Geoderma 261 (2016) 11–22

Contents lists available at ScienceDirect

Geoderma

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

Mapping soil organic carbon content by geographically weighted regression: A case study in the Heihe River Basin, China

Xiao-Dong Song [a], Dick J. Brus [a,b], Feng Liu [a], De-Cheng Li [a], Yu-Guo Zhao [a], Jin-Ling Yang [a], Gan-Lin Zhang [a,*]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] Alterra, Wageningen University and Research Centre, PO Box 32, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form 9 June 2015
Accepted 30 June 2015
Available online xxxx

Keywords:
Digital soil mapping
Kriging with an external drift (KED)
Restricted maximum likelihood (REML)
Map accuracy
Cross-validation
Effective sample size

ABSTRACT

In large heterogeneous areas the relationship between soil organic carbon (SOC) and environmental covariates may vary throughout the area, bringing about difficulty for accurate modeling of the regional SOC variation. The benefit of local, geographically weighted regression (GWR) coefficients was tested in a case study on soil organic carbon mapping across a 50,810 km^2 area in northwestern China. This area is composed of an alpine ecosystem in the upper reaches and oases in the middle reaches. The benefit was quantified by comparing the quality of the maps obtained by GWR and geographically weighted ridge regression (GWRR) on the one side and multiple linear regression (MLR) on the other side. In these methods spatial dependence of model residuals is ignored. The root mean squared error (RMSE) of predictions of natural log-transformed SOC obtained with GWR was smaller than with MLR: 0.565 versus 0.618 g/kg. The use of a local ridge parameter in GWRR did not lead to an increase in accuracy. Besides we compared the quality of maps obtained by geographically weighted regression followed by simple kriging of model residuals (GWRSK) and kriging with an external drift (KED) with global regression coefficients. In these methods the spatial dependence of model residuals is incorporated in the model. The RMSE with KED was smaller than with GWRSK: 0.515 versus 0.546 g/kg. We conclude that fitting regression coefficients locally as in GWR only paid when no spatial random effect was included in the model. When a spatial random effect was included, the flexibility of local, geographically weighted regression coefficients was not needed and even undesirable as it led to less accurate predictions than KED with global regression coefficients. In comparing the accuracy of prediction methods by leave-one-out cross-validation (LOOCV) of a non-probability sample it is important to account for possible autocorrelation of pairwise differences in the prediction errors. The effective sample sizes were substantially smaller than the total number of sampling points, so that most pairwise differences in MSE were not significant at a significance level of 10% in a two-sided paired *t*-test.

1. Introduction

Soil is one of the most important carbon stocks globally and maps showing soil organic carbon (SOC) can be used to guide practical soil management (Meersmans et al., 2008). Cost-efficient methods for mapping SOC content are therefore indispensable (Kheir et al., 2010).

A recent review of 90 papers on digital soil mapping and modeling revealed that 31% (28) of these papers focused on SOC (Grunwald, 2009). Amid the plethora of mapping methods exploiting the relation between SOC and covariates were regression kriging (Piccini et al., 2014), (boosted) regression trees (Vasques et al., 2008, 2009), random forest (Grimm et al., 2008), and neural network (Li et al., 2013).

Despite the significant progress based on these methods, there are still methodological challenges, especially in large, highly variable areas, with spatially varying relationships between soil properties like soil organic carbon and environmental covariates. Recently GWR has received increased attention because of its ability to account for local relationships between the study variable and covariates (Brunsdon et al., 1998; Fotheringham et al., 2002). In GWR modeling it is assumed that neighboring observations have a stronger effect on the regression at a target point than observations at a greater distance. In GWR a distance decay function is applied to obtain local estimates of the regression coefficients (Tu, 2011).

The potentials of GWR for SOC mapping have been explored in various regional studies. Mishra et al. (2010) compared GWR with multiple linear regression (MLR) and regression kriging (RK). In the latter two methods the regression coefficients were global, i.e., these were assumed constant throughout the area. GWR led to a reduction in RMSE of 22% over MLR, but only 2% over RK. Zhang et al. (2011) and Wang et al. (2013) compared GWR with MLR. In both studies GWR outperformed MLR: in the former study the RMSE as obtained with MLR was reduced by 5%, in the latter study by 11%.

* Corresponding author.
E-mail address: glzhang@issas.ac.cn (G.-L. Zhang).

http://dx.doi.org/10.1016/j.geoderma.2015.06.024
0016-7061/

Geographically weighted regression kriging (GWRK) (Harris et al., 2010; Kumar et al., 2012) and simple kriging with GWR-derived local means (GWRSK) (Harris and Juggins, 2011; Lloyd, 2010) are extensions of the GWR approach. In these approaches the spatial variation is modeled as the sum of a deterministic trend modeled by GWR, and spatially correlated residuals. Kumar et al. (2012) compared GWRK and RK, and found that RMSE as obtained by GWRK was reduced by 43% due to the local regression coefficients.

Other studies have demonstrated that GWR models not always outperformed geostatistical models assuming global regression coefficients (Harris and Juggins, 2011; Harris et al., 2010; Lloyd, 2010). Also, we are not aware of studies of SOC mapping in which GWR and GWRSK are compared with kriging with an external drift (KED) using restricted maximum likelihood (REML) estimation of the model parameters. In RK the model parameters are estimated in two separate steps. In the first step the regression coefficients are estimated by ordinary least squares assuming independent data. In the second step the variogram parameters are estimated by method-of-moments from the regression model residuals. Ideally these steps are repeated until convergence, with generalized least squares estimation of the regression coefficients. This estimation procedure is known to be suboptimal; the model parameters can best be estimated by REML (Lark and Webster, 2006). Suboptimal estimates of the model parameters may lead to suboptimal predictions. Therefore we prefer to compare GWR and GWRSK with KED–REML (hereafter shortly denoted as KED) in order to assess the benefit of local regression coefficients in mapping SOC.

The aim of this study was to quantify the benefit of using local, geographically weighted regression coefficients instead of global coefficients in mapping SOC in a study area with complex topographical conditions, under two modeling assumptions. In the first modeling assumption spatial dependence of data is ignored (data are assumed independent), whereas in the second assumption the spatial dependence of data is part of the model. The benefit of local regression coefficients assuming independent data is quantified by comparing various quality indices among which the RMSE of GWR (local coefficients) and MLR (global coefficients), whereas the benefit when accounting for spatial dependence of data is quantified by comparing the quality indices of GWRSK (local coefficients) and KED (global coefficients).

2. Study area and data

2.1. Study area

As the second largest inland river of China, the Heihe River is 821 km long, originating from the Qilian Mountains and flowing into the western Inner Mongolian Plateau (Lu et al., 2009; Wu, 2011). The study area, consisting of the upper and middle reaches of the Heihe River Basin, stretches for 340 km from the northwest to the southeast (Fig. 1), with coordinates of about 97°20′–101°51′ E and 37°41′–39°59′N. The major part of the study area is in Gansu Province and a small part in Qinghai Province. The upper reaches are in the southern part of this area (Fig. 1) with an average elevation of 3556 m a.s.l. Most peaks are higher than 4000 m a.s.l. The average elevation of middle reaches is about 1811 m a.s.l. This study area covers about 50,810 km^2, accounting for about 36% of the total area of the Heihe River Basin.

The mean temperature of this area during winter and summer is −3 and 7 °C, respectively. Annual precipitation varies spatially from 82.8 mm to 425.6 mm, with rainfall occurring mainly from June to September. The land covers of this area are heterogeneous due to the wide range of elevation and strong anthropogenic activities in terms of irrigation farming. The upper reaches are characterized by a humid and cold climate, whereas the climate of the middle reaches is a typical temperate arid environment with low precipitation and high evaporation (Li et al., 2012). While the annual precipitation is very limited in the middle section of the Heihe River Basin, water shortage is the major obstacle to the crops that rely mainly on irrigation (Kang et al., 2004).

2.2. Soil sampling

Part of the soil data (2010–2013), 144 soil points, were provided by "Heihe Plan Science Data Center, National Natural Science Foundation of China" (http://www.heihedata.org). To collect additional data we selected 404 sampling locations by purposive sampling, using the method of Zhu et al. (2008). In this method representative soil sites are selected from soil-scape units constructed by fuzzy *c*-means classification of pixels on the basis of the soil forming factors (covariates). The additional soil sampling was conducted in July 2012 and July 2013. All the sampling sites were located by handheld global positioning system (GPS) receivers. A 100–150 cm deep soil pit was dug at each site. In total 548 topsoil (0–20 cm) samples were collected and stored in a digital database (Fig. 1). The samples were subsequently dried, sieved at 2 mm and analyzed using the Walkley–Black procedure for SOC.

The observed SOC contents in surface soils varied from 0.70 to 132.19 g/kg, with a mean value of 31.33 g/kg (Table 1). The frequency distribution of SOC showed strong positive skew; the skewness was 1.50 (Fig. 2, Table 1). For all prediction methods we therefore transformed SOC measurements by taking the natural logarithms (LnSOC). The skewness dropped to 0.15 (Table 1).

2.3. Predictors

Predictors utilized in this study for the mapping of SOC content of 0–20 cm topsoil (g/kg) were slope, aspect, elevation, profile curvature, plan curvature, topographic wetness index (TWI), mean annual air temperature (MAAT), mean annual precipitation (MAP), solar radiation, soil type maps, land use maps, and normalized difference vegetation index (NDVI).

The terrain attributes slope, aspect, profile curvature, plan curvature and TWI were derived from a digital elevation model (DEM) which was obtained from the CIAT (International Centre for Tropical Agriculture) SRTM (Spaceshuttle Radar Topographical Mission) website (http://srtm.csi.cgiar.org). The SRTM DEM data (Jarvis et al., 2008) were georeferenced from three arc second resolution to 30 m × 30 m resolution. Slope, aspect and SAGA TWI (System of Automated Geoscientific Analyses Topographic Wetness Index) were extracted in SAGA GIS (SAGA Development Team, 2008). SAGA TWI was calculated based on the following equation (Moore et al., 1993):

$$\mathrm{TWI} = \ln\left(\frac{\alpha}{\tan\beta}\right) \quad (1)$$

where α is the accumulative upslope area per unit contour length (or specific catchment area) computed with the D8 algorithm (O'Callaghan and Mark, 1984), and β the local slope gradient. SAGA TWI will assign a more realistic, higher potential soil wetness than the TWI_{D8} to grid cells situated in valley floors with a small vertical distance to a channel (Boehner et al., 2002). The cosine function was selected to transform aspect data to the range from −1 to 1 indicating the degrees of north. Potential insolation (incoming solar radiation), expressed in kW m^{-2}, mainly depends on elevation, slope and aspect, and thus can be derived directly from DEMs. Average annual potential insolation was calculated for the study area by means of the Solar Radiation function in ArcGIS 10.0 (ESRI, 2012). One year instead of a long-term average potential insolation was calculated, as the potential insolation only changes based on the scales of obliquity (Kunkel et al., 2011).

Beijing-1 multispectral data were obtained from Watershed Allied Telemetry Experimental Research (Li et al., 2009) at 32 m spatial

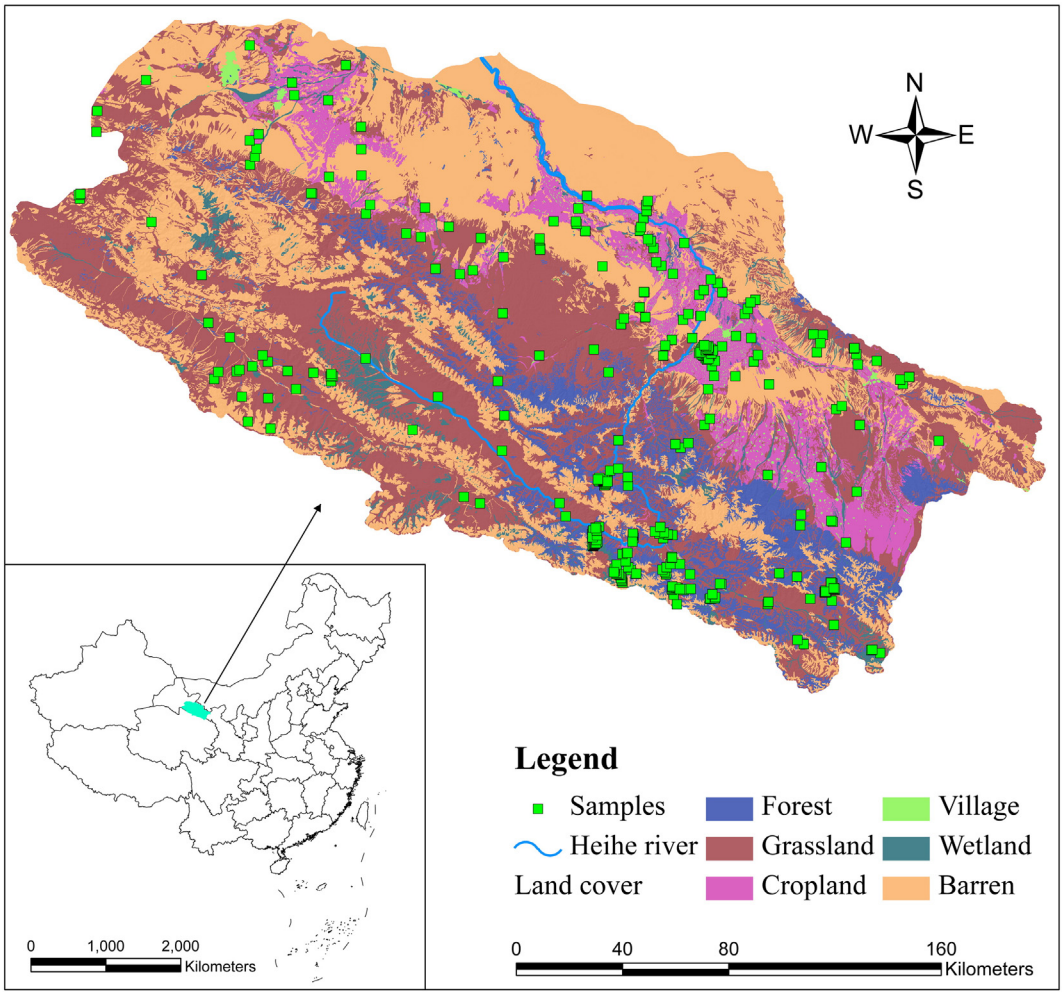

Fig. 1. Land use and sampling sites in the Heihe River Basin, China.

resolution in the year 2008. The NDVI was computed in the ENVI software environment described as:

$$\mathrm{NDVI} = \frac{(NIR - red)}{(NIR + red)} \quad (2)$$

where *NIR* is the near-infrared band and *red* the red band (Rouse et al., 1973).

Mean annual air temperature (MAAT) and mean annual precipitation (MAP) data were extracted from the database of the National Meteorological Information Center, China Meteorological Administration (CMA, 2011). Land cover and soil type data were acquired from Data Center for Resources and Environmental Sciences Chinese Academy of Sciences. Vector layers were converted to raster format, followed by a resampling of the entire variables to 30 m resolution.

All quantitative predictors were scaled, so that the mean values and standard deviations of the scaled predictors were 0 and 1, respectively. The absolute values of the regression coefficients of the scaled predictors reflect the importance of the predictors.

Based on the results obtained by ANOVA and Duncan's method for multiple comparison (Duncan, 1955) the number of land cover/land use categories was reduced to three: (1) cropland, village and barren, (2) forest, and (3) grassland and wetland. There were 14 soil types in the study area. The soil type is based on the Soil Map of the World (FAO90). Leptosols covered the largest area (39.7%). Eight primary soil types covered 89.5% of the study area, while the remaining 6 types of

Table 1
Statistical summary of the soil organic carbon and covariates.

	Minimum	Maximum	Mean	Median	Standard deviation	Skewness	Kurtosis
SOC (g/kg)	0.70	132.19	31.33	13.63	33.88	1.50	1.20
LnSOC (g/kg)	−0.36	4.88	2.90	2.61	1.05	0.15	−0.60
Elevation (m)	1351.00	4600.00	2417.14	2046.00	947.83	0.39	−1.41
Slp (%)	0.20	102.52	14.68	4.63	17.78	1.31	1.14
Cos(Asp)	−1.00	1.00	0.07	0.14	0.68	−0.18	−1.42
TWI	3.61	10.67	6.95	7.19	1.82	−0.27	−1.33
MAAT (°C)	−2.60	7.50	4.06	5.67	3.00	−0.48	−1.28
MAP (mm)	87.75	422.15	247.85	196.42	114.47	0.24	−1.67
Solar (10^3 kW m^{-2})	1.11	2.08	1.61	1.54	1.69	0.55	0.72
NDVI	−0.18	0.43	0.05	0.05	0.12	0.47	−0.12

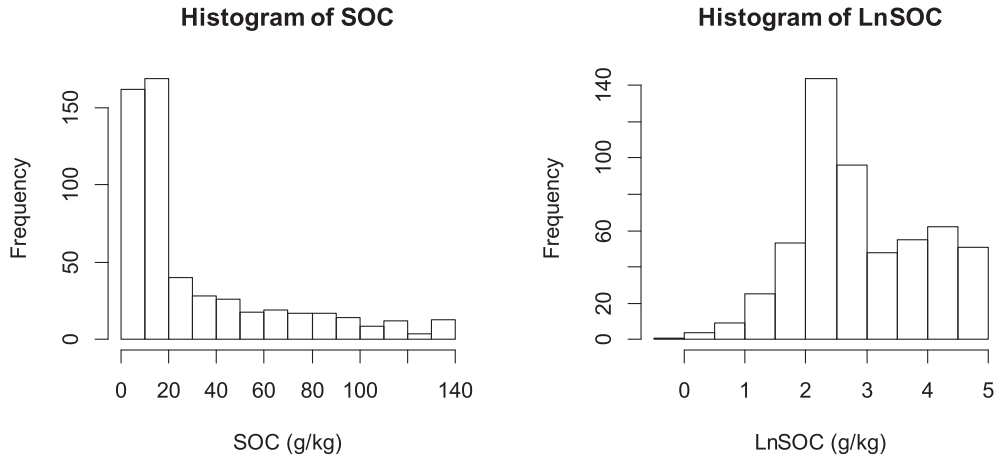

Fig. 2. Frequency distribution of raw SOC and LnSOC (transformed by natural logarithms).

soil covered 10.5% of the study area. Using again ANOVA and Duncan's method, the soil types were grouped into three classes: (1) Arenosols, Calcisols, Gypsisols, Phaeozems and Solonchaks, (2) Anthrosols, Fluvisols, Gleysols, Greyzems, and Kastanozems and (3) the others.

LnSOC showed the strongest correlation with elevation, MAP, MAAT, TWI and slope (Table 2). The strong positive correlation between LnSOC and elevation (correlation coefficient: 0.671) can be explained by the relation between elevation and both abiotic and biotic factors that directly affect the SOC content in the topsoil, such as MAP and MAAT. Also, elevation is associated with land cover. In upland areas land cover was mainly grassland and forest, whereas in lower regions this was mainly cropland and barren. Mutual correlation between predictors was strong for quite a few pairs. This collinearity of predictors may cause problems in fitting global and local regression models.

3. Regression modeling

In this section five prediction methods for LnSOC are described. Table 3 summarizes the modeling assumptions underlying these prediction methods.

3.1. MLR

In MLR the soil property of interest is modeled as a linear combination of predictors. It is assumed that the model residuals are independent and that the regression coefficients are global (Table 3). This implies that the soil property of interest is predicted by the simple formula:

$$\hat{y}(i) = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_k(i) \tag{3}$$

where $\hat{y}(i)$ is the predicted soil property at location i, $\hat{\beta}_0$ the estimated intercept, $\hat{\beta}_k$ the estimated regression coefficient for predictor k and $x_k(i)$ the value for the kth predictor at location i. The regression coefficients are estimated by ordinary least squares (OLS):

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

where T denotes the transposition of a matrix, $\mathbf{X}$ the design matrix formed by the values of predictors $x_k(i)$, and $\mathbf{y}$ the vector with measurements of the soil property of interest at the sampling locations. The relationship represented by Eq. (3) is assumed to be constant across the study area.

3.2. GWR

Unlike MLR, in GWR a local regression model is fitted at each prediction location. Similar to MLR in GWR it is assumed that the model-

Table 2
Pearson correlations between LnSOC and quantitative predictors.

	NDVI	Solar	MAP	MAAT	TWI	Cos(Asp)	Slope	Elevation
LnSOC	−0.123	0.176**	0.639**	−0.621**	−0.584**	0.069	0.582**	0.671**
Elevation	−0.359**	0.482**	0.883**	−0.886**	−0.828**	0.020	0.443**	1
Slope	−0.303**	0.047	0.697**	−0.621**	−0.878**	−0.016	1	
Cos(Asp)	0.050	0.059	0.035	−0.025	0.019	1		
TWI	0.353**	−0.264**	−0.772**	0.717**	1			
MAAT	0.166**	−0.423**	−0.914**	1				
MAP	−0.170**	0.384**	1					
Solar	−0.162**	1						
NDVI	1							

LnSOC: log-transformed SOC; Cos(Asp): cosine of aspect; TWI: system of Automated Geoscientific Analyses topographic wetness index; MAAT: mean annual air temperature; MAP: mean annual precipitation; Solar: incoming solar radiation; Soil: soil type indicator; Land: land use type indicator; NDVI: normalized difference vegetation index.

* Significant at the 0.05 level.

** Significant at the 0.01 level.

residuals are independent (Table 3). With GWR the soil property of interest is predicted by (Fotheringham et al., 2002):

$$\hat{y}(i) = \hat{\beta}_0(i) + \sum_{k=1}^{K} \hat{\beta}_k(i) x_k(i). \quad (5)$$

Note the argument (i) of the regression coefficients, which implies that they are a function of the prediction location i. The location-specific regression coefficients $\hat{\beta}_0(i)$ and $\hat{\beta}_k(i)$ $(k = 1...K)$ are estimated by weighted least squares, with weights related to the distance between the prediction location and the sampling locations. The closer a sampling location is to prediction location i, the larger its weight. The soil property of interest at a location i is predicted as follows: (i) draw a circle around prediction location i, (ii) weight a measurement of the soil property of interest at a sampling location in accordance with its proximity to prediction point i and (iii) estimate the regression coefficients by:

$$\hat{\boldsymbol{\beta}}(i) = \left(\mathbf{X}^T \mathbf{W}(i) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(i) \mathbf{y} \quad (6)$$

where $\mathbf{X}$ denotes the design matrix formed by the values of predictors $x_k(i)$ (all elements of the first column contain the value 1 for the intercept), and $\mathbf{W}(i)$ the diagonal weights matrix for prediction location i. The weights are determined by the weighting functions, also referred to as spatial kernels. The spatial kernel can be either fixed or adaptive. A fixed spatial kernel function keeps a one-sized kernel over the whole study area. An example is the Gaussian distance-decay function:

$$w_{ij} = \exp\left(-\frac{d_{ij}}{h}\right)^2 \quad (7)$$

where d_{ij} is the Euclidean distance between prediction location i and sampling location j, and h the kernel bandwidth. Note that the weights are nonzero for all data points, no matter how far they are from the center i (Fotheringham et al., 2002).

Conversely, the size of the adaptive kernel is defined by the number of the nearest neighborhoods. In areas with low sampling density the geometric size of the adaptive kernel is increased to capture enough observations. In a bi-square nearest neighborhood function the spatial weights (w_{ij}) are computed by:

$$w_{ij} = \begin{cases} \left[1 - (d_{ij}/h_i)^2\right]^2 & \text{if } d_{ij} < h_i \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where h_i is the nth nearest neighbor distance from prediction location i. When the distance from prediction location i to a sampling location is equal to or greater than the bandwidth h_i, the weight w_{ij} is zero. Previous studies showed that GWR is particularly sensitive to the representativeness and amount of sample data (Scull, 2010). In the case study hereafter, an adaptive spatial kernel was employed because the sampling density strongly varied throughout the study area. The optimal number of sampling locations in the search window, as obtained by cross-validation, was 79.

3.3. GWRR

It has been shown that in GWR the spatially varying coefficients are prone to collinearity even when there is no or weak collinearity among the predictors in the global regression model (Griffith, 2008; Wheeler, 2007; Wu and Zhang, 2013). This effect is caused by the weighting of the predictors. With increasing collinearity of predictors, the correlation of the regression coefficients will become stronger, leading to unreliable estimates of the local regression coefficients and unreliable model predictions. Therefore in GWR it is important to check for collinearity problems and to control the estimated regression coefficients.

Table 3
Assumptions in applied prediction methods.

Method	Regression coefficients	Model residuals
Multiple linear regression (MLR)	Global	Independent
Geographically weighted regression (GWR), geographically weighted ridge regression (GWRR)	Local	Independent
Kriging with an external drift (KED)	Global	Dependent
Simple kriging with GWR-derived local means (GWRSK)	Local	Dependent

As a remedy for collinearity of predictors leading to ill-conditioned design-matrices in multiple linear regression and unreliable estimates of regression coefficients Hoerl and Kennard (1970) proposed ridge regression. The ridge regression coefficient minimizes the residual sum of squares along with a penalty on the size of the squared coefficients as:

$$\hat{\boldsymbol{\beta}}^R = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{K} x_k(i) \beta_k \right)^2 + \lambda \sum_{k=1}^{K} \beta_k^2 \right\} \quad (9)$$

where λ is the ridge regression parameter that controls the amount of shrinkage in the regression coefficients and K the number of predictors. The constraint of the ridge regression can be explicitly defined as (Hastie et al., 2001):

$$\hat{\boldsymbol{\beta}}^R = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{K} x_k(i) \beta_k \right)^2, \quad \text{subject to } \sum_{k=1}^{K} \beta_k^2 \le s. \quad (10)$$

Note that the intercept is not constrained by the ridge parameter. The ridge regression coefficients are estimated by:

$$\hat{\boldsymbol{\beta}}^R = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^T \mathbf{y} \quad (11)$$

where $\mathbf{I}$ is the $K \times K$ identity matrix. In the above equations the ridge regression coefficients are global, but the same approach can be applied to constraint the local regression coefficients in GWR, leading to geographically weighted ridge regression (GWRR) (Brunsdon et al., 2012; Wheeler, 2007).

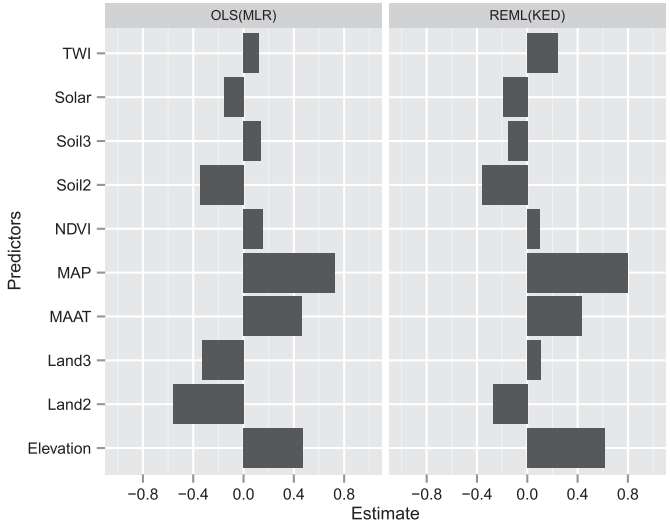

Fig. 3. The regression coefficients of multiple linear regression (MLR) estimated by ordinary least squares (OLS), and kriging with an external drift (KED) estimated by restricted maximum likelihood (REML).

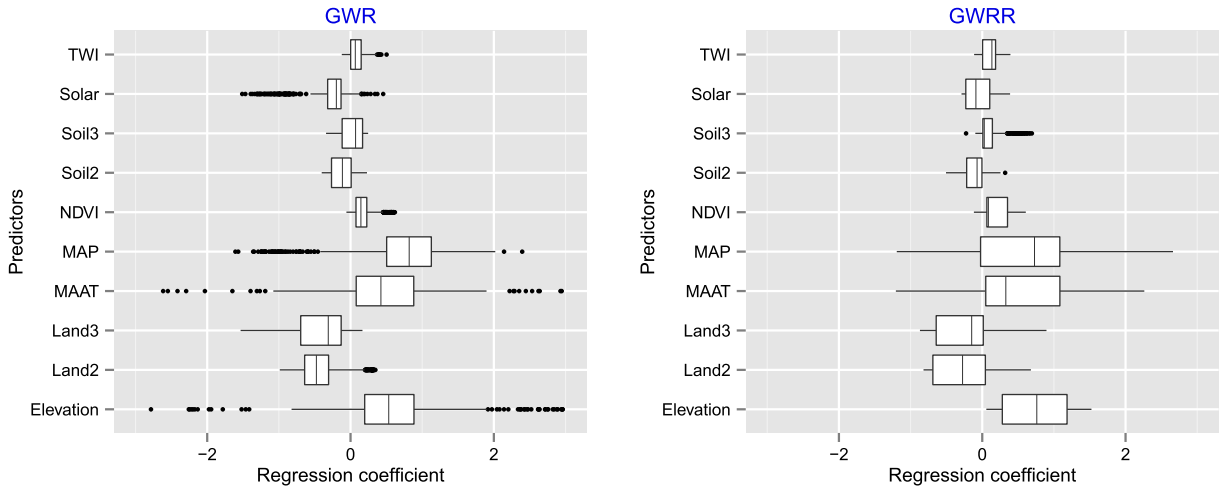

Fig. 4. The regression coefficients of geographical weighted regression (GWR) and geographical weighted ridge regression (GWRR).

For application of ridge regression in the GWR framework the intercept must be removed (Wheeler, 2007). In GWRR the intercept can be removed either globally or locally (Wheeler, 2006). With global centering the GWRR coefficients are estimated by:

$$\hat{\boldsymbol{\beta}}(i) = \left(\mathbf{X}^{*T}\mathbf{W}(i)\mathbf{X}^{*} + \lambda(i)\mathbf{I}\right)^{-1}\mathbf{X}^{*T}\mathbf{W}(i)\mathbf{y}^{*} \quad (12)$$

where $\mathbf{X}^*$ is the $n \times K$ matrix of standardized explanatory variables, and $\mathbf{y}^*$ the standardized response variable. If the local ridge parameter $\lambda(i)$ is 0, the estimated coefficients of GWR and GWRR are equal. We removed the intercept locally, as it introduces less bias in the coefficients than with the global centering. With local centering the predictors are first globally scaled, so that their means and variances are 0 and 1 respectively. Next for each local model the response variable is centered by subtracting local means, so that the means are 0 before estimating the ridge regression coefficients. These local means are computed as a weighted average, with weights equal to the square root of the kernel weights w_{ij}. The local regression coefficients are then estimated by:

$$\hat{\boldsymbol{\beta}}(i) = \left(\mathbf{X}_w^T\mathbf{X}_w + \lambda(i)\mathbf{I}\right)^{-1}\mathbf{X}_w^T\mathbf{y}_w \quad (13)$$

where $\mathbf{X}_w$ is the matrix of weighted, locally centered predictors, $\mathbf{y}_w$ the vector of weighted, locally centered responses, and $\mathbf{I}$ the identity matrix. The local ridge parameter $\lambda(i)$ is estimated by cross-validation, and so is the optimal value of the adaptive bandwidth, which was 244 in this case. After estimating the coefficients, the response variable predictions are calculated by adding the local mean $\bar{y}_w$ to $\hat{y}_w(i) = \mathbf{X}_w(i)\hat{\beta}(i)$.

3.4. KED

In multiple linear regression it is assumed that the model residuals are independent, so that the regression coefficients can be estimated by ordinary least squares. In kriging with an external drift the model residuals are assumed to be dependent. The covariance of the model-residuals at two locations was modeled as a function of the distance between two points, i.e., we assumed isotropy. The model parameters, i.e., the regression coefficients and the parameters of the covariance function (or semivariogram) were estimated by restricted maximum likelihood (REML) (Lark and Webster, 2006). With KED the soil property of interest at a location i is predicted by (Wackernagel, 2003):

$$\hat{y}_{\mathrm{KED}}(i) = \hat{m}_{\mathrm{KED}}(i) + \sum_{j=1}^{n}\eta_{\mathrm{KED}}(i,j)[y(j) - \hat{m}_{\mathrm{KED}}(j)] \quad (14)$$

where $\hat{m}_{\mathrm{KED}}(i)$ is the local mean estimated as a linear combination of predictors (computed with the REML estimates of the regression

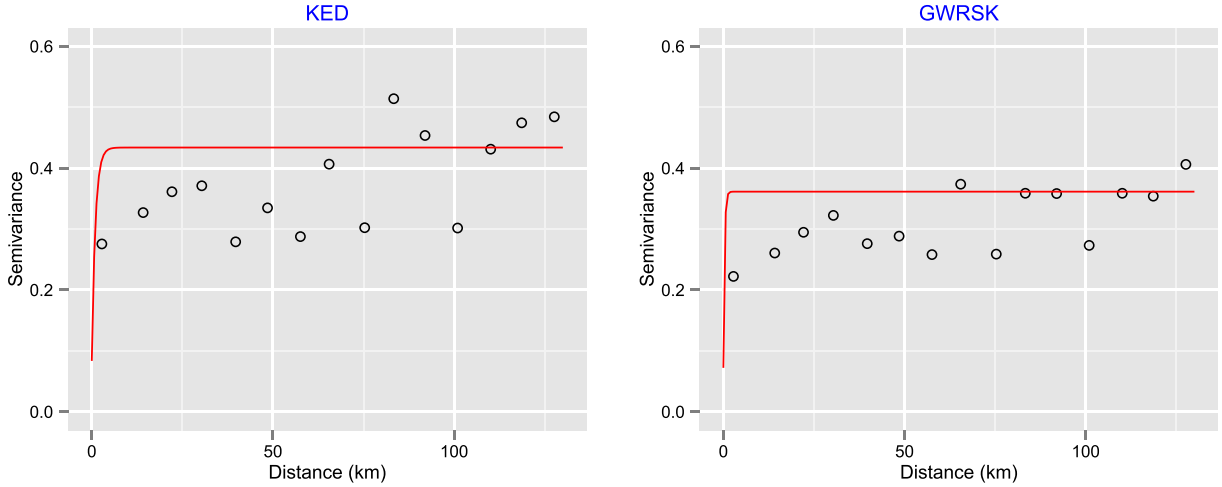

Fig. 5. Experimental residual variograms and models fitted by REML for KED (a) and GWRSK (b).

coefficients) and $\eta_{KED}(i,j)$ is the kriging weight attached to sampling location j. The kriging weights were obtained by solving the following system of equations:

$$\begin{cases} \sum_{j=1}^{n} \eta_{KED}(i,j) C_R(j,k) + \sum_{p=0}^{P} \mu_p(i) x_p(k) = C_R(i,k) \\ \sum_{j=1}^{n} \eta_{KED}(i,j) = 1 \\ \sum_{j=1}^{n} \eta_{KED}(i,j) x_p(k) = x_p(i) \end{cases} \quad (15)$$

where $C_R(k,j)$ is the spatial covariance of the residuals $R(.)$ ($R(.) = y(.) - \hat{m}_{KED}(.)$) at the sampling locations k and j, $C_R(i,k)$ the spatial covariance of $R(.)$ at prediction location i and sampling location k, and $\mu_p(i)$ Lagrange parameters. In prediction all sampling locations were used. This implies that we assumed that the regression coefficients were global (see Table 3). Note that if a local search window would have been used the regression coefficients would have been local, as in GWR and GWRSK. We used a global neighborhood so that by comparing the performance of GWR and KED we can be conclusive about whether local effects can best be modeled as fixed effects (as in GWR) or as a random effect (as in KED).

3.5. GWRSK

By relaxing the assumption of global regression coefficients as in MLR and KED with a global neighborhood, and the assumption of independent model residuals as in MLR, GWR and GWRR, our fifth prediction method is obtained, referred to as geographically weighted regression simple kriging (GWRSK). First the local regression coefficients at the sampling locations were estimated as described in Section 3.2 on GWR. Then the model residuals at these sampling points were estimated, followed by estimating a variogram of the model-residuals by restricted maximum likelihood. The soil property of interest was predicted by simple kriging with mean 0 of the GWR-model residuals, and adding the interpolated residual to the GWR-predicted value (Eq. 5):

$$\hat{y}_{GWRSK}(i) = \hat{y}_{GWR}(i) + \sum_{j=1}^{n} \eta_{SK}(i,j)[y(j) - \hat{y}_{GWR}(j)] \quad (16)$$

where $\hat{y}_{GWR}(i)$ is the trend as estimated by the GWR model and η_{SK} the simple kriging weights computed by:

$$\sum_{j=1}^{n} \eta_{SK}(i,j) C_R(j,k) = C_R(i,k) \quad \text{for } k = 1,\ 2,\ \ldots,\ n. \quad (17)$$

For all computations we used the statistical software R (version 3.1.1, http://cran.r-project.org/). For GWR we used R package spgwr (Fotheringham et al., 2002), for GWRR we used package GWmodel (Lu et al., 2014). For REML estimation of the variogram of KED and GWRSK we used package DEoptim (Ardia et al., 2011), for kriging we used package gstat (Pebesma, 2004).

3.6. Model selection

Predictors were selected by stepwise regression in both directions (forward and backward), using Akaike information criterion (AIC) as a selection criterion. In model selection we assumed that the model residuals were independent, i.e., the best subset of predictors was selected for the multiple linear regression (MLR) model. The same set of predictors was used in all other models, so that the estimated regression coefficients can be compared.

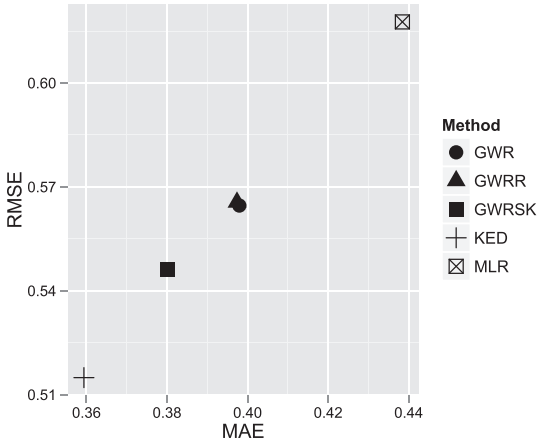

Fig. 6. Root mean square error (RMSE) vs. mean absolute error (MAE) for the five prediction methods as obtained by leave-one-out cross-validation.

All predictors were standardized. First their means were subtracted, then the centered values were divided by the standard deviations. In doing so the regression coefficients can be compared: the more extreme a coefficient is, either positive or negative, the stronger the effect of the associated predictor.

4. Cross-validation

Leave-one-out cross-validation (LOOCV) was performed to evaluate the prediction methods. Four quality indices were computed for model validation: mean error (ME), mean absolute error (MAE), root mean of squared error (RMSE), and ratio of performance to deviation (RPD) defined as:

$$RPD = \frac{STD}{RMSE} \quad (18)$$

where *STD* is the spatial standard deviation of SOC (g/kg). RPD quantifies the predictive ability of the model compared with using the sample average of SOC as a predictor. A prediction method with an RPD value smaller than 1.4, between 1.4 and 2.0, and larger than 2.0 is sometimes qualified as inaccurate, basically acceptable and accurate, respectively (Chang et al., 2001), although we realize that the prediction accuracy is largely determined by the residual variance, and that these boundary values are arbitrary.

To check whether we can use the paired *t*-test to test hypothesis on the difference in MSEs between methods, we computed Moran's I statistic standard deviate for paired differences in squared prediction errors (Moran, 1948). For all pairs of methods we rejected the null-hypothesis of no spatial autocorrelation of difference in squared prediction errors at a significance level of 0.01, i.e., a probability of rejecting the null-hypothesis of 1% when it is true. In other words the assumption of independent pairwise differences in the paired *t*-test is unrealistic. Spatial autocorrelation of the pairwise differences in a paired *t*-test

Table 4
Quality indices of predictions of natural log of soil organic carbon as obtained with the five methods.

Method	ME (g/kg)	MAE (g/kg)	RMSE (g/kg)	RPD	Correlation coefficient
MLR	0.00000035	0.438	0.618	1.704	0.80
GWR	0.029	0.398	0.565	1.864	0.84
GWRR	0.021	0.397	0.566	1.861	0.84
KED	0.017	0.359	0.515	2.044	0.87
GWRSK	0.037	0.380	0.546	1.928	0.85

Table 5
The *p*-values of two-sided paired *t*-tests of mean differences in squared prediction errors for six pairs of methods, as obtained by ignoring and accounting for spatial autocorrelation of differences.

Autocorrelation	MLR/GWR	KED/GWRSK	KED/GWR	KED/MLR	GWR/GWRR	GWR/GWRSK
Ignored	0.013	0.026	0.0032	0.00016	0.88	0.060
Accounted for	0.19	0.21	0.10	0.057	0.96	0.22
Effective sample size	152	171	166	140	60	230

can be accounted for by computing the effective sample size (Dale and Fortin, 2009):

$$n' = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{cor}\left(d_i^2, d_j^2\right)} \tag{19}$$

where $\mathrm{cor}(d_i^2, d_j^2)$ is the spatial autocorrelation of the difference in squared prediction errors obtained with two methods at point *i* and point *j*. The correlogram of the pairwise differences in squared prediction errors was estimated, as before, by REML using differential evolution as an optimization method.

The effective sample size was used to compute the *t*-statistic:

$$t = \frac{\overline{d} - \overline{d}_0}{\sqrt{\mathrm{Var}(d)/n'}} \tag{20}$$

where $\overline{d}_0$ is the mean of the difference in squared errors (difference in MSEs) under the null-hypothesis. We tested the null-hypothesis of zero difference in MSE ($\overline{d}_0 = 0$). The two-sided *p*-value of *t* was computed from a Student's *t* distribution with $n' - 1$ degrees of freedom.

5. Results

5.1. Model parameters

5.1.1. Regression coefficients

In stepwise regression all predictors except slope and aspect were selected. For most predictors the regression coefficients as estimated by OLS (MLR) and REML (KED) were very similar (Fig. 3). Exceptions were the coefficients for Soil3 and Land3. Estimated with REML, Land3 had a small positive effect, whereas estimated with OLS (MLR) this predictor had a somewhat stronger negative effect. Also for Soil3 the signs were opposite. The coefficients for MAP, elevation, Land2, and MAAT

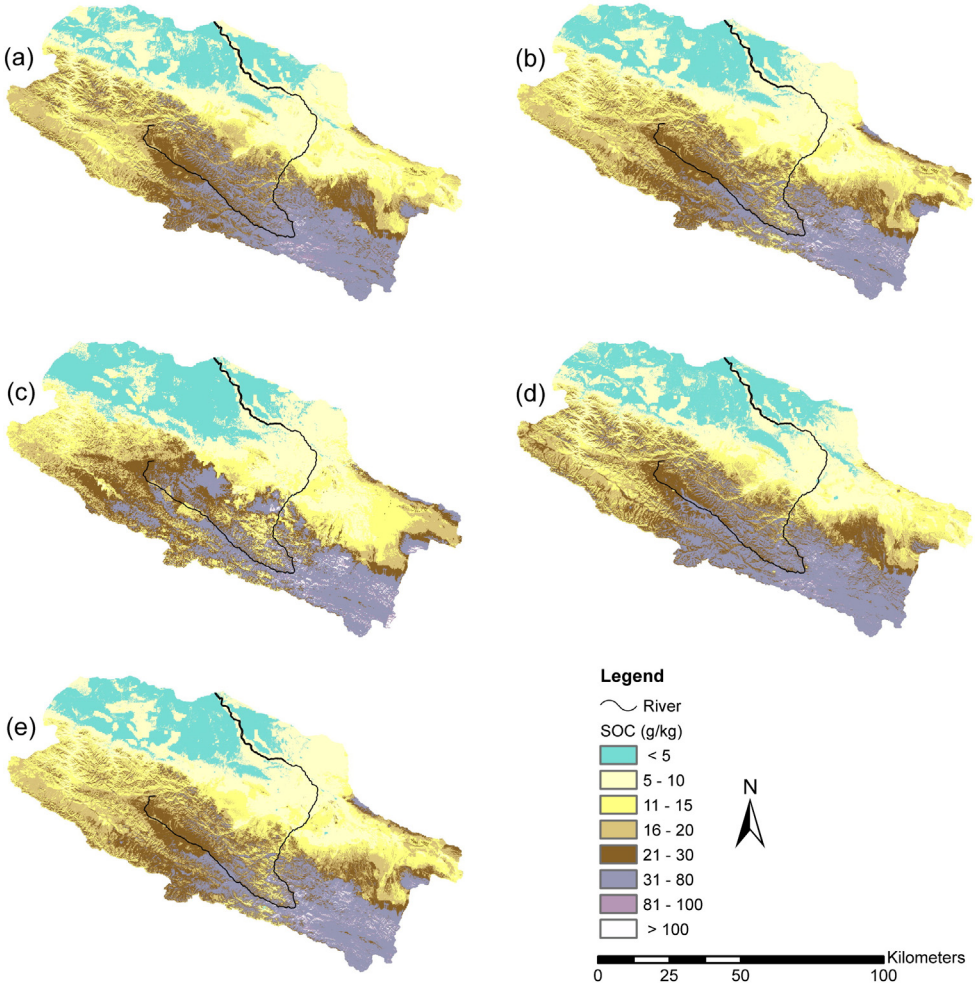

Fig. 7. Maps of soil organic carbon predicted by (a) MLR, (b) GWR, (c) GWRR, (d) KED and (e) GWRSK.

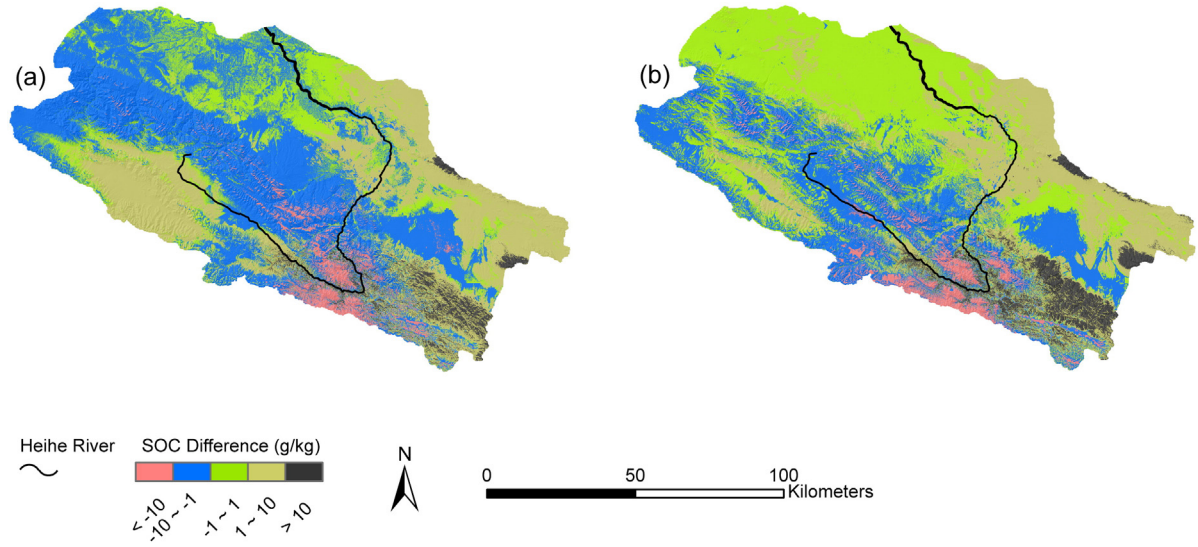

Fig. 8. Difference in predictions of SOC obtained with GWR and MLR ($\hat{y}_{GWR} - \hat{y}_{MLR}$) (a), and obtained with GWRSK and KED ($\hat{y}_{GWRSK} - \hat{y}_{KED}$) (b). Note: the values are draped over the DEM they are based on.

were most extreme with both estimation methods, and thus had the strongest effect on LnSOC (explained most of the spatial variation of LnSOC). These predictors also showed relatively strong correlation with LnSOC (Table 2).

The GWR coefficients showed substantial variation across the study area, especially those of MAAT, MAP, solar and elevation (Fig. 4). The effects of elevation, MAAT and MAP varied from strongly positive to strongly negative. The average values of the regression coefficients for elevation and NDVI were close to the values of the regression coefficients as estimated by OLS (MLR) and REML (KED). For all predictors the signs of the average regression coefficients were consistent with those obtained by OLS (MLR).

The GWRR coefficients showed less variation compared to the GWR coefficients (Fig. 4). The very extreme regression coefficients for elevation, MAAT and MAP as obtained with GWR were avoided. The averages of the GWRR coefficients were close to the GWR coefficients.

5.1.2. Variogram parameters

The fitted residual variograms of KED and GWRSK are shown in Fig. 5. Note that both fitted variograms were not fitted to the method-of-moments experimental variograms, but by REML. For KED an exponential model with a 0.08 nugget was fitted, and for GWRSK a spherical model with 0.07 nugget was fitted. We also fitted nested spherical + exponential models, but based on AIC these nested models were inferior. Both fitted variograms had a very short distance parameter of about 100 m. The nugget-to-sill ratio was 18.6% for KED and 19.4% for GWRSK. The variogram parameters indicated that there was not much spatial structure in the residuals of the spatial trend, neither for KED nor for GWRSK. The larger sill (nugget + partial sill) of the KED variogram compared to that of the GWRSK variogram showed that more variation was explained by the GWR model compared to the spatial trend part of the KED model with global regression coefficients.

5.2. Cross-validation results

Details of the performance of the five models are given in Fig. 6 and Table 4. For all five methods the cross-validation ME was very small. MAE and RMSE with GWR were smaller than with MLR, showing that when spatial dependence of model residuals is ignored, the accuracy of predictions can be increased by local, geographically weighted regression coefficients. However, when spatial dependence is incorporated in the model as in KED and GWRSK, the use of local, geographically weighted regression coefficients did not pay. KED outperformed all other methods, GWRSK included, in terms of RMSE and MAE. The RMSE and MAE of GWR were reduced by simple kriging of the GWR residuals as in GWRSK, but GWRSK performed not as good as KED. Surprisingly, GWRR failed to improve the performance of GWR. The cross-validation RMSE and MAE of GWRR were nearly equal to those of GWR.

The correlation between predicted SOC (as obtained on natural log-scale) and observed SOC ranged from 0.80 for MLR to 0.87 for KED (Table 4). Overall, all these methods showed acceptable performance; RPD values were larger than 1.4 for all methods.

Spatial autocorrelation of the pairwise differences in squared prediction errors at the sampling points caused the effective sample sizes to be much smaller than 548 (Dale and Fortin, 2009), the total number of sampling points (Table 5). As a consequence, the *p*-values of the *t* statistics increased considerably. Accounting for spatial autocorrelation, only the difference in MSEs between KED and MLR was significant at a significance level of 10%. The difference in MSEs between GWR and KED was very close to being significant at this significance level. All other tested pairs of methods had no significant difference in MSEs.

5.3. Maps of predicted soil organic carbon content

Maps of predicted SOC as obtained by the five methods are shown in Fig. 7. Note that all models were on the natural log scale. Here we back-transformed the predictions on the log-scale by exponentiation, so we did not add half the prediction error variance to the exponent (Webster and Oliver, 2007). If we assume a log-normal distribution for the model-residuals, then the back-transformed prediction thus obtained is the median of the distribution, not the mean.

Broadly speaking, the maps were very similar. All maps showed strong spatial variation of SOC in topsoil. All maps showed a clear northwest/southeast boundary dividing this area into two regions

Table 6
Summary statistics of the predicted soil organic carbon content.

Method	Minimum (g/kg)	Median (g/kg)	Maximum (g/kg)	Mean (g/kg)	Standard deviation (g/kg)	Coefficient of variation (%)
MLR	2.79	14.94	233.50	19.39	16.44	84.80
GWR	2.09	14.35	398.18	18.35	16.77	91.39
GWRR	1.21	14.59	377.79	18.76	19.34	103.09
KED	1.08	14.34	281.20	18.97	16.17	85.25
GWRSK	0.73	14.65	312.06	19.04	17.94	94.22

Table 7
Soil organic carbon content under different land covers using the KED method.

Land cover	Area (km^2)	Minimum (g/kg)	Maximum (g/kg)	Mean (g/kg)	Standard deviation (g/kg)	Coefficient of variation (%)
Forest	4864	2.15	229.31	37.78	18.03	47.72
Grassland	19,218	1.81	152.69	20.17	13.11	65.00
Cropland	5083	1.92	74.54	11.51	7.96	69.16
Village	686	2.94	64.89	10.28	8.25	80.25
Wetland	1829	2.36	104.19	26.1	16.23	62.18
Barren	19,130	0.94	191.24	13.27	13.89	104.67

with high and low SOC contents respectively. High predicted SOC values occurred in the southern part of the study area where elevation was high and the land was mainly covered by plateau grassland and forest. The high SOC contents in this part of the study area can be explained by the higher rainfall and lower mean temperature in the alpine environment. In the middle reaches with a low annual precipitation, relatively high predicted values of SOC occurred in the northeastern part where the typical ecosystem was human controlled agro-ecosystem and many artificial oases occurred, supplied with water from the Heihe River.

To highlight differences between the maps, we plotted a map showing differences in predicted SOC as obtained with GWR and MLR (GWR minus MLR), and similarly for GWRSK and KED methods (GWRSK minus KED) (Fig. 8). Absolute differences between GWR and MLR predictions were smaller than 2 g/kg for 47% of the pixels; for GWRSK-KED absolute differences these were 55% of the pixels. For an absolute difference of 20 g/kg these percentages were 2% and 3%, respectively. All the differences were relatively small in the northwestern part, where land was mainly covered by Gobi and SOC content in topsoil was low. These differences between GWR and MLR were smaller than between GWRSK and KED. The differences between GWRSK and KED were relatively large in the mountainous regions, despite that elevation was used as a predictor in both models, showing the complex spatial variation of SOC contents.

Summary statistics of the predicted SOC as obtained with the five methods are presented in Table 6, so as to make a further comparison. For all five methods the mean of the predicted SOC was smaller than the mean of the observed SOC value (compare Tables 1 and 6), whereas the median of the predicted SOC exceeded the observed median. A map of SOC contents produced by MLR showed the smallest range (2.79 g/kg to 233.5 g/kg), whereas this range on the GWR map was the largest (2.09 g/kg to 398.18 g/kg). Also for all five methods the maximum predicted SOC exceeded the observed maximum. For MLR the difference between predicted and observed maximum was the smallest, for GWR the largest. This is as expected as the use of global regression coefficients will lead to a stronger smoothing effect and can lead to exceptionally large predicted values.

Summary statistics of the SOC contents as predicted by the best-performing method KED under various land covers are shown in Table 7. The mean value of predicted SOC was the largest for forest, followed by wetland and grassland, all three land cover types mainly occurring in the alpine areas. High SOC contents could be explained by the denser vegetation coverage and mainly natural ecosystems, including cold desert, montane, alpine meadows and steppe. This distribution was also consistent with the spatial variation of SOC under different land covers (Fig. 1).

6. Discussion

The central research question was whether the quality of predictions can be improved by local regression coefficients as implemented in GWR and GWRSK instead of global regression coefficients as in MLR and KED with a global search neighborhood. Ignoring the spatial dependence of the model residuals, the use of local regression coefficients in GWR reduced the RMSE of MLR by 8.6%. However, by incorporating the spatial dependence of the model residuals into the model, the use of local geographically weighted regression coefficients did not pay. We found that KED outperformed GWR (RMSE of KED was 8.8% smaller) and GWRSK, although the difference in MSEs with GWRSK was not significant ($p = 0.21$) and with GWR only marginally significant ($p = 0.10$) in a two-sided paired t-test if we accounted for spatial autocorrelation of the pairwise differences in squared errors. This result shows that in this case local effects can better be modeled as a spatial random effect than as local fixed effects.

An alternative for GWR, suggested by one of the reviewers, is to account for spatially varying regression coefficients in a more standard regression approach. For instance, we may account for different regression coefficients for MAAT, MAP et cetera between land covers or soil types. This can be done either by adding interaction terms between the categorical predictors land covers/soil types and the other (quantitative) predictors (fixed effects only), or by adding a random effect for the other predictors with land covers/soil types as a blocking factor, leading to a linear mixed model. Accounting for different coefficients between land cover types for MAP and elevation, the residual variance was 0.602 for both the linear model with interaction (fixed effects only) and the linear mixed model. Comparing these residual variances with the cross-validation RMSE of GWR (Table 4) shows that these alternatives most likely will not perform as good as GWR.

The somewhat disappointing performance of GWR can possibly be explained by collinearity among the predictors. We checked for collinearity problems by computing variance inflation factors (VIF) (Wheeler and Tiefelsdorf, 2005).

The VIF of a predictor can be computed by regressing this predictor on all remaining predictors. The VIF is then computed as:

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2} \tag{21}$$

where R_i^2 is the coefficient of determination of predictor i. As a rule of thumb, a VIF > 10 indicates a collinearity problem. The VIF values for each predictor in global (MLR) and local (GWR) regressions are summarized in Table 8. In global regression elevation was strongly related to the remaining predictors, indicating a slight collinearity problem with MLR. This collinearity problem was not severe, as the standard error of

Table 8
Variance inflation factors for MLR model and GWR model.

	MAAT	MAP	Solar	Elevation	TWI	NDVI	Land2	Land3	Soil2	Soil3
MLR										
	8.787	9.174	1.484	11.614	4.050	1.538	3.398	3.398	3.431	3.431
GWR										
Minimum	1.365	1.293	1.155	1.914	1.273	1.199	1.231	1.403	1.146	1.179
25%	3.954	3.431	1.243	2.386	1.853	1.318	1.368	1.535	1.322	1.331
Median	18.019	14.699	1.748	5.695	2.279	1.482	3.781	3.872	1.726	1.388
Mean	18.052	17.388	2.348	7.322	2.552	1.674	8.305	8.550	1.718	1.654
75%	23.603	25.305	2.474	9.499	3.053	1.714	7.030	7.245	2.097	1.673
Maximum	190.983	158.198	8.014	49.555	7.951	3.136	106.096	107.256	2.433	4.891

the estimated regression coefficient for elevation was reasonable. With local regression the mean VIFs of MAAT and MAP were much larger than 10 indicating somewhat stronger collinearity problems in GWR compared to MLR. For five predictors the maximum VIF values were extremely large, much larger than 10, however the 75th percentile of the distributions of VIF exceeded 10 for only MAAT and MAP. But these maximum values were outliers as the 75th third quartiles seemed to be outliers whereas only two of them had VIFs larger than 10 at 75th percentiles (MAAT and MAP).

Collinearity of predictors may lead to unrealistic estimates of regression coefficients (Wheeler, 2006), and as a consequence possibly unrealistic predictions with large errors. The use of the local ridge parameter in GWRR may have avoided unrealistic estimates of the regression coefficients (Fig. 4), but does not lead to an increase of accuracy compared to GWR.

We used the same data set for calibration of the various models and for validation by LOOCV. This is not ideal, as calibration and validation ask for different sampling designs (Brus et al., 2011). Preferably, for validation additional locations are selected by probability sampling. This probability sample is neither used for calibration nor for prediction. With probability samples model-free, there is valid estimation of the quality indices and their variances. Here such validation data set was not available. Common practice is to report simply the cross-validation MSE, without testing the statistical significance of the differences in MSE between methods. We attempted to improve on this practice by computing effective sample sizes which account for spatial autocorrelation of the pairwise differences of squared errors. These effective sample sizes were substantially smaller than the total number of pairwise difference values (number of sampling locations), so that most differences in MSE became insignificant at a significance level of 10% in a two-sided paired *t*-test. The effective sample sizes were based on a calibrated model of the spatial autocorrelation. This makes our conclusions on the significance of differences in MSE model-based, and as a consequence to some extent debatable.

In this study we did not validate the prediction error variances as computed by the models, for instance by computing means of standardized squared errors. R packages spgwr for GWR and GWmodel for GWRR do not provide these variances. Validation of these variances is recommended for future research.

7. Conclusions

1. Fitting regression coefficients locally as in GWR only paid when no spatial random effect was included in the model.
2. When a spatial random effect was included in the model, the flexibility of local, geographically weighted regression coefficients was not needed and even undesirable as it led to less accurate predictions than KED with global regression coefficients.
3. In comparing the accuracy of prediction methods by leave-one-out cross-validation of a non-probability sample it is important to account for possible autocorrelation of pairwise differences in the prediction errors.
4. In this case the effective sample sizes were substantially smaller than the total number of sampling points, so that most pairwise differences in MSE were not significant at a significance level of 10% in a two-sided paired *t*-test.

Acknowledgements

The study was supported financially by the National Natural Science Foundation of China (grant No. 41130530, No. 91325301, No. 41401237, and No. 41201207) and partly by the Jiangsu Province Science Foundation for Youths (No. BK20141053). Acknowledgements are also extended to Dr. Chaosheng Zhang, National University of Ireland, Ireland, for his technical help during the soil mapping. We kindly acknowledge the three anonymous reviewers for their valuable comments on a first version of this paper, which have greatly improved the paper.

References

Ardia, D., Boudt, K., Carl, P., Mullen, K.M., Peterson, B.G., 2011. Differential evolution with DEoptim: an application to non-convex portfolio optimization. R J. 3 (1), 27–34.
Boehner, J., Koethe, R., Conrad, O., Gross, J., Ringeler, A., Selige, T., 2002. Soil regionalisation by means of terrain analysis and process parameterisation. In: Micheli, E., Nachtergaele, F., Montanarella, L. (Eds.), Soil Classification 2001. European Soil Bureau, Research Report No. 7, EUR 20398 EN, Luxembourg, pp. 213–222.
Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. R. Stat. Soc. Ser. D (Stat.) 47, 431–443.
Brunsdon, C., Charlton, M., Harris, P., 2012. Living with collinearity in local regression models. Accuracy 2012 — 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 10th–13th July, 2012, Florianópolis, SC, Brazil.
Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407.
Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R.J., 2001. Near-infrared reflectance spectroscopy—principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65 (2), 480–490.
CMA (China Meteorological Administration), 2011. China Meteorological Data Daily Value. China Meteorological Data Sharing Service System, Beijing, China (http://cdc.cma.gov.cn/index.jsp).
Dale, M.R.T., Fortin, M.J., 2009. Spatial autocorrelation and statistical tests: some solutions. J. Agric. Biol. Environ. Stat. 14 (2), 188–206.
Duncan, D.B., 1955. Multiple range and multiple F tests. Biometrics 11 (1), 1–42.
ESRI, 2012. ArcGIS Desktop: Release 10. Environmental Systems Research Institute, Redlands, CA.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.
Griffith, D., 2008. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). Environ. Plan. A 40 (11), 2751–2769.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forest analysis. Geoderma 146 (1–2), 102–113.
Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152 (3–4), 195–207.
Harris, P., Juggins, S., 2011. Estimating freshwater acidification critical load exceedance data for Great Britain using space-varying relationship models. Math. Geosci. 43 (3), 265–292.
Harris, P., Fotheringham, A.S., Crespo, R., Charlton, M., 2010. The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. Math. Geosci. 42 (6), 657–680.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12 (1), 55–67.
Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for globe version 4 Available from CGIAR-CSI SRTM 90m database http://srtm.csi.cgiar.org.
Kang, S., Su, X., Tong, L., Shi, P., Yang, X., Abe, Y.K., Du, T., Shen, Q., Zhang, J., 2004. The impacts of human activities on the water–land environment of the Shiyang River basin, an arid region in northwest China. Hydrol. Sci. J. 49 (3), 413–427.
Kheir, R.B., Greve, M.H., Bøcher, P.K., Greve, M.B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: the case study of Denmark. J. Environ. Manag. 91 (5), 1150–1160.
Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189–190, 627–634.
Kunkel, M.L., Flores, A.N., Smith, T.J., McNamara, J.P., Benner, S.G., 2011. A simplified approach for estimating soil carbon and nitrogen stocks in semi-arid complex terrain. Geoderma 165 (1), 1–11.
Lark, R.M., Webster, R., 2006. Geostatistical mapping of geomorphic variables in the presence of trend. Earth Surf. Process. Landf. 31, 862–874.
Li, X., Li, X., Li, Z., Ma, M., Wang, J., Xiao, Q., Liu, Q., Che, T., Chen, E., Yan, G., Hu, Z., Zhang, L., Chu, R., Su, P., Liu, Q., Liu, S., Wang, J., Niu, Z., Chen, Y., Jin, R., Wang, W., Ran, Y., Xin, X., Ren, H., 2009. Watershed allied telemetry experimental research. J. Geophys. Res. 114 (D22), 1–19.
Li, X., Lu, L., Yang, W., Cheng, G., 2012. Estimation of evapotranspiration in an arid region by remote sensing—a case study in the middle reaches of the Heihe River Basin. Int. J. Appl. Earth Obs. Geoinf. 17, 85–93.
Li, Q., Yue, T., Wang, C., Zhang, W., Yu, Y., Li, B., Yang, J., Bai, G., 2013. Spatially distributed modeling of soil organic matter across China: an application of artificial neural network approach. Catena 104, 210–218.
Lloyd, C.D., 2010. Nonstationary models for exploring and mapping monthly precipitation in the United Kingdom. Int. J. Climatol. 30 (3), 390–405.
Lu, L., Li, X., Veroustraete, F., Kang, E., Wang, J., 2009. Analysing the forcing mechanisms for net primary productivity changes in the Heihe River Basin, north-west China. Int. J. Remote Sens. 30 (3), 793–816.
Lu, B., Charlton, M., Harris, P., Fotheringham, A.S., 2014. Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. Int. J. Geogr. Inf. Sci. 28 (4), 660–681.
Meersmans, J., De Ridder, F., Canters, F., De Baets, S., Van Molle, M., 2008. A multiple regression approach to assess the spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 143 (1–2), 1–13.

Mishra, U., Lal, R., Liu, D., Van Meirvenne, M., 2010. Predicting the spatial variation of the soil organic carbon pool at a regional scale. Soil Sci. Soc. Am. J. 74 (3), 906–914.
Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57 (2), 443–452.
Moran, P.A.P., 1948. The interpretation of statistical maps. J. R. Stat. Soc. Ser. B Methodol. 10 (2), 243–251.
O'Callaghan, J.F., Mark, D.M., 1984. The extraction of drainage networks from digital elevation data. Comp. Vision Graph. Image Process. 28 (3), 323–344.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691.
Piccini, C., Marchetti, A., Francaviglia, R., 2014. Estimation of soil organic matter by geostatistical methods: use of auxiliary information in agricultural and environmental assessment. Ecol. Indic. 36, 301–314.
Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS. In: Freden, S.C., Mercanti, E.P., Becker, M. (Eds.), Third Earth Resources Technology Satellite—1 Symposium—Technical Presentations, Section A 1. Goddart Space Flight Center, National Aeronautics and Space Administration, Washington, DC, pp. 309–317.
SAGA Development Team, 2008. System for Automated Geoscientific Analyses (SAGA GIS). Germany. URL http://www.saga-gis.org/.
Scull, P., 2010. A top-down approach to the state factor paradigm for use in macroscale soil analysis. Ann. Assoc. Am. Geogr. 100 (1), 1–12.
Tu, J., 2011. Spatially varying relationships between land use and water quality across an urbanization gradient explored by geographically weighted regression. Appl. Geogr. 31 (1), 376–392.
Vasques, G.M., Grunwald, S., Sickman, J.O., 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146 (1–2), 14–25.
Vasques, G.M., Grunwald, S., Sickman, J.O., 2009. Modeling of soil organic carbon fractions using visible/near-infrared spectroscopy. Soil Sci. Soc. Am. J. 73 (1), 176–184.
Wackernagel, H., 2003. Multivariate Geostatistics. 3rd completely revised ed. Springer, Berlin.
Wang, K., Zhang, C., Li, W., 2013. Predictive mapping of soil total nitrogen at a regional scale: a comparison between geographically weighted regression and cokriging. Appl. Geogr. 42, 73–85.
Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. Wiley, Chichester.
Wheeler, D.C., 2006. Diagnostic Tools and Remedial Methods for Collinearity in Linear Regression Models with Spatially Varying Coefficients. The Ohio State University, pp. 65–97.
Wheeler, D.C., 2007. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. Environ. Plan. A 39 (10), 2464–2481.
Wheeler, D.C., Tiefelsdorf, M., 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J. Geogr. Syst. 7 (2), 161–187.
Wu, J., 2011. The effect of ecological management in the upper reaches of Heihe River. Acta Ecol. Sin. 31 (1), 1–7.
Wu, W., Zhang, L., 2013. Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. Appl. Geogr. 37, 52–62.
Zhang, C., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. 26 (7), 1239–1248.
Zhu, A.X., Yang, L., Li, B., Qin, C., English, E., Burt, J.E., Zhou, C.H., 2008. Purposive sampling for digital soil mapping for areas with limited data. In: Hartemink, A.E., McBratney, A.B., Mendonca Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer-Verlag, New York, pp. 233–245.