are independent and identically distributed, allowing different base learners to have globally fixed weights. As an example, the linear regression model (LinReg) is used as a meta-model, which globally regresses the output results F_{GWR}, F_{GOS} and F_{RF} of the three base learners (GWR, GOS and RF). The ensemble prediction result $\hat{Y}$ is calculated as:

$$\hat{Y} = \beta_0 + \beta_1 F_{GWR} + \beta_2 F_{GOS} + \beta_3 F_{RF} \tag{7}$$

where β_0 is the constant term coefficient and β_l is the global average regression coefficient of the output results of the lth base learner, which can be calculated by ordinary least squares method, where $l = 1, 2, 3$.

Due to the spatial heterogeneity of the geographic elements, the relationship between the outputs of the base learners and the final ensemble result shows variability in different spatial locations. The regression coefficients obtained from the solution of the ordinary linear regression model are the optimal unbiased estimates based on all data samples, which can be considered as the average level of the relationship between the outputs of the base learners and the final ensemble result over the entire study area. The difference in this relationship in different spatial locations can be considered as the fluctuation of the average level produced by spatial heterogeneity. Therefore, this study adds a spatial heterogeneous ensemble weight matrix W_i to the global average regression coefficient β_l^{OLS} to measure the extent of this fluctuation, $W_i = [w_{i1}, w_{i2}, w_{i3}]$. The ensemble prediction result $\hat{Y}_i$ is calculated as:

$$\hat{Y}_i = w_{i1}\beta_1^{OLS}F_{i1} + w_{i2}\beta_2^{OLS}F_{i2} + w_{i3}\beta_3^{OLS}F_{i3} \tag{8}$$

where β_k^{OLS} denotes the global average ensemble weight of the lth base learner, which can be computed by ordinary least squares method. F_{i1}, F_{i2} and F_{i3} are the output results of GWR, GOS and RF for the sample p_i, respectively. w_{ik} is the spatial heterogeneous ensemble weight of the kth base learner for the sample p_i. Obviously, the key

to the implementation of the above model is to accurately solve the spatial heterogeneous ensemble weight matrix W_i.

3.3.2. Design of spatially weighted ensemble neural network

In geographical modelling of spatial heterogeneity, it is often necessary to construct a weighting kernel function parameterized by a measure of spatial proximity and then solve for the regression coefficients using the local least squares method. Therefore, the accurate solution of the kernel function is the key to achieving accurate modelling of spatial heterogeneity. Although there are many such functions, their simple does not always adequately describe the complex nonlinear effects of spatial proximity on the ensemble weights, leading to difficulties in the representation of spatial heterogeneity for complex geographical relationships. Indeed, the exact construction of weight kernel functions essentially belongs to the solution of complex nonlinear problems. Considering the strong ability of neural network models to solve complex nonlinear problems, this paper designs the SWENN to accurately solve the W_i (Figure 2).

The SWENN takes the spatial proximity vectors as inputs and utilizes the neural network structure with high-dimensional topology and a gradient descent algorithm based on differential equations to achieve the expression of the complex nonlinear relationship between spatial proximity and ensemble weights:

$$W_i = SWENN([d_{i1}, d_{i2} \ldots d_{in}]^T) \tag{9}$$

where $[d_{i1}, d_{i2} \ldots d_{in}]$ is the spatial proximity vector for the sample p_i, n is the number of observed locations, d_{ij} is the Euclidean distance between the sample p_i and p_j.

The SWENN uses the multi-layer perceptron as the backbone network, with two hidden layers between the input and output layers. The layers of the neural network are concatenated with full connections and the dropout technique is introduced to enhance the generalization of the network. In addition, a parametric rectified linear

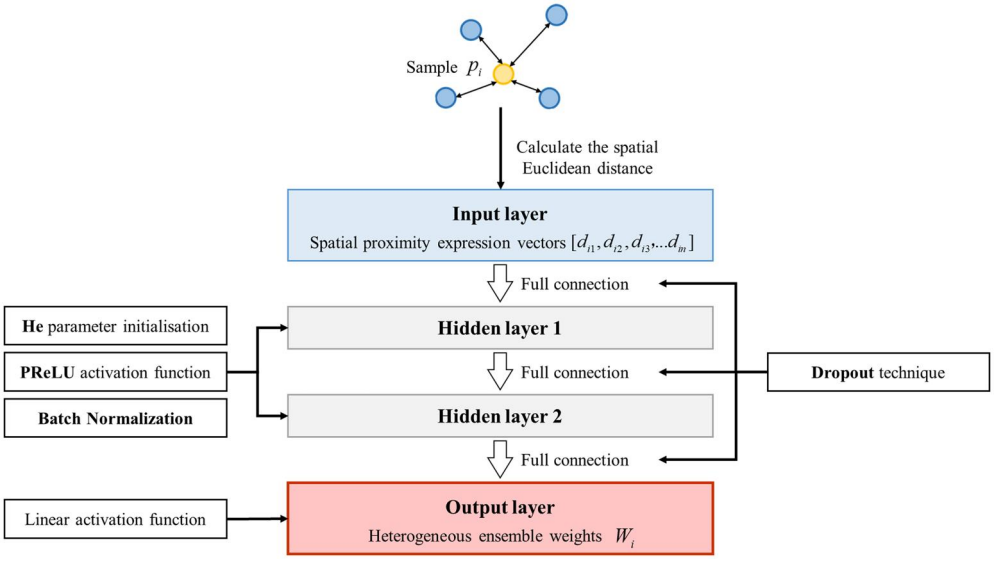

Figure 2. The structure of the SWENN module.

unit (PReLU) is used as the activation function in the hidden layer and the initial parameters are set using the He parameter initialization method. Meanwhile, the Batch Normalization layer is set to normalize the data to improve the training speed of the model. The dimensions of the output layer are consistent with the number of base learners, resulting in a spatial heterogeneous ensemble weight W_i of dimension 3.

3.4. Training framework for GSH-EL

Due to the use of neural networks to solve the spatially heterogeneous ensemble weights in GSH-EL, neural networks are prone to problems such as underfitting or overfitting, gradient vanishing or explosion during training. To improve the training ability of the GSH-EL, a training framework is used for the GSH-EL (Figure 3). The main steps are:

1. According to the principles of meta-learning method, three base learners are trained using observed locations in a cross-validation manner, and the predictions of the base learners are used as features to generate new datasets to estimate the following ensemble weights.
2. The new dataset is randomly divided into training, validation, and testing datasets with TR_n, V_n, TS_n samples respectively.
3. The global average ensemble weight β_l^{OLS} for each base learner is calculated based on the training dataset samples by ordinary least squares method.
4. The spatial proximity vector is constructed as an input to the SWENN by calculating the Euclidean distance from the unobserved location p_i to all the training samples. In the forward propagation process, the ensemble weight coefficient of the base learner is obtained by multiplying the spatial heterogeneity ensemble weight W_i output by SWENN with the global average ensemble weight β_l^{OLS}, and are further multiplied by the output features of base learners to obtain the prediction results. In the backpropagation process, the loss function is calculated by comparing the predicted values $\hat{Y}_i$ with the true values Y_i, and the model parameters are iteratively updated until convergence.

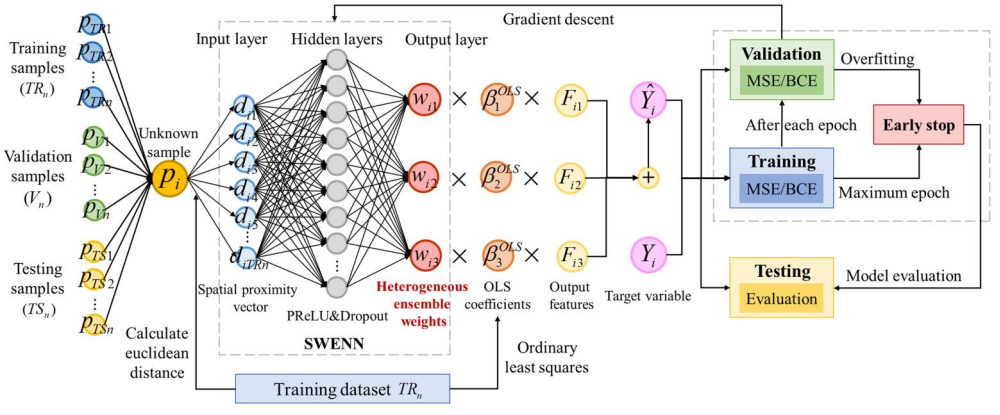

Figure 3. Training framework of the GSH-EL.

5. The overfitting problem is considered to occur if the accuracy metrics of the validation dataset show a continuous upward or flat trend during the training process, then the early stopping strategy is used to stop training.

4. Experiments

The proposed method is evaluated using two real tasks, including the regression task of continuous variables on the $PM_{2.5}$ air quality dataset in China and the binary classification task of discrete variables on the landslide dataset in Hong Kong, China.

4.1. Case one: Spatial prediction for $PM_{2.5}$ distributions

4.1.1. Study area and data

The observed annual average $PM_{2.5}$ concentration for 2018 was downloaded from the China National Environmental Monitoring Center (http://www.cnemc.cn). Considering that $PM_{2.5}$ concentration is affected by factors such as aerosol optical depth (AOD), surface elevation (DEM), temperature (TEMP), precipitation (TP), wind speed (WS), wind direction (WD), and relative humidity (RH), these factors are used as explanatory variables for spatial prediction of $PM_{2.5}$ concentration. The AOD concentration is derived from LAADS DAAC (https://ladsweb.modaps.eosdis.nasa.gov) and the other variables are derived from the data products of the ECMWF global climate reanalysis model ERA5 (https://cds.climate.copernicus.eu). The dataset (Table 1, Figure 4) contains a total of 1456 observation records, which are randomly divided into training (815), validation (204), and testing (437) datasets.

4.1.2. Experimental implementation

4.1.2.1. Comparative methods. Four existing ensemble strategies are used as comparison models to evaluate the proposed method, including average ensemble (SA-EL), weighted ensemble (WA-EL), linear regression ensemble (LinReg-EL), and geographically weighted regression ensemble (GWR-EL). WA-EL uses the R^2 metrics of each base learner as weights. LinReg-EL adopts the least square method to calculate the global weighting coefficient. GWR-EL uses the mgwr2.1.2 Python package to compare the results of the Gaussian and Bi-square kernel functions with fixed bandwidth and adaptive bandwidth. The adaptive Gaussian kernel function is finally determined, and the optimal bandwidth is set to 257 samples by AICc.

Table 1. Descriptive statistics of $PM_{2.5}$ dataset.

	Mean	Max	Min	Std
$PM_{2.5}/(ug/m^3)$	41.070	129.110	8.037	13.653
AOD	530.23	1200.82	64.86	186.97
DEM/m	396.61	4525.00	−6.00	660.73
TEMP/K	287.702	297.213	271.645	5.270
TP/m	9.4×10^{-5}	2.1×10^{-4}	2.0×10^{-6}	4.7×10^{-5}
WS/(m/s)	1.512	16.228	3.5×10^{-7}	2.074
WD/°	150.481	238.715	81.689	44.255
RH/%	62.217	86.685	24.962	12.406

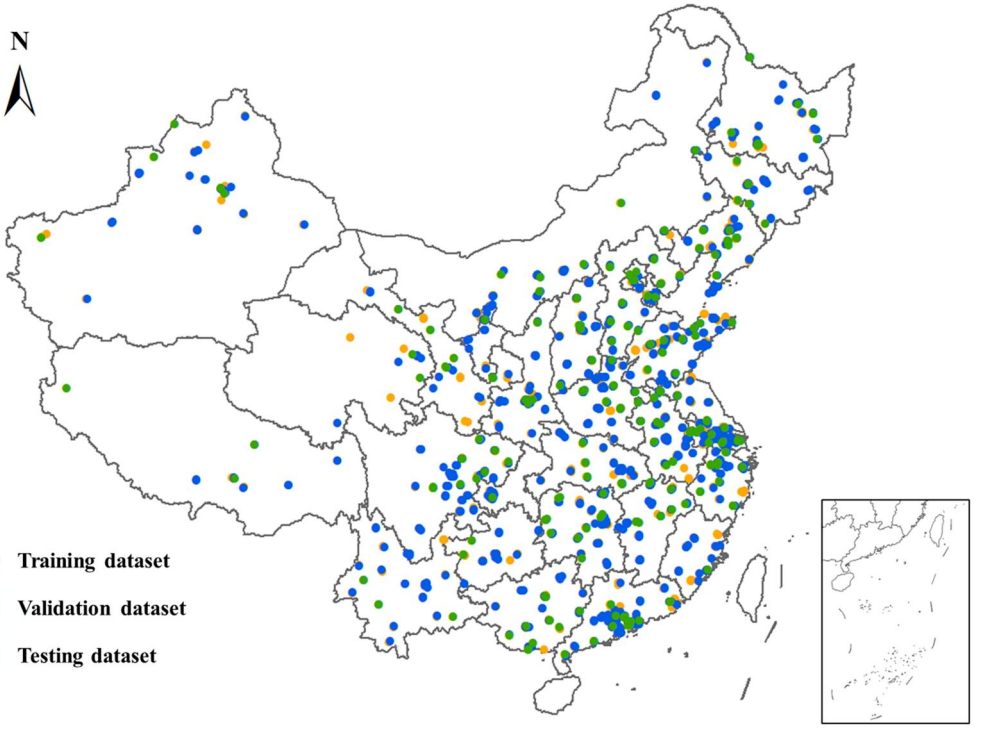

Figure 4. Spatial distribution of $PM_{2.5}$ dataset. IJGIS remains strictly neutral with respect to jurisdictional claims on disputed territories and the naming conventions used in the maps included in the figure.

4.1.2.2. Evaluation metrics. Six accuracy evaluation metrics commonly used in regression tasks are selected to quantitatively evaluate the proposed method, including mean absolute error (*MAE*), mean absolute percentage error (*MAPE*), root mean squared error (*RMSE*), determination coefficient (R^2), adjusted determination coefficient ($Adj.R^2$) and Akaike information criterion (*AIC*):

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

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} \tag{13}$$

$$Adj.R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} \times \frac{n-1}{n-p-1} \tag{14}$$

$$AIC = n\log_e\left(\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\right) + 2p \tag{15}$$

Table 2. Settings of architectures and hyper-parameters for GSH-EL on the $PM_{2.5}$ dataset.

GSH-EL	Input	Hidden 1	Hidden 2	Output	
	815	256	32	3	
Hyper-parameters	Learning rate	Maximum epoch	Optimizer	Batch size	Dropout
	0.01	200	Adam	64	0.2

where p is the total number of explanatory variables. *MAE*, *MAPE* and *RMSE* reflect the prediction error of the model such that smaller values correspond to better predictive performance. R^2 is between 0 and 1, with the closer to 1 the better predictive performance of the model. $Adj.R^2$ and *AIC* reflect the fitting degree between the actual and predicted values considering model complexity.

4.1.2.3. Parameter tuning. For the parameter configuration of the base learner, the GWR uses the adaptive Bi-square kernel function, and the optimal bandwidth is determined to be 87 samples. The similarity threshold in the GOS is determined to be 0.003 by the CV method. The parameters of the RF are determined by the grid search, and 250 sub-regression trees are set up, each containing at least 1 sample in the leaf nodes and at least 2 samples in the non-leaf nodes. For the GSH-EL, the SWENN consists of an input layer, two hidden layers, and an output layer. The cross-search strategy is adopted to determine the optimal number of neurons in the two hidden layers. The network structure and hyper-parameter for GSH-EL are shown in Table 2.

In addition, the mean squared error (MSE) is used as a loss function for the model training process and as the overfitting evaluation indicator for the validation dataset. The maximum overfitting tolerance is set to 20 epochs, and if the continuous upward or flat trend of the indicator exceeds this value, we stop training and return to the optimal model parameters previously recorded. Figure 5 shows the performance variations of training and validation datasets for GSH-EL. The MSE value of the training dataset keeps decreasing and converges after the 150th epoch. However, the MSE value of the validation dataset started to maintain an upward or flat trend after dropping to the lowest value at the 163rd epoch, which can be considered overfitted. Therefore, the model at epoch $=$ 163 is taken as the optimal model.

4.1.3. Results and discussion

4.1.3.1. Quantitative analysis of prediction accuracy. The experimental results of the three base learners and five ensemble learning methods on the $PM_{2.5}$ testing dataset are shown in Table 3. In the base learners, the GWR performs best on the RMSE, R^2, Adj. R^2, and AIC metrics, and the RF performs best on the MAE and MAPE metrics. The GOS has the lowest prediction accuracy of the three base learner models due to its sensitivity to outlier samples.

The five ensemble learning methods use different strategies to integrate the base learners, and their prediction accuracy is higher than that of a single base learner. This demonstrates that the ensemble of multiple models can integrate the advantages of diverse models for the prediction task, thereby achieving better accuracy than a single model. SA-EL, WA-EL, and LinReg-EL assign globally fixed ensemble weights to the base learners, resulting in minimal differences in prediction accuracy. GWR-EL considers spatial heterogeneity in the ensemble process and outperforms SA-EL, WA-EL, and

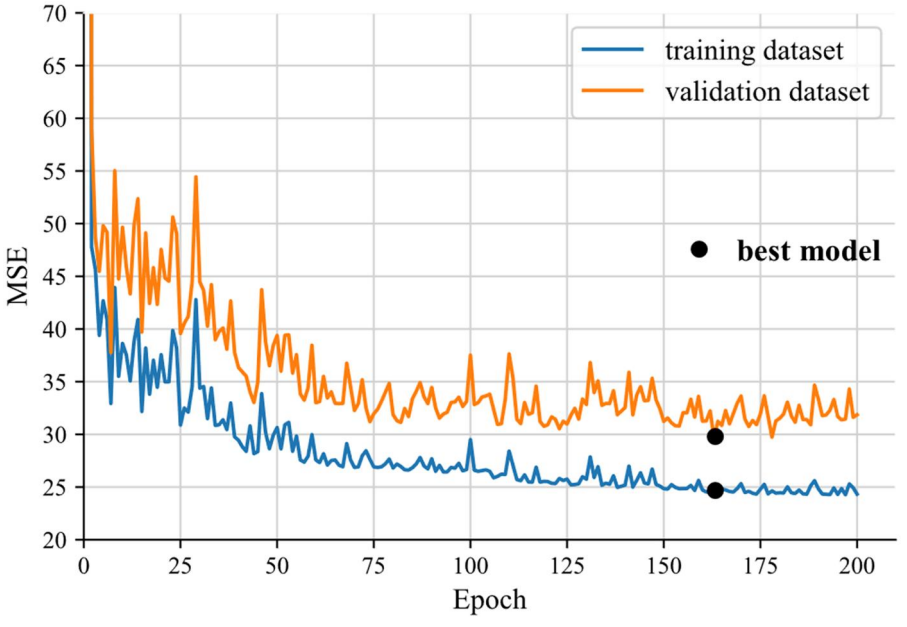

Figure 5. Performance variations of training and validation datasets for GSH-EL on the $PM_{2.5}$ dataset.

Table 3. Experimental results of different methods on the $PM_{2.5}$ testing dataset.

	Model	MAE	MAPE	RMSE	R^2	Adj.R^2	AIC
Base learner	GOS	4.430	0.1216	6.060	0.8032	0.8000	1588.7
	GWR	4.147	0.1138	5.770	0.8217	0.8188	1545.8
	RF	4.029	0.1118	5.810	0.8191	0.8161	1551.9
Ensemble method	SA-EL	3.960	0.1079	5.517	0.8369	0.8358	1498.7
	WA-EL	3.959	0.1078	5.516	0.8370	0.8359	1498.5
	LinReg-EL	3.931	0.1055	5.553	0.8348	0.8337	1504.3
	GWR-EL	3.915	0.1054	5.540	0.8356	0.8344	1502.2
	GSH-EL	**3.863**	**0.1052**	**5.433**	**0.8419**	**0.8408**	**1485.3**

LinReg-EL in MAE metrics. However, the simple structure of the adaptive Gaussian kernel function used in the GWR limits the predictive ability of the GWR-EL, resulting in no evidence improvement in other metrics compared to SA-EL, WA-EL, and LinReg-EL. The proposed GSH-EL re-solves the weight kernel function, and all indicators are better than the existing models, indicating that the accurate expression of spatial heterogeneity in the ensemble strategy can effectively improve the prediction accuracy of the ensemble model.

4.1.3.2. Qualitative analysis of prediction accuracy. To qualitatively analyse the predictive performance of different ensemble models, we present the scatter plots of predicted and true values of WA-EL, LinReg-EL, GWR-EL, and GSH-EL on the testing dataset (Figure 6). The GSH-EL values are mainly distributed around the 1:1 line with a correlation coefficient of $r = 0.9177$, which is the highest among the three compared methods. In particular, for outlier samples on the testing dataset, such as data samples with $PM_{2.5}$ concentrations above 75ug/m^3, the GSH-EL has relatively strong predictive power, which proves that the GSH-EL has strong adaptability to outlier samples.

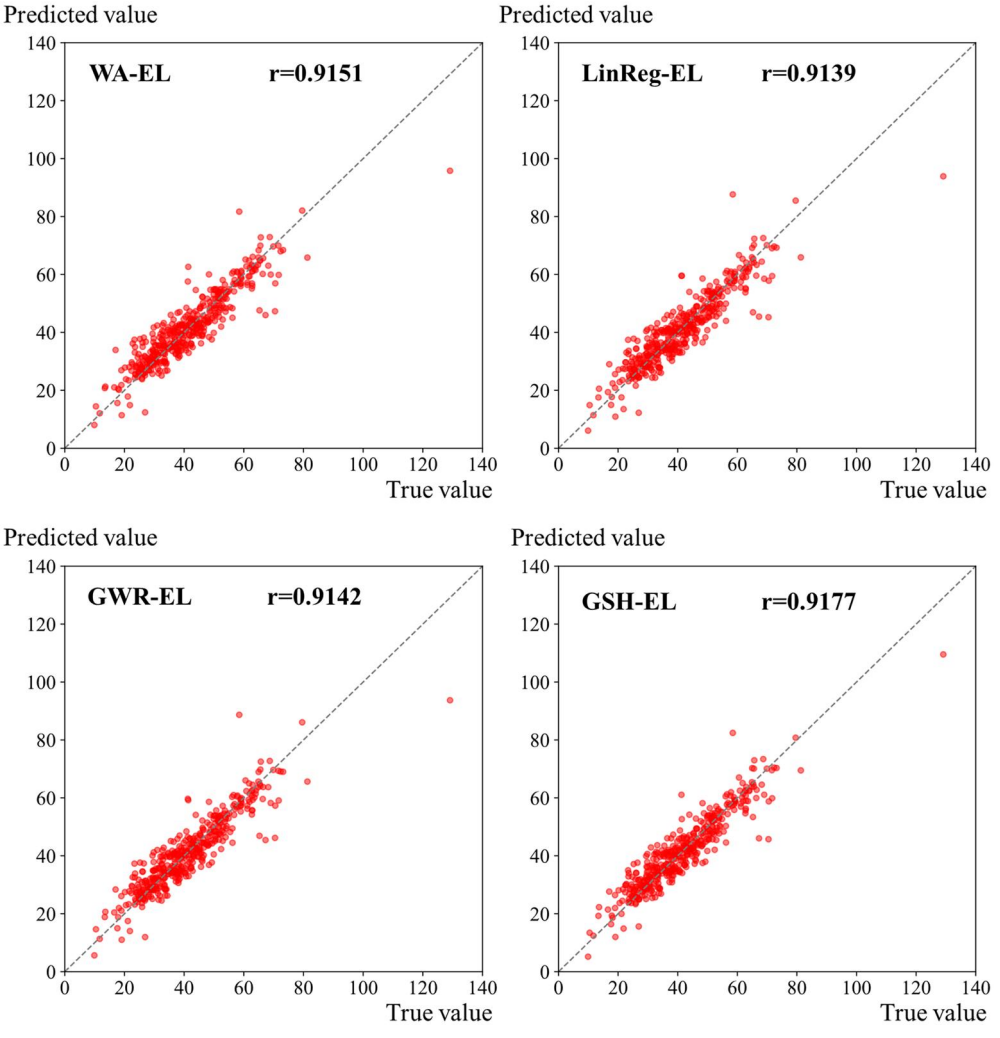

Figure 6. Scatter plots of predicted and true values of different ensemble models on the $PM_{2.5}$ dataset.

By comparing the spatial distribution and percentage of prediction errors of the different methods (Figure 7), it is found that the number of samples with an absolute error greater than 12 in the prediction results of the GSH-EL accounts for 1.15%, which is lower compared to the WA-EL (3.20%), LinReg-EL (2.98%), and GWR-EL (3.21%), and that 52.17% of the samples had an absolute error of 3 or less in the prediction results of the GSH-EL, more than WA-EL (49.20%), LinReg-EL (49.88%), and GWR-EL (50.34%), which explains the better overall prediction results of the GSH-EL model.

4.1.3.3. Effect of base learner on prediction accuracy. Considering that the three base learners describe geographical relationships from different perspectives, this study combines them in pairs to train three different versions of GSH-EL, in order to further explore the effect of different base learners on its prediction accuracy. Different base learners contribute differently to the final prediction result (Table 4). The removal of

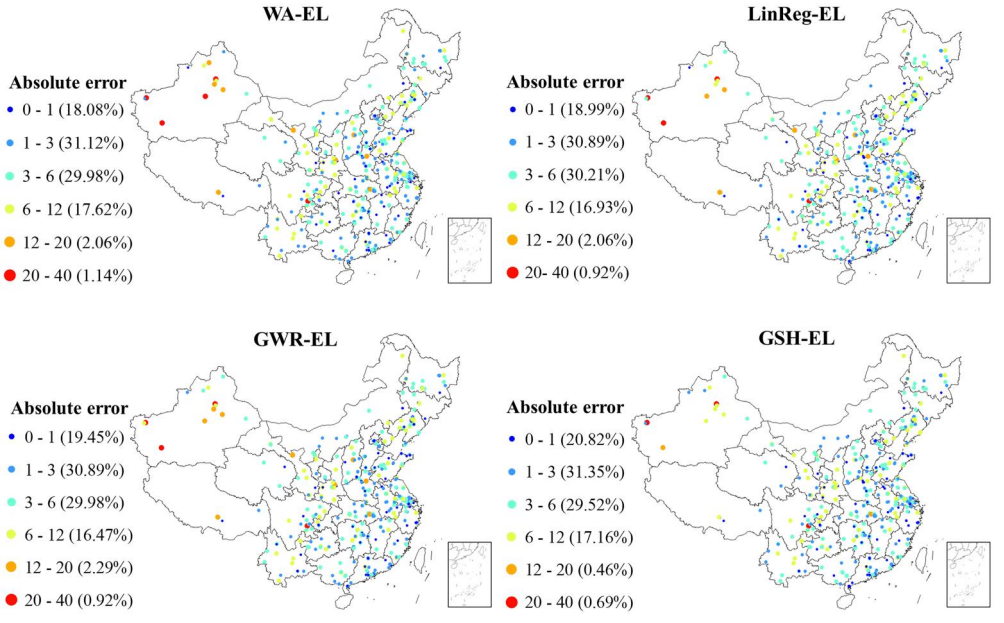

Figure 7. Comparison of the absolute errors of the WA-EL, LinReg-EL, GWR-EL and GSH-EL on the $PM_{2.5}$ dataset. IJGIS remains strictly neutral with respect to jurisdictional claims on disputed territories and the naming conventions used in the maps included in the figure.

Table 4. Comparison of accuracy of ablation experiments for GSH-EL on the $PM_{2.5}$ dataset.

Model	MAE	MAPE	RMSE	R^2
GOS + RF	3.951	0.1091	5.673	0.8276
GWR + GOS	3.978	0.1090	5.550	0.8350
GWR + RF	3.902	0.1062	5.511	0.8373
GWR + GOS + RF (GSH-EL)	**3.863**	**0.1052**	**5.433**	**0.8419**

the GWR has a large effect on the overall prediction accuracy, while the removal of the GOS has a relatively small effect. In addition, we found that even if the specific base learners are removed, the pairwise ensemble model can still achieve better prediction accuracy than the single model, which proves the superiority of the ensemble model in spatial prediction tasks. Furthermore, we found that better performance of the base learning model led to better prediction accuracy of the ensemble model. For example, the GWR and RF perform better among the three base learners, as a result, the ensemble of GWR and RF produces better prediction results than the ensemble of GOS and RF, as well as the ensemble of GWR and GOS. Overall, these results demonstrate that the reasonable expression of spatial dependencies by base learners is crucial for improving the prediction accuracy of ensemble models.

4.2. Case two: Spatial prediction for landslide susceptibility

4.2.1. Study area and data

The Hong Kong landslide dataset was randomly sampled to obtain 1000 historical landslide sample sites and 1000 non-landslide sample sites. The land use (LU), lithology (LITH), elevation (DEM), slope (SLO), curvature (CUR), aspect (ASP), normalised

difference vegetation index (NDVI), stream power index(SPI), topographic wetness index (TWI), distance to nearest road (Road_D), distance to nearest drainage (DRA_D), distance to nearest catchment (CAT_D), distance to the most faulted road (Fault_D), and deformation velocity (DV) are used as explanatory variables to predict the landslide susceptibility. These datasets were downloaded from the Computer Network Information Center, Chinese Academy of Science (https://www.cnic.cn), and United States Geological Survey (https://www.usgs.gov). The dataset (Table 5, Figure 8) was randomly divided into training (1120), validation (280), and testing (600) datasets.

4.2.2. Experimental implementation

4.2.2.1. Comparative methods. The voting method, logistic regression, and GWR are used as ensemble strategies to build comparative models, including Voting-EL, LogReg-EL, and GWR-EL. Voting-EL uses a minority-majority strategy to process the binary classification results of three base learners. GWR-EL uses a similar methodology as in Section 4.1.2 to determine the optimal bandwidth as 130.

4.2.2.2. Evaluation metrics. The area under the Receiver Operating Characteristic (ROC) curve (AUC) is a summary metric that quantifies the overall performance of the model across all possible thresholds. The closer the AUC value is to 1, the better the performance. In addition, six precision evaluation indexes including Overall Accuracy (OA), Precision, Recall, F1 Score, Matthews Correlation Coefficient (MCC) and Intersection over Union (IOU) are selected to evaluate the prediction precision of the proposed method:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{19}$$

Table 5. Descriptive statistics of the landslide susceptibility dataset.

	Mean	Max	Min	Std
LU/cat	2.063	3	1	0.744
LITH/cat	2.422	3	1	0.742
DEM/m	136.475	930	−23	136.486
SLO/°	23.213	68.744	0.000	15.324
CUR/°	31.331	79.936	0.000	20.205
ASP/°	159.514	358.264	−1	113.541
NDVI	6093.883	8361	0	1986.475
SPI	0.005	4.621	−9.903	2.502
TWI	3.384	18.865	0.319	2.491
Road_D/m	105.711	2116.337	0.000	166.874
DRA_D/m	464.809	4698.640	0.003	519.348
CAT_D/m	2630.313	10739.570	0.000	2335.380
Fault_D/m	814.692	7281.344	0.533	827.112
DV/(mm/d)	−0.653	14.775	−13.098	2.357

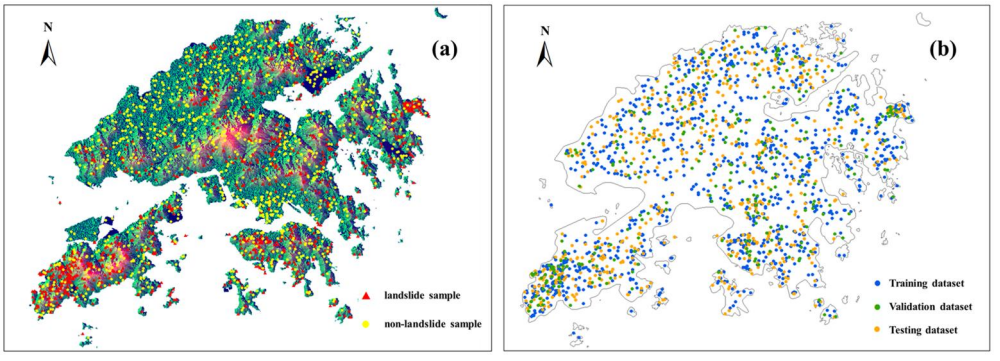

Figure 8. Spatial distribution of the landslide susceptibility dataset as (a) landslide and non-landslide and (b) training, validation and testing locations.

Table 6. Settings of architectures and hyper-parameters for GSH-EL on the landslide susceptibility dataset.

GSH-EL	Input	Hidden 1	Hidden 2	Output	
	1120	256	48	3	
Hyper-parameters	Learning rate	Maximum epoch	Optimizer	Batch size	Dropout
	0.01	120	Adam	64	0.15

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{20}$$

$$IOU = \frac{TP}{TP+FP+FN} \tag{21}$$

where TP, FP, and FN represent true examples, false positive examples, and false negative examples, respectively.

4.2.2.3. Parameter tuning. Regarding the parameter configuration of the base learner, a similar method as in Section 4.1.2 is used to determine the parameters of each model. The GWR sets an adaptive Bi-square kernel function with an optimal bandwidth of 218 samples. The GOS has a similarity threshold of 0.01 and the RF is set up with 300 sub-classification trees.

In addition to setting the input layer, two hidden layers, and the output layer, SWENN also adds a sigmoid function to the output layer, which transforms the model output features into a probability value from 0 to 1 to determine whether landslides are occurring or not. The settings of architectures and hyper-parameters for GSH-EL are shown in Table 6 and the performance variations of training and validation datasets for GSH-EL are shown in Figure 9.

4.2.3. Results and discussion

The experimental results of different methods on the test dataset are shown in Table 7, and the ROC curves are plotted based on the landslide susceptibility prediction results (Figure 10). The order of overall accuracy of the three base learners, from high to low, is RF, GWR, and GOS. The four ensemble methods are better than the three base learners in all evaluation metrics. Voting-EL directly processes the binary

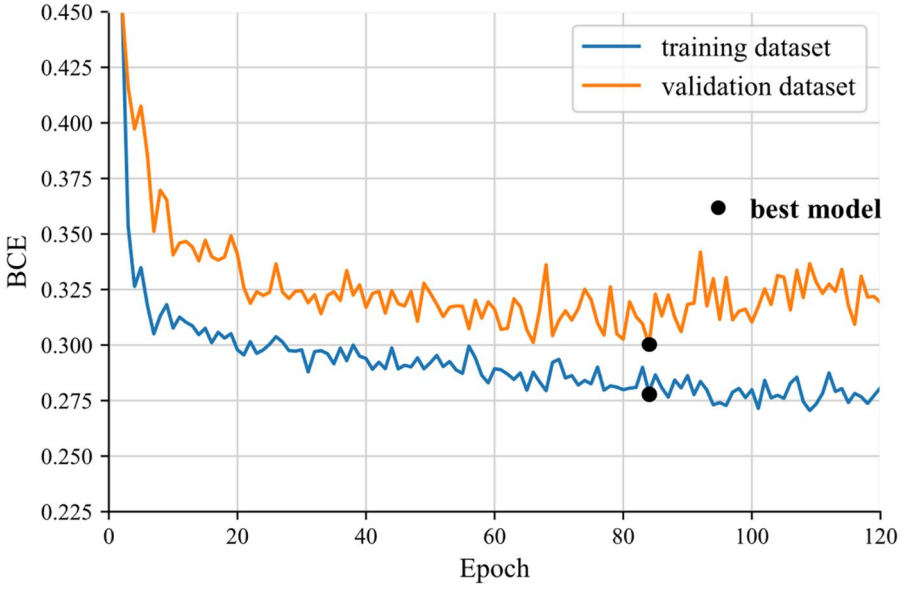

Figure 9. Performance variations of training and validation datasets for GSH-EL on the landslide susceptibility dataset.

Table 7. Experimental results of different methods on the landslide susceptibility testing dataset.

	Model	OA	Precision	Recall	F1	MCC	IOU	AUC
Base learner	GOS	0.820	0.804	0.847	0.825	0.641	0.702	0.9040
	GWR	0.852	0.863	0.863	0.849	0.704	0.738	0.9337
	RF	0.855	0.823	0.903	0.862	0.713	0.757	0.9503
Ensemble method	Voting-EL	0.865	0.850	0.887	0.868	0.731	0.767	–
	LogReg-EL	0.870	0.868	0.873	0.870	0.740	0.771	0.9546
	GWR-EL	0.882	0.871	0.897	0.883	0.764	0.791	0.9554
	GSH-EL	**0.895**	**0.881**	**0.913**	**0.897**	**0.791**	**0.813**	**0.9563**

classification results of three types of base learners. Although the strategy of a minority following the majority can avoid the discrimination error of a certain base learner, this method cannot obtain the correct results when the majority of base learners predict errors, resulting in lower overall accuracy. Compared to LogReg-EL performing global regression on the output probabilities of the three base learners to obtain globally fixed average parameters, GWR-EL calculates ensemble weights through local weighting, thus obtaining better prediction results than Voting-EL and LogReg-EL. GSH-EL can more accurately express spatial heterogeneity, resulting in optimal performance for all evaluation metrics, with an OA indicator of 89.5% and an AUC indicator of 95.63%.

5. Conclusions

Current research in spatial prediction tends to use heterogeneous ensemble learning methods for modelling geospatial data. However, existing studies usually adopt

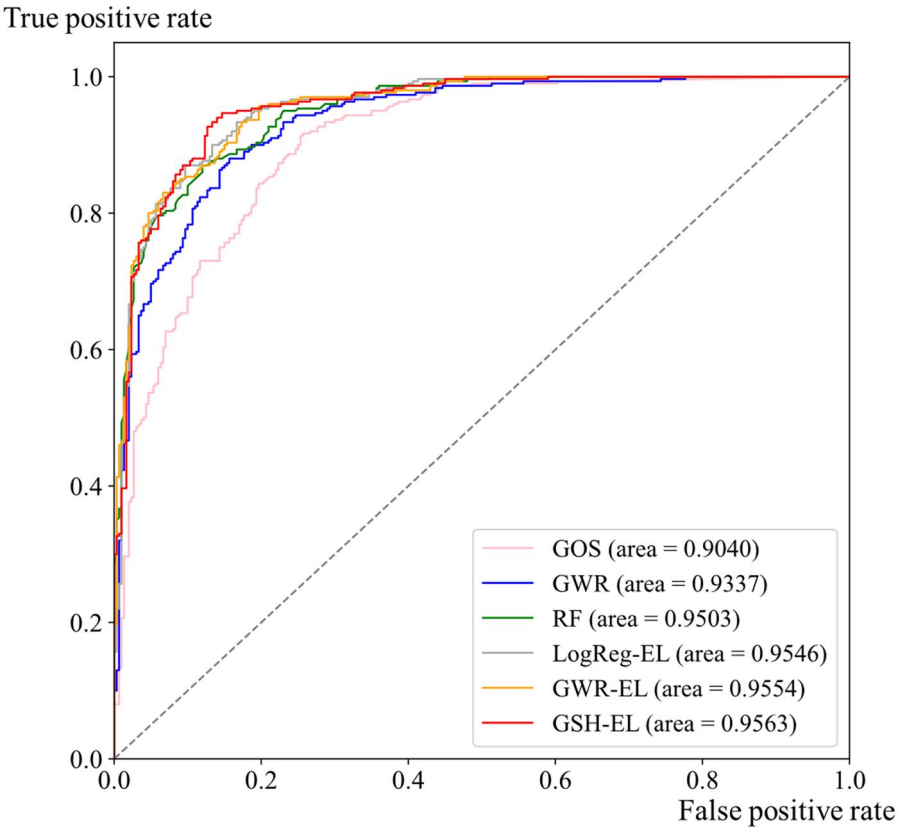

Figure 10. ROC curves of different methods on the landslide susceptibility testing dataset.

globally fixed ensemble weights when designing learning-based ensemble strategies, ignoring the statistical constraints of spatial heterogeneity on the ensemble process. While the simple kernel function structure used in the ensemble strategy considering spatial heterogeneity is unable to fully describe the complex nonlinear effect of spatial proximity on ensemble weights, resulting in the inability to accurately solve the spatial heterogeneity of complex geographic relationships, which can severely restrict the predictive ability of the model.

To address the above issues, this study proposes a novel ensemble spatial prediction method considering spatial heterogeneity. From the perspective of different geographic relationship expressions, the method fully considers the local spatial heterogeneity, global feature correlation, and nonlinear relationship of geographic elements by designing three base learners (GWR, GOS, and RF). In addition, the SWENN with adaptive learning capability is used to achieve an accurate expression of spatial heterogeneity in the ensemble strategy. This model utilizes the highly abstract expression ability and high-dimensional dynamic learning ability of neural networks to establish a complex nonlinear relationship between spatial proximity and ensemble weights, allowing for adaptive integration of base learners based on spatial patterns.

Finally, an ensemble learning framework considering spatial heterogeneity is designed to integrate the prediction results of the three base learners into the SWENN to yield more accurate prediction results.

The proposed method was evaluated using two datasets, a regression prediction task with continuous variables using the Chinese $PM_{2.5}$ air quality dataset, and a binary classification task with discrete variables using the Chinese Hong Kong landslide dataset. The experimental results show that the proposed method considering spatial heterogeneity achieves more accurate prediction results than the current ensemble learning strategies, which verifies the effectiveness and applicability of the proposed method. Therefore, we can conclude that accurate expression of spatial heterogeneity during model ensemble can effectively improve the prediction ability of ensemble models. In addition, we find that a reasonable representation of spatial dependencies by the base learner is crucial for improving the prediction accuracy of the ensemble model.

Although the proposed method shows excellent prediction performance, there are some limitations. Firstly, only the traditional Euclidean distance is used in our model to characterize spatial proximity. Indeed, spatial proximity is generated by the interaction of multiple distances, such as topological network distance, azimuthal distance, Minkowski distance, etc. Therefore, the unified expression of spatial proximity needs to be investigated to further improve the accuracy of solving the weight kernel function. Secondly, this paper designs an ensemble learning strategy considering spatial heterogeneity with the powerful learning capability of neural networks. However, the black box process of neural networks to some extent limits the interpretability of the ensemble model. Further research is needed to explore the mechanism of geographic spatial heterogeneity in the ensemble process of base learners, and investigate the intensity and response rules of different base learners in the ensemble process. Finally, our study focuses mainly on the expression of spatial heterogeneity, ignoring the temporal non-stationarity of the geospatial data. Therefore, further work can be carried out on ensemble learning methods considering spatiotemporal heterogeneity to further expand the application of ensemble learning in the field of geosciences.

Acknowledgement

The authors would also like to thank the anonymous referees for their helpful comments and suggestions.

Disclosure statement

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

Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences [grant numbers XDB0740100-02], the National Natural Science Foundation of China [grant number 42371469, 42101423], the Innovation Project of LREIS [grant number KPI003 and YPI002] and China National Postdoctoral Support Program for Innovative Talents [grant number BX20230360].

Notes on contributors

Shifen Cheng is an Associate Professor with the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. His research interests include spatiotemporal data mining, urban computing, and intelligent transportation.

Lizeng Wang is a Ph. D. candidate in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His research interests include spatiotemporal data mining, spatiotemporal prediction, and urban computing.

Peixiao Wang is a Postdoctoral Fellow from State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences. His research topics include spatiotemporal data mining, and spatiotemporal prediction, especially focus on spatiotemporal prediction of transportation systems.

Feng Lu is a Professor in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His research interests cover trajectory data mining, computational transportation science and location-based services.

Data and codes availability statement

The data and code that support the findings of this study are available in 'figshare.com' with the identifier https://doi.org/10.6084/m9.figshare.23947284.v2.

References

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

Cai, Z., *et al.*, 2023. A traffic data interpolation method for IoT sensors based on spatio-temporal dependence. *Internet of Things*, 21, 100648.

Chen, J., *et al.*, 2019. Stacking machine learning model for estimating hourly PM2.5 in China based on Himawari 8 aerosol optical depth data. *The Science of the Total Environment*, 697, 134021.

Chen, X., *et al.*, 2022. Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 23 (8), 12301–12310.

Cheng, S., and Lu, F., 2017. A two-step method for missing spatio-temporal data reconstruction. *ISPRS International Journal of Geo-Information*, 6 (7), 187.

Cheng, S., Peng, P., and Lu, F., 2020. A lightweight ensemble spatiotemporal interpolation model for geospatial data. *International Journal of Geographical Information Science*, 34 (9), 1849–1872.

Dai, Z., *et al.*, 2022. Geographically convolutional neural network weighted regression: a method for modeling spatially non-stationary relationships based on a global spatial proximity grid. *International Journal of Geographical Information Science*, 36 (11), 2248–2269.

Deng, M., *et al.*, 2016. A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets. *ISPRS International Journal of Geo-Information*, 5 (2), 13.

Du, Z., *et al.*, 2020. Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. *International Journal of Geographical Information Science*, 34 (7), 1353–1377.

Du, Z., *et al.*, 2021. A spatially weighted neural network based water quality assessment method for large-scale coastal areas. *Environmental Science & Technology*, 55 (4), 2553–2563.

Fang, Z., *et al.*, 2021. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. *International Journal of Geographical Information Science*, 35 (2), 321–347.

Feng, L., *et al.*, 2020. Estimating hourly and continuous ground-level PM2.5 concentrations using an ensemble learning algorithm: The ST-stacking model. *Atmospheric Environment*, 223, 117242.

Ge, Y., *et al.*, 2014. Upscaling sensible heat fluxes with area-to-area regression kriging. *IEEE Geoscience and Remote Sensing Letters*, 12 (3), 656–660.

Ge, Y., *et al.*, 2019. Principles and methods of scaling geospatial Earth science data. *Earth-Science Reviews*, 197, 102897.

Hagenauer, J., and Helbich, M., 2022. A geographically weighted artificial neural network. *International Journal of Geographical Information Science*, 36 (2), 215–235.

Hao, Y., and Tian, C., 2019. A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting. *Applied Energy*, 238, (December 2018), 368–383.

Jiang, Z., 2019. A survey on spatial prediction methods. *IEEE Transactions on Knowledge and Data Engineering*, 31 (9), 1645–1664.

Karpatne, A., *et al.*, 2019. Machine learning for the geosciences: challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31 (8), 1544–1554.

Lei, M., *et al.*, 2022. Bayesian kernelized matrix factorization for spatiotemporal traffic data imputation and kriging. *IEEE Transactions on Intelligent Transportation Systems*, 23 (10), 18962–18974.

Li, L., 2019. Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. *Remote Sensing*, 11 (11), 1378.

Li, L., *et al.*, 2017. Constrained mixed-effect models with ensemble learning for prediction of nitrogen oxides concentrations at high spatiotemporal resolution. *Environmental Science & Technology*, 51 (17), 9920–9929.

Li, L., *et al.*, 2019. Missing value imputation for traffic-related time series data based on a multi-view learning method. *IEEE Transactions on Intelligent Transportation Systems*, 20 (8), 2933–2943.

Liu, P., and Biljecki, F., 2022. A review of spatially-explicit GeoAI applications in urban geography. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102936.

Liu, W., Du, P., and Wang, D., 2015. Ensemble learning for spatial interpolation of soil potassium content based on environmental information. *PLoS One*, 10 (4), e0124383.

Luo, P., *et al.*, 2023. A generalized heterogeneity model for spatial interpolation. *International Journal of Geographical Information Science*, 37 (3), 634–659.

Murray, N.L., *et al.*, 2019. A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation. *Environmental Research*, 178, 108601.

Narendra Babu, C., Sure, P., and Bhuma, C.M., 2021. Sparse Bayesian Learning Assisted Approaches for Road Network Traffic State Estimation. *IEEE Transactions on Intelligent Transportation Systems*, 22 (3), 1733–1741.

Pham, B.T., *et al.*, 2021. Ensemble machine learning models based on Reduced Error Pruning Tree for prediction of rainfall-induced landslides. *International Journal of Digital Earth*, 14 (5), 575–596.

Qin, M., *et al.*, 2019. A matrix completion-based multiview learning method for imputing missing values in buoy monitoring data. *Information Sciences*, 487, 18–30.

Requia, W.J., *et al.*, 2020. An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. *Environmental Science and Technology*, 54 (18), 11037–11047.

Saez, M., and Barceló, M.A., 2022. Spatial prediction of air pollution levels using a hierarchical Bayesian spatiotemporal model in Catalonia, Spain. *Environmental Modelling and Software*, 151, 105369.

Sarker, S., *et al.*, 2021. Spatial prediction of seaweed habitat for mariculture in the coastal area of Bangladesh using a Generalized Additive Model. *Algal Research*, 60, 102490.

Shafizadeh-Moghadam, H., *et al.*, 2018. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of Environmental Management*, 217, 1–11.

Song, X.Y., *et al.*, 2021. A machine learning approach to modelling the spatial variations in the daily fine particulate matter (PM2.5) and nitrogen dioxide (NO2) of Shanghai, China. *Environment and Planning B: Urban Analytics and City Science*, 48 (3), 467–483.

Song, Y., 2022a. Geographically optimal similarity. *Mathematical Geosciences*, 55 (3), 295–320.

Song, Y., 2022b. The second dimension of spatial association. *International Journal of Applied Earth Observation and Geoinformation*, 111, 102834.

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.

Su, H., *et al.*, 2022. Estimating regional PM2.5 concentrations in china using a global-local regression model considering global spatial autocorrelation and local spatial heterogeneity. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, 14 (18), 4545.

Wang, J.-F., *et al.*, 2013. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. *Environment and Planning A: Economy and Space*, 45 (10), 2515–2534.

Wang, P., *et al.*, 2022. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. *International Journal of Geographical Information Science*, 36 (6), 1231–1257.

Wang, Y., *et al.*, 2020. A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. *CATENA*, 188, 104425.

Wu, C., *et al.*, 2023. A hybrid deep learning model for regional O3 and NO2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network. *Environmental Pollution (Barking, Essex: 1987)*, 320, 121075.

Wu, S., *et al.*, 2021. Geographically and temporally neural network weighted regression for modeling spatiotemporal non-stationary relationships. *International Journal of Geographical Information Science*, 35 (3), 582–608.

Xu, C.D., *et al.*, 2013. Interpolation of missing temperature data at meteorological stations using P-BSHADE. *Journal of Climate*, 26 (19), 7452–7463.

Yang, Y., *et al.*, 2019. Local-scale landslide susceptibility mapping using the B-GeoSVC model. *Landslides*, 16 (7), 1301–1312.

Yao, S., and Huang, B., 2023. Spatiotemporal interpolation using graph neural network. *Annals of the American Association of Geographers*, 113 (8), 1856–1877.

Yi, Z., *et al.*, 2021. Inferencing hourly traffic volume using data-driven machine learning and graph theory. *Computers, Environment and Urban Systems*, 85, 101548.

Zeng, H., *et al.*, 2022. Graph neural networks with constraints of environmental consistency for landslide susceptibility evaluation. *International Journal of Geographical Information Science*, 36 (11), 2270–2295.

Zhou, Z.H., 2012. *Ensemble methods: foundations and algorithms*. New York: Chapman and Hall/CRC.

Zhu, A., *et al.*, 2018. Spatial prediction based on third law of geography. *Annals of GIS*, 24 (4), 225–240.

Zhu, D., *et al.*, 2020. Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science*, 34 (4), 735–758.