Natural Resources Science Faculty Publications          Natural Resources Science

2019

# A Comparative Assessment of Geostatistical, Machine Learning, and Hybrid Approaches for Mapping Topsoil Organic Carbon Content

Lin Chen

Chunying Ren

Lin Li

Yeqiao Wang
*University of Rhode Island*, yqwang@uri.edu

Bai Zhang

*See next page for additional authors*

Follow this and additional works at: https://digitalcommons.uri.edu/nrs_facpubs

Authors

Lin Chen, Chunying Ren, Lin Li, Yeqiao Wang, Bai Zhang, Zongming Wang, and Linfeng Li

*Article*

# A Comparative Assessment of Geostatistical, Machine Learning, and Hybrid Approaches for Mapping Topsoil Organic Carbon Content

**Lin Chen** [1,2,3] ![ID], **Chunying Ren** [1,*], **Lin Li** [4], **Yeqiao Wang** [3], **Bai Zhang** [1], **Zongming Wang** [1] and **Linfeng Li** [5]

1   Key Laboratory of Wetland Ecology and Environment, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China; chenlin@iga.ac.cn (L.C.); zhangbai@iga.ac.cn (B.Z.); zongmingwang@iga.ac.cn (Z.W.)
2   University of Chinese Academy of Sciences, Beijing 100049, China
3   Department of Natural Resources Science, University of Rhode Island, Kingston, RI 02881, USA; yqwang@uri.edu
4   Department of Earth Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA; ll3@iupui.edu
5   Nong'an Senior High School, Changchun 130200, China; 13634347821@163.com
*   Correspondence: renchy@iga.ac.cn; Tel.: +86-431-8554-2297

![check for updates]

**Abstract:** Accurate digital soil mapping (DSM) of soil organic carbon (SOC) is still a challenging subject because of its spatial variability and dependency. This study is aimed at comparing six typical methods in three types of DSM techniques for SOC mapping in an area surrounding Changchun in Northeast China. The methods include ordinary kriging (OK) and geographically weighted regression (GWR) from geostatistics, support vector machines for regression (SVR) and artificial neural networks (ANN) from machine learning, and geographically weighted regression kriging (GWRK) and artificial neural networks kriging (ANNK) from hybrid approaches. The hybrid approaches, in particular, integrated the GWR from geostatistics and ANN from machine learning with the estimation of residuals by ordinary kriging, respectively. Environmental variables, including soil properties, climatic, topographic, and remote sensing data, were used for modeling. The mapping results of SOC content from different models were validated by independent testing data based on values of the mean error, root mean squared error and coefficient of determination. The prediction maps depicted spatial variation and patterns of SOC content of the study area. The results showed the accuracy ranking of the compared methods in decreasing order was ANNK, SVR, ANN, GWRK, OK, and GWR. Two-step hybrid approaches performed better than the corresponding individual models, and non-linear models performed better than the linear models. When considering the uncertainty and efficiency, ML and two-step approach are more suitable than geostatistics in regional landscapes with the high heterogeneity. The study concludes that ANNK is a promising approach for mapping SOC content at a local scale.

**Keywords:** soil organic carbon mapping; methods comparison; hybrid approaches; machine learning; geographically weighted regression

## 1. Introduction

Soil organic carbon (SOC) is the organic carbon component of soil, consisting of living soil biota and dead biotic material derived from biomass. SOC tends to be concentrated in the topsoil with the primary source from vegetation [1]. Globally, the soil biotic C pool, second only to the oceanic, is of

importance for carbon sequestration [2]. Small changes in the SOC could result in significant impacts on the atmospheric carbon concentration [3]. This makes SOC an important ecological indicator of the greenhouse effect, as well as a major global change driver, due to its high sensitivity to human disturbance [4,5]. Soils thus should be managed to sequester more organic carbon and discharge fewer greenhouse gases by maintaining sustainable land use [6,7]. As a complex mixture, SOC is also involved in soil quality and ecosystem services like food production. It plays a significant role in supplying nutrients and the formation of improving soil structure [8]. Accurate SOC content mapping is critically important for monitoring the baseline of the carbon pool, as well as its role in climate change and food security [9,10].

Geographic information systems (GIS) and remote sensing (RS) have been practiced acquiring spatial distribution and pattern of SOC content at different scales. Digital soil mapping (DSM) is to develop a geographically referenced database by integration of field observations with environmental variables through quantitative relationships [11]. Nonetheless, SOC content shows strong spatial heterogeneity and dependence because of influences of natural factors and human disturbances vary by locations and scales [12,13]. Climate and relief affect surface runoff and transport of soil along the surface, modifying the spatial distribution of SOC [14,15]. Land use, fertilizer application and agricultural production also play important roles in influencing the SOC dynamics [16,17]. Conventionally, DSM is conducted using environmental variables, including soil properties, climatic, relief, and spectral indices of remote sensing data. Due to the above-mentioned complexity and uncertainty, mapping SOC is still a challenging subject [18].

Generally, the main types of DSM techniques include traditional statistical, geostatistical, machine learning (ML) and hybrid approaches. Traditional statistical methods are used to determine the correlation between environmental variables and SOC content [19]. Commonly used non-spatial models include multiple linear regression [20], partial least square regression [21], generalized linear models [22] and linear mixed models [23]. Although easily applicable, their prerequisites of independent and identical distribution with large sample demands for SOC content observations are among challenges. Those methods are also known as lack of spatial information, making them less stable and unsuitable for delineating local changes [24].

In geostatistical techniques, such as ordinary kriging (OK) and geographically weighted regression (GWR), spatial autocorrelation or mutual correlation is considered [25,26]. OK, as a kriging-based on the assumption of unknown mean, is one of the commonly used methods for spatial interpolation [27–29]. However, this method overlooks the influence of environmental variables [30], and is not able to achieve the desired accuracy, due to the stationarity assumption [31]. GWR is based on the local smooth idea and relationships among different environmental variables within a local space. It is capable of embedding the spatial location of samples into a regression through locally weighted least square method [25,32]. Scull (2010) found that GWR performed better than multiple linear regression for prediction of SOC with precipitation or temperature as an environmental variable [33]. Using simulated data sets, Harris et al. (2010) compared spatial prediction performances of GWR and OK methods [34]. The comparison concluded that GWR provided extra information on the spatial processes, but the spatial dependence of residuals still existed. Kriging methods are typically used to interpolate geographical characteristics with significant spatial autocorrelation, such as climatic factors, natural soil properties, and geological elements [12]. Whereas GWR is proposed to predict highly random factors with a significant degree of spatial heterogeneity, such as socio-economic indices.

ML methods can accommodate non-linearity and multicollinearity, and they can overcome overfitting with limited soil observations and auxiliary environmental information [35,36]. They are typified by support vector machines and artificial neural networks (ANN), which are widely applied in DSM [37–39]. Support vector machines can construct an optimal hyperplane by projecting the data onto a new hyperspace by the means of kernel functions. The hyerplane separates classes and creates the widest margin in the classification. It also can fit data (support vector machines for regression, SVR) with a modeling function that minimizes empirical risk and complexity when representing

non-linear patterns [40]. ANN simulate the human learning processes. The linkages between the input and output data in ANN are established and reinforced by non-linear and interconnected nodes [41]. Li et al. (2017) used SVR and ANN for the prediction of *L. chinensis* carbon, nitrogen and phosphorus contents [42]. Both SVR and ANN demonstrated high prediction accuracy. Xu et al. (2018) reported that SVR outperformed ANN for mapping soil organic matter with a visible and near infrared spectral dataset [43]. The ML methods do not need explicit assumptions about data distributions, and they allows for modelling of complex relationships [44]. However, their major shortcoming is that SOC content at a particular grid is estimated only from environmental variables of that node, without considering its spatial autocorrelation at the node with surrounding data [45].

Recently, hybrid approaches have drawn attentions. Two hybrid methods are geographically weighted regression kriging (GWRK) and artificial neural networks kriging (ANNK). GWRK and ANNK can incorporate the spatial autocorrelation of measured variables to achieve better predictions and lower errors [46,47]. Both GWRK and ANNK can account for the spatial parametric non-stationarity and the relationship between SOC and other environmental variables. Additionally, residuals can be interpolated through kriging and they are added to the estimated trend as a spatially stochastic variable. Guo et al. (2017) reported that GWRK outperformed partial least squares regression kriging for predicting soil organic matter with visible and near-infrared spectra [48]. Mirzaee et al. (2016) evaluated the ability for geostatistical methods and ANNK to predict soil organic matter [47]. Despite a variety of DSM methods have been used in mapping SOC [49], there is a lack of systematic comparison among geostatistical, machine learning and hybrid approaches for mapping SOC.

With the above, the aim of this study is to perform such a methodological comparison based on soil samples and environmental variables derived from remote sensing data. In particular, this study compared methods of two geostatistical (OK and GWR), two machine learning (SVR and ANN), and two hybrid approaches (GWRK and ANNK) for their performance in SOC mapping. The comparison of intra and extra classes of three DSM techniques and their suitable situations for the application were also discussed.

## 2. Materials and Methods

### 2.1. Materials

#### 2.1.1. Study Area

This study was conducted in the hinterland of the Northeast Plain of northeastern China (43° 15′ N–44° 05′ N, 125° 03′ E–126° 01′ E, Figure 1). The total area was about 3592 km$^2$. The study area, including Changchun the capital city of Jilin Province, has a total urban population of approximately 5.20 million in 2016. As a main grain-production area in China, the fertile lands are prevalent and are vital for the country's food security. This area has a temperate continental monsoon climate. The mean annual temperature is 4.8 °C, with mean temperatures of −15.1 °C in January and a mean temperature of 23.1 °C in July, respectively. The annual precipitation ranges from 522 to 615 mm, and more than 60% of which occurs from June to August. The dominant soil types are black soil, dark-brown soil and chernozem.
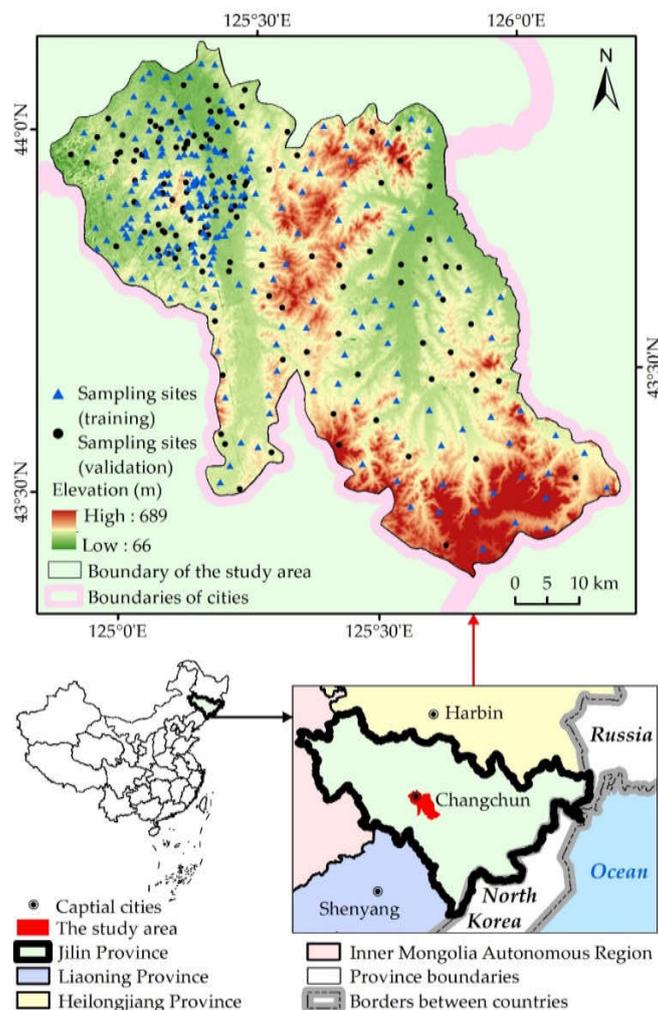
**Figure 1.** Location of study area in China and soil sampling sites.

### 2.1.2. Soil Observations

The field campaign was conducted from September to October 2016. The distribution of sampling points was generated with water bodies and impervious surfaces being masked out. The 30 × 30 m plot was applied at each sampling site. The location, and land use and land cover classification of the sampling sites were recorded. Among the 395 sampling sites, 156 were located in human-managed farmland, 232 in forest and 7 in grassland (Figure 1). The sampling sites were randomly divided into training (n = 263) for modeling and validation sets (n = 132) (Figure 1) to evaluate performances of models in comparison [50–52]. At each sampling site, SOC content and bulk density (BD, g·cm$^{-3}$) were collected in topsoil throughout a depth range of 0–20 cm in a radial sampling scheme using an auger [53]. A core ring sampler (5 cm in diameter, 5 cm in height) was used to collect samples at the center of each plot for BD determination [54]. The soil samples for measuring SOC content (g·kg$^{-1}$) were air-dried, grounded, and passed through a 2 mm mesh with the Walkley-Black wet oxidation method [55].

### 2.1.3. Environmental Variables

In DSM, the variability of SOC content is essentially explained by its relationships with soil-forming factors, including soil properties, climate, organisms, relief, parent material, time and space. Based on data accessibility, 27 environmental variables were retrieved from soil properties, climate and remote sensing datasets (Table 1).

**Table 1.** Environmental variables for soil organic carbon (SOC) mapping.

| Sources | Variables | Description |
|---|---|---|
| Field Observation | BD | Bulk Density |
| Weather stations | W | Mean relative humidity |
| | T | Average temperature |
| | P | Average precipitation |
| Landsat 8 OLI | b2 | Blue, 0.450–0.515 μm |
| | b3 | Green, 0.525–0.600 μm |
| | b4 | Red, 0.630–0.680 μm |
| | b5 | NIR, 0.845–0.885 μm |
| | b6 | SWIR1, 1.560–1.660 μm |
| | b7 | SWIR2, 2.100–2.300 μm |
| | EVI | $2.5 \times (b5 - b4)/(b5 + 6 \times b4 - 7.5 \times b2 + 1)$ |
| | NDVI | $(b5 - b4)/(b5 + b4)$ |
| | MSAVI | $\{2 \times b5 + 1 - sqrt[(2 \times b5 + 1)\,\hat{}\,2 - 8 \times (b5 - b4)]\}/2$ |
| ASTER GDEM | H | Altitude |
| | β | Slope |
| | α | Aspect |
| | sinα | Sine of aspect, the extent of the location toward the east |
| | cosα | Cosine of aspect, the extent of the location toward the north |
| | C | Curvature |
| | $C_v$ | Vertical curvature |
| | $C_h$ | Horizontal curvature |
| | SOS | Slope of the slope, the curvature of the surface |
| | SOA | Slope of aspect, the curvature of the contour line |
| | RLD | Relief of land surface, Hmax − Hmin |
| | M | Surface roughness, $1/\cos\beta$ |
| | TWI | Topographic wetness index, $Ln[Ac/\tan\beta]$, Ac is the catchment area directed to the vertical flow |
| | SPI | Stream power index, $Ln[Ac \times \tan\beta \times 100]$ |

BD represented a parameter of soil properties. Climate data of monthly average values of September and October 2016, i.e., mean relative humidity (W, %), average temperature (T, °C), and precipitation (P, mm) were recorded at 60 weather stations of Jilin Province (http://data.cma.cn/). The inverse distance weighted method, a useful and common approach to acquiring spatially continuous soil and climate factors [56–58], with the power value of two in ArcGIS was applied to data of BD, W, T and P.

A Landsat 8 Operational Land Imager (OLI) image (http://glovis.usgs.gov/) of the study area was acquired on 22 September 2016 at 30 m spatial resolution. This image was processed for the conversion of the digital numbers to top-of-atmosphere reflectance and atmospheric calibration, using the FLAASH (Fast Line-of-Sight Atmospheric Analysis of Spectral Hypercubes) module embedded in ENVI (version 5.1, Exelis Visual Information Solutions, Broomfield, Colorado, USA). Enhanced Vegetation Index (EVI), Normalized Difference Vegetation Index (NDVI), and Modified Soil Adjusted Vegetation Index (MSAVI) [59] were computed and were used as inputs into the models, along with OLI bands 2-7. Relief data, altitude (H, m), slope (β), aspect (α, °), curvature (C), vertical curvature ($C_v$), horizontal curvature ($C_h$), slope of slope (SOS), slope of aspect (SOA), relief of land surface (RLD), sinα, cosα, surface roughness (M), topographic wetness index (TWI) and stream power index (SPI) [60–62] were computed based on Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model (ASTER GDEM V1) (http://glovis.usgs.gov/) at 30 m spatial resolution.

The above environmental predictors were transformed to the Albers equal-area conic WGS84 coordinate system. The BD and climatic grids were resampled to 30 m resolution using the nearest neighbor method, to match with the raster grids of other variables. Eventually, the attribute values of all grids were extracted for all the sampling points, and they were used as inputs for spatial modeling.

## 2.2. Modeling and Prediction

### 2.2.1. Statistical Analysis

The descriptive statistical values of BD and SOC were estimated for the soil samples. The pairwise Pearson's product-moment correlation analysis was conducted to determine the collinearity between explanatory variables. Predictor variables that were highly correlated to each other ($r \geq 0.8$), and had high variance inflation factors (VIFs $\geq 10$) in regression analysis were excluded [63,64]. The analysis steps were performed using SPSS (version 21.0, IBM, Armonk, NY, USA). All of the explanatory variables were transformed to Z-score to eliminate the effects of the index dimension and the quantity of data [65]. The procedures were illustrated in Figure 2.



**Figure 2.** The flow chart of the comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content.

2.2.2. Geostatistical Models

OK is a widely used geostatistical technique that generates an optimal unbiased estimated surface by means of a semivariogram based on regionalized variables [66]. The semi-variance is defined using Equation (1):

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)|d_{ij}=h} (Z_j - Z_i)^2,$$

(1)

where $\gamma(h)$ is the semi-variance of data values for points of $h$ distance apart, $N(h)$ is the number of point pairs with $h$ distance apart, $Z_i$ is the data value at point $i$, $(i,j)|d_{ij} = h$ represent a pair of points (i.e., $i,j$) which are separated by $h$ distance.

The parameters of OK were listed in Table 2. The semivariogram can be modeled in GS+ (version 9.0, Leland Stanford Junior University, Stanford, CA, USA) by three typical functions: Spherical, Exponential and Gaussian. All of these functions are defined by three parameters: Nugget, range, and sill. Nugget represents the spatial variance of measurement errors at the infinite small distance. Range delineates the effective distance of the spatial autocorrelation. Sill is the maximum value of the semivariogram when the spatial distance between two locations reaches the value of the range [17]. The partial sill is defined as sill-nugget, and a stronger spatial autocorrelation is denoted by higher values of partial sill/sill. Meanwhile, the spatial variation is characterized by the basal effect defined as nugget/sill. In other words, a larger value of nugget/sill shows that the spatial variation among samples is more strongly caused by stochastic factors [62].

**Table 2.** Algorithms used in the study.

| Algorithms | Software | Necessary Parameters |
|:---:|:---:|:---:|
| Ordinary kriging (OK) | GS+ | Model type, nugget, sill, range |
| Geographically weighted regression (GWR) | GWR | Kernel type, bandwidth selection criteria (AICc) |
| Artificial neural network (ANN) | MATLAB | Learning algorithm, hidden layers, learning rate, training time |
| Support vector machine for regression (SVR) | | C (the regularization parameter), kernel (Gaussian radial basis kernel) and its $\sigma$ (the bandwidth parameter) |
| Geographically weighted regression kriging (GWRK) | GWR, GS+ | Kernel type, bandwidth selection criteria (AICc), model type, nugget, sill, range |
| Artificial neural network kriging (ANNK) | MATLAB, GS+ | Learning algorithm, hidden layers, learning rate, training time, model type, nugget, sill, range |

Based on the semivariogram modeling, an error variance model is defined as the objective function for which a set of weights ($w_i$) are selected to minimize the error [66]. After that, an interpolation via OK in ArcGIS follows. The interpolation is expressed by Equation (2):

$$\hat{Z}_0 = \sum_{i=1}^{n} w_i \cdot z_i \ , \quad \sum_{i=1}^{n} w_i = 1 \ ,$$

(2)

where $\hat{Z}_0$ is the estimated SOC content at point $i$, $z_i$ is the measured SOC content data, $w_i$ is the weight associated with the measured SOC, which is estimated by the stationary OK system, and $n$ is the number of measured values within a neighborhood.

GWR is another commonly applied geostatistical approach for DSM, and it is based on the local smoothing idea. It takes the spatial locations of samples into consideration, and uses the locally

weighted least square method to model the observations of soil parameters [32]. It can be written as Equation (3):

$$\hat{y}_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x^*_{ik} + \varepsilon_i, \tag{3}$$

where $(u_i, v_i)$ are the coordinates of point $i$, $\beta_0(u_i, v_i)$ is the intercept, $\beta_k(u_i, v_i)$ is the coefficient of different explanatory variables, $x^*_{ik}$ is the value of explanatory variable $k$ at point $i$, $p$ is the total number of explanatory variables, $\varepsilon_i$ is the error term that is generally assumed to be explanatory and normally distributed with zero mean and constant variance, and the values of the above parameters vary with the location. The parameters can be estimated using a weighting function as in Equation (4):

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y, \tag{4}$$

where $X$ is the matrix formed by $x^*_{ik}$, $Y$ is the vector formed by values of SOC content, $W(u_i, v_i)$ is a weight matrix to ensure that those observations near the point $i$ have more influence on the results than those that are farther away. The parameters of the GWR were listed in Table 2. GWR was conducted using GWR (version 4.0, Ritsumeikan University, Kyoto, Japan) by which the weight function (a geographic kernel type) and the minimum value of the corrected Akaike Information Criterion (AICc, small sample bias corrected AIC) are determined to find the optimal bandwidth [67].

### 2.2.3. Machine Learning Methods

SVR, a non-linear machine learning method, derives a model hyperplane to describe the empirical data as correctly as possible and to minimize the distances from the hyperplane to the training data [68]. The parameters of SVR were listed in Table 2. In this study, SVR was conducted in MATLAB software (version 2017a, MathWorks, Natick, MA, USA). The SMO (sequential minimal optimization) algorithm was used to solve the quadratic programming optimization problem step-by-step. It updated the SVR function, as shown in Equation (5), to reflect the new values until the Lagrange multipliers converged [69].

$$f(x) = \sum_{k=1}^{P} (\alpha_k - \alpha_k^*) K(x_k, x_j) + b, \tag{5}$$

where $x$ is a vector of the input predictors (environmental variables), $f(x)$ is an optimal function developed by SVR, $b$ is a constant threshold, $K(x_k, x_j)$ is the Gaussian radial basis kernel function with the best bandwidth parameter $\sigma$, and $\alpha_k$ and $\alpha_k^*$ are the weights (Lagrange multipliers) with the constraints given in Equation (6):

$$\begin{cases} \sum_{k=1}^{p} (\alpha_k + \alpha_k^*) = 0 \\ 0 \le \alpha_k, \alpha_k^* \le C \end{cases}, \tag{6}$$

where $C$ is the regularization parameter for balancing between the training error and model complexity.

ANN attempt to mimic how the human brain processes and stores information. Among numerous proposed ANN algorithms, the multi-layer perceptron neural network (MLPNN) with back-propagation algorithm was used in this study. The MLPNN was constructed in MATLAB. The architecture of the MLPNN consists of an input layer containing the predictor variables, one or more hidden layers, and an output layer containing the response variable, along with interconnection weights characterizing the connection strength between these layers (Figure 3). By developing the network with training samples, the interconnection weights of the network are adjusted to minimize the prediction error. Meta-parameters to be specified are the number of hidden nodes, the learning algorithm, learning rate, as well as the number of training iterations in those processes [70,71], as Table 2 list.
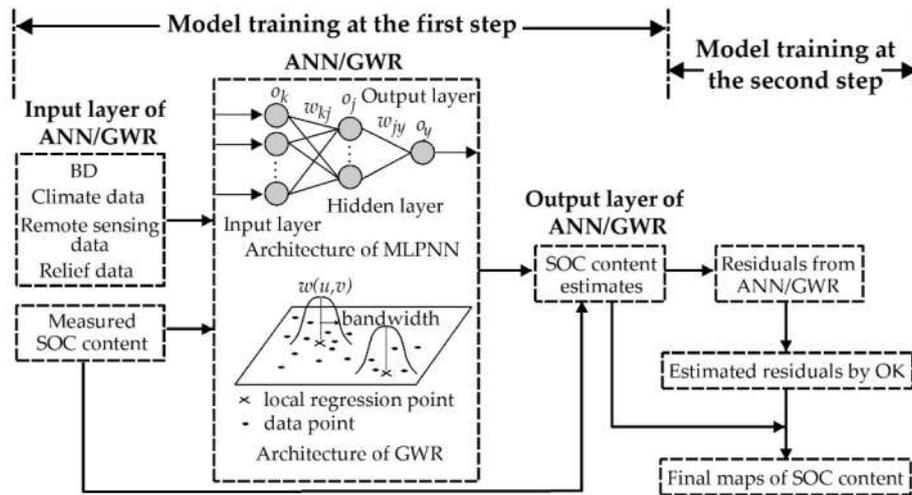
**Figure 3.** Flow diagram of soil organic carbon (SOC) content mapping using two-step hybrid approaches of geographically weighted regression kriging (GWRK) and artificial neural networks kriging (ANNK).

### 2.2.4. Two-Step Hybrid Approaches

GWRK and ANNK are the extension of GWR and ANN, respectively. They fully consider spatial parametric non-stationarity and the effects of environmental variables derived from the benefits of GWR and ANN. They also add spatial dependence of the residuals interpolated though OK to the estimated trend, as part of the spatial autocorrelation. Their implementation includes two steps (Figure 3). First, the GWR and ANN models are built based on Equation (3) and 2.2.3 to model the relationship between SOC content and environmental variables. Second, the residuals resulting from GWR and ANN are estimated using the OK approach (Equations (1) and (2)).

The estimates obtained by the two-step hybrid approaches GWRK and ANNK are the sum of the estimates $\hat{y}_{GWR/ANN}(u_i, v_i)$ and OK estimates of the residuals $\hat{\varepsilon}_{OK}(u_i, v_i)$, as shown in Equation (7):

$$\hat{y}_{GWRK/ANNK}(u_i, v_i) = \hat{y}_{GWR/ANN}(u_i, v_i) + \hat{\varepsilon}_{OK}(u_i, v_i), \tag{7}$$

### 2.2.5. Model Testing and Comparison

The validation set (n = 132) was used to test and compare the performances of OK, GWR, SVR, ANN, GWRK and ANNK models based on the root mean squared error (RMSE, Equation (8)), mean error (Equation (9)) and coefficient of determination ($R^2$, Equation (10)) [66,72].

$$\text{RMSE} = \sqrt{\sum_{1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}, \tag{8}$$

$$\text{ME} = \sum_{1}^{n} \frac{(y_i - \hat{y}_i)}{n}), \tag{9}$$

$$R^2 = 1 - \frac{\sum_{1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{1}^{n}(y_i - \overline{y})^2} \;,\;\; \overline{y} = \frac{1}{n}\sum_{1}^{n} y_i \;, \tag{10}$$

where $\hat{y}_i$ is the estimated value of each model, $y_i$ is the measured value, and $n$ is 132 in this study. The mean error and $R^2$ should be close to zero and one, respectively, while RMSE should be as small as possible.

## 3. Results

### 3.1. Statistics Analysis

The descriptive statistics values of SOC and BD were presented in Table 3. SOC content and BD ranged from 2.54 to 68.94 g·kg$^{-1}$ and from 0.81 to 1.89 g·cm$^{-3}$, respectively. The coefficient of variation (CV) indicated a strong variability of SOC content (CV > 0.35) and a moderate variability of BD (0.15 < CV < 0.35), which can be attributed to random factors like environmental factors and measurement errors [53].

Among 27 explanatory variables, $C_v$ ($r_{Cv,C}$ = −0.90, VIFs = 20.5), $C_h$ ($r_{Ch,C}$ = 0.87, VIFs = 28.8), M ($r_{M,RLD}$ = 0.93, VIFs = 10.3), b4 ($r_{b4,3}$ = 0.98, VIFs = 40.8), b6 ($r_{b6,7}$ = 0.88, VIFs = 15.9) and b7 ($r_{b7,4}$ = 0.83, VIFs = 20.6) were excluded from model building because their *r* and VIFs exceeded the above-mentioned threshold ($r \geq 0.8$ and VIFs $\geq 10$). It reduced the number of explanatory variables from 27 to 21.

**Table 3.** Descriptive statistics values of SOC and bulk density (BD) at the sampling sites.

| Statistics | Mean | SD [1] | CV [2] | Skewness | Kurtosis | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| SOC (g·kg$^{-1}$) | 18.41 | 9.61 | 0.52 | 1.62 | 3.61 | 2.54 | 68.94 |
| BD (g·cm$^{-3}$) | 1.35 | 0.19 | 0.14 | −0.21 | −0.18 | 0.81 | 1.89 |

[1] SD is standard deviation; [2] CV represents coefficient of variation.

### 3.2. Models Training of GWR, SVR and ANN

The weight function for GWR was an adaptive spatial kernel type to avoid border effects [32,73]. The optimal bandwidth was 0.54, with the minimum value of AICc as 1886.96, and the adjusted $R^2$ was 0.31. For a given environmental variable, its coefficient from GWR varied across the study area. The absolute mean values of coefficients for BD (3.49), b5 (1.99), α (1.50), b2 (1.35), and SOA (1.01) showed the stronger capacity of explaining the relationship with SOC than other variables, while b3 (0.16) and H (0.20) had a relative lack of explanatory ability. The CV values showed that coefficients of T and b3 had a strong variability, and coefficients of W, C, and TWI had moderate variabilities.

In the SVR model, the best parameters *C* and *σ* obtained using the training data were 1000 and 0.0001, respectively, with the minimum RMSE being 8.61. As for the MLPNN model, the accuracy for various numbers of nodes in the hidden layer is shown in Figure 4. The results revealed that the optimized MLPNN architecture with an RMSE value of 7.30 was in 21-3-1 structure, i.e., indicating 21 input nodes in the input layer, three in the hidden layer with the unipolar sigmoid as the transfer function, and one in the output layer. Using the Levenberg-Marquardt learning algorithm, the best learning rate and training time (iterations) obtained were determined to be 0.01 and 13, respectively.
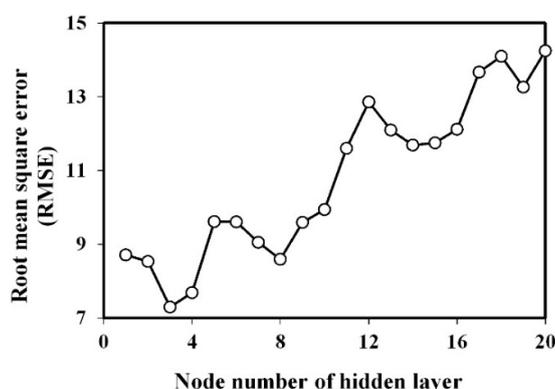


**Figure 4.** The root mean squared error (RMSE) of a multi-layer perceptron neural network (MLPNN) under different node number in the hidden layer.

### 3.3. OK, GWRK and ANNK Training and Six Models Validation

The results from Kolmogorov-Smirnov test (K-S) ($p < 0.05$) showed that the probability distributions for SOC of the training samples (skewness = 1.57, kurtosis = 3.56), and the residuals from GWR (skewness = 0.84, kurtosis = 2.63) and ANN (skewness = 0.90, kurtosis = 1.28) possessed normal distribution. The training set and residuals were directly used to calculate experimental semivariograms for OK interpolation. Table 4 showed the basic information on the training samples, GWR residuals ($R_{GWR}$), and ANN residuals ($R_{ANN}$) in the OK. The equivalent strong performances of semivariogram models for training samples, $R_{GWR}$ and $R_{ANN}$ were evident, based on the low residual sum of squares (RSS) and high $R^2$, and the Exponential model for $R_{ANN}$ performed the best among the three. A greater spatial field fluctuation of training samples than $R_{ANN}$ and $R_{GWR}$ was explained by its highest range. The values of the nugget indicated that the stationarity assumption was more valid for $R_{GWR}$, than for $R_{ANN}$ and training samples. The highest nugget value of training samples suggested that OK was more sensitive to the sampling design for SOC content at a smaller distance. While environment variables were considered as GWR and ANN, the randomness or stochastic showed the reduction according to lower values of the nugget. The smallest nugget/sill for $R_{ANN}$ displayed its stronger basal effect than $R_{GWR}$ and training samples. It also revealed that $R_{ANN}$ obtained the higher spatial autocorrelation and more suitability for OK than the other two. Due to the strong spatial autocorrelation and good modeling performance of $R_{GWR}$ and $R_{ANN}$, GWRK and ANNK, which additionally predicted $R_{GWR}$ and $R_{ANN}$ by OK, were more reasonable than GWR and ANN.

**Table 4.** Semivariogram parameters for SOC observations and the residuals of geographically weighted regression (GWR) and artificial neural network (ANN).

| Parameters (g·kg$^{-1}$) | Model | Range (km) | Nugget | Nugget/Sill | $R^2$ | RSS |
|---|---|---|---|---|---|---|
| Training samples | Exponential | 0.09 | 32.2 | 0.30 | 0.92 | 8.60 |
| $R_{GWR}$ | Spherical | 0.04 | 18.6 | 0.26 | 0.85 | 4.03 |
| $R_{ANN}$ | Exponential | 0.05 | 23.4 | 0.23 | 0.97 | 0.99 |

Table 5 presented the accuracy of different models for estimating the SOC of the validation set. The OK model resulted in a mean error of −0.71 g·kg$^{-1}$ had the highest tendency for overestimation. The ANNK model with a mean error of −0.38 g·kg$^{-1}$ showed the lowest tendency for overestimation. The GWR model with a mean error of 1.63 g·kg$^{-1}$ was the only one for underestimation. ANNK gave rise to the lowest RMSE and closest-to-zero mean error values, as well as the highest $R^2$ value. Hence, it was the best predictive method for estimating the SOC. Besides, the accuracy ranking of six methods from high to low was ANNK, SVR, ANN, GWRK, OK and GWR, which was consistent with the ranking of the consistency between the predicted and measured ranges.

**Table 5.** Performance evaluation of six models.

| Evaluation Index | ANNK | ANN | GWRK | GWR | SVR | OK |
|---|---|---|---|---|---|---|
| Mean error (g·kg$^{-1}$) | −0.38 | −0.59 | −0.61 | 1.63 | −0.53 | −0.71 |
| RMSE(g·kg$^{-1}$) | 8.89 | 9.47 | 9.54 | 9.92 | 9.13 | 9.81 |
| $R^2$ | 0.60 | 0.48 | 0.47 | 0.30 | 0.51 | 0.32 |

### 3.4. Spatial Prediction and Mapping of SOC Content

The spatial distribution of SOC produced by OK, GWR, SVR, ANN, GWRK and ANNK are illustrated in Figure 5. All the six maps showed that a high SOC region was on the northwest section of the study area with content values ranging from 29.29 to 70.93 g·kg$^{-1}$, where urban parks and green space occupied. While low SOC regions were located near the outskirt of Shuangyang District with values ranging from 4.18 to 9.55 g·kg$^{-1}$. Relatively high SOC contents were present in the northern side of Xinlicheng Reservoir, and the areas between Jingyuetan and the northwestern side of Shitoukoumen

Reservoir. The map resulting from GWRK was characterized by more explicit spatial variation than those from GWR and OK. Comparing the range of the estimated SOC content and that of measured values (2.54–68.94 g·kg$^{-1}$) resulted in the following performance ranking for the six algorithms from strong to weak: ANNK, SVR, ANN, GWRK, OK and GWR.



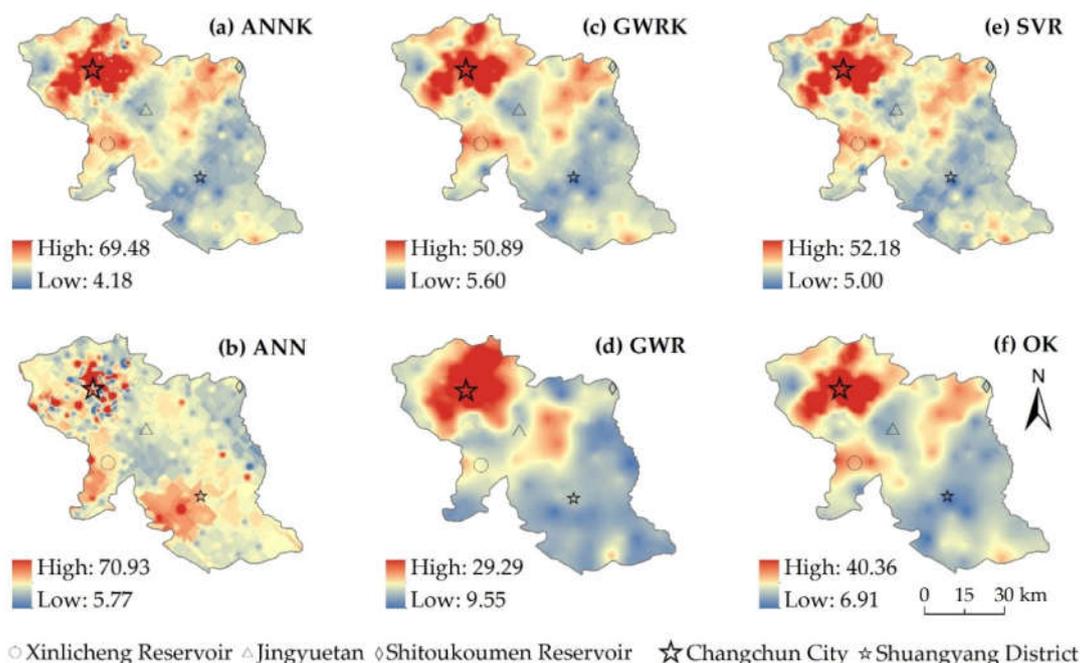**Figure 5.** Maps of predicted SOC content (g·kg$^{-1}$) produced by ANNK (**a**), artificial neural networks (ANN) (**b**), GWRK (**c**), geographically weighted regression (GWR) (**d**), support vector machines for regression (SVR) (**e**) and ordinary kriging (OK) (**f**).

## 4. Discussion

### 4.1. Comparison between Intra-Classes of Three DSM Techniques

The study suggested that, for geostatistical techniques, OK performed better than GWR. A weaker performance of GWR compared to OK in this research can be attributed to the spatial autocorrelation and local multicollinearity of environmental variables [74], as VIFs can only detect the global collinearity. Specifically, the study area has fairly flat terrain, so that environmental variables, such as BD and relief grids used in this study were measured at a rather coarser resolution than the requirement for explicitly reflecting the variability of SOC content. Besides, GWR is an extension of the linear regression, so that the number of variables included could have disproportionately negatively impacted the results when the number of variables is excessive. Liu et al. (2015) implemented GWR and OK for predicting SOC stocks with terrain factors, distance factors, land cover type, and spectral indices, and they obtained more accurate estimates, as RMSE$_{GWR}$ was 43.4% of the mean value of SOC and RMSE$_{OK}$ was 34.4%, than those reported in this study [50]. Taking a town as the study area, they considered man-made disturbances in addition to natural factors with fairly good resolution, leading to the better performance of GWR on SOC mapping. However, Zhang et al. (2011) applied OK and GWR onto the environmental factors of rainfall, land cover and soil type to map soil organic carbon, and the accuracy was lower than that which was obtained by this study [73]. Theoretically speaking, OK considers the spatial autocorrelation of the soil samplings, whereas GWR considers the spatial heterogeneity of environmental variables. In other words, it was found by this study that the validation data were closer to the measuring data in the soil sampling locations, due to a better performance of OK than GWR. Whereas, the spatial heterogeneity of environmental variables and its relationship with soil sampling were more significant and matched at the study scale of Zhang et al. (2011) based on better

performance of GWR than OK. Overall, due to the different mechanism of GWR and OK, it is suggested that both models should be calibrated, and then the best result is applied for the spatial prediction of target soil attributes.

The results for the two ML models showed that SVR performed slightly better than ANN, which was in agreement with the conclusions by Wijewardane et al. (2016), Ottoy et al. (2017) and Xu et al. (2018) [43,71,75]. It was reported by prior studies that SVR can eliminate the local minimum issue of the error function of ANN [76,77]. Nevertheless, these evaluations were different from those that were obtained by Taghizadeh-Mehrjardi et al. (2016, 2017), who showed better performances of ANN than SVR for soil attributes prediction [78,79]. Were et al. (2015) input nine variables of soil properties into SVR and ANN, one climate variable, land cover data, four relief factors, as well as two spectral indices to map SOC stocks, and obtained higher values of $R^2$, but lower values of RMSE than did in this study [64]. This may be because of different extents of the study areas, sampling sizes, the number and quality of environmental variables used. ANN is capable of learning complex patterns in the noisy environment, but it requires representative training samples and appropriate values of network parameters [80,81]. SVR is able to generalize the model well with limited training samples, but it suffers from parameter assignment issues [82]. Since each model has its pros and cons, studies all strongly suggested the application of both ML models, ANN and SVR, due to their close performances.

As for the hybrid approaches, ANNK outperformed GWRK for the prediction of SOC content in this study. The better performance of ANNK compared to GWRK is related to the non-linear relationship between environmental variables and SOC content. It was overcome by the ability of ANN to solve non-linear problems [83]. However, GWRK is much simpler to operate, and more mature for DSM applications than ANNK. Because these two approaches have not been compared in any previous research, this finding can provide a reference for mapping SOC in future.

Except that the two ML techniques had close performance, the two-step hybrid approaches performed differently, as well as the geostatistical techniques. It also indicated that choosing a suitable model at the first step is crucial for estimating the SOC via hybrid approaches.

## 4.2. Comparison among Geostatistical, ML and Hybrid Techniques

Comparison among three DSM techniques (Figure 5 and Table 5), showed that the geostatistical models had the lowest accuracy. The poor performance of OK was found by Mirzaee et al. (2016) when compared with ANNK [47], by Emamgholizadeh et al. (2017) who showed a performance order of ANN, GWR and OK [84], and by Ye et al. (2017) when compared with GWRK [85]. ANNK showed the highest accuracy, but GWRK performed worse than the two ML models. It showed that the spatial cross correlation between SOC content and environmental variables in this study was depicted definitely by non-linear regression like SVR and ANN, rather than GWR as a local linear regression. ML in this study, i.e., ANN and SVR have built in methods for dimension reduction, so that the number of variables and local multicollinearity of predictors influences less on ML than GWR. However, additionally considering the spatial autocorrelation of SOC residuals, ANNK apparently outperformed SVR, in spite of a close performance of ANN and SVR. Obviously, two-step hybrid approaches performed better than the corresponding one-step models, which was also revealed by comparisons made in Kumar (2015), Zeng et al. (2016) and Guo et al. (2017) about GWRK and GWR [48,86,87], and in Dai et al. (2014) and Song et al. (2017) about ANNK and ANN [72,88]. Dai et al. (2014) used ANNK and ANN for mapping soil organic matter content with auxiliary variables, including climate data, relief data and spectral indices of the remote sensing data, and obtained a higher accuracy than that of this study [88]. Due to the high variability of environmental variables and the fairly simple relationship between variables and the SOM content in the Tibetan Plateau, ANNK and ANN showed stronger performances than reported by this study. The relative improvement of the prediction accuracy of ANNK compared to ANN was also more obvious.

The uncertainty is a key issue for SOC mapping, based on these three DSM techniques. The processing of soil samples collection and inputs are the same for six models. While, the distribution

and number of soil samples and predictor inputs have a greater influence on geostatistics than ML methods. It also results in the higher uncertainty of GWRK than ANNK. As for different modeling methods, two-step hybrid approaches would increase uncertainty than corresponding one-step models. However, two-step hybrid approaches greatly improved accuracy than geostatistics and ML according to reported studies and our findings as abovementioned. For a regional area like this study, the six methods cost a similar amount of time when considering a trade-off between accuracy and speed. Thus, the two-step hybrid approach resulted in better predictions and lower errors than corresponding one-step models, which were recommended for a regional SOC mapping.

## 5. Conclusions

To map the spatial distribution of a vital ecological indicator of soil quality, SOC content, six DSM methods were developed with soil properties, climatic data, relief factors and spectral indices being environmental variables. These methods were geostatistical (OK and GWR), ML (SVR and ANN) and two-step hybrid approaches (GWRK and ANNK). The performance ranking of six methods from high to low was ANNK, SVR, ANN, GWRK, OK and GWR based on the prediction accuracy measured by RMSE, ME, $R^2$ and the consistency with measured values. Additionally, the two-step hybrid approaches, ANNK and GWRK, gave rise to low RMSE, close-to-zero mean error and high $R^2$ for the validation set as compared with GWR and ANN. Because of the non-linear relationship between environmental variables and SOC content, two ML methods performed better than linear models. The strong performances of the two-step hybrid approaches were attributed to their capability of explicitly addressing the spatial dependency and non-stationarity of SOC content. It led to apparent outperformances of ANNK compared to SVR, in spite of the close performance of ANN and SVR.

When considering uncertainty and efficiency, the suitable situations of these six methods on SOC mapping were also discussed. For a regional study area with high heterogeneity, such as urban and mountainous landscapes, ML and two-step approach are more suitable than geostatistics. Selection of a suitable model at the first step is crucial for two-step hybrid approaches. Limited by the number of samples, SVR performs better than ANN. Influenced by predictor inputs, GWR performs worse and has a larger uncertainty than OK. Therefore, ANNK combining ANN with OK is considered to be the most promising approach to estimating regional SOC.

**Author Contributions:** Lin Chen, Chunying Ren, and Bai Zhang designed this research. Lin Chen and Lin Li conducted field sampling, performed the experiments. Lin Chen conducted the analysis and drafted the manuscript. Lin Chen, Chunying Ren, Linfeng Li, Yeqiao Wang, Bai Zhang, and Zongming Wang revised and finalized the manuscript.

## References

1. Jobbágy, E.G.; Jackson, R.B. The vertical distribution of soil organic C and its relation to climate and vegetation. *Ecol. Appl.* **2000**, *10*, 423–436. [CrossRef]
2. Stockmann, U.; Adams, M.A.; Crawford, J.W.; Field, D.J.; Henakaarchchi, N.; Jenkins, M.; Minasny, B.; McBratney, A.B.; de Courcelles, V.D.; Singh, K.; et al. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agric. Ecosyst. Environ.* **2013**, *164*, 80–99. [CrossRef]

3. Davidson, E.A.; Janssens, I.A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **2006**, *440*, 165–173. [CrossRef] [PubMed]

4. Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **2004**, *304*, 1623–1627. [CrossRef]

5. Smith, P.; House, J.I.; Bustamante, M.; Sobocká, J.; Harper, R.; Pan, G.; West, P.; Clark, J.; Adhya, T.; Rumpel, C. Global change pressures on soils from land use and management. *Glob. Chang. Biol.* **2015**, *22*, 1008–1028. [CrossRef]

6. Wiesmeier, M.; Munro, S.; Barthold, F.; Steffens, M.; Schad, P.; Kögel-Knabner, I. Carbon storage capacity of semi-arid grassland soils and sequestration potentials in northern China. *Glob. Chang. Biol.* **2015**, *21*, 3836–3845. [CrossRef] [PubMed]

7. Zhang, X.; Xu, X.; Liu, Y.; Wang, J.; Xiong, Z. Global warming potential and greenhouse gas intensity in rice agriculture driven by high yields and nitrogen use efficiency. *Biogeosciences* **2016**, *13*, 2701–2714. [CrossRef]

8. Tiessen, H.; Cuevas, E.; Chacon, P. The role of soil organic matter in sustaining soil fertility. *Nature* **1994**, *371*, 783–785. [CrossRef]

9. Milne, E.; Adamat, R.A.; Batjes, N.H.; Bernoux, M.; Bhattacharyya, T.; Cerri, C.C.; Cerri, C.E.P.; Coleman, K.; Easter, M.; Falloon, P. National and sub-national assessments of soil organic carbon stocks and changes: The GEFSOC modelling system. *Agric. Ecosyst. Environ.* **2007**, *122*, 3–12. [CrossRef]

10. Zhang, H.; Wu, P.B.; Yin, A.J.; Yang, X.H.; Zhang, M.; Gao, C. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. *Sci. Total Environ.* **2017**, *592*, 704–713. [CrossRef] [PubMed]

11. McBratney, A.B.; Santos, M.L.M.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]

12. Cambardella, C.; Moorman, T.; Parkin, T.; Karlen, D.; Novak, J.; Turco, R.; Konopka, A. Field-scale variability of soil properties in central Iowa soils. *Soil Sci. Soc. Am. J.* **1994**, *58*, 1501–1511. [CrossRef]

13. Liu, D.; Wang, Z.; Zhang, B.; Song, K.; Li, X.; Li, J.; Li, F.; Duan, H. Spatial distribution of soil organic carbon and analysis of related factors in croplands of the black soil region, Northeast China. *Agric. Ecosyst. Environ.* **2006**, *113*, 73–81. [CrossRef]

14. Zhang, S.; Zhang, X.; Huffman, T.; Liu, X.; Yang, J. Influence of topography and land management on soil nutrients variability in Northeast China. *Nutr. Cycl. Agroecosyst.* **2011**, *89*, 427–438. [CrossRef]

15. Umali, B.P.; Oliver, D.P.; Forrester, S.; Chittleborough, D.J.; Hutson, J.L.; Kookana, R.S.; Ostendorf, B. The effect of terrain and management on the spatial variability of soil properties in an apple orchard. *Catena* **2012**, *93*, 38–48. [CrossRef]

16. Song, C.; Wang, E.; Han, X.; Stirzaker, R. Crop production, soil carbon and nutrient balances as affected by fertilisation in a Mollisol agroecosystem. *Nutr. Cycl. Agroecosyst.* **2011**, *89*, 363–374. [CrossRef]

17. Ou, Y.; Rousseau, A.N.; Wang, L.X.; Yan, B.X. Spatio-temporal patterns of soil organic carbon and pH in relation to environmental factors-A case study of the Black Soil Region of Northeastern China. *Agric. Ecosyst. Environ.* **2017**, *245*, 22–31. [CrossRef]

18. Kumar, S.; Lal, R. Mapping the organic carbon stocks of surface soils using local spatial interpolator. *J. Environ. Monit.* **2011**, *13*, 3128–3135. [CrossRef] [PubMed]

19. Burrough, P.A.; McDonnell, R.A. *Principles of Geographical Information Systems*; Oxford University Press: New York, NY, USA, 1998.

20. Meersmans, J.; de Ridder, F.; Canters, F.; de Baets, S.; van Molle, M. A multiple regression approach to assess the spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma* **2008**, *143*, 1–13. [CrossRef]

21. Amare, T.; Hergarten, C.; Hurni, H.; Wolfgramm, B.; Yitaferu, B.; Selassie, Y.G. Prediction of soil organic carbon for Ethiopian highlands using soil spectroscopy. *ISRN Soil Sci.* **2013**, *2013*, 720589. [CrossRef]

22. Yang, Y.; Fang, J.; Tang, Y.; Ji, C.; Zheng, C.; He, J.; Zhu, B. Storage, patterns and controls of soil organic carbon in the Tibetan grasslands. *Glob. Chang. Biol.* **2008**, *14*, 1592–1599. [CrossRef]

23. Doetterl, S.; Stevens, A.; van Oost, K.; Quine, T.A.; van Wesemael, B. Spatially explicit regional scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. *Geoderma* **2013**, *204–205*, 31–42. [CrossRef]

24. Lian, G.; Guo, X.D.; Fu, B.J.; Hu, C.X. Prediction of the spatial distribution of soil properties based on environmental correlation and geostatistics. *Trans. Chin. Soc. Agric. Eng.* **2009**, *25*, 112–122.

25. Brunsdon, C.; Fotheringham, A.S.; Charlton, M.E. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geogr. Anal.* **1996**, *28*, 281–298. [CrossRef]

26. Webster, R.; Oliver, M. *Geostatistics for Environmental Scientists*; John Wiley & Sons: Chichester, UK, 2001.

27. Elbasiouny, H.; Abowaly, M.; Alkheir, A.A.; Gad, A.A. Spatial variation of soil carbon and nitrogen pools by using ordinary Kriging method in an area of north Nile Delta, Egypt. *Catena* **2014**, *113*, 70–78. [CrossRef]

28. Oliver, M.A.; Webster, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* **2014**, *113*, 56–59. [CrossRef]

29. Pająk, M.; Halecki, W.; Gąsiorek, M. Accumulative response of Scots pine (*Pinus sylvestris* L.) and silver birch (*Betula pendula Roth*) to heavy metals enhanced by Pb-Zn ore mining and processing plants: Explicitly spatial considerations of ordinary kriging based on a GIS approach. *Chemosphere* **2017**, *168*, 851–859. [CrossRef] [PubMed]

30. Mishra, U.; Lal, R.; Slater, B.; Calhoun, F.; Liu, D.; Van Meirvenne, M. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. *Soil Sci. Soc. Am. J.* **2009**, *73*, 614–621. [CrossRef]

31. Eldeiry, A.A.; Garcia, L.A. Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using Landsat images. *J. Irrig. Drain. Eng.* **2010**, *136*, 355–364. [CrossRef]

32. Fotheringham, A.S.; Brunsdon, C.; Charlton, M.E. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*; Wiley: Chichester, UK, 2002.

33. Scull, P. A top-down approach to the state factor paradigm for use in macroscale soil analysis. *Ann. Assoc. Am. Geogr.* **2010**, *100*, 1–12. [CrossRef]

34. Harris, P.; Fotheringham, A.; Crespo, R.; Charlton, M. The use of geographically weighted regression for spatial prediction: An evaluation of models using simulated data sets. *Math. Geosci.* **2010**, *42*, 657–680. [CrossRef]

35. Drake, J.M.; Randin, C.; Guisan, A. Modelling ecological niches with support vector machines. *J. Appl. Ecol.* **2006**, *43*, 424–432. [CrossRef]

36. Gautam, R.; Panigrahi, S.; Franzen, D.; Sims, A. Residual soil nitrate prediction from imagery and non-imagery information using neural network technique. *Biosyst. Eng.* **2011**, *110*, 20–28. [CrossRef]

37. Khlosi, M.; Alhamdoosh, M.; Doualk, A.; Gabriels, D.; Cornelis, W.M. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. *Eur. J. Soil Sci.* **2016**, *67*, 276–284. [CrossRef]

38. Nguyen, P.M.; Haghverdi, A.; Pue, J.D.; Botula, Y.D.; Le, K.V.; Waegeman, W.; Cornelis, W.M. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. *Biosyst. Eng.* **2017**, *153*, 12–27. [CrossRef]

39. Krishna, G.; Sahoo, R.N.; Singh, P.; Bajpai, V.; Patra, H.; Kumar, S.; Dandapani, R.; Gupta, V.K.; Viswanathan, C.; Ahmad, T.; et al. Comparison of various modelling approaches for water deficit stress monitoring in rice crop through hyperspectral remote sensing. *Agric. Water Manag.* **2019**, *213*, 231–244. [CrossRef]

40. Gunn, S.R. *Support Vector Machines for Classification and Regression*; University of Southampton: Southampton, UK, 1998.

41. Haykin, S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall PTR: Upper Saddle River, NJ, USA, 1998.

42. Li, Y.F.; Liang, S.; Zhao, Y.Y.; Li, W.B.; Wang, Y.J. Machine learning for the prediction of *L. chinensis* carbon, nitrogen and phosphorus contents and understanding of mechanisms underlying grassland degradation. *J. Environ. Manag.* **2017**, *192*, 116–123. [CrossRef]

43. Xu, S.X.; Zhao, Y.C.; Wang, M.Y.; Shi, X.Z. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis-NIR spectroscopy. *Geoderma* **2018**, *310*, 29–43. [CrossRef]

44. Garcia, M.; Saatchi, S.; Ustin, S.; Balzter, H. Modelling forest canopy height by integrating airborne LiDAR samples with satellite Radar and multispectral imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *66*, 159–173. [CrossRef]

45. Takata, Y.; Funakawa, S.; Akshalov, K.; Ishida, N.; Kosaki, T. Spatial prediction of soil organic matter in northern Kazakhstan based on topographic and vegetation information. *Soil Sci. Plant Nutr.* **2007**, *53*, 289–299. [CrossRef]

46. Kumar, S.; Lal, R.; Liu, D. A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma* **2012**, *189*, 627–634. [CrossRef]

47. Mirzaee, S.; Ghorbani-Dashtaki, S.; Mohammadi, J.; Asadi, H.; Asadzadeh, F. Spatial variability of soil organic matter using remote sensing data. *Catena* **2016**, *145*, 118–127. [CrossRef]

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

49. Karunaratne, S.B.; Bishop, T.F.A.; Baldock, J.A.; Odeh, I.O.A. Catchment scale mapping of measureable soil organic carbon fractions. *Geoderma* **2014**, *219–220*, 14–23. [CrossRef]

50. Liu, Y.L.; Guo, L.; Jiang, Q.H.; Zhang, H.T.; Chen, Y.Y. Comparing geospatial techniques to predict SOC stocks. *Soil Tillage Res.* **2015**, *148*, 46–58. [CrossRef]

51. Akpa, S.I.C.; Odeh, I.O.A.; Bishop, T.F.A.; Hartemink, A.E.; Amapu, I.Y. Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma* **2016**, *271*, 202–215. [CrossRef]

52. Keskin, H.; Grunwald, S.; Harris, W.G. Digital mapping of soil carbon fractions with machine learning. *Geoderma* **2019**, *339*, 40–58. [CrossRef]

53. Wilding, L.G. *Spatial Variability: Its Documentation, Accommodation and Implication to Soil Surveys*; Soil Spatial Variability: Wageningen, The Netherlands, 1985.

54. Blake, G.R. *Bulk Density*; American Society of Agronomy: Madison, WI, USA, 1965.

55. Nelson, D.W.; Sommers, L.E. *Total Carbon, Organic Carbon and Organic Matter*; American Society of Agronomy: Madison, WI, USA, 1982.

56. Li, J.W.; Richter, D.D.; Mendoza, A.; Heine, P. Effects of land-use history on soil spatial heterogeneity of macro- and trace elements in the Southern Piedmont USA. *Geoderma* **2010**, *156*, 60–73. [CrossRef]

57. Wu, Q.; Li, Q.L.; Gao, J.B.; Lin, Q.Y.; Xu, Q.Y.; Groffman, P.M.; Yu, S. Non-algorithmically integrating land use yype with spatial interpolation of surface soil nutrients in an urbanizing watershed. *Pedosphere* **2017**, *27*, 147–154. [CrossRef]

58. Barrios, A.; Trincado, G.; Garreaud, R. Alternative approaches for estimating missing climate data: Application to monthly precipitation records in South-Central Chile. *For. Ecosyst.* **2018**, *5*, 28. [CrossRef]

59. Ma, C.Y.; Wang, J.L.; Chen, Z.; Chen, Z.F.; Liu, Z.D.; Huang, X.Q. An assessment of surface soil moisture based on in situ observations and Landsat 8 remote sensing data. *Fresenius Environ. Bull.* **2017**, *26*, 6848–6856.

60. Wilson, J.P.; Gallant, J.C. *Terrain Analysis: Principles and Applications*; John Wiley & Sons: New York, NY, USA, 2000.

61. Zhang, S.M.; Wang, Z.M.; Zhang, B.; Song, K.S.; Liu, D.W.; Li, F.; Ren, C.Y.; Huang, J.; Zhang, H.L. Prediction of spatial distribution of soil nutrients using terrain attributes and remote sensing data. *Trans. Chin. Soc. Agric. Eng.* **2010**, *25*, 188–194.

62. Tang, G.A.; Yang, X. *ArcGIS Experimental Course for Spatial Analysis*; Science Press: Beijing, China, 2013.

63. O'brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [CrossRef]

64. Were, K.; Bui, D.T.; Dick, Ø.B.; Singh, B.R. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol. Indic.* **2015**, *52*, 394–403. [CrossRef]

65. Cheadle, C.; Vawter, M.P.; Freed, W.J.; Becker, K.G. Analysis of microarray data using Z score transformation. *J. Mol. Diagn.* **2003**, *5*, 73–81. [CrossRef]

66. Isaaks, E.H.; Srivastava, R.M. *An Introduction to Applied Geostatistics*; Oxford University Press: Oxford, UK, 1989.

67. Nakaya, T.; Charlton, M.; Lewis, P.; Brunsdon, C.; Yao, J.; Fotheringham, S. *GWR4 User Manual, Windows Application for Geographically Weighted Regression Modelling*; Ritsumeikan University: Kyoto, Japan, 2014.

68. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2000.

69. Platt, J. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*; MIT Press: Cambridge, MA, USA, 1999.

70. Lee, S.; Evangelista, D.G. Earthquake-induced landslide susceptibility mapping using an artificial neural network. *Nat. Hazards Earth Syst. Sci.* **2006**, *6*, 687–695. [CrossRef]

71.    Ottoy, S.; De Vos, B.; Sindayihebura, A.; Hermy, M.; Van Orshoven, J. Assessing soil organic carbon stocks under current and potential forest cover using digital soil mapping and spatial generalisation. *Ecol. Indic.* **2017**, *77*, 139–150. [CrossRef]

72.    Song, Y.Q.; Yang, L.A.; Li, B.; Hu, Y.M.; Wang, A.L.; Zhou, W.; Cui, X.S.; Liu, Y.L. Spatial prediction of soil organic matter using a hybrid geostatistical model of an extreme learning machine and ordinary kriging. *Sustainability* **2017**, *9*, 754. [CrossRef]

73.    Zhang, C.S.; Tang, Y.; Xu, X.L.; Kiely, G. Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. *Appl. Geochem.* **2011**, *26*, 1239–1248. [CrossRef]

74.    Yang, Q.Y.; Jiang, Z.C.; Li, W.J.; Li, H. Prediction of soil organic matter in peak-cluster depression region using kriging and terrain indices. *Soil Tillage Res.* **2014**, *144*, 126–132. [CrossRef]

75.    Wijewardane, N.K.; Ge, Y.F.; Morgan, C.L.S. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. *Geoderma* **2016**, *267*, 92–101. [CrossRef]

76.    Abraham, A. Meta learning evolutionary artificial neural networks. *Neurocomputing* **2004**, *56*, 1–38. [CrossRef]

77.    Sakizadeh, M.; Mirzaei, R.; Ghorbani, H. Support vector machine and artificial neural network to model soil pollution: A case study in Semnan Province, Iran. *Neural Comput. Appl.* **2017**, *28*, 3229–3238. [CrossRef]

78.    Taghizadeh-Mehrjardi, R.; Nabiollahi, K.; Kerry, R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma* **2016**, *266*, 98–110. [CrossRef]

79.    Taghizadeh-Mehrjardi, R.; Neupane, R.; Sood, K.; Kumar, S. Artificial bee colony feature selection algorithm combined with machine learning algorithms to predict vertical and lateral distribution of soil organic matter in South Dakota, USA. *Carbon Manag.* **2017**, *8*, 277–291. [CrossRef]

80.    Mas, J.F.; Flores, J.J. The application of artificial neural networks to the analysis of remotely sensed data. *Int. J. Remote Sens.* **2008**, *29*, 617–663. [CrossRef]

81.    Zhang, C.Y.; Denka, S.; Cooper, H.; Mishra, D.R. Quantification of sawgrass marsh aboveground biomass in the coastal Everglades using object-based ensemble analysis and Landsat data. *Remote Sens. Environ.* **2018**, *204*, 366–379. [CrossRef]

82.    Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]

83.    Rossel, R.A.V.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [CrossRef]

84.    Emamgholizadeh, S.; Shahsavani, S.; Eslami, M.A. Comparison of artificial neural networks, geographically weighted regression and Cokriging methods for predicting the spatial distribution of soil macronutrients (N, P, and K). *Chin. Geogr. Sci.* **2017**, *27*, 747–759. [CrossRef]

85.    Ye, H.C.; Huang, W.J.; Huang, S.Y.; Huang, Y.F.; Zhang, S.W.; Dong, Y.Y.; Chen, P.F. Effects of different sampling densities on geographically weighted regression kriging for predicting soil organic carbon. *Spat. Stat.* **2017**, *20*, 76–91. [CrossRef]

86.    Kumar, S. Estimating spatial distribution of soil organic carbon for the Midwestern United States using historical database. *Chemosphere* **2015**, *127*, 49–57. [CrossRef]

87.    Zeng, C.Y.; Yang, L.; Zhu, A.X.; Rossiter, D.G.; Liu, J.; Liu, J.Z.; Qin, C.Z.; Wang, D.S. Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method. *Geoderma* **2016**, *281*, 69–82. [CrossRef]

88.    Dai, F.Q.; Zhou, Q.G.; Lv, Z.Q.; Wang, X.M.; Liu, G.C. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. *Ecol. Indic.* **2014**, *45*, 184–194. [CrossRef]