Local attribute-similarity weighting regression algorithm for interpolating soil property values

Zhou Jiaogen[1*], Dong Daming[2], Li Yuyuan[1]

(1. *Key Laboratory of Agro-ecological Processes in Subtropical Region, Institute of Subtropical Agriculture, Chinese Academy of Sciences, Hunan 410125, China*; 2. *National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China*)

Abstract: Existing spatial interpolation methods estimate the property values of an unmeasured point with observations of its closest points based on spatial distance (SD). However, considering that properties of the neighbors spatially close to the unmeasured point may not be similar, the estimation of properties at the unmeasured one may not be accurate. The present study proposed a local attribute-similarity weighted regression (LASWR) algorithm, which characterized the similarity among spatial points based on non-spatial attributes (NSA) better than on SD. The real soil datasets were used in the validation. Mean absolute error (MAE) and root mean square error (RMSE) were used to compare the performance of LASWR with inverse distance weighting (IDW), ordinary kriging (OK) and geographically weighted regression (GWR). Cross-validation showed that LASWR generally resulted in more accurate predictions than IDW and OK and produced a finer-grained characterization of the spatial relationships between SOC and environmental variables relative to GWR. The present research results suggest that LASWR can play a vital role in improving prediction accuracy and characterizing the influence patterns of environmental variables on response variable.

Keywords: attribute similarity, geographically weighted regression, local regression, spatial interpolation

DOI: 10.25165/j.ijabe.20171005.2209

Citation: Zhou J G, Dong D M, Li Y Y. Local attribute-similarity weighting regression algorithm for interpolating soil property values. Int J Agric & Biol Eng, 2017; 10(5): 95–103.

1 Introduction

Spatial interpolation is an important field in spatial analysis and is used in many cases where interested property values at some points are required but not measured[1-3]. Conceptually, spatial interpolation is a process that estimates the interested property values at unmeasured points with observed values of measured points around the target points. The interested property values at the unmeasured points are generally estimated with values from k neighboring points closest to the target points among n ($n>k$) spatial points that have observed values for the interested property. This estimation is based on the assumption that the target points are spatially closer to the observed points, the more similar the property values at the target points are as the property values at the observed points, and the more accurate the estimation will be.

The proximity among spatial points is generally assessed using spatial distance (SD), commonly defined as Euclidean distance based on their geographical coordinates[4,5]. SD is widely used to search the k-nearest neighbors for unmeasured spatial point whose property values are to be estimated with spatial interpolation methods. Common used interpolation methods include OK[1], GWR[6], regression kriging (RK)[7], local RK[8,9] and geographically weighted regression kriging (GWRK)[10]. These methods are also

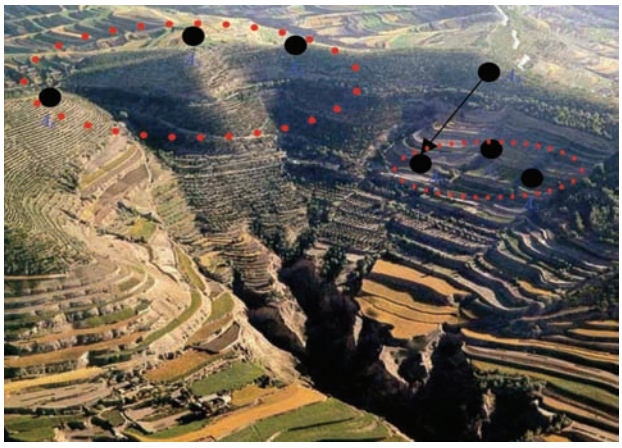
Received date: 2016-10-28 **Accepted date:** 2017-03-01

Biographies: Dong Daming, PhD, Associate Professor, research interests: biochemical sensor, Email: Dongdm@nercita.org.cn; **Li Yuyuan,** PhD, Professor, research interests: agriculture non-source pollution, Email: liyy@isa.ac.cn.

*** Corresponding author: Zhou Jiaogen,** PhD, Associate Professor, research interests: precise agriculture, spatial statistics, agriculture non-source pollution. Institute of Subtropical Agriculture, Chinese Academy of Sciences, Changsha 410125, Hunan, China. Tel: +86-731-84615224, Fax: +86-731-84619736; Email: zhoujg@isa.ac.cn.

based on the aforementioned assumption, which is applicable to estimation of property values with good spatial dependency. For properties with less spatially dependent, other factors may play a more important role than SD in determining the interested properties value of spatial points. In other words, the property values among spatial points may not be similar even though they are spatially close to each other[11-13].

Figure 1 presents a case when spatial interpolation methods based on spatial distance may not work well under a complex geographical environment. The three measured points (A_1-A_3) are located at the middle slope, and the other three measured points (A_4-A_6) and the unmeasured point (A_0) at the upper slope. Generally, the points of A_1 to A_3 are more spatially close, and the points of A_4 to A_6 are not spatially close but closer in terms of properties due to environmental factors like soil erosion. Thus, the estimation of soil property at A_0 with its three spatially closer neighbors of A_1-A_3 is less accurate and reliable than estimation made with A_4-A_6. This illustrates that spatial distance is not good choice to characterizing the similarity between a pair of points under complex geographical environments.



Note: A_0 represents the soil location to be estimated, and A_1-A_6 for six soil measured locations. The points of A_1-A_3 are more spatially close, and points of A_4-A_6 are not spatially close but closer in terms of properties.

Figure 1 Challenge for spatial interpolation methods based on spatial distance under a complex geographical environment

Inspired by the observation above, we proposed a local attribute-similarity weighted regression algorithm to predict the properties of unmeasured locations. LASWR searches the measured points closest to an unmeasured point based on non-spatial attribute similarity (NSA), and is expected to provide more accurate predictions than based on SD. The following sections will discuss the LASWR algorithm in detail.

2 Proposed algorithm

2.1 Description of LASWR algorithm

NSA was assumed to be better for characterizing the proximity of different points and provided more accurate predictions than SD. The GWR method was extended to LASWR method by using NSA to predict the values of response variables for unmeasured locations. The LASWR method modeled the correlation between the environmental variables and response variable, which helped to understand spatial relationships between environmental variables and the response variable. The LASWR method was different from GWR in terms of having a different neighborhood for each of unmeasured points. A LASWR neighborhood was composed of the spatial points whose environmental attributes were very similar to those of the point to be estimated, and who might be located at any possible places within the whole study area. The GWR neighborhood only contained those points which were spatially close to the points to be estimated, while their non-spatial properties might not be similar.

The basic steps to apply LASWR for spatial prediction included: (1) determine the environmental variables correlating well with the property of interest (response variable); (2) find its k nearest neighbors based on NSA for each of unmeasured spatial points; (3) design weighting function and compute regression coefficients with the weighted least-squares technique; and (4) estimate the values of the response variable for all unmeasured points. The steps (2) and (3) are the key steps in LASWR and are discussed below.

2.2 Search k nearest neighbors based on NSA

Given n number of measured spatial points and each of unmeasured points x_0, LASWR calculated all distances of x_0 to the n measured spatial points based on NSA and selected the k number of measured points closest to x_0 as its k-nearest neighbors. The resulted k nearest neighbors was used to estimate the property values at x_0.

LASWR used NSA to measure the similarity between any two spatial points in their non-spatial attribute space.

Given an unmeasured spatial point x_0 and any measured spatial point x_i, NSA was defined as the Euclidean distance of the point x_0 to x_i in non-spatial attribute space rather than in geographical space as follows:

$$d_{i0} = \sqrt{[P(x_0) - P(x_i)][P(x_0) - P(x_i)]^T} \quad (1)$$

where, $P(x_0)$ is the row vector of the predictors at x_0; $P(x_i)$ is the row vector of the predictors at x_i, and T for matrix transpose operation.

Noting a spatially measured point normally contains both spatial attributes (geographical coordinates) and non-spatial attributes (such as environmental factor and other properties determined by physical or chemical analysis methods). Some constraints were firstly on the definition of attributes in this research. For the measured spatial points, their geographical coordinates, environmental attributes and physical or chemical properties of interest are considered as the spatial attributes, predictors and response variables, respectively. The unmeasured spatial points are defined as spatial objects whose spatial attributes and predictors are known, and whose response variables are yet to be predicted. Considering that not all environmental variables are related well with the response variable, step-regression method with forward selection was used to reduce the variables not correlating well with the targeted response variable in the paper.

2.3 Estimate regression coefficients

In LASWR, the attribute value $\hat{z}(x_0)$ at an unmeasured spatial point x_0 is estimated as follows:

$$\hat{z}(x_0) = \beta_0(x_0) + \sum_{j=1}^{m} \beta_j(x_0) P_j(x_0) \quad (2)$$

where, $\beta_0(x_0)$, $\beta_j(x_0)$, $P_j(x_0)$ and m are the regression intercept coefficient, the regression coefficient, the value of the j^{th} predictor at the point x_0, and the number of the predictors at x_0, respectively.

Let $\beta(x_0) = [\beta_0(x_0), \beta_1(x_0), \beta_2(x_0), ..., \beta_m(x_0)]$ and $P(x_0) = [1, P_1(x_0), P_2(x_0), ..., P_m(x_0)]^T$, then Equation (2) can be rewritten as follows:

$$\hat{z}(x_0) = \beta(x_0) P_0(x_0) \quad (3)$$

In fact, the regression coefficient $\beta(x_0)$ is generally unknown, and the estimation $\hat{\beta}(x_0)$ of $\beta(x_0)$ depends on the k nearest neighbors of x_0 in non-spatial attribute space. Thus, LASWR first performed a k-nearest neighbor query to estimate $\beta(x_0)$. The regression coefficient $\beta(x_0)$was estimated by using a weighted least squares technique to minimize the weighted sum of residual squares (WS) as follows:

$$WS(\beta) = \sum_{i=1}^{k} W(x_0^i)[Z_i - \sum_{j=1}^{m} \beta_j(x_0) P_{ij}(x_0)]^2 \quad (4)$$

where, $P_{ij}(x_0)$ is the value of the j^{th} predictor of the i^{th} nearest neighbor of x_0; $W(x_0^i)$ is the weight of the i^{th} nearest neighbor on x_0, and Z_i is the value of the response variable of the i^{th} nearest neighbor of x_0.

Differentiating with regarding to β in Equation (4), the unique solution can be obtained in the matrix notation from Equation (5).

$$\hat{\beta}(x_0) = [P^T W(x_0) P]^{-1} P^T W(x_0) Z \quad (5)$$

where, P is a $k\times(m+1)$ predictor matrix of the k neighbors around the location x_0; $W(x_0)$ is a $k\times k$ diagonal matrix whose off-diagonal elements are zero and diagonal elements denote the weighting of the neighbors for the location x_0; Z is the column vector of response variables of the k neighbors around location x_0. As a result, the estimation $\hat{z}(x_0)$ at location x_0 is as follows:

$$z(x_0) = P_0[P^T W(x_0) P]^{-1} P^T W(x_0) Z \quad (6)$$

where, P_0 is the m+1 vector of m predictors at x_0.

2.4 Construct weighting function

In LASWR, the weights are in accordance with the closeness of the neighbors to the point x_0 in non-spatial attribute space. The closer the neighbors to x_0 are, the higher their weights are. Various weighting functions were discussed and used to calculate their weights[14]. A Gaussian kernel function was used in the present study to calculate the weights. The weight $W(x^i_0)$ of the i^{th} neighbor for the point x_0 is calculated in Equation (7).

$$W(x_0^i) = \exp[-0.5 \times (d_{io} / \alpha)^2] \quad (7)$$

where, d_{i0} is the Euclidean distance between the point x_0 and its i^{th} neighbor in non-spatial attribute space rather than in geographical space and calculated as Equation (1), and α is the kernel bandwidth for x_0.

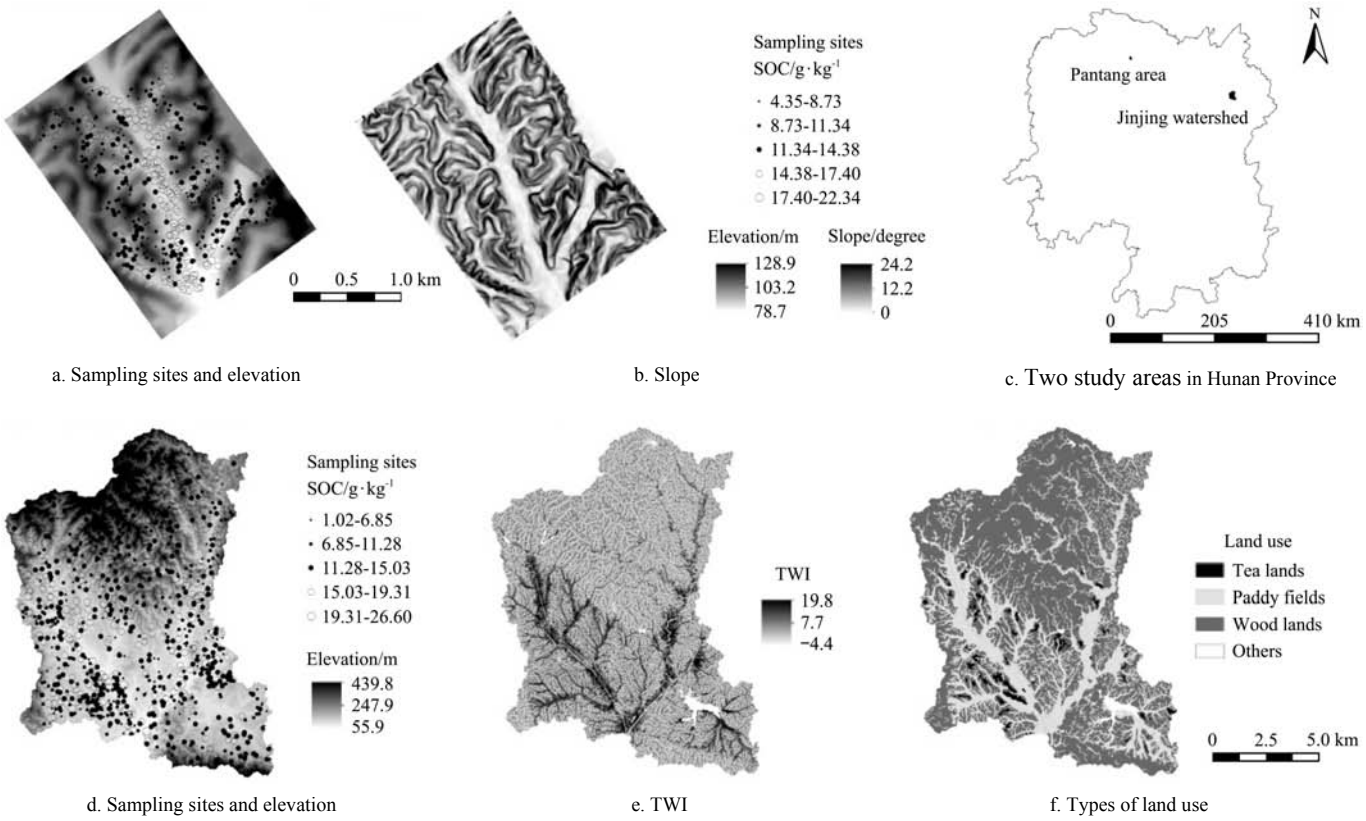
The bandwidth α can be simply considered as the radius of this influence region, and the parameter α controls the kernel bandwidth at location x_0[14]. The

distribution of neighbors may be different across the study region, and therefore the selected bandwidth should be flexible to satisfy all locations. To accomplish this, an adaptive strategy was used to automatically set the parameter. Specifically, the parameter used the median value of the distances of each unmeasured location to its neighbors.

3 Experiments and results

3.1 Data acquisition and pre-processing

In this research, the real soil data sets were collected from two study areas for evaluating the performance of LASWR. These two study area were Pantang area (PTA) in Taoyuan County with an area of 4.5 km^2 and the Jinjing watershed (JJW) in Changsha County with an area of 134.4 km^2. They are both located in the typically subtropical and hilly red soil region of Hunan province, China (Figure 2c). In the PTA study area, 523 surface (0-20 cm) soil samples were collected with averaged sampling density of 116.2 samples/km^2, and analyzed for SOC and TSN contents in 2003 (Figure 2a). In the JJW study area, 1033 topsoil samples were collected with averaged sampling density of 7.7 samples/km^2, and analyzed for SOC in 2010 (Figure 2d). The contents of SOC and TN were determined with a C/N elemental analyzer (Vario MAX, Elementary, Hanau, Germany). The most recent digital elevation model (DEM) with a spatial resolution of 5 m for the two study areas and a land use map with a spatial resolution of 5 m for the JJW study area were acquired from the land surveying department in Hunan province, China.



a. Sampling sites and elevation b. Slope c. Two study areas in Hunan Province

d. Sampling sites and elevation e. TWI f. Types of land use

Figure 2 Location of the sampling sites in the two study areas

The topographical factors of elevation, slope and topographical wetness index (TWI) were derived from the DEM. Elevation and slope were used as co-variables in the PTA study area, while elevation, TWI and land use were used as co-variables in the JJW study area. There were three main land use types (paddy field, wood land and tea land) in JJW, which had average SOC values of 14.51 g/kg, 13.22 g/kg and 9.43 g/kg, respectively.

To extract unmeasured points, grid maps of 5mspatial resolution were generated for PTA and JJW. The centroids of these grids were used as the locations for the unmeasured points. The co-variables at each measured and unmeasured point were obtained by overlaying the

point layer on the corresponding raster data of each co-variable. To alleviate the problem of skewness in the raw data, logarithmic transformation (log2) was applied to the data of SOC and TSN before applying the prediction models. The pre-processing of spatial data was carried out with ArcGIS9.0 software.

3.2 Model calibration and validation

To assess the performances of NSA and SD, the observed datasets containing SOC or TSN in the two study areas were randomly split into two parts: 75% as training data and 25% as validation data, respectively. The mean similarity index (MSI) was calculated to assess how close the SOC or TSN value at each location in validation dataset was to ones of their neighbors in training data in Equation (8):

$$MSI = \frac{1}{p \times k} \sum_{j=1}^{p} \sum_{i=1}^{k} \left| Z_{ij} - Z_j \right| \tag{8}$$

where, k is the given number of the nearest neighbors; p is the number of validation samples; Z_j is the observed value of the j^{th} validation sample, and Z_{ij} is the value of its i^{th} nearest neighbor of the j^{th} validation sample. Generally, the smaller the MSI value is, the more similar their soil properties (SOC or TSN) are.

The ten-fold cross-validation method was used to evaluate the prediction performance of the four models (IDW, OK, GWR and LASWR). The two real data sets were randomly split into ten equally sized sub-datasets, respectively. For each round, nine of the 10 sub-datasets were used as training data and the rest one sub-dataset was used as validation data to test the models. The process was repeated 10 times with each of the 10 subsamples used exactly once as the validation data. In order to improve the reliability of the validation for a given method, the parameters that obtain the minimum root mean square error (RMSE) value are considered optimal when using leave-one-out cross-validation[15]. Leave-one-out cross-validation was first performed to obtain the optimal parameters of the four methods for the observed data for both SOC and TSN. These optimal parameters were then applied in cross-validation.

The performance of IDW, OK, GWR and LASWR were evaluated by determining how close the predicted value $\hat{z}(x_0)$ is to the measured value at each location x_i. Two indices were used to evaluate the performance: mean absolute error (MAE) and RMSE, defined as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{z}(x_i) - z(x_i) \right| \tag{9}$$

$$RSME = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[\hat{z}(x_i) - z(x_i) \right]^2} \tag{10}$$

where, n, $\hat{z}(x_i)$ and $z(x_i)$ are the number of validation samples, the predicted and measured values of location x_i, respectively.

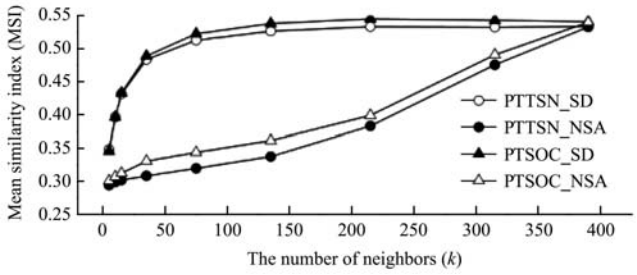
3.3 Programming and implementation

The four methods of IDW, OK, GWR and LASWR were programmed and implemented using Matlab9.0 in Windows XP. For OK, the software package of mGstat was used to obtain the parameters for the semivariogram. mGstat provides an interface for Gstat, which is commonly used for geostatistical modeling. The interface makes it straightforward to call Gstat using Matlab scripting language. The four datasets of the interpolated SOC across the JJW by IDW, OK, GWR and LASWR were firstly all saved in a excel format, and finally produced to raster maps with ArcGIS9.0 software.

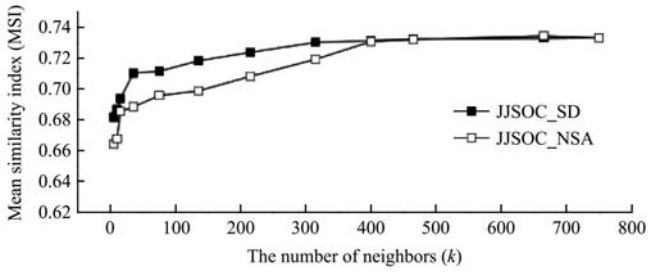
3.4 Performance comparison of NSA with SD

To compare the performance of NSA and SD, The MSI values of sampled locations was computed in the PTA and JJW study areas. The smaller the MSI value was, the larger the similarity between the observations for a soil sample and its k nearest neighbors was.

Given a value of k, the MSI value was computed using Equation (8). The results are shown in Figure 3, where the value of k varies from 5 to 390 for PTA and from 5 to 770 for JJW. In the PTA study area (Figure 3a), the MSI values based on NSA (MSI values of 0.28-0.47 g/kg) are lower than those based on SD (MSI values of 0.35-0.54 g/kg) for both SOC and TSN, for all k values less than 390 considered. In the JJW study area (Figure 3b), the MSI values based on NSA are less than those based on SD when $k<400$, but are close to those based on SD when $400<k<770$.



a. Pantang area (PTA)



b. Jinjing watershed (JJW)

Note: PTA study area, PTSOC_NSA and PTTSN_ NSA are the MSI values for SOC and TSN using NSA, and PTSOC_SD and PTTSN_SD are the MSI values for SOC and TSN using SD. For the JJW study area, JJSOC_NSA and JJSOC_SD are the values of MSI for SOC using NSA and SD, respectively.

Figure 3 MSI values of NSA and SD for validation datasets of SOC and TSN in the two study areas

For both PTA and JJW, the MSI values generally increased as the *k* value increased. This was a result of including more heterogeneous points in the neighborhood of each point with the *k* value increasing and the larger differences of soil properties between the set of *k* nearest neighbors. This also resulted in the little differences of MSI values between SD and NSA when there were a large number of soil points in the neighborhood of each location. This explained why the MSI values based on NSA were always close to those based on SD when *k* value close to 390 for the PTA study area and $400<k<770$ for the JJW study area. The above experimental results demonstrated that NSA characterized the proximity between soil points better than SD.

In this experiment, NSA consistently resulted in smaller MSI values compared to SD, and moreover the differences of MSI values between NSA and SD for SOC and TSN in the PTA study area are significantly more than those for SOC in the JJW study area. This may be controlled by sampling density. The averaged sampling density in PTA (116.2 samples/km^2) is higher than that in JJW (7.7 samples/km^2). In the unit space, the higher the sampling density, the higher the similarity of environment variables between spatial points. This is why the performance of NSA for SOC in PTA is better than that of NSA for SOC in JJW.

NSA has another advantage over SD, which can extend the scope of data utilization, because it is not limited by geographic distance. Specifically, spatial prediction using NSA can use soil samples that are geographically isolated from each other and at a substantial distance. Overall, NSA is not intended to replace SD, but instead may provide a good alternative for certain circumstances. A hybrid of NSA and SD may be more effective in predicting soil properties under complex geographical conditions.

3.5 Comparison among interpolate on methods

The performance of LASWR was compared to those of IDW, OK and GWR using MAE and RMSE to evaluate the prediction errors of the four methods through cross-validation. Smaller values of MAE and RMSE indicated better prediction. The results were shown in Table 1. In the cases for both the PTA and the JJW study area, the LASWR resulted in a smaller MAE (0.25-0.45) and RMSE (0.31-0.60) values compared to IDW (MAE values of 0.31-0.58 and RMSE values of 0.39-0.79), GWR (MAE values of 0.28-0.50 and RMSE values of 0.36-0.71) and OK (MAE values of 0.28-0.53 and RMSE values of 0.39-0.74).

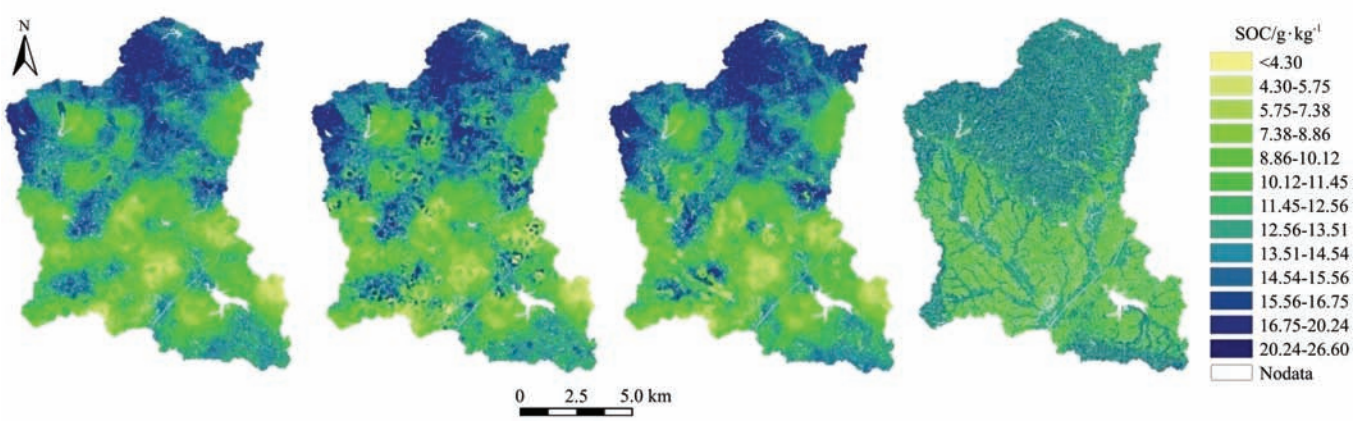
Table 1 MAE and RMSE for SOC and TSN of validation sets in the two study areas

Soil data	Metric	Prediction methods			
		IDW	OK	GWR	LASWR
PTA_SOC	MAE	0.31	0.29	0.28	0.25
	RMSE	0.39	0.39	0.39	0.31
PTA_TSN	MAE	0.29	0.28	0.26	0.23
	RMSE	0.40	0.40	0.36	0.30
JJW_SOC	MAE	0.58	0.53	0.50	0.45
	RMSE	0.79	0.74	0.71	0.60

Note: Logarithmic transformation (log2) was applied to the SOC and TSN data. PTA_SOC, PTA_TSN and JJW_SOC are for SOC and TSN in the PTA study, and SOC in the JJW study area, respectively.

The comparison of the maps predicted with IDW, OK, GWR and LASWR, was made on the SOC dataset from the JJW study area. The predicted SOC maps by IDW, OK, GWR and LASWR were illustrated in Figure 4. The maps of the regression coefficients between SOC and the co-variables by LASWR and GWR were presented in Figure 5. The spatial distributions of the SOC values

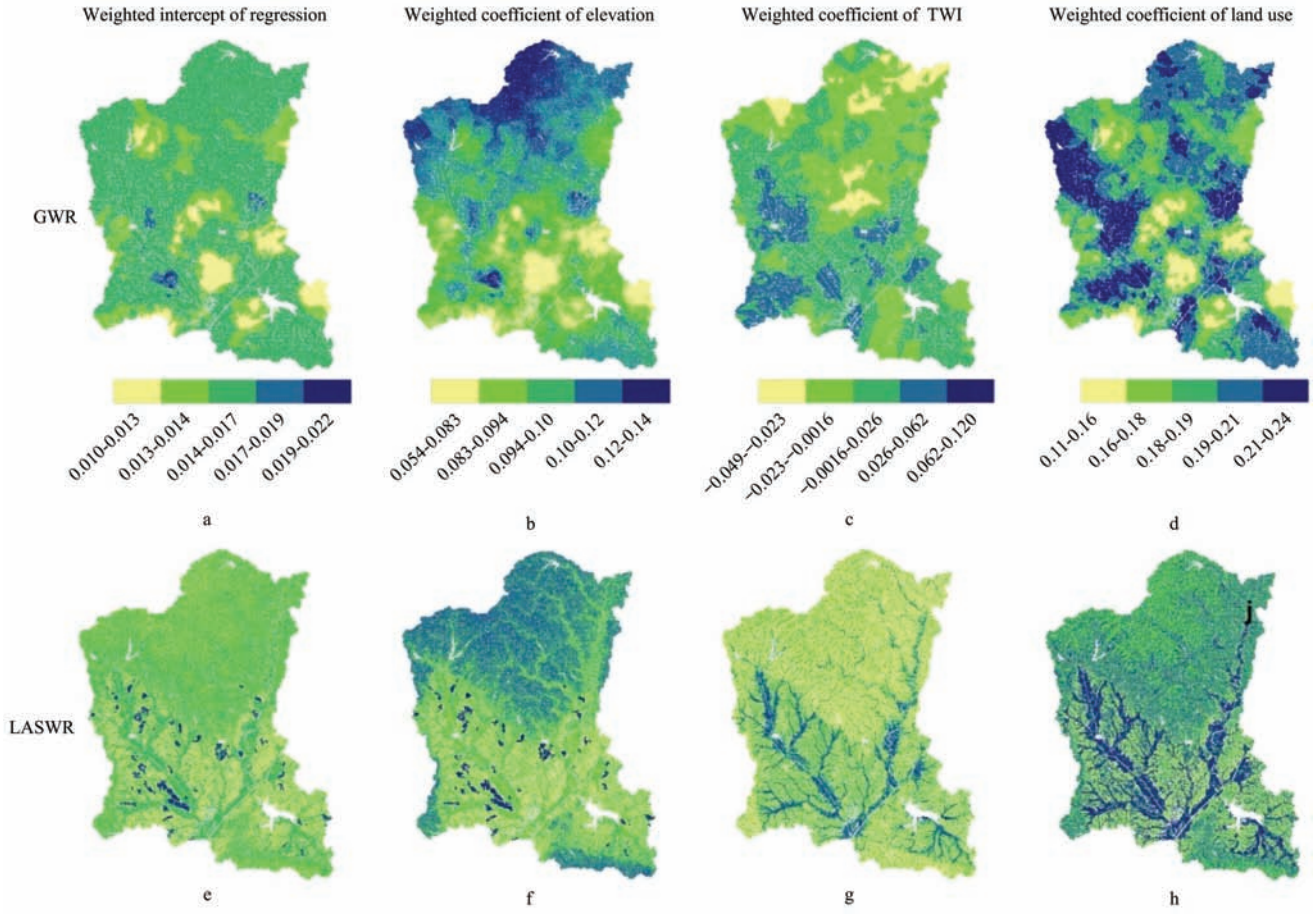
predicted by the four methods generally follow the expected distribution. The SOC values was relatively high in the northern regions with forested land and in some regions with paddy field, and relatively low SOC values in the middle regions with less forested lands. Compared with IDW, OK and GWR, LASWR (Figure 4d) produced a finer-grained SOC map, which showed the influence of terrain factors (i.e., elevation and topographical wetness index) and land use on prediction of SOC contents.



a. IDW b. OK c. GWR d. LASWR

Note: The 'Nodata' in all legends stands for a null value.

Figure 4 Maps of SOC predicted by (a) IDW, (b) OK, (c) GWR and (d) LASWR in the JJW study area



Note: The 'Nodata' in all legends stands for a null value.

Figure 5 Spatial distributions of regression coefficients between SOC and co-variables as predicted by (a-d) GWR and (e-h) LASWR for the JJW study area

LASWR and GWR could represent correlation between environmental variables and response variable, but IDW and OK could not. The regression coefficients predicted by LASWR and GWR spatially varied across the area of JJW, and reflected the influences of environmental variables on SOC (Figure 5). There were

the higher coefficients of elevation, TWI and land use appearing in the regions with higher elevation, and TWI, and lands with higher SOC contents. Compared to GWR (Figures 5a-5d), LASWR (Figures 5e-5h) produced clearer and finer-grained maps of regression coefficients. The regression coefficients by LASWR and GWR spatially varied across the area of PTA (Figure 6). The regression coefficients between SOC and elevation by LASWR (Figures 6d-6f) presented better spatial patterns of the influencing of elevation and slope on SOC than GWR (Figures 6a-6c) in the case of PTA. The above experimental results demonstrated that LASWR characterized spatial distributions of regression coefficients between SOC and co-variables better than GWR.

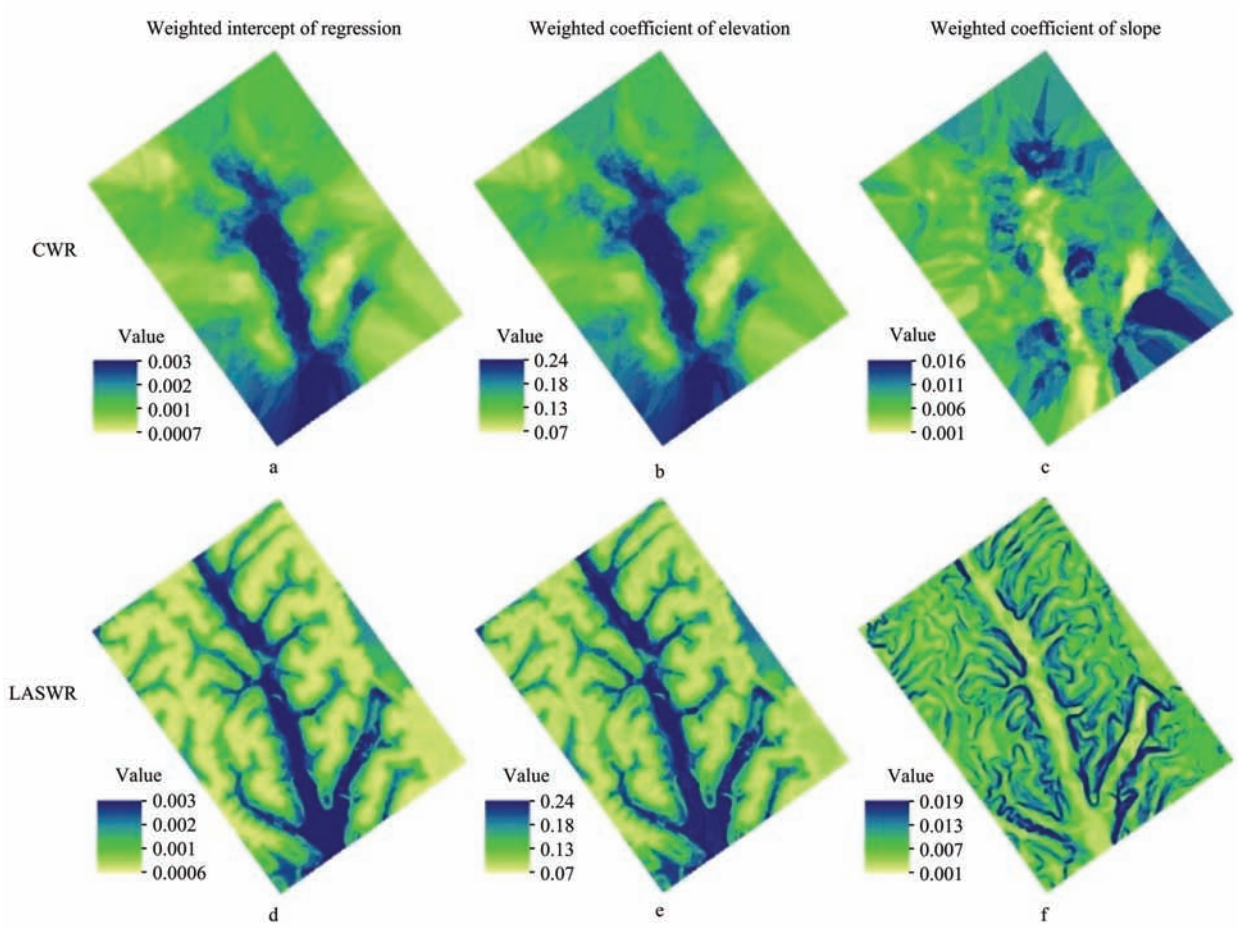


Figure 6 Spatial distributions of regression coefficients between SOC and co-variables as predicted by (a-c) GWR and (d-f) LASWR for the PTA study area

4 Conclusions

The proposed LASWR method based on NSA generally achieved better prediction performance, compared to the methods based on SD (IDW, GWR and OK). One major advantage of LASWR is that it can produce more fine-grained maps that depict the relationships between environmental co-variables and soil conditions. This can assist in better understanding how environmental factors influence soil conditions. Therefore, LASWR can play a vital role in improving the prediction accuracy and reflecting the influencing patterns of environmental variables on soil conditions. Considering that the results were obtained on the datasets of SOC and TSN at the landscape and watershed scales, datasets of different soil conditions at larger spatial scales may be necessary to further validate NSA and LASWR.

Acknowledgment

This work was supported by National Natural Science Foundation (41201299) and the Ministry of Water Resources Public Welfare Industry Scientific Research Special (201501055).

[References]

[1] Goovaerts P. Geostatistics in soil science: state of the art and perspectives. Geoderma, 1999; 89: 1–45.

[2] McBratney A B, Odeh I O A, Bishop T F A, Dunbara M S, Shatar T M. An overview of pedometric techniques for use in soil survey. Geoderma, 2000; 97(3-4): 293–327.

[3] Bostan P A, Heuvelink G B M. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. Int. J. Appl. Earth Obs. Geoinf, 2012; 19: 115–126.

[4] Ester M. Spatial data mining: databases primitives, algorithms and efficient DBMS support. Data Mining and Knowledge Discovery, 2000; 4: 193–216.

[5] Han J, Kamber M. Data mining: concepts and techniques, Academic Press, San Francisco, 2001.

[6] Fotheringham A S, Charlton M E. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environ. Plann. A, 1998; 30: 1905–1927.

[7] Hengl T, Heuvelink G B M, Rossiter D. About regression-kriging: From equation to case studies. Comput. Geosci, 2007; 33: 1301–1315.

[8] Sun W, Minasny B, McBratney A B. Analysis and prediction of soil properties using local regression-Kriging. Geoderma, 2012; 171-172: 16–23.

[9] Sun W, Whelan B M, Minasny B, McBratney A B. Evaluation of a local regression Kriging approach for mapping apparent electrical conductivity of soil (ECa) at high resolution. Journal of Plant Nutrition and Soil Science, 2012; 175(2): 212–220.

[10] Kumar S, Lal R, Liu D S. A geographically weighted regression Kriging approach for mapping soil organic carbon stock. Geoderma, 2012; 189-190: 627–634.

[11] Zhou J G, Guan J H, Bian F L, Li P X. DCAD: a dual clustering algorithm for distributed spatial databases. Geo-spatial Information Science, 2007; 10(2): 137–144.

[12] Lin C R, Liu K H. Dual clustering: integrating data clustering over optimization and constraint domains. IEEE Trans. Knowl. Data Eng, 2005: 17(5): 628–637.

[13] Jiao L M, Liu Y L, Zou B. Self-organizing dual clustering considering spatial analysis and hybrid distance measures. Sci. China Ser. D, 2011; 54(8): 1268–1278.

[14] Hastie T, Tibshiani R. The elements of statistical leaning: Data mining, inference and prediction, second ed. Springer, New York, 2009.

[15] Harris P, FotheringhamA S, CrespoP, Charlton M. The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. Mathematical Geosciences, 2010; 42(6): 657–668.