Environmental Science and Pollution Research (2022) 29:33205–33217
https://doi.org/10.1007/s11356-021-17513-3

RESEARCH ARTICLE

Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model

Yaowen Luo[1] · Jianguo Yan[2] · Stephen C. McClure[2] · Fei Li[1,2]

Received: 3 June 2021 / Accepted: 9 November 2021 / Published online: 13 January 2022

Abstract

Correlations between socioeconomic factors and poverty in regression models do not reflect actual relationships, especially when data exhibit patterns of spatial heterogeneity. Spatial regression models can estimate the relationships between socioeconomic factors and poverty in defined geographical areas, explaining the imbalanced distribution of poverty, but the relationships between these factors and poverty are not always linear however, and conventional simple linear local regression models do not accurately capture these nonlinear relationships. To fill this gap, we used a local regression method, geographically weighted random forest regression (GW-RFR), that integrates a spatial weight matrix (SWM) and random forest (RF). The GW-RFR evaluates the spatial variations in the nonlinear relationships between variables. A county-level poverty data set of China was employed to estimate the performance of the GW-RFR against the random forest (RF). In this poverty application, the value of R^2 was 0.128 higher than that of the RF, the NRMSE value was 1.6% lower than the RF, and the MAE value was 0.295 lower than the RF. These results showed that the relationship between poverty factors and poverty varies with space at the county level in China, and the GW-RFR was suitable for dealing with nonlinear relationships in local regression analysis.

Keywords Poverty · Spatial variation · Nonlinear · Variable importance · Random forest

Introduction

Eradicating poverty is one of the sustainable development goals (SDG) of the United Nations. More than 700 million people (ten percent of the world's population) live in extreme poverty by now. The COVID-19 crisis posed a huge challenge to global economic and social development, especially for developing countries, which could push half a billion people into poverty (Sumner et al. 2020). Previous studies (M. Liu et al. 2020; Tong et al. 2019; Zandi et al. 2019) showed the spatial distribution of poverty and its driver factors including environmental and socio-economic factors, etc. are uneven. Thus, the spatial visualization of poverty and the geographical difference in the relationship between poverty and multidimensional factors are helpful to understand the spatial distribution pattern of poverty, so as to help policy-makers to formulate precise poverty alleviation measures.

A sorting of independent (predictor) variables according to their degree of correlation with the dependent (response) variable reveals variable importance, relevant in many research fields, such as medical genetics, ecology, and the humanities. The results from variable importance analysis support research related to prediction, theory testing, and explanation (Tonidandel et al. 2011). Estimating the variable importance of poverty factors can reflect the relative importance of these factors to poverty, which is important for a better understanding of the nature of poverty. Some statistical models are frequently used to estimate variable importance in the social sciences (Nathans et al. 2012), such as multivariate linear regression (MLR) analysis and principal component regression (PCR). In scenarios with a limited

Responsible editor Philippe Garrigues.

✉ Jianguo Yan
jgyan@whu.edu.cn

✉ Fei Li
fli@whu.edu.cn

[1] Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan 430070, China

[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430070, China

set of independent variables, the selection of variables and setting their weights is easier but often complicated by nonlinear relationships between the independent variables and a dependent variable. Thus the conventional regression models that often perform well in dealing with linear regression are not suitable when nonlinear relationships exist between independent and dependent variables.

Many methods, such as machine learning (ML), have been developed in recent years to solve both linear and nonlinear problem in regression analysis. A set of ML methods can be used for regression analysis to generalize a nonlinear relationship between variables and is applied in data mining in regression (Vries et al. 2016) and classification (Kaczmarczyk et al. 2009; Liu et al. 2009) tasks. Given a set of independent and dependent variables, a trained model will generalize the interactions between the variables and predict the associated dependent variables using a new set of independent variables (Zhu et al. 2020). One of the ML is artificial neural networks (ANN) which are widely used to model complex relationships between inputs and outputs (Ardestani et al. 2014; Bataineh et al. 2016; Liu et al. 2009). However, an ANN often overfit the data, and the capability to generalize weakens when dealing with small data samples (Ardestani et al. 2014). Ensemble approaches usually deliver relatively accurate results because of their fast and robust responsiveness to noise in the data (Kontschieder et al. 2011). We focused on one of the most widely used ensemble machine-learning methods, the random forest (RF) method that combines a large number of decision trees. The RF proposed by Breiman is nonparametric and one of the most popular supervised machine-learning algorithms (Breiman 2001). A RF combines hundreds of weak decision trees into one strong forest. Each decision tree performs a regression or classification, and the algorithm selects the outcome with the most votes as the result of the RF model. The accuracy of the model is improved by the overall decision as complemented by weak decision trees. The RF method has been widely used in image classification (Akcay et al. 2018; G. Cai et al. 2019). It is also suitable for regression analysis to identify nonlinear relationships between the variables even in high-dimension settings with complex interactions; it ranks the variables and determines the impact of each variable on the result (Breiman 2015). The RF model is not easily susceptible to overfitting and has higher stability than an artificial neural network (ANN) (W.-c. Wang et al. 2015). The RF model is more tolerant to noise and outliers in data and has a higher fitting accuracy than a support vector machine (SVM) (Yaseen et al. 2019). In addition, the RF model has fewer adjustment parameters and is easier to operate relative to other methods such as particle swarm optimization (PSO) (D. Liu et al. 2020). The RF model has been widely used in variable importance studies by researchers from various fields (Li et al. 2020a, 2020b; Yi Li et al. 2018; D. Liu et al. 2020; Yu et al. 2017)and considered one of the most accurate model for regression and classification (Ardestani et al. 2014). Niu et al. (2020) used RF to construct the index of urban poverty in Guangzhou. Nevertheless, a RF model is a global model when used as a spatial statistics method, which assumes that the relationship between independent and dependent variables is globally stable. The RF does not account for the importance of variables across geographies and thus cannot reflect the imbalanced space distribution of the variables. In the real world, the distribution of multiple things is uneven, so the relationship between independent variables also varied over space. Therefore, it is necessary to consider the spatial variation when estimating the relationship between variables.

The exploratory and confirmatory nature of spatial data analysis aroused the attention of researchers as the availability of large spatial data sets and the capabilities to visualize, rapidly retrieve, and manipulate data in geographic information systems (GIS) expanded (Anselin 1988, 1990). Technologies focused on the spatial aspect of data developed rapidly (Anselin 2010, 2019; Anselin et al. 1992). Spatial heterogeneity and correlation often coexist, because the spatial distribution of natural resources and socioeconomic factors is imbalanced (Y. Wang et al. 2013). Georganos et al. (Georganos et al. 2019) used geographical random forests in remote sensing image classification, but did not focus on regression analysis. The study of the relationship between variables incorporating spatial factors more accurately reflects the distribution of things in the real world. Although the spatial error model (SEM) and spatial lag model (SLM) do consider spatial factors, they focus more on the analysis of spatial correlation and do not pay attention to analyzing spatial heterogeneity and the spatial variation of the relationships between variables (Wu 2020). Furthermore, spatial heterogeneity also encompasses unbalanced distributions of events, traits, and their relationship across a region (Anselin 2010; Dutilleul 2011). Therefore, it is impossible to explain a situation in different local areas using global parameters. Geographically weighted regression model (GWR) as proposed by Brunsdon et al. (Brunsdon et al. 2010) and further developed by Fotheringham et al. (Fotheringham et al. 2002) considers spatial heterogeneity and extends the ordinary least square (OLS) method by using a spatial weight to estimate local parameters. Because the GWR model can accurately generate a local spatial variation coefficient in regression (Ke et al. 2016), it has been widely used in ecological (Galli et al. 2012; Sheng et al. 2017; Wu 2020), atmospheric (Hu et al. 2014; Zhang et al. 2019), and water resource evaluations (Huang et al. 2015). As the GWR is based on multiple local linear regression models, it is not suitable in scenarios featuring nonlinear relationships between independent and dependent variables.

In this study, we used an approach that can measure the variable importance in each local area, called geographically weighted random forest regression (GW-RFR). The GW-RFR combines a spatial weight matrix (SWM) and random forest (RF), suitable for dealing with local high-dimensional variables, and can identify nonlinear relationships between variables. This method was used to estimate the spatial variation of the relationship between the geographical and socioeconomic factors and poverty in China at the county level. In general, poverty can be classified into absolute poverty and relative poverty. In our work, we focus on the per capita savings, one of the most important indicators of the poverty, which is also highly related to the absolute poverty. The absolute poverty indicates that people are unable to meet the basic physical or material needs, which was commonly used in the developing countries. We stress that the per capita savings, which was selected as the target object of this work, is not exactly equal to poverty, but it can represent one important aspect of absolute poverty in China.

This paper is organized as follows: In Sect. 2, a real-world poverty dataset, the methods including the SWM, RF, the proposed GW-RFR, the parameter settings, and the evaluation metrics are introduced. In Sect. 3, GW-RFR is demonstrated in a real-world poverty scenario. We employ data set from 2056 counties of China to validate our model. Finally, conclusions and future research are discussed in Sect. 4.

Data and method

In order to estimate the spatial variation of the nonlinear relationship between poverty factors and poverty, the local model GW-RFR was used in a poverty dataset and compared with the global model RF. In this section, we introduce the data (Sect. 2.1) and describe the proposed GW-RFR method. The first step of GW-RFR is to make a SWM for each local area using the spatial information of the data. The process for constructing the SWM is introduced in Sect. 2.2. The local RF is applied to each local area. The RF and the variable importance measurement (VIM) of the RF are discussed in Sect. 2.3. The process for constructing the proposed GW-RFR model is described in Sect. 2.4. The parameter settings are introduced in Sect. 2.5, and the evaluation metrics in Sect. 2.6.

Data

The distribution of poverty in China is extremely imbalanced (T. Li et al. 2020a, 2020b; Yansui Liu et al. 2016), and the factors affecting poverty including environmental factors and socioeconomic factors also have spatial characteristics. As poverty data for some counties were not available, we selected 2056 counties of China as the study area (see Fig. 1) in this study. The selected 2,056 counties as study areas accounted for 93% of the total area

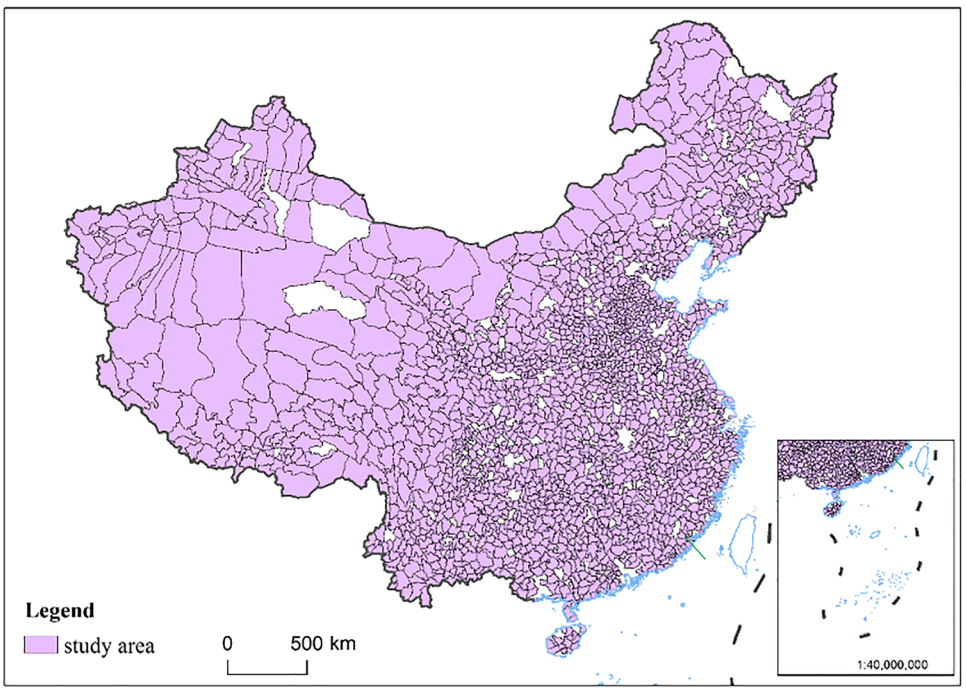

Fig. 1 Study area

of the Chinese mainland. Each county had a geographical location attribute. The location information for each county came from the National Geomatics Center of China.

In this study, we used the per capita savings which is one of the most related indicators with poverty as the target object and took the indicator of per capita savings (average value of resident savings) as the dependent variable (Y) for the regression model. The factors that lead to poverty can be broadly divided into two categories: geographical and socioeconomic (Barbier et al. 2018; Decancq et al. 2019; Zhou et al. 2020). We selected 28 poverty indicators (Table 1) according to the previous study as independent variables (X, ($X = X_1, X_2, \cdots, X_{28}$)). Elevation and slope image data at 30-m resolution were obtained from Google Earth Engine. Data of railway, highway, and river networks were obtained from the National Geomatics Center of China. X_6, X_7,$\cdots$, X_{28} are socioeconomic data and were extracted from the China County Statistical Yearbook (2016).

The units of these 28 poverty indicators are different; thus, the poverty indicators must be normalized before

Table 1 Definition of poverty indicators

Poverty indicators	Indicator meaning	Unit
Elevation X_1	The average elevation of each county based on a 30-m resolution elevation image	km
Relief X_2	The difference between the maximum and minimum elevation of each county	km
Slope X_3	The average slope of each county based on a 30 m resolution slope image	degree
Railway density X_4	The length of railways per square kilometer of land area	km/km^2
Highway density X_5	The length of highways per square kilometer of land area	km/km^2
Rivers density X_6	The length of rivers per square kilometer of land area	km/km^2
Proportion of secondary industry employees X_7	The proportion of secondary industry employees per 10,000 population	%
Proportion of tertiary industry employees X_8	The proportion of tertiary industry employees per 10,000 population	%
Per capita GDP X_9	Gross domestic product/total population	10^4 Yuan
Proportion of landline subscribers X_{10}	Proportion of landline subscribers per 10,000 population	%
Public revenue X_{11}	Public revenue of each county	10^4 Yuan
Public financial expenditure X_{12}	Public financial expenditure of each county	10^4 Yuan
Per loan amount X_{13}	Average value of loan amount of local residents	10^4 Yuan
Per capita total power of agricultural machinery X_{14}	Total power of agricultural machinery/total population	10^3 w
Per capita area of agricultural machine harvesting X_{15}	Total area of agricultural machine harvesting area/total population	km^2/individual
Per capita area of facility agriculture X_{16}	Total area of facility agriculture/total population	km^2/individual
Per grain production X_{17}	Grain production/total population	10^3 kg/individual
Per oil production X_{18}	Gil production/total population	10^3 kg/individual
Per meat production X_{19}	Meat production/total population	10^3 kg/individual
Number of units of large-scale industrial enterprises X_{20}	Number of units of large-scale industrial enterprises of each county	individual
Gross industrial output of large-scale industry X_{21}	Gross industrial output of large-scale industry of each county	10^4 Yuan
Fixed asset investment X_{22}	Fixed asset investment of each county	10^4 Yuan
Number of social welfare receiving units X_{23}	Number of social welfare receiving units of each county	individual
Proportion of students in regular secondary schools X_{24}	Number of students in regular secondary schools per 10,000 population	%
Proportion of students in regular vocational secondary schools X_{25}	Number of students in secondary vocational schools per 10,000 population	%
Proportion of primary school students X_{26}	Number of primary school students/total population per 10,000 population	%
Per capita hospital beds X_{27}	Number hospital beds per 10,000 population	individual
Per capita beds of various social welfare receiving units X_{28}	Number of beds of various social welfare receiving units per 10,000 population	individual

regression to eliminate the influence of variation in the units of measurement of the poverty indicators as follows:

$$X_{ki} = \frac{X_{ki} - \overline{X}_k}{\sigma_k} (i \in 1, 2, \cdots 2056; k \in 1, 2, \cdots, 28) \quad (1)$$

where X_{ki} represents the normalized value of the k th poverty indicator in the i th county, X_{ki} represents the original value of the k th poverty indicator in the i th county, $\overline{X}_k$ represents the average value of the k th poverty indicator, and σ_k represents the standard deviation of k th poverty index.

The dataset was divided into two parts, 70% of it was defined as training data set, and 30% of it was defined as validation data set. All the regression models including RF and GW-RF were implemented using the R software (version 3.5.3, http://cran.r-project.org), and the results were mapped in ArcGIS 10.5 (https://www.esri.com/zh-cn/home).

Spatial weight matrix (SWM)

Tobler's first law of geography notes that "everything is related to everything else, but near things are more related than distant things (Decancq et al. 2019; Tobler 1970)". Therefore, the closer things are in space, the smaller the difference in their attributes. In spatial analysis, spatial samples closed to the target sample at location i are generally considered to have a greater impact on the parameter estimates of the sample at location i than those far from it. Nearness refers to a central organizing principle of geographic space, but there is no standard definition for it (Miller 2004; Zhou et al. 2020). Here, we introduce two choices for the weight matrix of the county at location i, which are distance-based and adjacent with common edges. Setting the weight on county j to 1 if county j is a "neighbor" of county i, otherwise 0. Based on Q counties from a study area, the distance-based and common edge–based spatial weight matrix at location i is expressed as $W(L)_j(i)$ and $W(E)_j(i)$:

$$W(L)_j(i) = \begin{cases} 1, if d_{ij} \le L, j = 1, 2, \cdots, Q \\ 0, if d_{ij} > L. \end{cases} \quad (2)$$

where L is the distance threshold and d_{ij} is the distance between county i and county j. If the distance between county i and county j is less than L, the county j is considered to be a neighbor of the county i:

$$W(E)_j(i) = \begin{cases} 1, if county i and county j have commen edges, j = 1,2, \cdots, Q \\ 0, if county i and county j have no commonedge. \end{cases} \quad (3)$$

where determining whether j is a neighbor of i is based on whether there is a common edge between i and j.

Random forest for variable importance

Random forest (RF)

The RF proposed by Breiman (Breiman 2001; G. Cai et al. 2019) is a machine learning method ensembled with multiple decision trees for regression and classification. The RF is nonparametric and can easily learn nonlinear relationships between multiple variables without explicitly modeling them and works well when estimating the variable importance of each variable (Grömping 2009). The algorithm flow of the RF is as follows:

(1) The n sub-data sets $D_1, D_2, \cdots, D_n$ are randomly extracted from the whole data set D, and n decision trees $H_1, H_2, \cdots, H_n$ are generated according to n sub-data sets.
(2) Each decision tree has q variables, $m(m < q)$ variables were randomly selected for a node of the tree, and each node of the decision tree is split by the optimal segmentation criterion. Each decision tree can grow to its largest extent without pruning, until all the nodes cannot be split.

When constructing decision trees, about 36.8% of the data counties were not used. These counties are the out-of-bag (OOB) data for the decision tree. The accuracy of the RF model is estimated from the OOB data as in Eq. (4):

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \overline{\hat{y}}_i \right)^2 \quad (4)$$

where N is the number of counties of OBB data, y_i is the actual value of the dependent variable of the i th county, and $\overline{\hat{y}}_i$ is the average prediction for the dependent variable of the i th county from all trees in the RF.

The variable importance measurement (VIM)

Average impurity reduction (Gini importance) and mean squared error (MSE) reduction are two methods used to estimate the variable importance in a RF, but variable importance by impurity reduction is biased (Miller 2004; Strobl et al. 2007). The MSE reduction method is suggested when permuting the variables (Grömping 2009; Ishwaran 2007; Strobl et al. 2008, 2007). The MSE reduction method estimates the variable importance using the MSE value from the OOB data (H. Cai et al. 2018; Strobl et al. 2008). It is determined as follows:

(1) Calculate the MSE of the OBB data for each decision tree. The MSE of OOB data of the decision tree t is calculated by Eq. (5):

$$\mathrm{MSE}_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(y_i - \hat{y}_{i,t}\right)^2 \quad (5)$$

where N_t is the number of counties of the OBB data in the tree t and $\hat{y}_{i,t}$ is the prediction of the dependent variable of the i th county for the tree t.

(2) The target variable l is randomly replaced, and then the corresponding value of the MSE for the new tree t is calculated by Eq. (6):

$$\mathrm{MSE}_t(l) = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(y_i - \hat{y}_{i,t}(l)\right)^2 \quad (6)$$

where $\hat{y}_{i,t}(l)$ is the prediction of the dependent variable for the i th county of the new tree t with the target variable l randomly replaced.

(3) Calculate the MSE reduction between MSE_t and $\mathrm{MSE}_t(l)$. The variable importance for variable l of the decision tree t can be obtained from the MSE reduction. The variable importance of variable l of the RF is the average over MSE reduction of all n trees as expressed in Eq. (7):

$$\mathrm{VIM}(l) = \mathrm{MSE}(l) = \frac{1}{n}\sum_{t=1}^{n}\left(\mathrm{MSE}_t - \mathrm{MSE}_t(l)\right) \quad (7)$$

Geographically weighted random forest regression (GW-RFR)

In this section, we introduced the GW-RFR, a local nonlinear machine learning method. The GW-RF was proposed by Luo et al. (2021) and was applied in many studies especially in the spatial analysis about COVID-19 epidemic (Maiti et al. 2021; Quiñones et al. 2021). The GW-RFR integrates the SWM and RF into a local regression model. The GW-RFR is a variation of the RF, which is applicable to local systems. It can estimate the nonlinear relationship between the high-dimensional variables even for high correlated variables (Archer et al. 2008). The variable importance for each county can be obtained from the GW-RFR. The process of the GW-RFR model is as follows:

(1) The SWM for each county of the whole data set should be made according to a specified spatial weight rule such as a distance-based or common edge-based spatial weight rule. The SWM for the whole study area with p spatial counties can be expressed as in Eq. (8):

$$W = \begin{bmatrix} W(1) \\ W(2) \\ \vdots \\ W(i) \\ \vdots \\ W(p) \end{bmatrix} = \begin{bmatrix} w_{11}w_{12}\cdots w_{1p} \\ w_{21}w_{22}\cdots w_{2p} \\ \vdots\vdots\vdots\vdots \\ w_{i1}w_{i2}\cdots w_{ip} \\ \vdots\vdots\vdots\vdots \\ w_{p1}w_{p2}\cdots w_{pp} \end{bmatrix} \quad i \in (1,2,\cdots,p) \quad (8)$$

Because the local random forest for a county needs to consider the county itself, the value of w_{ii} is set to 1 ($w_{ii} = 1$). According to the pre-set spatial weight rule, for county i, if the county j ($j \in (1,2,\cdots,p) \wedge i \neq j$) is a "neighbor" of the county i, the value of spatial weight w_{ij} between them is set to 1. While county j is not a "neighbor" of county i, $w_{ij} = 0$.

(2) Select all the "neighbors" of each spatial county according to the SWM. For county i, the "neighbors" of it can be obtained from the SWM W where $w_{ij} \neq 0, (j \in (1,2,\cdots,p) \wedge i \neq j)$.

(3) The county i and its "neighbors" are as the inputs of a local RF for county i (RF (i)). Then the variable importance for spatial county i can be abtained from the RF (i).

(4) Repeat steps (2) and (3) to construct the local RF for each spatial county in the whole study area. The local variable importance for each county can be estimated from the local RF.

Parameter settings

The poverty data set were employed in the RF and GW-RFR. In the implementation of the GW-RFR, the SWM is the key to implement GW-RFR model. The SWM of each county was generated by a distance-based rule, k-nearest neighbors (KNN). The neighbors of the target county are defined as the K counties closest to the target county using KNN. We set K=125 according to the test of the performance of multiple GW-RFR models with a different number of the local samples (Table S4); that is, the value of L in Eq. (2) is defined as the distance between the target county and the 125th closest county.

In the implementation of the GW-RFR, the number of decision trees n tree and the number of candidate split variables of the tree node m try are the two main parameters that influence the performance of the model. Referring to the relevant research about the RF (G. Cai et al. 2019) and experimental tuning of the GW-RFR, the parameters for each local RF of the whole GW-RFR were set as follows: the number of decision trees n tree = 500 and the number of candidate split variables of the tree node m try = 10.

Evaluation metrics

We used three metrics to evaluate the performance of the GW-RFR. For the dependent variable y, we computed the coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2} \tag{9}$$

the normalized root mean square error (*NRMSE*),

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}}{y_{max} - y_{min}} \tag{10}$$

and the mean absolute error (*MAE*):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{11}$$

where $\hat{y}_i$ denotes the predicted dependent variable by a regression model, y_i is the actual value of the denpendent variable,$\bar{y}$ is the mean value of y_i, N is the total number of the counties, y_{max} is the maximum value of the actual dependent variable, and y_{min} is the minimum value of the actual dependent variable.

Results and analysis

China is the largest developing country with a large area and uneven economic development among different regions. The distribution of poverty and poverty factors in China varies in space. On a large spatial scale, China has 14 concentrated contiguous zones of extreme poverty. At the small and medium scale, the poverty and poverty factors at the county level in China are also uneven. To explore the spatial variation characteristics of county poverty, the GW-RFR and RF were used to explore the causes of poverty from the socio-economic and geographical perspectives at county level in China. A strong correlation between the independent variables may cause multicollinearity during the regression analysis. Multicollinearity occurs when there are several high linear relationships between regressors, leading to the statistically insignificant outcomes for the individual t test (Ishwaran 2007; Pesaran 2015). Thus, before performing regression analysis, the Pearson correlation coefficient was used to evaluate the correlation between independent variables and the correlation between X and Y. A Pearson correlation coefficient value greater than 0.8 indicates a significant correlation between variables. Table S1 shows the Pearson correlation coefficient values between the poverty indicators (independent variables). The value of the Pearson correlation coefficient between X_{11}, X_{12}, X_{20}, and X_{21} is greater than 0.8, indicating that there is a significantly strong correlation between them. Therefore, they cannot be put together as input independent variables into the regression analysis. Table S2 shows the Pearson correlation coefficient values between each poverty indicator and Y. The higher the value of the Pearson correlation coefficient between a poverty indicator and Y, the stronger its correlation with poverty. To avoid the multicollinearity, X_{12}, X_{20}, and X_{21} cannot be selected for the regression analysis according to the analysis result from Table S1 and Table S2. The independent variables including X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{22}, X_{23}, X_{24}, X_{25}, X_{26}, X_{27}, and X_{28} were the variables input to the RF and the GW-RFR.

The selected 25 poverty indicators were used as independent variables and per capita savings as dependent variables, and they were used to conduct regression analysis in GW-RFR and RF, respectively. We used three metrics introduced in Sect. 2.6 including R^2, NRMSE, and MAE to measure the performance of the GW-RFR and RF. Table 2 provides the evaluation of the RF and GW-RFR with the evaluation metrics R^2, NRMSE, and MAE.

As shown in Table 2, the value R^2 of the GW-RFR was 0.918 and 0.16 higher than the RF. Compared with global regression, the goodness of fit of local regression is improved obviously. The value of the NRMSE was 2.0%, and MAE was 0.312, they were both lower than the RF. This indicates that the performance of GW-RFR in regression analysis was improved compared with the RF. Table 2 shows the overall performance of the model, but the performance of the GW-RFR was not consistent in space. To show the local fitting performance of GW-RFR, we mapped the local R^2 of GW-RFR through Arcgis 10.7. Figure 2 shows the spatial distribution of local R^2 of the GW-RFR.

In Fig. 2, we visualized the value of local R^2 in five ranges (≤ 0.2, (0.2, 04], (0.4, 06], (0.6, 08], > 0.8), and we calculated the percentage of counties in these five ranges (Table 3). The values of local R^2 were high in the majority counties, especially the counties in the southwest, eastern coastal areas and the western areas. The percentage of counties where local R^2 was higher than 0.6 was 70.0% and higher than 0.4 was 94.65%. Only a few counties in the central and south central areas had lower values of

Table 2 The coefficient of determination (R^2), the normalized root mean square error (NRMSE), and the mean absolute error (MAE) of the RF and GW-RFR in the application example

	Value of R^2	Value of NRMSE	Value of MAE
RF	0.758	4.0%	0.612
GW-RFR	0.918	2.0%	0.312

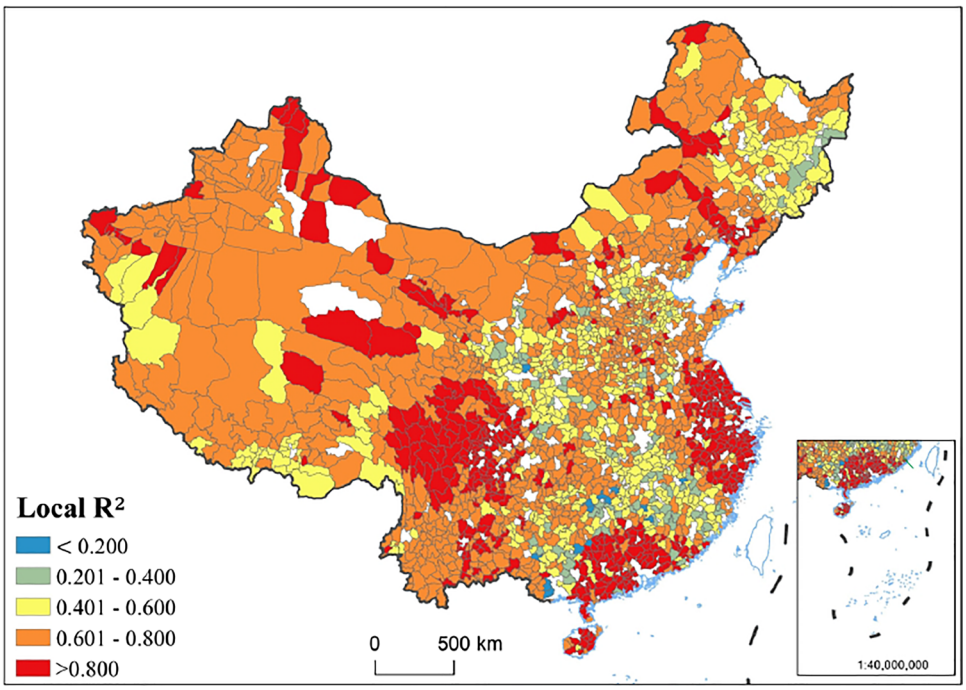

Fig. 2 The distribution of local R^2 of the GW-RFR in the application example

Table 3 The statistic of local R^2 of the GW-RFR, we calculated the average value of local R^2 and the percentage of counties in five local R^2 range (≤ 0.2, (0.2, 04], (0.4, 06], (0.6, 08], > 0.8)

The value of local R^2	Percentage of counties
≤ 0.2	0.73%
(0.2, 04]	4.62%
(0.4, 06]	24.66%
(0.6, 08]	46.06%
> 0.8	23.93%

R^2. This indicates the proposed local model GW-RFR obtained an accurate result in most of the local regions at a finer scale.

To estimate the correlation between each poverty indicator and poverty, we estimated the variable importance and did a significance test for each poverty indicator in both RF and GW-RFR model. We ranked the poverty indicators based on the variable importance marked them as vip i ($i \in (1, 2, \cdots, 25)$). Table 4 shows poverty indicators of the RF and the variable importance and the P value.

As shown in Table 4, the per loan amount (X_{13}), per capita GDP (X_{9}), elevation (X_1), proportion of landline subscribers (X_{10}), proportion of secondary industry employees (X_7), and per capita beds of various social welfare receiving units (X_{28}) had a significant correlation with poverty with a P value lower than 0.05. According to the RF regression analysis, in the whole study area, the poverty indicator with the strongest correlation with poverty (vip1) was per loan amount (X_{13}), followed by per capita

Table 4 Poverty indicators based on vip of the RF in the application example

Variable importance order	Poverty indicator	Variable importance	P value
vip1	X_{13}	47.972	0.020
vip2	X_9	21.672	0.020
vip3	X_1	19.872	0.020
vip4	X_{10}	17.778	0.020
vip5	X_7	13.096	0.020
vip6	X_{28}	12.211	0.020
vip7	X_{11}	11.342	0.216
vip8	X_{22}	9.688	0.353
vip9	X_{26}	8.350	0.098
vip10	X_{15}	7.710	0.431
vip11	X_5	7.704	0.157
vip12	X_3	7.186	0.784
vip13	X_2	7.017	0.765
vip14	X_{25}	6.645	0.431
vip15	X_{27}	6.324	0.235
vip16	X_{23}	6.151	0.451
vip17	X_{17}	5.975	0.804
vip18	X_8	5.600	0.510
vip19	X_{19}	5.347	0.294
vip20	X_{18}	4.407	0.373
vip21	X_{24}	3.465	0.725
vip22	X_{14}	2.978	0.922
vip23	X_4	2.783	0.922
vip24	X_{16}	2.525	0.745
vip25	X_6	2.238	0.863

GDP (X_9) and elevation (X_1). The variable importance and the *P* value of each poverty indicators from the GW-RFR were shown in Table S3 in the online supplementary file. The relationship between each poverty indicator and poverty was different from the overall relationship obtained by RF. The variable importance of each poverty indicator varies from region to region, which was consistent with the spatial heterogeneity characteristics of poverty factors found in previous studies. The three poverty indicators with the highest local average variable importance are per loan amount (X_{13}), per capita GDP (X_9), and proportion of landline subscribers (X_{10}). The poverty indicators that are most correlated with each county were also different. Figure 3 provides a detailed spatial distribution of the poverty indicators with the most correlated poverty indicator (the poverty indicator with the highest value of variable importance (vip1)) in each county.

In Fig. 3, among the poverty indicators of vip1 in the GW-RF, per loan amount (X_{13}) accounts for the largest proportion, followed by per capita GDP (X_9), proportion of landline subscribers (X_{10}), and per capita hospital beds (X_{27}). The poverty indicator that had the strongest correlation with poverty was not the same in different counties. The per loan amount (X_{13}) was the primary poverty indicator in most of the eastern areas, the western part such as Xinjiang Province, the west of Xizang Province, Qinghai Province, and Gansu Province, and in the norther part such as Inner Mongolia Province, and the southeastern part such as Yunnan Province. The per capita GDP (X_9) was the primary poverty indicator in the center of Xizang Province, the north of Sichuan Province, Chongqing Province, and the west of Guangxi Province. The proportion of landline subscribers X_{10} has the strongest correlation with the poverty in the southeastern of Tibet, the south of Chongqing Province, and the south of Jilin Province. The per capita hospital beds X_{27} was the primary poverty indicator in the northeastern part such as Heilongjiang Province. To explore the distribution of the variable importance of each poverty indicator, Fig. 4 displays a detailed spatial distribution of the value of variable importance for the first three relatively important poverty indicators per loan amount (X_{13}), per capita GDP (X_9), and proportion of landline subscribers (X_{10}).

As shown in Fig. 4, each poverty indicator had a different level of correlation with poverty in different regions. The geographical distribution of the per loan amount, per capita GDP, and proportion of landline subscribers is concentrated and extensive. In the western region, these three poverty indicators show strong correlation with poverty. But the distribution of these three factors of poverty was different in other regions. This may be explained by the various geographical environment and socioeconomic development patterns of local regions, leading to the different impact of each poverty factor on poverty in local regions. The poverty indicator of per loan amount (X_{13}) had a greater correlation with poverty in the northwestern regions and eastern areas, but it has a smaller impact in the northeastern regions. The variable importance of per capita GDP and proportion of landline subscribers shows a distribution pattern of "high in the west and low in the east" on the whole, which is similar to the distribution of poor counties in China (H. Cai et al.

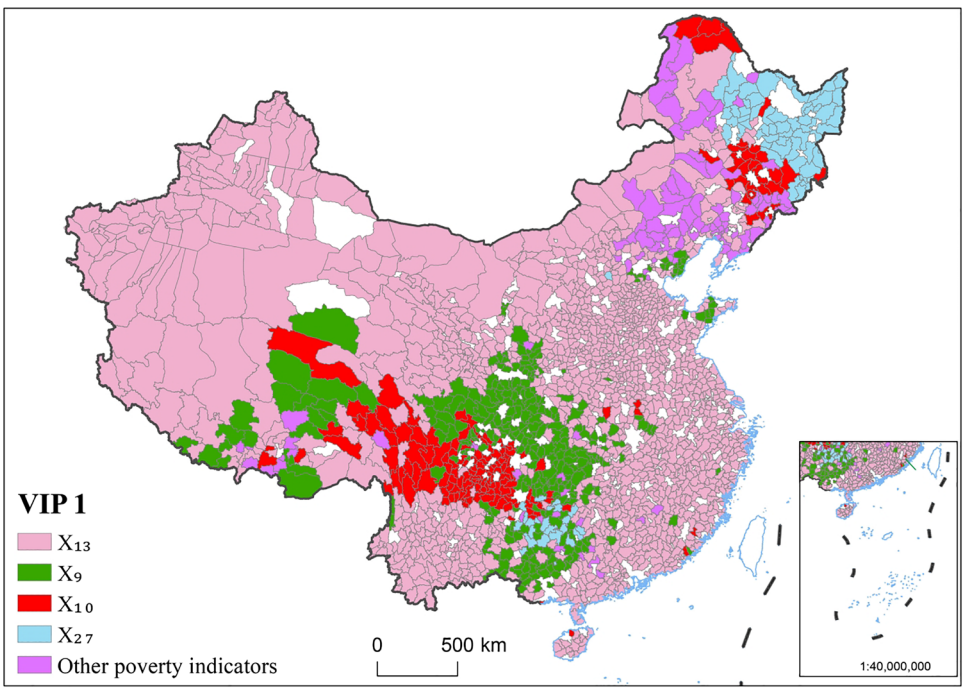

Fig. 3 The distribution of the poverty indicators with the highest value of variable importance (vip 1) in each county

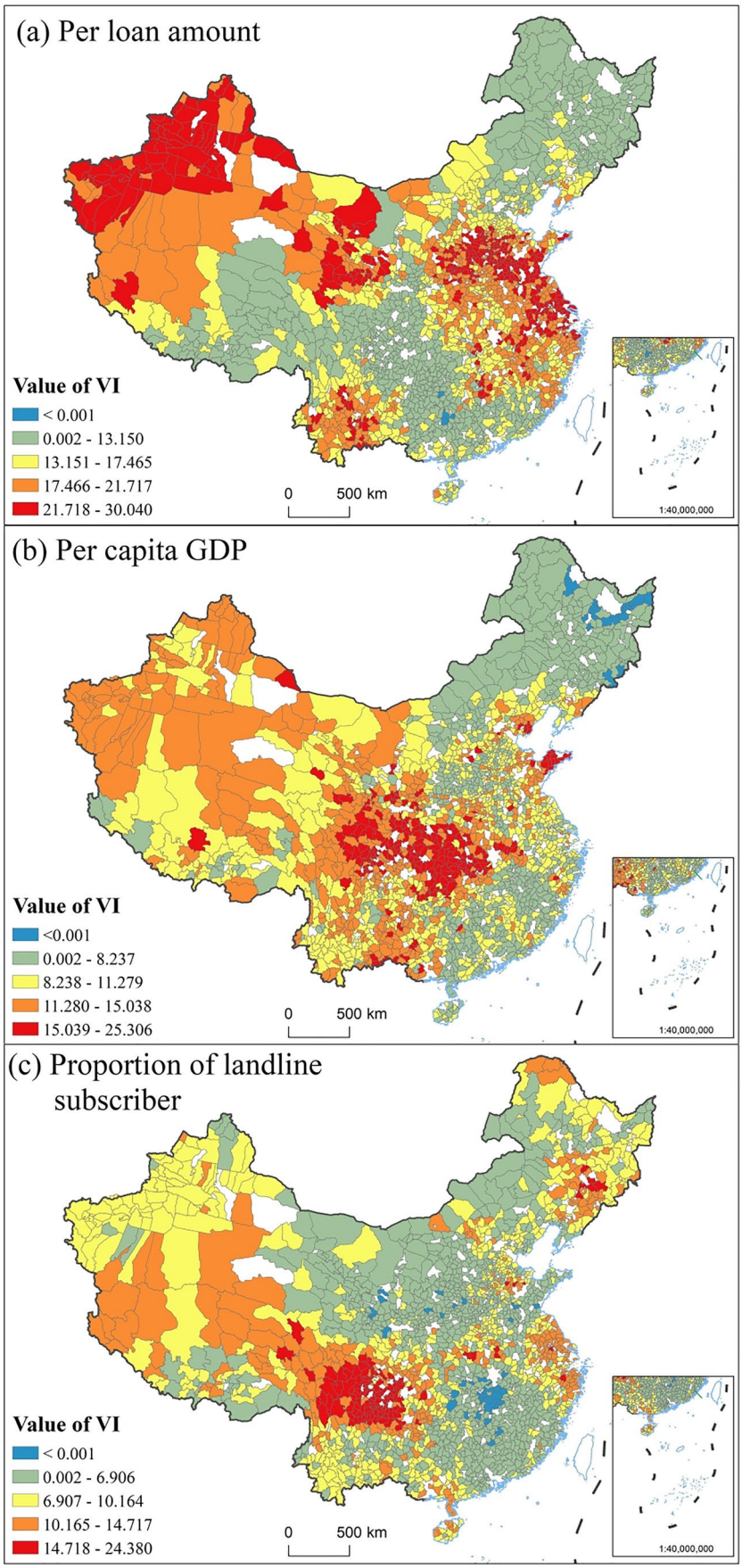

Fig. 4 The distribution of the value of variable importance (VI) for poverty indicator per loan amount (a), per capita GDP (b), and proportion of landline subscribers (c) in the application example

2018; Yanhua Liu et al. 2016). The per capita GDP (X_9) had a greater impact on poverty in the western and central regions especially in Sichuan Province, Chongqing Province, and Hubei Province. The proportion of landline subscribers (X_{10}) was an influential poverty in Jiangsu Province, Sichuan Province, the west of Xinjiang Province, and the west of Tibet. Policy-makers should pay attention to the variation of these poverty indicators over space to tailor measures for different regions.

Discussion and Conclusion

The distribution of poverty in space is not balanced. Thus, the variables (poverty and multidimensional factors) and their relationships varied across geographical locations. The relationship between poverty and its factors is not always linear in real-world data sets. The RF is a machine learning model, which can explain the nonlinear relationship between variables, but its results are consistent in the global research area and cannot explain the geospatial differences of variables in the local area. In order to explore the nonlinear relationships between variables at various spatial locations, it is necessary to deal with nonlinearity in local regression models. Thus, we used a local regression model GW-RFR to handle nonlinear relationships between poverty and multiple factors across various locations.

In this paper, the method GW-RFR and RF (G. Cai et al. 2019) were employed to analyze a real-world poverty data set of China. In the poverty application example, the value of R^2 was 0.128 higher than that of the RF, the NRMSE value was 1.6% lower than the RF, and the MAE value was 0.295 lower than the RF. It indicates that the proposed local model GW-RFR can conduct a more accurate regression analysis of the poverty dataset compared with the global model RF. Our results showed that per loan amount, per capita GDP, and proportion of landline subscribers and per capita hospital beds are the main poverty factors in most areas of China. And the relative importance of these factors to poverty varied over space. The result was consistent with the other research findings on poverty in China (Yuheng Li et al. 2016; Yanhua Liu et al. 2016; Pesaran 2015; Tian et al. 2018). The value of local R^2 of the GW-RF was imbalanced in the study area, high in the majority of the counties, but low in a few of them. Previous studies found that the RF performed effectively in a global regression when dealing with nonlinear systems. Our results show, however, that the GW-RFR method inherits the merits of RF and can analyze the nonlinear regression at different locations in space.

Although the proposed GW-RFR can effectively deal with the nonlinearity at various locations in a regression analysis, it also has limitations. The R^2, NRMSE, and MAE metrics indicate that the local GW-RFR model outperformed a global model RF, but the local performance was imbalanced across the whole study area. While the GW-RFR performed effectively in a majority local areas, a few local areas were outliers. In the future work, we will improve the performance of the GW-RFR by increasing the number of local samples. In order to highlight the local features of each sample in the improved GW-RFR with increased local sample size, we will assign different weights to the neighbors of the target local sample according to their distance from the target, for example, setting the weights using an inverse distance weighted rule.

In this paper, a local nonlinear regression method GW-RFR is used, which consists of several local RFs. Through the space weight matrix, the GW-RFR will find adjacent space units for each space unit, thus constructing a local RF for each space unit. The GW-RFR improves the goodness of fit of RF by local analysis. This method also provides the correlation between independent variables and the dependent variable for each local spatial unit, as well as the prediction of local dependent variables. We used the GW-RFR to estimate the spatial variation of the relationships between poverty and socioeconomic factors. The result showed that the relationship between each factor and poverty presented a unique spatial pattern. In addition to being applied to the analysis of spatial poverty, this improved GW-RFR could also help users select the most important factors and predict processes affected by complex multiple factors at a finer scale.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-021-17513-3.

Acknowledgements This research is supported by a grant provided by the National Scientific Foundation of China (Grant No. U41874010 and 42030110), Pre-research Project on CivilAerospace Technologies (No. D020103), and the fundamental research funds for the central universities (2042018KF0231).

Author contribution Conceptualization, Yaowen Luo and Jianguo Yan; methodology, Yaowen Luo; software, Yaowen Luo; validation, Jianguo Yan and Yaowen Luo; formal analysis, Yaowen Luo; investigation, Yaowen Luo; resources, Yaowen Luo; data curation, Yaowen Luo; writing—original draft preparation, Yaowen Luo; writing—review and editing, Jianguo Yan and Stephen C. McClure; visualization, Yaowen Luo; supervision, Jianguo Yan and Fei Li; project administration, Jianguo Yan and Fei Li; funding acquisition, Jianguo Yan. All authors have read and agreed to the published version of the manuscript.

Funding National Scientific Foundation of China (Grant No. 41874010 and 42030110), Pre-research Project on CivilAerospace Technologies (No. D020103), and the fundamental research funds for the central universities (2042018KF0231).

Data availability The data we used in this study was obtained from Google Earth Engine, the National Geomatics Center of China, and the China County Statistical Yearbook (2016).

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All the co-authors consent to the publication of this work.

Competing interests The authors declare no competing interests.

References

Akcay, O., Avsar, E. O., et al. (2018). Assessment of segmentation parameters for object-based land cover classification using color-infrared imagery. *isprs international journal of geo information, 7*(11). doi:https://doi.org/10.3390/IJGI7110424

Anselin, L. (1988). Spatial econometrics: methods and models. *journal of the american statistical association, 85*(411), 905–907. doi:https://doi.org/10.1007/978-94-015-7799-1

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *journal of regional science, 30*(2), 185–207. doi:https://doi.org/10.1111/J.1467-9787.1990.TB00092.X

Anselin, L. (2010). Local indicators of spatial association—LISA. *geographical analysis, 27*(2), 93–115. doi:https://doi.org/10.1111/J.1538-4632.1995.TB00338.X

Anselin, L. (2019). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In (pp. 111–126).

Anselin, L., & Getis, A. (1992). Spatial statistical analysis and geographic information systems. *annals of regional science, 26*(1), 19–33. doi:https://doi.org/10.1007/978-3-642-01976-0_3

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *computational statistics & data analysis, 52*(4), 2249–2260. doi:https://doi.org/10.1016/J.CSDA.2007.08.015

Ardestani, M. M., Zhang, X., et al. (2014). Human lower extremity joint moment prediction: a wavelet neural network approach. *expert systems with applications, 41*(9), 4422–4433. doi:https://doi.org/10.1016/J.ESWA.2013.11.003

Barbier EB, Hochard JP (2018) Poverty, rural population distribution and climate change. Environ Dev Econ 23(3):234–256. https://doi.org/10.1017/s1355770x17000353

Bataineh, M., Marler, T., et al. (2016). Neural network for dynamic human motion prediction. *expert systems with applications, 48*, 26–34. doi:https://doi.org/10.1016/J.ESWA.2015.11.020

Breiman, L. (2001). Random Forests. In (Vol. 45, pp. 5–32).

Breiman L (2015) Random forest: Breiman and Cutler's random forests for classification and regression. R Package Version 4:6–12

Brunsdon, C., Fotheringham, A. S., et al. (2010). Geographically weighted regression : a method for exploring spatial nonstationarity. *geographical analysis, 28*(4), 281–298. doi:https://doi.org/10.1111/J.1538-4632.1996.TB00936.X

Cai, G., Ren, H., et al. (2019). Detailed urban land use land cover classification at the metropolitan scale using a three-layer classification scheme. *sensors, 19*(14). doi:https://doi.org/10.3390/S19143120

Cai, H., Lam, N. S. N., et al. (2018). A synthesis of disaster resilience measurement methods and indices. *international journal of disaster risk reduction, 31*, 844–855. doi:https://doi.org/10.1016/J.IJDRR.2018.07.015

Decancq K, Fleurbaey M et al (2019) Multidimensional poverty measurement with individual preferences. Journal of Economic Inequality 17(1):29–49. https://doi.org/10.1007/s10888-019-09407-9

Dutilleul, P. R. L. (2011). *Spatio-temporal heterogeneity: concepts and analyses*.

Fotheringham, A. S., Brunsdon, C., et al. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*.

Galli, A., Kitzes, J., et al. (2012). Assessing the global environmental consequences of economic growth through the ecological footprint: a focus on China and India. *ecological indicators, 17*, 99–107. doi:https://doi.org/10.1016/J.ECOLIND.2011.04.022

Georganos, S., Grippa, T., et al. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *geocarto international*, 1–16. doi:https://doi.org/10.1080/10106049.2019.1595177

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *the american statistician, 63*(4), 308–319. doi:https://doi.org/10.1198/TAST.2009.08199

Hu, X., Waller, L. A., et al. (2014). Estimating ground-level PM(sub 2.5) concentrations in the southeastern united states using MAIAC AOD retrievals and a Two-stage model. *remote sensing of environment, 140*, 220–232. doi:https://doi.org/10.1016/J.RSE.2013.08.032

Huang, J., Huang, Y., et al. (2015). Geographically weighted regression to measure spatial variations in correlations between water pollution versus land use in a coastal watershed. *ocean & coastal management, 103*, 14–24. doi:https://doi.org/10.1016/J.OCECOAMAN.2014.10.007

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *electronic journal of statistics, 1*, 519–537. doi:https://doi.org/10.1214/07-EJS039

Kaczmarczyk, K., Wit, A., et al. (2009). Gait classification in post-stroke patients using artificial neural networks. *gait & posture, 30*(2), 207–210. doi:https://doi.org/10.1016/J.GAITPOST.2009.04.010

Ke, S., & Zhongmin, X. U. (2016). The impacts of human driving factors on grey water footprint in China using a GWR model. *geographical research, 35*(1), 37–48. doi:https://doi.org/10.11821/DLYJ201601004

Kontschieder, P., Bulo, S. R., et al. (2011). *Structured class-labels in random forests for semantic image labelling*. Paper presented at the International Conference on Computer Vision.

Li, L., Chen, S., et al. (2020). Prediction of plant transpiration from environmental parameters and relative leaf area index using the random forest regression algorithm. *journal of cleaner production, 261*. doi:https://doi.org/10.1016/J.JCLEPRO.2020.121136

Li, T., Cao, X., et al. (2020). Exploring the spatial determinants of rural poverty in the interprovincial border areas of the loess plateau in China: a village-level analysis using geographically weighted regression. *isprs international journal of geo information, 9*(6). doi:https://doi.org/10.3390/IJGI9060345

Li Y, Su B et al (2016) Realizing targeted poverty alleviation in China: people's voices, implementation challenges and policy implications. China Agricultural Economic Review 8(3):443–454. https://doi.org/10.1108/CAER-11-2015-0157

Li, Y., Zou, C., et al. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. *applied energy, 232*, 197–210. doi:https://doi.org/10.1016/J.APENERGY.2018.09.182

Liu, D., Fan, Z., et al. (2020). Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm. *journal of cleaner production, 250*. doi:https://doi.org/10.1016/J.JCLEPRO.2019.119468

Liu M, Hu S et al (2020) Using multiple linear regression and random forests to identify spatial poverty determinants in rural China. Spatial Statistics 42:100461. https://doi.org/10.1016/j.spasta.2020.100461

Liu, Y., Shih, S.-M., et al. (2009). Lower extremity joint torque predicted by using artificial neural network during vertical jump. *journal of biomechanics, 42*(7), 906–911. doi:https://doi.org/10.1016/J.JBIOMECH.2009.01.033

Liu Y, Xu Y (2016) A geographic identification of multidimensional poverty in rural China under the framework of sustainable livelihoods analysis. Appl Geogr 73:62–76. https://doi.org/10.1016/J.APGEOG.2016.06.004

Liu Y, Zhou Y et al (2016) Regional differentiation characteristics of rural poverty and targeted poverty alleviation strategy in China. Bull Chin Acad Sci 31(3):269–278

Luo Y, Yan J et al (2021) Distribution of the environmental and socio-economic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. Environ Sci Pollut Res 28(6):6587–6599

Maiti, A., Zhang, Q., et al. (2021). Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States. *68*, 102784.

Miller, H. J. (2004). Tobler's first Law and spatial analysis. *annals of the association of american geographers, 94*(2), 284–289. doi:https://doi.org/10.1111/J.1467-8306.2004.09402005.X

Nathans, L. L., Oswald, F. L., et al. (2012). Interpreting multiple linear regression: a guidebook of variable importance. *practical assessment research and evaluation, 17*(9), 1–19.

Niu T, Chen Y et al (2020) Measuring urban poverty using multi-source data and a random forest algorithm: a case study in Guangzhou. Sustain Cities Soc 54:102014. https://doi.org/10.1016/j.scs.2020.102014

Pesaran, M. H. (2015). *Time series and panel data econometrics*.

Quiñones, S., Goyal, A., et al. (2021). Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. *11*(1), 1–13.

Sheng, J., Han, X., et al. (2017). Spatially varying patterns of afforestation/reforestation and socio-economic factors in China: a geographically weighted regression approach. *journal of cleaner production, 153*, 362–371. doi:https://doi.org/10.1016/J.JCLEPRO.2016.06.055

Strobl, C., Boulesteix, A.-L., et al. (2008). Conditional variable importance for random forests. *bmc bioinformatics, 9*(1), 307–307. doi:https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., et al. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *bmc bioinformatics, 8*(1), 25–25. doi:https://doi.org/10.1186/1471-2105-8-25

Sumner, A., Hoy, C., et al. (2020). *Estimates of the Impact of COVID-19 on Global Poverty*: United Nations University World Institute for Development Economics Research.

Tian, Y., Wang, Z., et al. (2018). A geographical analysis of the poverty causes in China's contiguous destitute areas. *sustainability, 10*(6). doi:https://doi.org/10.3390/SU10061895

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *economic geography, 46*, 234–240. doi:https://doi.org/10.2307/143141

Tong X, Kim JH (2019) Concentration or diffusion? Exploring the emerging spatial dynamics of poverty distribution in Southern California. Cities 85:15–24. https://doi.org/10.1016/j.cities.2018.11.022

Tonidandel, S., & Lebreton, J. M. (2011). Relative importance analysis: a useful supplement to regression analysis. *journal of business and psychology, 26*(1), 1–9. doi:https://doi.org/10.1007/S10869-010-9204-3

Vries, W. H. K. d., Veeger, H. E. J., et al. (2016). Can shoulder joint reaction forces be estimated by neural networks. *journal of biomechanics, 49*(1), 73–79. doi:https://doi.org/10.1016/J.JBIOMECH.2015.11.019

Wang, W.-c., Chau, K.-w., et al. (2015). Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. *environmental research, 139*, 46–54. doi:https://doi.org/10.1016/J.ENVRES.2015.02.002

Wang, Y., Kang, L., et al. (2013). Estimating The environmental Kuznets curve for ecological footprint at the global level: a spatial econometric approach. *ecological indicators, 34*(34), 15–21. doi:https://doi.org/10.1016/J.ECOLIND.2013.03.021

Wu, D. (2020). Spatially and temporally varying relationships between ecological footprint and influencing factors in China's provinces using geographically weighted regression (GWR). *journal of cleaner production, 261*. doi:https://doi.org/10.1016/J.JCLEPRO.2020.121089

Yaseen, Z. M., Sulaiman, S. O., et al. (2019). An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *journal of hydrology, 569*, 387–408. doi:https://doi.org/10.1016/J.JHYDROL.2018.11.069

Yu, F. W., Ho, W. T., et al. (2017). Critique of operating variables importance on chiller energy performance using random forest. *energy and buildings, 139*, 653–664. doi:https://doi.org/10.1016/J.ENBUILD.2017.01.063

zandi, R., Zanganeh, M., et al (2019) Zoning and spatial analysis of poverty in urban areas (Case Study: Sabzevar City-Iran). Journal of Urban Management 8(3):342–354. https://doi.org/10.1016/j.jum.2019.09.002

Zhang, W., Jiang, L., et al. (2019). Effects of urbanization on airport CO2 emissions: a geographically weighted approach using nighttime light data in China. *resources conservation and recycling, 150*. doi:https://doi.org/10.1016/J.RESCONREC.2019.104454

Zhou Y, Li YR et al (2020) The nexus between regional eco-environmental degradation and rural impoverishment in China. Habitat Int 96:15. https://doi.org/10.1016/j.habitatint.2019.102086

Zhu, Y., Xu, W., et al. (2020). Random forest enhancement using improved artificial fish swarm for the medial knee contact force prediction. *artificial intelligence in medicine, 103*. doi:https://doi.org/10.1016/J.ARTMED.2020.101811