*The Journal of Agricultural Science*

**cambridge.org/ags**

# Spatial modelling of soil organic carbon stocks with combined principal component analysis and geographically weighted regression

Long Guo, Mei Luo, Chengsi Zhangyang, Chen Zeng, Shanqin Wang and Haitao Zhang

College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China

**Crops and Soils Research Paper**

**Cite this article:** Guo L, Luo M, Zhangyang C, Zeng C, Wang S, Zhang H (2018). Spatial modelling of soil organic carbon stocks with combined principal component analysis and geographically weighted regression. *The Journal of Agricultural Science* **156**, 774–784. https://doi.org/10.1017/S0021859618000709

Received: 7 June 2017
Revised: 21 July 2018
Accepted: 19 August 2018
First published online: 18 October 2018

**Key words:**
Data dimensionality; dimensionality reduction; geostatistical approaches; soil organic carbon; spatial variation

**Author for correspondence:**
Haitao Zhang, E-mail: hzau_zht@163.com

## Abstract

With the development of remote sensing and geostatistical technology, complex environmental variables are increasingly easily quantified and applied in modelling soil organic carbon (SOC). However, this emphasizes data redundancy and multicollinearity problems adding to the difficulty in selecting dominant influential auxiliary variables and uncertainty in estimating SOC stocks. The current paper considers the spatial characteristics of SOC density (SOCD) to construct prediction models of SOCD on the basis of reducing the data dimensionality and complexity using the principal component analysis (PCA) method. A total of 260 topsoil samples were collected from Chahe town, China. Eight environmental variables (elevation, aspect, slope, normalized difference vegetation index, normalized difference moisture index, nearest distance to construction area and road, and land use degree comprehensive index) were pre-analysed by PCA and then extracted as the main principal component variables to construct prediction models. Two geostatistical approaches (ordinary kriging and ordinary co-kriging) and two regression approaches (ordinary least squares and geographically weighted regression (GWR)) were used to estimate SOCD. Results showed that PCA played an important role in reducing the redundancy and multicollinearity of the auxiliary variables and GWR achieved the highest prediction accuracy in these four models. GWR considered not only the spatial characteristics of SOCD but also the related valuable information of the auxiliary attributes. In summary, PCA-GWR is a promising spatial method used here to predict SOC stocks.

## Introduction

The spatial distribution and storage of soil organic carbon (SOC) represent the basic knowledge needed to understand soil hydrological properties and the cycling of global carbon, and play important roles in precision agricultural management (Six *et al*., 2000; Lal, 2003). Natural factors (e.g. soil parent material, climate, topography and vegetation) and human activities (e.g. land management, land use and degradation) can influence soil moisture, ventilation conditions and soil temperature, which can then collectively influence the decomposition and transformation of SOC via soil microbes. These environmental factors and such transformation processes may result in spatial instability and non-uniformity of SOC in different landscapes (Song *et al*., 2016). Therefore, the current paper proposes a test methodology for accurate estimation of SOC stocks that extracts valuable information inherent in data on complex environmental factors.

The spatial analysis of soil distribution patterns is an important field in soil science. Adjacent soil patterns share similar natural environments and tillage methods. This condition results in the spatial dependence of SOC in nearby geographical locations. In addition, SOC has spatial heterogeneity because the environment varies with scale and geographical location (Mishra *et al*., 2010). Natural and human-induced factors, which influence SOC content, interact and affect each other resulting in complex inter-relationships and multicollinearity. Multicollinearity generates data redundancy and contradicts the standard hypothesis regression model that explanatory variables should be independent from one another (Liu *et al*., 2013; Conforti *et al*., 2015). Therefore, multicollinearity among impact factors must be eliminated before constructing a SOC model (Kumar *et al*., 2013; Sun *et al*., 2015). The conventional method is to extract the valuable variables by Pearson's correlation coefficient, stepwise linear regression or analysis of variance, and then construct predictive models (Kumar *et al*., 2012*a*; Guo *et al*., 2017). These methods ignore the valuable information from secondary environmental factors. Principal component analysis (PCA) is a widely used technique because it can decrease the basic dimensions of input parameters and reduce data redundancy (Song *et al*., 2016). However, conventional PCA does not consider spatial relations and is not designed specifically to identify spatial structures.

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

Several prediction methods, ranging from simple regression models to geostatistical models, have been suggested for mapping the SOC density (SOCD) from sparse soil samples to continuous surfaces. In terms of geostatistical approaches, recent studies have been devoted to the utilization of a variety of environmental variables for enhanced SOCD modelling (Kumar *et al.*, 2012*b*). Ordinary kriging (OK) is a method used commonly for estimating untested locations by calculating the weighted averages of explained variables from observed samples (Jobbágy and Jackson, 2000). Specifically, when the correlated environmental variables are valuable, they can provide great help in constructing the prediction models of SOCD with high prediction accuracy and limited observations (Keser *et al.*, 2012). The ordinary co-kriging (OCK) model can be used to interpolate the spatial distribution of soil attributes based on the related and appropriate regionalized variables, if the main soil attribute is sparse but the related auxiliary information is abundant (Wang *et al.*, 2013*b*). As the representative of traditional regression models, ordinary least squares (OLS) can minimize the sum of the squared vertical distances between the predicted data and the observed data through linear approximation (Evrendilek *et al.*, 2004), and the relationships between the soil properties and environmental variables can be used to predict the soil properties in unknown geographical locations by OLS. However, traditional regression methods only use the available data at target locations and ignore existing spatial autocorrelation of SOCD and its auxiliary variables (Liu *et al.*, 2015). Geographically weighted regression (GWR) considers the spatial weights between the independent and dependent factors relative to traditional regression models (Harris *et al.*, 2010). Also, the coefficients of the GWR model can respond to the spatial non-stationary characteristics of explanatory variables to the study object in geographical locations (Keser *et al.*, 2012; Wang *et al.*, 2013*a*).

Chahe Town, located in the middle of Jianghan Plain, China, was chosen as the study region. Jianghan Plain is an important agricultural region in China as it is a typical alluvial plain. In the current paper, PCA was performed to capture extensive explanatory variable information via orthogonal transformation. Two geostatistical approaches (OK and OCK) and two regression approaches (OLS and GWR) were used to estimate the spatial distribution of SOCD with the help of the principal components (PCs) which were processed via PCA. The aims of the current work are (1) to extract the valuable information from the complicated and various environmental factors by PCA method, (2) to construct a high-precision and efficient prediction models of SOCD, and (3) to draw the spatial distribution characteristics of SOCD by environmental factors.

## Materials and methods

### Study area

Chahe Town is located at the centre of Jianghan Plain, China (29.39–30.13° N, 113.6–114.05° E). The elevation ranges from 2 to 35 m asl, and the geographical area covers 153 km$^2$. Jianghan Plain is a typical alluvial plain and it is an important agricultural region in China that provides cotton, commodity grains and edible oil. The mean annual precipitation is 1154 mm and the average air temperature is 16.1 °C. A total of 260 topsoil samples were collected in June 2013 by random sampling. However, the minimum distance between two soil samples was >100 m. The potassium dichromate method was used to measure soil organic matter (SOM) content (Viscarra Rossel and McBratney, 1998). The soil types are paddy soil, moisture soil, dark yellow-brown soil and yellow-brown soil based on the Chinese soil taxonomy classifications (Shi *et al.*, 2006). The approximate classifications based on the World Reference Base of Soil Resources (Deckers *et al.*, 1998) are as follows: Typical Haplaquept, Dystrochrept, Eutroboralf and Hapludalf. The spatial distribution of the sampling sites is shown in Fig. 1.

### The auxiliary variable data

The spatial distribution characteristics of SOC is affected by multitudinous environmental factors in different geographical locations. The terrain factors (e.g. elevation, slope and aspect), distance factors (e.g. nearest distance to construction area (TRA) and road (TRD)), and remote vegetation indices such as the normalized difference vegetation index (NDVI) and normalized difference moisture index (NDMI) were chosen as the auxiliary variables (Wilson and Sader, 2002). These were calculated as follows:

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{RED})/(\mathrm{NIR} + \mathrm{RED})$$
$$\mathrm{NDMI} = (\mathrm{NIR} - \mathrm{MIR})/(\mathrm{NIR} + \mathrm{MIR})$$

where NIR denotes the near-infrared band, RED denotes the red band and MIR denotes the middle-infrared band of the Landsat 8 OLI imagery. The elevation, slope and aspect were calculated with the Global Digital Elevation Model Version 2 using ArcGIS 10.3 (ESRI Inc., Redlands, CA, USA), and the Euclidean distance tool of ArcGIS was used to calculate TRD and TRA. Aspect is expressed in positive degrees from 0 to 359.9, measured clockwise from north. If the input raster is flat with zero slope, the aspect is assigned −1. The land use types in the study region were classified by Landsat 8 OLI image on 26 June 2013 with 30 m spatial resolution. The atmospheric rectification and geometric rectification of the OLI images were first processed by ENVI 4.6 (ITT, Boulder, CO, USA). The NDVI and NDMI were calculated based on this remote-sensing image. Subsequently, land use types were interpreted by man–machine collaboration based on the OLI images, and the woodland, farmland, residential area (construction area), wetland and unused land were classified according to the Chinese Academy of Sciences (Wang *et al.*, 2001). Finally, a field survey was applied to improve the interpretation accuracy until it was >80%. The TRA and TRD were calculated based on this interpretation result. The land use types were thereafter quantified with the land use degree comprehensive index (LDCI). The equation is as follows:

$$\mathrm{LDCI}_a = 100 \times \sum_{i=1}^{n} A_i \times C_i \tag{1}$$

Where $\mathrm{LDCI}_a$ is the land use degree index; $A_i$ is the land use classification index and the quantitative values are 4, 3, 2, 2 and 1 for construction area, farmland, woodland, wetland and unused land; and $C_i$ is the area percentage of different land use types in one unit. High $A_i$ values indicate a large effect of human activities at one region (Zhuang and Liu, 1997). The LDCI has been used as the auxiliary variable in predicting SOCD and it has been proved that LDCI has significant correlation with SOCD (Liu *et al.*, 2015).

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

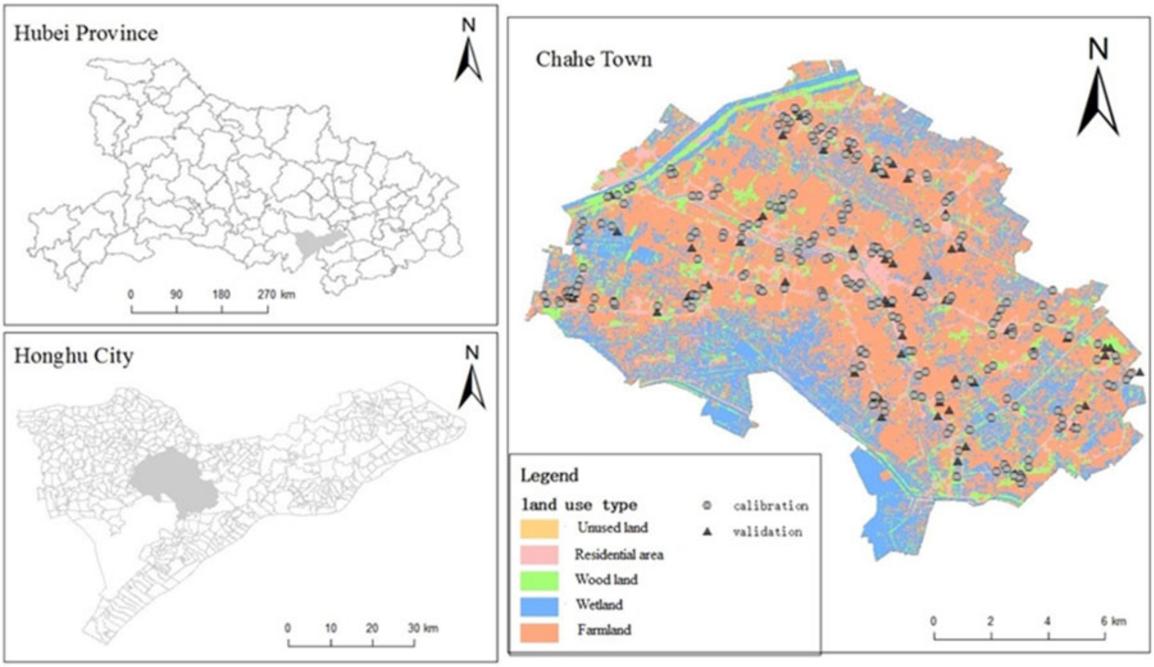

**Fig. 1.** Location of the study area in Chahe Town and the spatial distribution of samples for model calibration ($n = 173$) and validation ($n = 87$). Colour online.

### *Coefficient of variation*

The coefficient of variation (CV) shows the dispersion degree of data sets and is defined as the ratio of the standard deviation value to the mean value. High CV values indicate a high degree of variation for a particular variable. The variability of variable is classified based on the research of Wilding (1985), where properties with CV < 0.15 are least variable, properties with 0.15 < CV < 0.35 are moderately variable and those with CV > 0.35 are most variable.

### *Calculation of soil organic carbon density*

Soil organic carbon density is calculated as follows:

$$\rho_{\text{SOC}} = \sum_{i=1}^{n} (1 - \theta_i\%) \times p_i \times C_i \times T_i/100 \qquad (2)$$

where $\rho_{\text{soc}}$ is SOCD (kg/m$^2$) of the top soil (0–20 cm), $i$ is the soil horizon, $\theta_i\%$ is the gravel concentration (>2 mm) in the $i$th horizon, $p_i$ is the soil bulk density in the $i$th horizon (g/cm$^3$), $C_i$ is SOC content (g/kg) obtained by multiplying SOM by 0.58 (Bemmelen conversion fraction) and $T_i$ is the soil thickness (cm) in the current study.

### *Model calibration*

Topsoil samples (260) were divided into a calibration data set ($n = 173$, 2/3) and validation data set ($n = 87$, 1/3) on the basis of the Kennard–Stone algorithm (De Groot *et al.*, 1999). Eight environmental factors (TRA, TRD, NDMI, NDVI, elevation, slope, aspect and LDCI) were used as auxiliary variables. Principal component analysis was performed as the pre-processing method to reduce the multicollinearity and the dimension of these variables. A suitable number of PCs were chosen as the explanatory variables for constructing the prediction models of SOCD via OK, OCK, OLS and GWR.

### *Principal component analysis*

Principal component analysis, a widely used traditional statistical procedure, can explore trends in multiple variables. The procedure transforms correlated variables into the number of independent variables which are uncorrelated, namely, PCs, which are linear combinations of the original variables. The first principal component (PC1) includes the largest possible variance, under the constraint that the preceding component is orthogonal, and each subsequent component has the highest possible variance in turn (Jeyabharathi and Suruliandi, 2013). Other details on PCA are found in the work of Johnson and Wichern (2002), which provides a good overview of PCA.

### *Ordinary kriging and ordinary co-kriging*

Kriging as an advanced geostatistical procedure can generate a continuous surface from sparse soil samples based on their attributes. Assume $Z(Xi)$ is a regionalized variable with a variogram $\gamma$ $(h)$, which is a function describing the spatial aggregation field or stochastic process $Z(u)$. Exponential and spherical methods are used as the semi-variance model of the variation function. The spherical function is defined as:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left(\dfrac{3h}{2h} - \dfrac{h^3}{a^3}\right) & 0 < h \le a \\ C_0 & h > a \end{cases} \qquad (3)$$

The function of a Gaussian model can be estimated as follows:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C(1 - \mathrm{e}^{-(r^2/a^2)}) & h > 0 \end{cases} \qquad (4)$$

In these two equations, $a$ is the range of the soil samples; $h$ is the spatial lag; $C_0$ is the nugget; and $C_0 + C$ is the partial sill.

The traditional OK can provide unbiased estimates with minimum errors. The function of OK is expressed as

$$Z^{*}(x_0) = \sum_{i=1}^{n} \lambda_i(x_0) Z(x_i) \tag{5}$$

Here, $\sum_{i=1}^{n} \lambda_i(x_0) = 1$; $Z^{*}(x_0)$ is the predicted value of the variable $z$ at location $x_0$; $Z(x_i)$is the measured data; $\lambda_i(x_0)$ refers to the weights associated with the measured values; and $n$ is the number of predicted values within some neighbour soil samples.

Ordinary co-kriging is used to incorporate an auxiliary variable in process and is usually used in cases when two or more auxiliary variables are exhaustively available. The OCK estimator is written as follows:

$$Z^{*}_{\mathrm{OCK}}(u) = \sum_{\alpha_1=1}^{n_1(u)} \lambda_{\alpha_1}^{\mathrm{ock}}(u) Z_1(u_{\alpha_1}) + \sum_{i=2}^{N_v} \lambda_i^{\mathrm{ock}}(u)[Z_i(u) - m_i + m_1] \tag{6}$$

where the single constraint is that the sum of all weights must be equal to 1.

$$\sum_{\alpha_1=1}^{n_1(u)} \lambda_{\alpha_1}^{\mathrm{ock}}(u) + \sum_{i=2}^{N_v} \lambda_i^{\mathrm{ock}}(u) = 1 \tag{7}$$

where, $Z^{*}_{\mathrm{OCK}}(u)$ is the predicted value of the original variable $Z_1$ at unknown location $u$; $\lambda_i^{\mathrm{ock}}(u)$ is the weight of measured values in different geographical locations; $Z_1(u_{\alpha_1})$ denotes the primary values; $N_v$ is the total number of auxiliary variables; $Z_1(u_{\alpha_1})$ is the correlate data of the $i$th auxiliary variable; $\lambda_i^{\mathrm{ock}}(u)$ is the weight of the correlate data of the $i$th variable; and $m_1$ is the mean of the primary variables and $m_i$ is the mean of the $i$th auxiliary variable.

### *Geographically weighted regression model*

An extension of the traditional regression model, GWR considers the geographical locations and spatial weights of auxiliary variables (Brunsdon *et al.*, 1998; Fotheringham *et al.*, 2002). It is calculated as follows:

$$\hat{C}_{\mathrm{gwr}}(s_0) = \beta_0 + \sum_{k=0}^{p} \beta_k(s_0) \times X_k(s_0) + \varepsilon(u) \tag{8}$$

where $\hat{C}_{\mathrm{gwr}}(s_0)$ represents the predicted value of SOCD at the location of $S_0$; $X_k(s_0)$ is the independent variables at the geographical location of $S_0$; $\beta_0$ means the intercept; $\beta_k(s_0)$ is the coefficient of GWR which considers the relationship between the SOCD and the auxiliary variables; $p$ is the number of the soil samples; and $\varepsilon(u)$ is the error term.

The corrected Akaike information criterion (AICc) was used to determine the optimal bandwidth in the current paper. The AICc is defined as:

$$\mathrm{AICc} = -2 \ln L(\hat{\theta}_{\mathrm{L}}, X) + 2q \tag{9}$$

where $\hat{\theta}_{\mathrm{L}}$ is the maximum likelihood estimator, and $q$ is the number of unknown parameters. A high likelihood function indicates an accurate estimator. Thus, a minimum value of AICc is suitable to optimize the model.

### *Model evaluation*

Four SOCD models were constructed using calibration data sets based on OK, OCK, OLS and GWR. The predicted accuracy of different models is evaluated by the mean absolute estimation error (MAEE), root mean square error (RMSE) and Pearson's correlation coefficient ($r$) in validation data sets.

$$\mathrm{MAEE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2} \tag{11}$$

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - \left(\sum x_i\right)^2} \sqrt{n \sum y_i^2 - \left(\sum y_i\right)^2}} \tag{12}$$

where $x_i$ is the estimated SOCD at location $i$, $y_i$ is the observed SOCD at geographical location $i$, and $n$ is the number of sample observations. An accurate model should have the lowest RMSE and MAEE values.

## Results

### *Descriptive statistics*

The statistical summary of SOCD and environmental factors is shown in Fig. 2. The observed SOCD varied from 0.33 to 11.23 kg/m$^2$, with an arithmetic mean of 5.05 kg/m$^2$ and a range of 10.91 kg/m$^2$. The values of skewness and kurtosis were 0.39 and 3.62, respectively, which implied that the samples fitted a normal distribution. Thus, the original SOCD data can be used to construct spatial models without any transformation. The CV of SOCD was 32.50%, which indicated that SOCD in the study region was moderately variable. For the other environmental variables, elevation had the least variability; NDVI, NDMI and LDCI were moderately variable; and all the other variables were highly variable. The mean values of TRA, TRD, NDMI, NDVI, elevation, slope, aspect and LDCI were 176.64 m, 813.15 m, 0.27, 0.41, 21.59 m, 1.25°, 109.41 and 2.57, respectively. The topography of the study area is flat, with slope ranging from 0 to 6.75°. The LDCI ranged between 1 and 3.46 (mean 2.57). The values of NDVI were from 0.06 to 0.58, which showed that most of the land surface was covered with vegetation. Figure 2 also shows the basic statistics of other environmental variables.

The formation, decomposition and transformation of SOC were influenced by many environmental factors, which were complex and varied greatly across different natural landscapes (Zhi *et al.*, 2013). The degree of influence of these factors on SOCD should be distinguished and decided. Principal component analysis was used to classify these factors into several categories and to eliminate data redundancy. The major information of the impact factors was captured by PCA (Table 1). The first three PCs were chosen as the explanatory variables, as their eigenvalues were >1, thus explaining 60% of the observed variance in the environmental factors. The first component (PC1) explained approximately 26.7% of the total variance and had a significant positive correlation with NDMI (0.90) and NDVI (0.93). Both

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

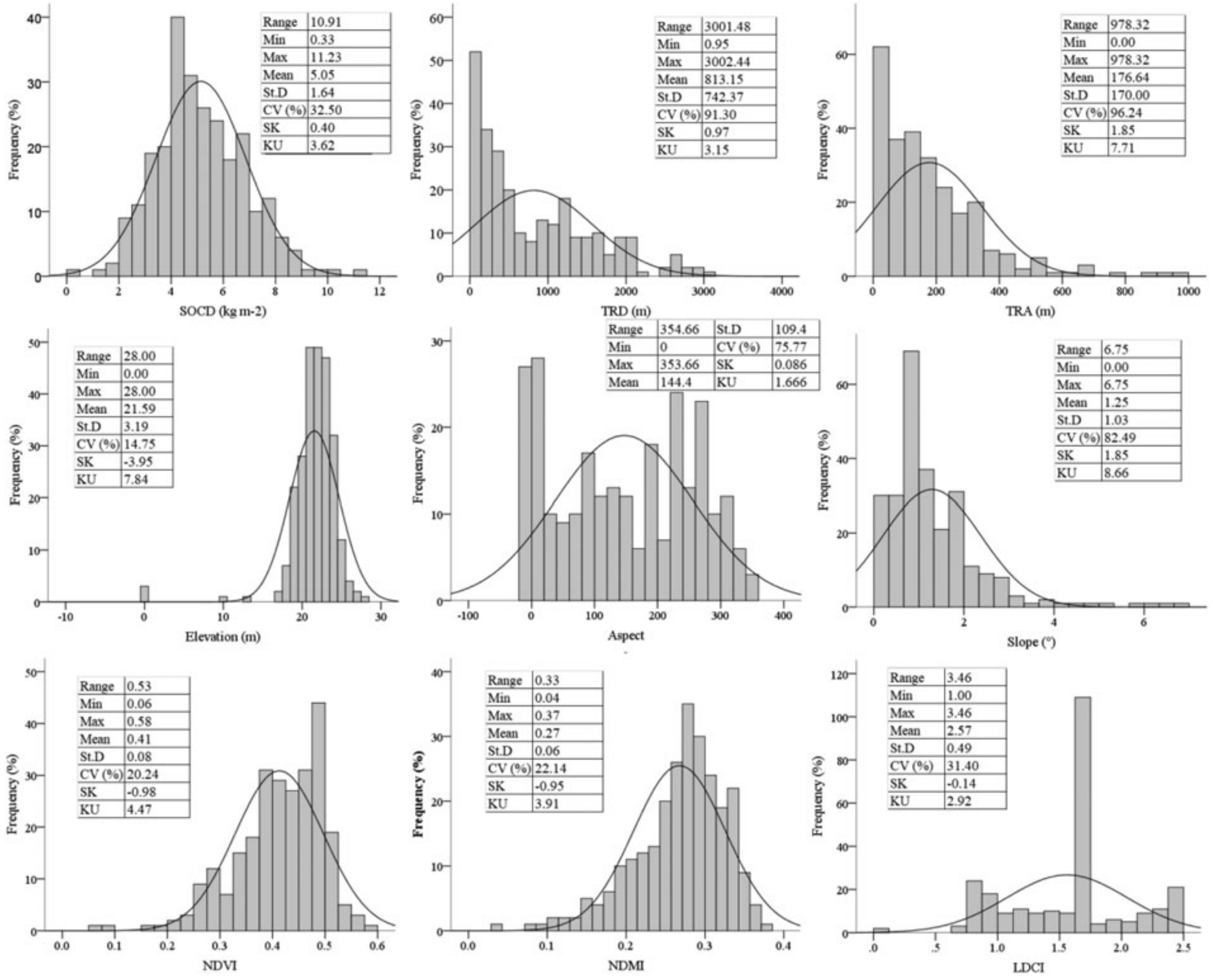

**Fig. 2.** Descriptive statistics of measured soil organic carbon density and other environmental variables of the study area (*n* = 260). SOCD, soil organic carbon density; NDVI, normalized difference vegetation index; NDMI, normalized difference moisture index; TRD, recent distance to road; TRA, recent distance to city; CV, coefficient of variation; LDCI, land use degree comprehensive index.

are spectral indices calculated with the spectrum of remote-sensing images and reflected soil moisture and vegetation growth. The second component (PC2) explained about 18.0% of the observed variance in the data set. This component was correlated negatively with TRA (−0.44) and TRD (−0.68) but positively with elevation (0.72) and LDCI (0.44). It was the summary of the elevation and human-induced factors. The third component (PC3) accounted for an additional 15.5% of the observed variance in the data set and had a strong positive correlation with aspect (0.70) and slope (0.80): it represented terrain information. The relationships observed from the PCA suggested that various environmental factors were intricately connected in the study region.

Figure 3 shows the spatial distribution characteristics of the first three PCs, which represent different environmental factors according to the eigenvectors of the correlation matrix. PC1, which represented vegetation growth and soil moisture, was related positively to NDVI and NDMI (Table 1). The PC1 value ranged from −5.76 to 3.14, with large values indicating high ratios of vegetation and soil moisture. High PC1 values were observed in the southwest of the study area, where many farmland area and woodland could be found. Low PC1 values were observed in the south, where wetland was the main land use type. PC2 represented the elevation and human-induced factors. PC2 was negatively related to TRA and TRD but was positively related to elevation and LDCI (Table 1). PC2 also showed some zonal characteristics because of the influence from the road and residential area. Thus, the PC2 values were the comprehensive results of elevation, TRA, TRD and LDCI, and human activities were the main influence factors on PC2. Low PC2 values were observed at the middle of the study area where residential area is the main land use type. High PC2 values were found at the edge of the study area, which was located far away from areas with human activities. PC3 represented terrain information because of its positive relationship with slope and aspect (Table 1). High PC3 values indicated high slopes and the proximity of a specific region to the south of the aspect. Although the impact factors were transformed by PCA, these three PCs could reflect the original information of these factors according to the eigenvectors of the correlation matrix. The values of the PCs in different geographical locations could be interpreted on the basis of natural or human-induced factors: PC1 representing the remote-sensing vegetation index, PC2 representing the human activities and PC3 representing the topographic features.

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

**Table 1.** Eigenvalues and the eigenvectors of the correlation matrix in principal component analysis

| | Component variables | | | |
|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC4 |
| Items | | | | |
| Eigenvalues | 2.1 | 1.4 | 1.2 | 0.9 |
| Percentage of variance (%) | 26.7 | 18.0 | 15.5 | 11.5 |
| Cumulative percentage (%) | 26.7 | 44.7 | 60.2 | 71.8 |
| | Eigenvectors | | | |
| Elevation | 0.14 | 0.72 | 0.04 | 0.26 |
| Aspect | 0.03 | 0.35 | 0.70 | 0.19 |
| Slope | 0.03 | 0.10 | 0.80 | −0.30 |
| NDVI | 0.93 | −0.02 | −0.00 | −0.14 |
| NDMI | 0.90 | −0.29 | 0.02 | −0.02 |
| TRD | 0.14 | −0.44 | 0.20 | 0.71 |
| TRA | −0.23 | −0.68 | 0.30 | −0.33 |
| LDCI | 0.17 | 0.44 | −0.14 | −0.34 |

NDVI, normalized difference vegetation index; NDMI, normalized difference moisture index; TRD, recent distance to road; TRA, recent distance to city; LDCI, land use degree comprehensive index; PC, principal component.

### Diagnostic information of the geostatistical and regression models

Geostatistical models (OK and OCK) and regression models (OLS and GWR) were used to simulate SOCD in the study region. Table 2 shows the diagnostic parameters for the calibration data set. The major ranges of the empirical semi-variable function were 194.0 and 493.2 m. They were used in the fitting procedure, which involved modelling with OK and OCK approaches. The ratio of $C_0/(C_0 + C)$ reflects the spatial structure of the estimated factors, with 0–25% indicating strongly structured spatial dependence, 25–75% indicating moderate dependence and >75% indicating weak dependence (Jobbágy and Jackson, 2000). The ratio for SOCD in OK was 11.1%, which indicated the strong dependence of SOCD. By contrast, the ratio for SOCD in OCK was 61.7%, which indicated the moderate dependence of SOCD. Such a difference can be attributed mainly to the fact OCK considered the PCs as auxiliary variables and that part of the spatial distribution of SOCD was influenced by their spatial characteristics. The effective number (8.3) and sigma (0.48) of GWR are important parameters in GWR model, and these figures are always used to choose the suitable bandwidth. The AICc of GWR was 307.34, which was smaller than that of OLS (309.5) and indicated that GWR has better simulation precision than OLS. The RMSE can be used to evaluate simulation precision between different models. In the current study, the regression models yielded better results than the geostatistical models in terms of RMSE values, and also the spatial characteristics and the auxiliary variables played an important role in improving the model performance.

RMSE, MAEE and $r$ were used to evaluate the performance of these four models based on the validation data set, namely, OK, OCK, OLS and GWR. The RMSE values of these models were 1.2, 1.1, 1.2 and 1.1 kg/m$^2$; MAEE values were 0.91, 0.92, 0.93 and 0.85 kg/m$^2$; and $r$ values were 0.30, 0.39, 0.35 and 0.46. The RMSE *v.* the MAEE points are plotted in Fig. 4 for a comprehensive comparison of the model performance. As shown in Fig. 4, a long distance from the point to the origin would result in poor prediction accuracy. The distances of the OK, OCK, OLS and GWR models were 1.50, 1.46, 1.49 and 1.39, respectively. The highest prediction accuracy was seen in GWR, followed by OCK, OLS and OK. In these models, OLS only considers relationships between the auxiliary variables and SOCD, and OK only considers the spatial characteristics of SOCD. Moreover, GWR could account for both the spatial trend and local variations of the relationships between the environmental factors and SOCD, whereas OCK could not easily capture the valuable information regarding environmental factors in unknown soil samples. Thus, environmental factors and local spatial variation of SOCD played important roles in predicting SOCD, and GWR is a very promising spatial interpolation method for predicting soil properties.

### Analysing the spatial distribution of soil organic carbon density

Figure 5 illustrates the spatial distribution maps of SOCD estimated by OK, OCK, OLS and GWR. Generally, the spatial distribution characteristics of SOCD are the same across the four images. Low estimated values of SOCD were concentrated in the middle region and high estimated values were distributed at the edge of the study region. Between these two kinds of models, the SOCD values ranged from 0.32 to 11.23 kg/m$^2$ in the geostatistical models and from 2.76 to 9.97 kg/m$^2$ in the regression models. The geostatistical models had the bigger numerical ranges than the regression models because the extremums of SOCD were smoothed by the regression models. Meanwhile, in the middle area, the area of low (0.33–4.98 kg/m$^2$) SOCD values in geostatistical models were bigger than regression models. The spatial distribution of SOCD under these two kinds of models obviously differed. The shape was similar to the rings in geostatistical models, but the lines were similar to those in the regression models. These results can be attributed to the fact that geostatistical models are spatial interpolation models and that regression models are multiple linear regression models. The measured SOCD and environmental variables played different roles in estimating SOCD from different models.

Although these two models shared many similarities, their interiors revealed detailed differences. Relative to OK, the spatial distribution of SOCD in OCK can reflect more detailed information in the local spatial, and the different values of SOCD were divided more accurately. This phenomenon resulted from the fact that OK only considered the spatial variation and dependency of SOCD, whereas OCK considered the spatial information of environmental factors that influenced the spatial distribution of SOCD. In Fig. 5, the spatial distribution features of SOCD were similar between OLS and GWR. High values were found in the west and the northeast portion of the study region, and low values were concentrated in the middle part of the study area. The spatial patterns of SOCD were different in many patches because the values of SOCD were estimated through auxiliary variables and their weights. Table 3 shows the global coefficients and the important parameters of PCs in the OLS model, while Fig. 6 presents the local coefficients of PCs in different locations in GWR model. The spatial weights of the auxiliary variables were constant in different geographical locations in the OLS model but were different in the GWR model.

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

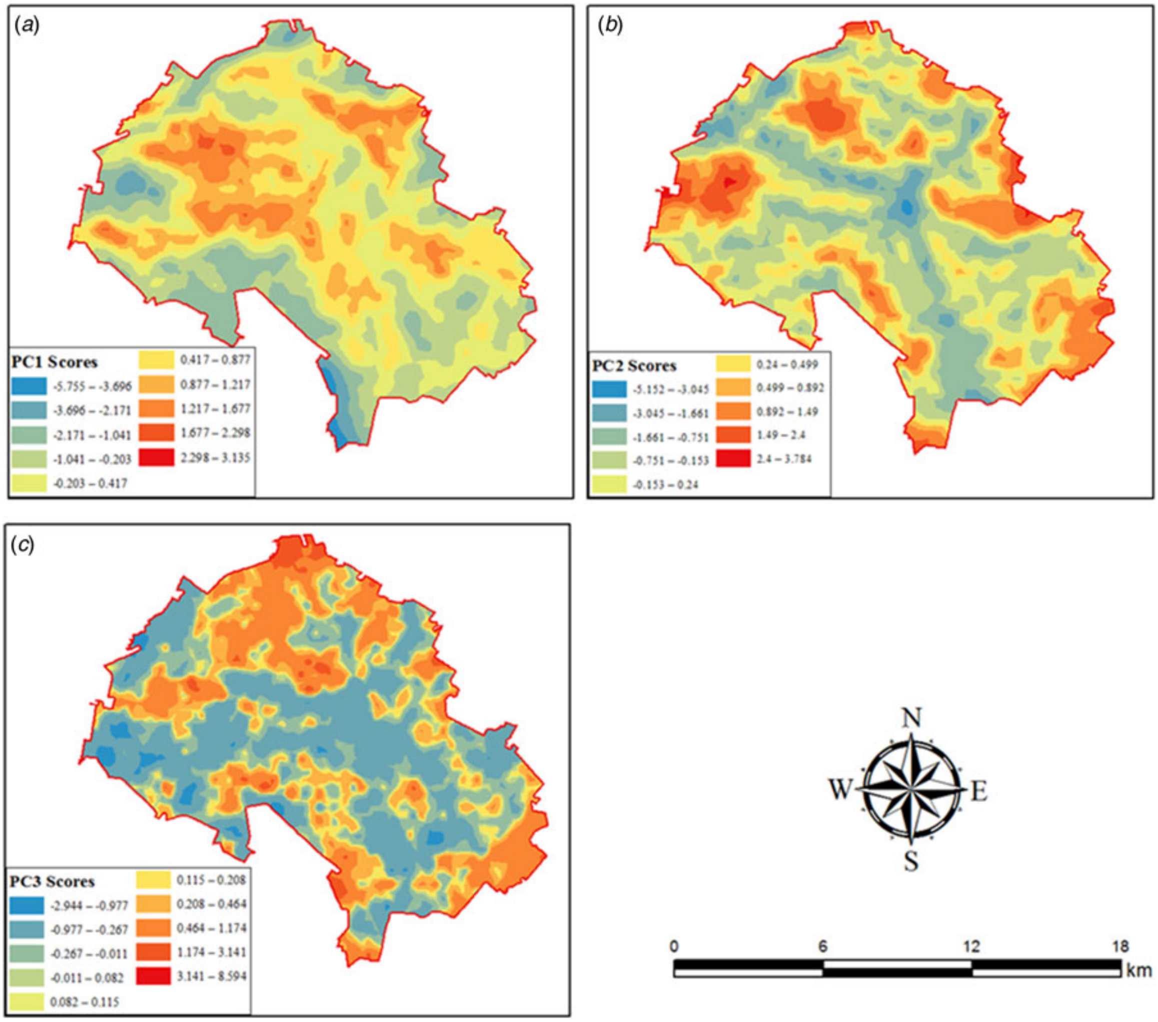

**Fig. 3.** Spatial distribution of the principal component scores: (a) first principal component scores (26.697%), (b) second principal component scores (18.026%) and (c) third principal component scores (15.507%). Colour online.

**Table 2.** Diagnostic parameters of the geostatistical and regression models

| Geostatistical models | | | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Semivariance | Variables | $C_0$ | $C$ | Range (m) | $C_0/(C_0+C)$ (%) | RMSE |
| OK | Exponential | SOCD | 0.30 | 2.4 | 194.0 | 11.13 | 1.90 |
| OCK | Gaussian | SOCD | 1.9 | 1.2 | 493.3 | 61.72 | 1.87 |
| | | PC1 | 0 | 2.0 | | 0.00 | |
| | | PC2 | 0 | 1.0 | | 0.00 | |
| | | PC3 | 0.41 | 0.89 | | 31.7 | |
| Regression models | | | | | | | |
| | | Effective number | Sigma | AICc | RMSE | | |
| OLS | | – | – | 309.5 | 0.54 | | |
| GWR | | 8.39 | 0.48 | 307.3 | 0.53 | | |

SOCD, soil organic carbon density; OK, ordinary kriging; OCK, ordinary co-kriging; OLS, ordinary least squares; GWR, geographically weighted regression; $C_0$, nugget; $C$, partial sill; $C_0+C$: total sill; RMSE, root mean square error; Range, major range of the empirical semi-variable function, the range of exponential is $3a$ and the range of Gaussian is $\sqrt{3}a$; AICc, corrected Akaike information criterion.

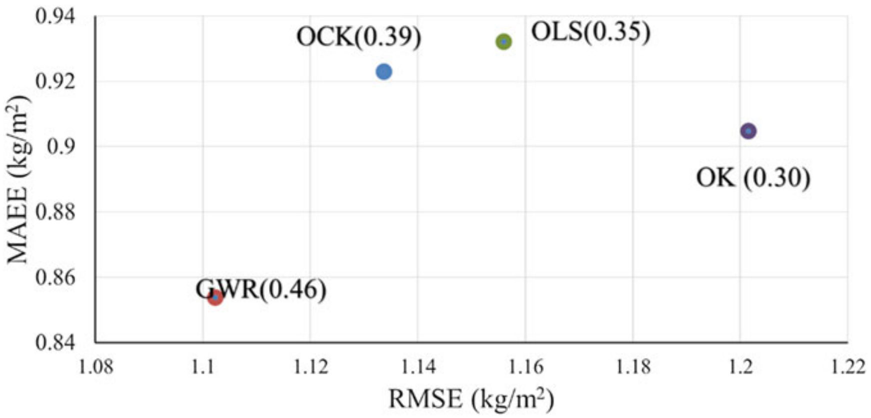

**Fig. 4.** Root mean square error (RMSE) *v.* mean absolute estimation error (MAEE) plots for the four models and their Pearson's *r* values, namely, ordinary kriging (OK), ordinary co-kriging (OCK), ordinary least squares (OLS) and geographically weighted regression (GWR). Colour online.

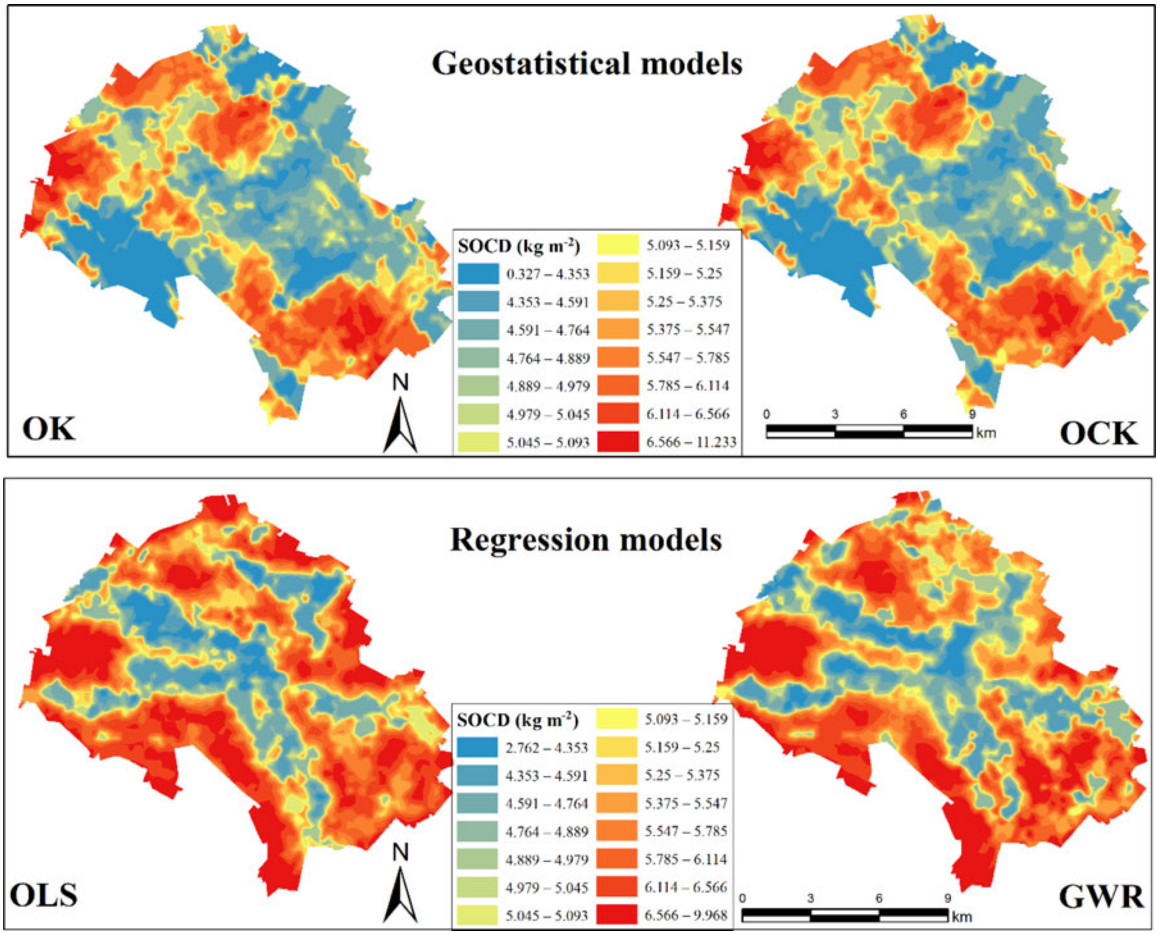

**Fig. 5.** Spatial distribution of soil organic carbon density (SOCD) by ordinary kriging (OK), ordinary co-kriging (OCK), ordinary least squares (OLS) and geographically weighted regression (GWR). Colour online.

## Discussion

### *Influence of different independent variables on soil organic carbon density*

Soil organic carbon density is influenced by different environmental variables, and such influence results in strong spatial heterogeneity and spatial dependence; these variables vary with the environment and geographical locations (Zhan *et al.*, 2013). Kumar *et al.* (2012*a*) integrated natural variables (e.g. temperature, precipitation, elevation, slope, geology and NDVI) with human variables (e.g. land use) in their spatial modelling of SOCD in the state of Pennsylvania, USA. Zhang *et al.* (2011) investigated and used environmental factors (e.g. rainfall, land cover and soil types) as explanatory variables to establish SOC models. Therefore, the influence of different environmental variables and human activities should be explored when mapping SOCD. The constant of a regression model can guarantee that the residuals have a zero mean. PC1 had a coefficient of 0.06 in

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

**Table 3.** Principal component variables in the ordinary least squared (OLS) model

| Variable | Coefficients | SE | *Z*-value | Probability | VIF |
|---|---|---|---|---|---|
| Intercept | 5.1 | 0.13 | 37.7 | <0.001 | – |
| PC1 | 0.06 | 0.10 | 0.63 | 0.491 | 1.0 |
| PC2 | −0.35 | 0.11 | −3.2 | 0.005 | 1.0 |
| PC3 | 0.01 | 0.12 | 0.07 | 0.945 | 1.0 |

S.E., standard error of mean; VIF, variance inflation factor.

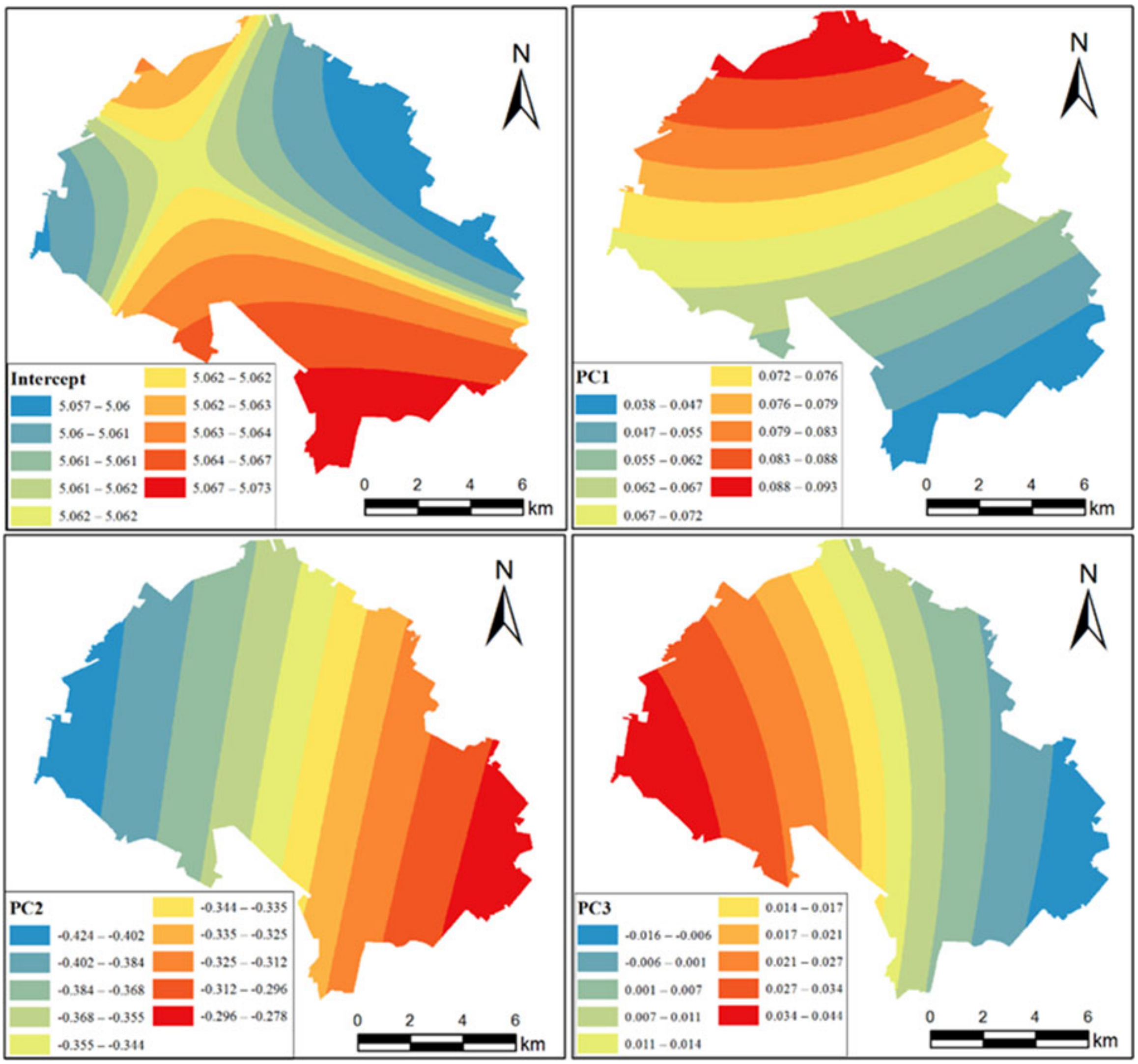

**Fig. 6.** Spatial distribution of the coefficients of different auxiliary variables in the geographically weighted regression (GWR) model. Colour online.

the OLS model but a coefficient of 0.04–0.09 in the GWR model. PC1 was also positively correlated with SOCD and high PC1 values equate to high SOCD. This result is reasonable because PC1 was representative of NDVI and NDMI, which were usually positively related to SOCD (Ruiz-Colmenero *et al.*, 2013). PC1 had a stronger influence in the north than in the south, possibly because of the differences in land use types, soil moisture and land cover. Normalized difference moisture index was an indicator of soil moisture, and a higher soil moisture could improve the living environment of soil microbes, accelerate the transformation and decomposition of SOC, and prevent the net loss of organic soils from oxidation (Liu *et al.*, 2015). PC2 had a coefficient of −0.35 in OLS model and a coefficient of −0.02 to −0.28 in GWR model. PC2, which represented elevation and human-induced factors, had a negative correlation with SOCD. On the basis of the eigenvectors and coefficients of the models, SOCD had a positive correlation with TRA and TRD but a negative correlation with elevation and LDCI. Soil organic carbon density decreased along with an increase in elevation because the soil could be easily transferred from highly elevated areas to

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

low elevated areas by rains or winds. Moreover, SOC tends to accumulate in low-lying areas (Liu *et al*., 2015). The LDCI is a comprehensive index that indicates the influence of human activities on land use. A strong influence of human activities would result in a low SOCD value because human activities could destroy the soil structure that provides a favourable living environment for soil microbes (Wu *et al*., 2003). The coefficients of TRA and TRD reflected the influence of the residential area and road on SOCD. A short distance would result in a low SOCD because closeness to residential areas and roads would strengthen the influence of human disturbance on SOCD. The vegetation and soil structure in natural landscapes are usually destroyed during the processes of urban development and road construction, thereby accelerating soil erosion, reducing soil nutrients and eventually reducing SOCD (Shubin, 2006). PC3 had a coefficient of 0.01 in the OLS model and a coefficient of −0.02 to 0.04 in the GWR model. The relationships between the PC3 and the slope and aspect were positive, which means SOCD was positively related to the slope and aspect, except in a small area in the southeast. Slope could influence the accumulation of soil, while aspect could influence the time exposure of sunlight, which would subsequently influence soil moisture and the activities of soil microbes (Yimer *et al*., 2006). In summary, the coefficient maps of GWR can reflect the influence of different explanatory variables on SOCD across regions. This finding could help environmentalists distinguish the importance of different impact factors and help farmers manage their farmlands in a suitable and scientific way.

### *Evaluating the prediction accuracy of the ordinary kriging, ordinary co-kriging, ordinary least squares and geographically weighted regression models*

Previous studies have investigated a number of different estimation approaches considered here. For example, Zhang *et al.* (2011) evaluated the prediction accuracies of GWR, OK, inverse distance weighted and OLS in mapping SOC in Ireland by using rainfall, land cover and soil types as the explanatory variables. Wang *et al.* (2013*b*) used GWR and OCK to estimate the total nitrogen in soil by referring to elevation, slope, land use type, topographic wetness index and soil type. Kumar *et al.* (2013) compared the application of GWR and OLS in mapping the spatial distribution of SOCD in Ohio, USA. These studies showed the potential of GWR in predicting soil properties by using its excellent prediction accuracy and various weight coefficient maps. However, these studies considered all environmental variables as explanatory variables and ignored the multicollinearity and redundancy among them. In the current paper, PCA was used as a pre-processing method in the present work to address the aforementioned problems and constructed prediction models according to the PCs of the selected environmental variables. The prediction accuracy of GWR was the highest among all the tested models (OCK, GWR, OLS and OK) for the auxiliary variables and the spatial weights estimated were found to be highly beneficial in predicting SOCD. Ordinary least squares achieved poor accuracy because this model ignored the spatial characteristics of both SOCD and the auxiliary variables (PC1, PC2 and PC3) and simply used PCs as explanatory variables to construct SOCD models. Moreover, OLS could not account for the spatially varying relationships between the environmental factors and SOCD. Ordinary kriging also obtained a low accuracy because this model only considered the spatial dependence of SOCD but ignored the valuable information from the environmental variables. Therefore, OK and OLS achieved lower prediction accuracies than the other two models. Both GWR and OCK utilized the spatial correlations of the environmental variables and SOCD during the model construction, but the auxiliary variables of the soil samples cannot be used to predict SOCD at unknown geographical locations by OCK. Geographically weighted regression can construct a soil SOCD map based on the relationship between the SOCD and the auxiliary variables, and this is why the GWR has better prediction accuracy than OCK. Meanwhile, the SOCD map that was interpolated with GWR could provide additional information on SOCD within local regions. The coefficient maps could reflect the influence degree of environmental variables to SOCD in different geographical locations. Thus, PCA is one useful method in reducing the dimensionality and redundancy of the auxiliary variables. Geographically weighted regression has enormous potential in enabling prediction of soil properties as it provides more information in explaining the local spatial relationships between the environment variables and SOCD.

## Conclusion

A total of 260 soil samples from Chahe Town, which is located at the centre of Jianghan Plain, were analysed. Eight environmental variables (TRA, TRD, NDMI, NDVI, elevation, slope, aspect and land use type) were used as auxiliary variables to construct the prediction models of SOCD. The PCA method was used to reduce multicollinearity and redundancy of these environmental variables. The first three PC variables were chosen as explanatory variables. Two geostatistical models (OK and OCK) and two regression models (OLS and GWR) were used to map SOCD. As revealed by the evaluation indices (MAEE, RMSE and Pearson's *r*), GWR demonstrated the highest accuracy, followed by OCK, OLS and OK. The main reason was that GWR can capture more details on the local variation in SOCD and reflect the effects of auxiliary variables on SOCD. Therefore, the combination of PCA and GWR is a promising method in reducing the redundancy of environmental factors and in constructing prediction models of SOC stocks.

**Financial support.** This study was supported by the Natural Science Foundation of Hubei (2018CFB372); the Fundamental Research Funds for the Central Universities (2662016QD032); the Key laboratory of aquatic plants and watershed ecology of Chinese academy of sciences (Y852721s04), the Chinese National Natural Science Foundation (41371227); and the National Undergraduate Innovation and Entrepreneurship Training Program (201810504023, 201810504030).

**Conflict of interest.** None.

**Ethical standards.** Not applicable.

## References

**Brunsdon C, Fotheringham S and Charlton M** (1998) Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)* **47**, 431–443.

**Conforti M, Castrignano A, Robustelli G, Scarciglia F, Stelluti M and Buttafuoco G** (2015) Laboratory-based Vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. *Catena* **124**, 60–67.

**Deckers JA, Nachtergaele F and Spaargaren OC** (1998) *World Reference Base for Soil Resources: Introduction.* Leuven, Belgium: Acco.

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press

**De Groot P, Postma G, Melssen W and Buydens L** (1999) Selecting a representative training set for the classification of demolition waste using remote NIR sensing. *Analytica Chimica Acta* **392**, 67–75.

**Evrendilek F, Celik I and Kilic S** (2004) Changes in soil organic carbon and other physical soil properties along adjacent Mediterranean forest, grassland, and cropland ecosystems in Turkey. *Journal of Arid Environments* **59**, 743–752.

**Fotheringham AS, Brunsdon C and Charlton M** (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New York, USA: Wiley.

**Guo L, Zhao C, Zhang H, Chen Y, Linderman M, Zhang Q and Liu Y** (2017) Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology. *Geoderma* **285**, 280–292.

**Harris P, Fotheringham A, Crespo R and Charlton M** (2010) The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. *Mathematical Geosciences* **42**, 657–680.

**Jeyabharathi D and Suruliandi A** (2013) Performance analysis of feature extraction and classification techniques in CBIR. In *2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT)*. Piscataway, NJ, USA: IEEE, pp. 1211–1214.

**Jobbágy EG and Jackson RB** (2000) The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications* **10**, 423–436.

**Johnson RA and Wichern DW** (2002) *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA: Prentice Hall.

**Keser S, Duzgun S and Aksoy A** (2012) Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in Turkey. *Waste Management* **32**, 359–371.

**Kumar S, Lal R and Liu D** (2012a) A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma* **189–190**, 627–634.

**Kumar S, Lal R and Lloyd CD** (2012b) Assessing spatial variability in soil characteristics with geographically weighted principal components analysis. *Computational Geosciences* **16**, 827–835.

**Kumar S, Lal R, Liu DS and Rafiq R** (2013) Estimating the spatial distribution of organic carbon density for the soils of Ohio, USA. *Journal of Geographical Sciences* **23**, 280–296.

**Lal R** (2003) Global potential of soil carbon sequestration to mitigate the greenhouse effect. *Critical Reviews in Plant Sciences* **22**, 151–184.

**Liu Y, Wang C, Yue WZ and Hu YY** (2013) Storage and density of soil organic carbon in urban topsoil of hilly cities: a case study of Chongqing Municipality of China. *Chinese Geographical Science* **23**, 26–34.

**Liu Y, Guo L, Jiang Q, Zhang H and Chen Y** (2015) Comparing geospatial techniques to predict SOC stocks. *Soil and Tillage Research* **148**, 46–58.

**Mishra U, Lal R, Liu D and Van Meirvenne M** (2010) Predicting the spatial variation of the soil organic carbon pool at a regional scale. *Soil Science Society of America Journal* **74**, 906–914.

**Ruiz-Colmenero M, Bienes R, Eldridge D and Marques M** (2013) Vegetation cover reduces erosion and enhances soil organic carbon in a vineyard in the central Spain. *Catena* **104**, 153–160.

**Shi X, Yu D, Warner E, Sun W, Petersen G, Gong Z and Lin H** (2006) Cross-reference system for translating between genetic soil classification of China and soil taxonomy. *Soil Science Society of America Journal* **70**, 78–83.

**Shubin S** (2006) The changing nature of rurality and rural studies in Russia. *Journal of Rural Studies* **22**, 422–440.

**Six J, Paustian K, Elliott E and Combrink C** (2000) Soil structure and organic matter I. Distribution of aggregate-size classes and aggregate-associated carbon. *Soil Science Society of America Journal* **64**, 681–689.

**Song X-D, Brus DJ, Liu F, Li D-C, Zhao Y-G, Yang J-L and Zhang G-L** (2016) Mapping soil organic carbon content by geographically weighted regression: a case study in the Heihe River Basin, China. *Geoderma* **261**, 11–22.

**Sun W, Zhu YQ, Huang SL and Guo CX** (2015) Mapping the mean annual precipitation of China using local interpolation techniques. *Theoretical and Applied Climatology* **119**, 171–180.

**Viscarra Rossel RA and McBratney AB** (1998) Soil chemical analytical accuracy and costs: implications from precision agriculture. *Australian Journal of Experimental Agriculture* **38**, 765–775.

**Wang S-Y, Liu J-Y, Zhang Z-X, Zhou Q-B and Zhao X-L** (2001) Analysis on spatial-temporal features of land use in China. *Acta Geographica Sinica* **6**, 631–639.

**Wang JL, Kang SZ, Sun JS and Chen ZF** (2013*a*) Estimation of crop water requirement based on principal component analysis and geographically weighted regression. *Chinese Science Bulletin* **58**, 3371–3379.

**Wang K, Zhang C and Li W** (2013*b*) Predictive mapping of soil total nitrogen at a regional scale: a comparison between geographically weighted regression and cokriging. *Applied Geography* **42**, 73–85.

**Wilding LG** (1985) Spatial variability: its documentation, accommodation and implication to soil surveys. In Nielsen DR and Bouma J (eds), *Soil Spatial Variability. Workshop. Proceedings of a Workshop of the ISSS and the SSA, Las Vegas*. Wageningen, The Netherlands: Pudoc, pp. 166–194.

**Wilson EH and Sader SA** (2002) Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sensing of Environment* **80**, 385–396.

**Wu H, Guo Z and Peng C** (2003) Land use induced changes of organic carbon storage in soils of China. *Global Change Biology* **9**, 305–315.

**Yimer F, Ledin S and Abdelkadir A** (2006) Soil organic carbon and total nitrogen stocks as affected by topographic aspect and vegetation in the Bale Mountains, Ethiopia. *Geoderma* **135**, 335–344.

**Zhan C, Cao J, Han Y, Huang S, Tu X, Ping W and An Z** (2013) Spatial distributions and sequestrations of organic carbon and black carbon in soils from the Chinese Loess Plateau. *Science of the Total Environment* **465**, 255–266.

**Zhang C, Tang Y, Xu X and Kiely G** (2011) Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland. *Applied Geochemistry* **26**, 1239–1248.

**Zhi J-J, Jing C-W, Zhang C, Wu J-P, Ni Z-H, Chen H-J and Xu J** (2013) Estimation of soil organic carbon density and storage in Zhejiang Province of East China by using 1:50000 soil database. *Ying Yong Sheng Tai Xue Bao (Chinese Journal of Applied Ecology)* **24**, 683–689.

**Zhuang D-F and Liu J-Y** (1997) Study on the model of regional differentiation of land use degree in China. *Zi Ran Zi Yuan Xue Bao (Journal of Natural Resources)* **12**, 105–111.

https://doi.org/10.1017/S0021859618000709 Published online by Cambridge University Press