An improved similarity-based approach to predicting and mapping soil organic carbon and soil total nitrogen in a coastal region of northeastern China

Shuai Wang[1, 2, 3], Kabindra Adhikari[4], Qianlai Zhuang[3], Zijiao Yang[1], Xinxin Jin[1*], Qiubing Wang[1], and Zhenxing Bian[1]

[1] College of Land and Environment, Shenyang Agricultural University, Shenyang, Liaoning Province, China
[2] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[3] Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, IN 47907, United States of America
[4] Grassland, Soil and Water Research Laboratory, USDA-ARS, Temple, TX 76502, United States of America

Corresponding Author:
Xinxin Jin[1]
No. 120 Dongling Road, Shenhe District, Shenyang, Liaoning Province 110866, China
Tel: +86-24-8848-7155
Fax: +86-24-8848-7155
Email address: jinxinxin0218@163.com

Abstract

Soil organic carbon (SOC) and soil total nitrogen (STN) are major soil indicators for soil quality and fertility. Accurate mapping SOC and STN in soils would help both managed and natural soils and ecosystem management. This study developed an improved similarity-based approach (ISA) to predicting and mapping topsoil (0-20 cm soil depth) SOC and STN in a coastal region of northeastern China. Six environmental variables including elevation, slope gradient, topographic wetness index, the mean annual temperature, the mean annual temperature, and normalized difference vegetation index were used as predictors. Soil survey data in 2012 was designed based on the clustering of the study area into six climatic vegetation landscape units. In each landscape unit, 20-25 sampling points were determined at different landform positions considering local climate, soil type, elevation and other environmental factors, and finally 126 sampling points were obtained. Soil sampling from the depth of 0-20 cm were used for model prediction and validation. The ISA model performance was compared with the geographically weighted regression (GWR), regression kriging (RK), boosted regression trees (BRT) considering mean absolute prediction error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and maximum relative difference (RD) indices. We found that the ISA

40   method performed best with the highest R^2 and lowest MAE, RMSE compared to GWR, RK,
41   and BRT methods. The ISA method could explain 76 % and 83 % of the total SOC and STN
42   variability, respectively, 12-40% higher than other models in the study area. Elevation had the
43   largest influence on SOC and STN distribution. We conclude that the developed ISA model is
44   robust and effective in mapping SOC and STN, particularly in the areas with complex
45   vegetation-landscape when limited samples are available. The method needs to be tested for
46   other regions in our future research.
47
48   **Keywords**: Digital soil mapping, environmental variables, spatial variability, uncertainty
49

50   Introduction

51   Soil organic carbon (SOC) and soil total nitrogen (STN) influence soil physical, chemical,
52   biological properties and processes and determine soil quality that is related to food security by
53   affecting agricultural productivity (Batjes 1996; Wang et al. 2013; Elbasiouny et al. 2014). To a
54   certain extent, SOC and STN affect the concentration of greenhouse gases in the atmosphere.
55   The prediction of their sizes and changes has become a key climate change research area
56   (Jobbagy & Jackson 2000;   Post & Kwon 2000; Lal 2004; Davidson & Janssens 2006; Yang et
57   al. 2015). Previous studies indicated that the spatially explicit information of SOC and STN
58   plays an important role in quantifying global carbon and nitrogen cycles (Kieft et al. 1998; Yang
59   et al. 2015;   Wang et al. 2017).
60   Numerous studies have been conducted to understand the relationships between SOC and
61   STN with environmental factors including topography, climate and biology (e.g., Yang et al.,
62   2016). To map the spatial variations of SOC and STN based on field observations, Digital soil
63   mapping (DSM) technologies have proven as a rapid and inexpensive approach over large areas
64   using a limited amount of sample data (McBratney et al., 2003; Yang et al., 2016). Commonly
65   used DSM techniques include linear regression (Kunkel et al., 2011; Zhao et al., 2015), random
66   forest (RF) (Grimm et al., 2008; Yang et al., 2016), regression kriging (RK) (Odeh et al., 1995;
67   Zhao et al., 2015; Song et al., 2016), regression rules (Adhikari et al, 2014; Minasny and
68   McBratney, 2008), boosted regression trees (BRT) (Yang et al., 2016; Wang et al., 2016),
69   geographically weighted regression (GWR) (Kumar et al., 2012; Song et al., 2016), artificial
70   neural networks (Burke et al., 1989; Were et al., 2015), similarity-based method (Zhu, 1997;
71   Yang et al., 2015; Liu et al., 2016), and support vector machines (Kovačević et al., 2010; Stevens
72   et al., 2010) .
73   DSM conceptually applies the soil-landscape model of Jenny (1941) which describes the
74   changes in soil properties as a function of factors related to climate, organism, topography,
75   parent material and time. Therefore, the soil and its environment are specifically and spatially
76   related. Topography plays an important role in the process of soil development and formation. Its
77   impact on soil formation is mainly through the redistribution of water, nutrients, and energy,
78   affecting microclimate and biomes, thus indirectly influencing the spatial distribution of SOC,
79   and STN (Adhikari et al., 2018; Mondal et al., 2017; Garten Jr and Ashwood, 2002; Moore et al.,

80 1993). In the central highlands of Ethiopia, Tesfaye et al (2016) found topography as the main
81 factor affecting the spatial variation of SOC and STN content in the region. Similarly, climatic
82 influences on soil formation and development are mainly through temperature, and precipitation
83 and their interactive effects. Jobbagy and Jackson (2000) applied three global-scale soil profile
84 datasets to estimate the global SOC stocks in 3-m deep soil layer, and pointed out that
85 temperature and precipitation were the main climatic factors affecting the changes in SOC stocks.
86 Hudson (1992) considered that soil-forming factors in a landscape interact with each other in a
87 particular manner, indicating that the same soil-landscape presents with homogeneous soil types.
88 According to this theory, the relationship between soil and soil-forming factors can be used to
89 infer the spatial distribution of soil properties. Similarly, Zhu et al. (1997) suggested that soil-
90 forming factors indicate the spatial distribution of soil properties. Thus, similar soil-
91 environmental conditions can be assumed to exhibit similar soil types and properties. Based on
92 this hypothesis, the complex environment can be divided into several less heterogeneous
93 landscape units with potentially similar soil-forming environment. Soil properties in these
94 landscape units can be further represented by the characteristics of the sample points within the
95 landscape. The similarity-based method (Zhu 1997) has been applied and proven to be an
96 efficient method in predictive soil mapping. It has been widely used in the prediction of soil
97 properties and soil types (Liu et al. 2016; Liu 2010; Zhu et al., 2010). Yang et al. (2015) used the
98 similarity-based approach to map the spatial variation of soil salinity and obtained a higher
99 prediction accuracy. A higher performance of similarity-based mapping approach was also
100 confirmed by Shi et al. (2004), who generated a soil series map with 86% prediction accuracy.
101 Liu et al. (2010) indicated that the similarity-based approach exhibited notable predictive
102 performance compared to a kriging interpolation method applied in Ili of Xinjiang, China.
103 Similarly, Liu et al. (2016) integrated a similarity-based approach with depth functions to model
104 the three-dimensional (3D) distribution of SOM and obtained a lower global mean error (0.06 g
105 kg^{-1}). Their findings showed that the similarity-based approach combined with other models can
106 accurately predict soil properties.
107   The spatial distribution of SOC and STN in a landscape can be influenced by topography,
108 organic matter input, temperature, humidity, vegetation, parent material and soil management
109 (Sollins et al., 1996; Camino-Serrano et al., 2014), and can be better estimated by dividing the
110 landscape into sub-units. To achieve this, some previous studies used data segmentation
111 techniques to establish the relationship between SOC and STN with specific environmental
112 variables for smaller regions. For instance, Mulder et al. (2015) divided France into 10 soil-
113 landscape types using a model-based clustering technology based on climate, vegetation,
114 geology, and soil environmental variables and the regression models between SOC and
115 environmental variables were established for each landscape. They found that the classification
116 and prediction at each soil-landscape unit could better explain the SOC variations.
117   This study improved a similarity-based approach (ISA) to predict and map the spatial
118 distribution of SOC and STN by applying soil property data and environmental variables. In the
119 ISA model, the whole study area is clustered into several typical climatic-vegetation landscape

120 units by using gaussian mixture model for model-based clustering (GMMC). The spatial
121 prediction of SOC and STN was then carried out by using the similarity-based approach in this
122 region. In the GMMC model, the expectation maximization algorithm is adopted to estimate the
123 model parameters, whereas the Bayesian Information Criterion (BIC) is used to optimize the
124 clustering results, the optimal model and the numbers of clusters are then selected. This study
125 attempts to test the ISA model by comparing it with the commonly used DSM methods such as
126 geographically weighted regression, regression kriging and boosted regression trees models. The
127 proposed ISA model was tested to map SOC and STN in Montane ecosystems of Lushun City in
128 the northeastern coastal areas of China. Our research objectives to: (1) developed a GMMC
129 method to divide the study area into climate-vegetation landscape units; (2) model the effect of
130 environmental variables on SOC and STN variation; and (3) map the SOC and STN distribution
131 using ISA, GWR, RK, and BRT models and compared their performances.

Materials & Methods

Study area

134 The study area (38.72°–38.97°N, 121.08°–121.47°E) covers approximately 512 km^2 and is
135 located in Lushun City, Liaoning Province in the northeast coast of China (Figure 1). The area
136 lies in the southwest corner of Liaodong Peninsula and faces towards the Yellow Sea on the east
137 and Bohai Sea on the west. Therefore, this region presents typical continental and oceanic
138 climate characteristics forming unique climate-vegetation landscapes including coastal plain,
139 interior plain, medium-coverage grasslands, high-coverage grasslands, low-elevation forest, and
140 low-mountain shrub. The area is dominated by mountain landscape, which accounts for 53.1% of
141 the total area and is characterized by valleys and peaks, with an altitude ranging from 137 to 466
142 m above sea level. The mean annual temperature (MAT) in this region is nearly 10 °C, the
143 maximum and minimum temperatures are 27.5 °C in September and 8.2 °C in January. The
144 region experiences a total of 185 frost-free days with the mean annual precipitation (MAP)
145 ranging from 585 mm to 720 mm, 65–75% of which falls between June and September.
146 According to World Reference Base for Soil Resources (WRB), the dominant soil types are
147 Cambisols and Fluvisols (Schad et al. 2014).

148

149 (Figure 1 is about here)

150

Dataset

Field sample data

153 Almost 53% of the study area is under forests, and is densely covered with rivers and valleys
154 where the river system is rather complicated in a rugged terrain. To represent the spatial
155 characteristics of soil properties in such a complex landscape, we applied a stratified sampling
156 scheme following two steps.
157   First, we used a GMCC method to divide the study area into several typical climatic-
158 vegetation landscape units. The GMMC model is a hybrid model composed of the gaussian
159 mixture model (GMM) and the expectation maximization (EM) algorithm. The GMM model is

160   used to determine certain probability density functions (i.e., the probability of data points is
161   partitioned into each category) (Banfield et al., 1993) to achieve the division of datasets. The EM
162   algorithm is adopted to estimate the model parameters, whereas the BIC is used to optimize the
163   clustering results, and then the optimal model and the number of clusters are selected. The
164   clustering model is established in the R version 3.2.2 (R Development Core Team, 2013) using
165   the "mclust" package (Fraley & Raftery, 2002) for completion. Based on the pedogenetic
166   information of the study area, environmental factors such as elevation, MAT, MAP, and
167   normalized difference vegetation index (NDVI) were used as inputs (Zhu et al., 2008; Yang et al.
168   2013; Yang et al. 2016). The study area was divided into six climate-vegetation landscape units
169   (Figure 2). Within each climate-vegetation landscape, twenty to twenty-five sampling points
170   were identified at different landform positions considering local climate, soil type, elevation and
171   other environmental factors.
172   Although soil sampling in this study considered climate and vegetation properties of the
173   landscape for clustering, it was further verified by the local soil experts to determine whether the
174   units are typical and are representative of the study area. This data sampling was not based on
175   probability sampling strategy, thus it might underestimate or overestimate the SOC and STN
176   distribution in the region as reported in previous studies (Zhu, 1997; Yang L et al., 2016).
177   However, it was probably the best sampling option given that sampling was done in such a
178   densely forested terrain and with limited resources. Overall flowchart of the proposed
179   methodology is shown in Figure 3.
180
181   (Figure 2 is about here)
182   (Figure 3 is about here)
183
184   A total of 126 sample locations were established, and the geographic coordinates of each
185   point was recorded by a handheld global positioning system (GPS) (Table S1). From each
186   sample location, about 1 kg of topsoil (0-20 cm soil depth) sample was collected for laboratory
187   analysis. In the Testing & Analysis Center of Shenyang agricultural University, Shenyang,
188   Liaoning Provence, China, the litters were removed from samples. The samples were then air
189   dried, grinded, and passed through a 2-mm sieve. SOC and STN contents were measured by a
190   dry combustion method (Matejovic, 1993) using CN analyzer (Vario Max, Elementar Amerivas
191   Ins., Germany) . The soil depth considered in this study was limited to 0-20 cm because Wang et
192   al. (2017) found that in Liaoning Province, 69% of SOC and STN were stored in the topsoil (0-
193   30cm). It has been further confirmed by Liu et al. (2012) who reported about 43% of the SOC
194   stocks from the topsoil. In addition, 70% of all samples were randomly selected as training
195   dataset (n=88) and the rest 30% as independent verification dataset (n=38).
196   **Environmental variables**
197   Two topographic variables including slope gradient and topographic wetness index (TWI) were
198   used in addition to four other variables (elevation, MAP, MAT, and NDVI) for climate-
199   vegetation landscape partition (Table 1), to predict the spatial distribution of SOC and STN
200   content. The variables were collected from different sources and were converted to a raster grid

201 of 30m resolution. Measured data on SOC and STN, and all the predictors were brought into the
202 geographic information system (GIS) in a common projection system (Krasovsky_1940_Albers)
203 in ArcGIS 10.2 (ESRI Inc., USA) for further geospatial processing and analysis. Slope gradient
204 and TWI were derived from a 30-m resolution digital elevation model (DEM) captured from
205 Shuttle Radar Topography Mission (SRTM) (Farr and Kobrick, 2000). MAT and MAP as two
206 main climatic variables obtained from China Meteorological Data Service Center
207 (http://data.cma.cn/en) as 30-year annual average (1980-2010). In addition, we selected NDVI to
208 represent vegetation intensity, as determined by using two bands of Landsat 5, namely, band 3
209 and band 4 (Equation 1), and were downloaded from the United States Geological Survey
210 (USGS). The imagery covers from July to September of 2012 with <10% cloud cover. Image
211 processing included geometric correction, registration, image mosaic and clipping, cloud
212 removal, and shadow processing and was performed in ENVI 4.7 software (Huang et al., 2007).

213 $$NDVI = \frac{(band\,4 - band\,3)}{(band\,4 + band\,3)}$$ (1)

214

Model development
Improved similarity-based approach and its prediction consistency
217 An improved similarity-based approach was used to estimate the soil properties of unsampled
218 location by acquiring the environmental similarity with sampling points. Specifically, this
219 method was developed following three key steps:
220   Step 1 was to calculated the environmental similarity in each climate-vegetation landscape;

221 $$D = \sqrt{(S_1 - U_1)^2 + (S_2 - U_2)^2 + ... + (S_i - U_i)^2}$$ (2)

222   In this equation, D is Euclidean distance; Si and Ui are sampled point S and unsampled point
223 U, respectively, and the variable values in dimension i space. Suppose there are m sampling
224 points and n unsampled points, then a set of Euclidean distance (DE) matrix can be obtained as:

225 $$DE = \begin{bmatrix} D_{11}, & D_{21}, & ..., & D_{n1} \\ D_{12}, & D_{22} & ..., & D_{n2} \\ ... & & & \\ D_{1m} & D_{2m} & ..., & D_{nm} \end{bmatrix}$$

226 (3)

227   Euclidean distance only measures the degree of distance between two points. In order to
228 obtain the similarity between two points, it must be converted into the corresponding similarity.
229 We first applied equation (4) to obtain standard range of Euclidean distance (DE_{scale}) [0,1], and
230 then applied equation (5) to convert it into similarity (SE) with range [0, 1]
231

232 $$DE_{scale} = DE / \max(DE)$$ (4)

233 $$SE = 1 - DE_{scale}$$ (5)

234   So, the similarity matrix (*SE*) between the unsampled locations and the sampling points was
235   obtained using equation (6):

$$SE = \begin{bmatrix} S_{11}, & S_{21}, & ..., & S_{n1} \\ S_{12,} & S_{22} & ..., & S_{n2} \\ ... & & & \\ S_{1m} & S_{2m} & ..., & S_{nm} \end{bmatrix}$$

236   (6)

237   Step 2 was to predict soil properties for the landscape unit according to its environmental
238   similarity.
239   In each landscape, the soil property value of the unsampled locations was predicted by using
240   the environmental similarity where elevation, MAT, MAP, NDVI, TWI, and slope gradient were
241   used as predictors. The definite equation as follows:

242    $V_b = \sum_{a=1}^{m} S_{ab}.V_a / \sum_{a=1}^{m} S_{ab}$ (7)

243   where V_b is the predicted value of soil property at location b and V_a is the soil property value at
244   sampled location a; S_{ab} is the environmental similarity between location b and sampled location
245   a; m is the number of sampling points.
246   Step 3 was to calculate the prediction consistency using Equation (8):

247    $Consistency_m = 1 - \max(S_{1m}, S_{2m}, ..., S_{nm})$ (8)

248   where $Consistency_m$ is the prediction consistency at point m, and the lower the value, the better
249   the prediction consistency of the model; S_{nm} is the environmental similarity between point n and
250   sampled location m.
251   According to equation (8), the maximum value can be used to obtain the best representation
252   of the similarity of the environment between unsampled locations and sampling points (i.e., most
253   similar). With a low similarity, the sampling points cannot represent the unsampled locations
254   (Liu 2010) and the soil property values inferred from the sampling points will show a high
255   degree of speculative consistency.
256   **Geographically weighted regression**
257   Geographically weighted regression was first introduced into the study of geography by
258   Brunsdon et al. (1998). It embeds the spatial structure of the data into a regression model making
259   the regression parameters become a function of the geographical location of the observation
260   points. GWR is a non-parametric technique based on local weighted regression developed for
261   curve fitting and smoothing applications in statistics (Kumar et al., 2012). In this technique, local
262   regression parameters are estimated using a subset of data close to the estimated points of models
263   in variable spaces. The innovation of GWR is to use a subset of data in geographic space near the
264   calibration location of the model. In our GWR model, six environmental variables were used to
265   predict the spatial distribution of SOC and STN.
266   **Regression kriging**
267   Regression kriging is a spatial interpolation technique that combines a regression of the
268   dependent variable and auxiliary variables (such as terrain parameters, remote sensing imagery
269   and thematic maps) with kriging of the regression residuals (Odeh et al. 1995; Hengl et al. 2004).

270   We applied RK to interpolate the spatial distribution of SOC and STN following these five
271   steps: (1) determining the SOC and STN prediction model using multiple linear regressions
272   (MLR); (2) calculating the SOC and STN prediction model residuals at each calibration location;
273   (3) modeling the covariance structure of the SOC and STN residuals using a variogram GS+ 7.0
274   statistical software (Gamma Design Software, Plainwell, MI) was used to implement this
275   process; (4) spatially interpolating the SOC and STN residuals through the parameters of the
276   variogram model; and (5) adding the SOC and STN prediction model surface to the interpolated
277   residuals surface to get the final predicted map.

278   **Boosted regression trees**

279   Boosted regression trees model is a machine learning algorithm based on classification and
280   regression trees, which was proposed by Friedman (2000). The BRT model is similar to other
281   boosting models improving model performance by training multiple models and combining them
282   for prediction. It consists of two algorithms: regression trees and gradient boosting. In the model,
283   gradient lifting algorithm is used to linearly combine multiple regression trees of weak
284   regression to form an efficient and a strong regression model (e.g., Wang et al, 2016; Yang et al,
285   2016). The implementation of BRT model requires users to define four parameters: learning rate
286   (LR), tree complexity (TC), bag fraction (BF) and tree number (NT). LR represents the
287   contribution of each tree in the model to the final fitting model (Yang et al., 2016). TC is a direct
288   predictor of tree depth and maximum interaction level (Yang et al., 2015). BF represents the
289   scale of data used in each model (Wang et al., 2018). NT is determined by LR and TC. In order
290   to obtain the best prediction performance of BRT model, different parameter combinations LR
291   (0.0025, 0.025, 0.25, 0.50), TC (3, 6, 9), BF (0.20, 0.35, 0.50, 0.65), NT (600, 800, 1000, 1200)
292   were tested by 10-fold cross-validation. Finally, LR, TC, BF and NT values that achieved the
293   minimum prediction error through 10-fold cross-validation were set to 0.0025, 6, 0.65, and 800,
294   respectively, for SOC prediction, and 0.025, 6, 0.65 and 1000, respectively, for STN prediction.

295   **Model Validation**

296   In each landscape, 30% sampling points (38 observations in total) were randomly selected to test
297   the prediction performance of the ISA, GWR, RK, and BRT methods. Three commonly-used
298   indices including mean absolute prediction error (MAE), root mean square error (RMSE),
299   coefficient of determination (R^2) , and maximum relative difference (RD) (Equation 9 to 11)
300   were used to compare the model. All indices were calculated in the R version 3.2.2 (R
301   Development Core Team, 2013):

302

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - b_i|$$

(9)

303

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)^2}$$

(10)

304

$$R^2 = \frac{\sum_{i=1}^{n}(a_i - \bar{b}_i)^2}{\sum_{i=1}^{n}(b_i - b_i)^2}$$

(11)

$$RD = \frac{\bar{a}_i - \bar{b}_i}{\bar{a}_i}$$

305 (12)

306 where a_i, b_i, $\bar{a}_i$, and $\bar{b}_i$ are the observed values, predicted values, and mean values of soil property

307 at site i, and n is the number of the sampling point.

Results

Exploratory Data Analysis

310 Exploratory analysis of measured SOC and STN contents and values of environmental variables

311 at sampling locations are summarized in Table 1. Average SOC and STN in the topsoil were (±

312 3.78), and 0.85 (± 0.23) g kg^{-1}, respectively. The skewness and kurtosis coefficients were 1.07

313 and 0.06 for SOC and 0.81 and -0.31 for STN, respectively, indicating that the measured SOC

314 and STN approximately followed a normal distribution. In addition, both distribution passed

315 Kolmogorov–Smirnov (K-S) test (ρ = 0.11 and 0.09), respectively, suggesting that SOC and

316 STN data did not need to be transformed for subsequent analysis and modeling.

317

318   (Table 1 is about here)

319

320   SOC and STN were positively correlated with NDVI, MAP, elevation, and slope gradient but

321 negatively correlated with MAT and TWI (Figure 4). We observed multicollinearity among

322 environmental variables and believed that predicting SOC and STN with these variables using

323 traditional statistical methods such as simple multiple regression equations would be unreliable

324 (McBratney et al. 2003; Yang et al. 2016) and using ISA model could effectively overcome this

325 problem. In the process of GWR modeling, we eliminated the slope gradient variable as it was

326 highly correlated with MAT and elevation, to avoid the problem of multicollinearity, and used

327 the remaining five variables for spatial prediction.

328

329   (Figure 4 is about here)

Climate-vegetation landscape

331 SOC and STN usually vary in the long term and are influenced by various environmental factors

332 (i.e., rainfall, temperature, and vegetation). As a typical coastal ecosystem in Northeast China,

333 Lushun has both continental and marine climate characteristics with an absence of cold winter

334 and hot summer. Therefore, we selected the elevation, MAT, MAP, and NDVI to divide the

335 region into the climate-vegetation landscape units. The results showed that the study area can be

336 divided into six climate-vegetation landscapes (BIC, -678247.3) (Figure 2). Landscapes 1, 2 are

337 mainly distributed in the northern and western region of the study area; Landscapes 3 and 4 are

338 mainly for the central region; and Landscapes 5, 6 are mainly distributed in the northeastern and

339 southern regions. In terms of the area distribution area of landscapes, landscape 5 accounting for

340 25% of the total area of the study area, followed by landscape 3 (21%), landscape 2 (18%),

341 landscape 4 (18%), landscape 1 (12%), and landscape 6 (6%) (Figure 2).

342   The distribution of climate-vegetation landscapes and soil observations by elevation, MAP,

343 MAT, and NDVI is shown in Figure 5. The statistical characteristics of the four environmental

344 variables at sampling points were similar to those in the landscapes, indicating that the sampling
345 points could efficiently describe the characteristics of the major environmental factors in the
346 study area. However, some differences in the main environmental variables of each landscape
347 were observed. The Landscape 1 showed the lowest average elevation, a low MAP (608 mm),
348 and the highest MAT (10.6 °C), and the average NDVI value approached 0.4. Therefore, the
349 landscape type was coastal plain. Compared with landscape 1, landscape 2 was distributed at a
350 relatively low elevation but presented a better vegetation cover; therefore, the landscape type was
351 named interior plain. Landscape 6 showed the highest elevation (210 m), lowest MAT (10.0 °C),
352 and lowest value of NDVI; thus, the landscape type was low-mountain shrub. Landscapes 3, 4,
353 and 5 were mainly distributed in the middle elevation, but their MAT and NDVI values were
354 different. Landscapes 3 and 4 showed higher precipitation rates; however, given the limitation of
355 temperature, low vegetation cover was observed, thus landscapes 3 and 4 were classified as
356 medium-coverage grasslands and high-coverage grasslands. Landscape 5 presented a better
357 vegetation cover, and was classified as low-elevation forest.
358
359   (Figure 5 is about here)
360
361 **Model Performance**
362 In order to determine whether the study area should be divided into unique landscape units, and
363 then use the best prediction model to predict the spatial distribution of SOC and STN in each
364 landscape, the MAE, RMSE, and R^2 were compared between the observed and predicted values
365 (Table 2). The results showed that the model was significantly improved by dividing the study
366 area into climate-vegetation landscapes (Table 2).
367
368   (Table 2 is about here)
369
370   To obtain the best prediction model, four models (ISA, GWR, RK, and BRT) were compared
371 using 38 independent validation points. Results showed that the ISA model outperformed GWR,
372 RK, and BRT models with a lower MAE and RMSE, and a higher R^2 values. We also found that
373 the performance of BRT was better than the GWR and RK models. ISA model was efficient and
374 powerful in spatial prediction of SOC, and STN among six climatic vegetation landscape units
375 based on the accuracy verification. The prediction accuracy of ISA model in each landscape is
376 listed in Table 3. The mean predicted values of SOC and STN in the topsoil were comparable to
377 the mean observed values of the sampling points (Table 3). RD was obtained in the interior plain
378 (RD about 7.1% and 4.2%), followed by the coastal plain (RD about 3.4% and 3.3%) and low-
379 mountain shrub (RD about -4.2% and -4.1%), and the RD values of the other landscapes were
380 smaller (Table 3). The predicted R^2 ranged from 0.37 (Low-mountain shrub) to 0.77 (Interior
381 plain) for SOC and 0.39 (Low-mountain shrub) to 0.89 (Median-coverage grassland) for STN,
382 respectively. The lowest RMSE (0.13 g kg^{-1}) in SOC prediction was reported for Medium-
383 coverage grassland and the highest (2.53 g kg^{-1}) for Low-elevation forest. For the STN, RMSE

384 ranged from 0.01 to 0.06 depending on the landscape types. Overall, the lower MAE and RMSE,
385 and higher R^2 value for ISA model compared to the rest of the models indicated that the ISA
386 model could better predict SOC and STN distribution.
387
388   (Table 3 is about here)
389
390   For SOC prediction using the ISA model, MAE, RMSE, and R^2 values were 0.74, 1.01 g kg^{-1},
391 and 0.76, respectively, and for STN, the values were 0.03, 0.04 g kg^{-1}, and 0.83, respectively
392 (Figure 6). The ISA model overestimated the SOC and STN contents in the study area, but the
393 overall accuracy was high. The model explained approximately 76% and 83% of the total SOC
394 and STN variability in the study area, respectively. Predictive consistency maps of SOC and
395 STN were obtained using Equation 8. The higher the consistency value, the larger the difference
396 between the predicted value and measured value. Mean prediction consistency values were 0.14
397 for SOC and 0.13 for STN, respectively (Figure 7a, 7b), suggesting the ISA model had a reliable
398 predictability.
399
400   (Figure 6 is about here)
401   (Figure 7 is about here)
402
403 **Estimates of SOC and STN**
404 SOC and STN distribution was mapped using the ISA model where an average SOC and STN
405 content in the study area was 11.37 (±3) and 0.89 (±0.2) g kg^{-1}, respectively (Figure 7c, 7d). In
406 terms of spatial pattern, ISA model was more excellent than other models (Figure S1). Among
407 the climate-vegetation landscape units, soils under low-elevation forests contained more SOC
408 and STN than the rest of the landscape units (Table 4). Of all predictors, elevation was the main
409 predictor affecting the spatial distribution of SOC and STN in the study area (Figure 8).
410
411   (Table 4 is about here)
412   (Figure 8 is about here)
413
414 # Discussion
415 **Effects of ISA model on SOC and STN**
416 The study area divided into unique landscape units, and then used the best prediction model to
417 predict the spatial distribution of SOC and STN in each landscape. The results showed that the
418 model was significantly improved by dividing the study area into climate-vegetation landscapes.
419 This is consistent with previous studies using the same method. For instance, in Belgium, Lettens
420 et al. (2004) divided the whole study area into 289 landscape units and predicted the SOC stocks
421 for each landscape unit. They considered that SOC stocks were continuously influenced by a
422 number of external factors, mainly land-use history and current land management and climate.
423 The spatial distribution of SOC and STN varies closely with climate-vegetation-dominated

424 landscapes. By dividing a large-scale complex landscape into small units, the SOC and STN
425 could be better predicted. At the same time, compared to other SOC and STN prediction studies
426 in the coastal area, the ISA model we constructed performed better. For example, in Santa Cruz
427 Island, Galapagos, Rial et al. (2017) used terrain, climate and remote sensing variables combined
428 with a GWR method to predict topsoil SOC stocks, and their model could explain 66% of SOC
429 stocks variation in the region. In a separate study, Wang et al. (2018b) compared GWR and RK
430 models to predict topsoil SOC in the Northeast coastal area of China and found GWR a better
431 model that could explain 78% and 80% SOC variation during 1982 and 2012, respectively.
432   The ISA model could explain approximately 76% and 83% of the total SOC and STN
433 variability in the study area, respectively. In comparison with previous studies, we observed a
434 better prediction performance of the ISA method. Yang et al. (2016) developed a BRT model
435 that explained 50 to 58% of the SOC variability in the Qinghai Tibet Plateau, China. Using a
436 GWR approach, Wang et al. (2013) captured approximately 57% of the STN variability in Fujian
437 Province, China. In the Medinipur Block, Paschim Medinipur District and West Bengal in India,
438 Bhunia et al. (2019) used remote sensing techniques and multivariate regression model, which
439 explained 71% of SOC variability. Xu et al. (2018) used six different remote sensing spectral
440 indices of RK models to predict STN in two smallholder villages, Kothapally and Masuti in
441 South India, which helps explain 59% of the STN variability in the region.
442   Overall, the ISA model prediction consistency was higher towards the eastern part of the
443 study area compared to the rest area for both SOC and STN predictions. Although there were
444 other sources of uncertainty and error in mapping SOC and STN, including sampling error,
445 laboratory analysis errors, and low predictor precision (Wang et al. 2017; McBratney et al. 2003;
446 Yang et al. 2016; Grimm et al. 2008), quantification of these inevitable errors were not
447 considered in this study.
448 **Predicted distribution of SOC and STN content**
449 In general, the spatial distribution pattern for SOC and STN was similar and varies with different
450 climate-vegetation landscape units (Figure 3, 5, and 7). Spatially, higher SOC and STN were in
451 the northeastern and southern regions that were dominated by low-elevation forest and low-
452 mountain shrub. Coastal plain has the lowest SOC and STN. Overall, the lowest SOC and STN
453 contents were found in the northern and central regions with low MAT and vegetation coverage.
454 These results are consistent with the findings of Jobbagy & Jackson (2000) and Ding et al. (2018)
455 who reported for humid and rainy areas. Soil particles in high-elevation areas were easy to lose
456 and thick soil layers were formed in low-elevation areas, which were conducive to the
457 accumulation of SOC and STN. Of the six climate-vegetation landscape units, soil under coastal
458 plain had the lowest SOC and STN as also reported in Causarano et al. (2008). The low SOC and
459 STN in these landscape units might be caused by increased organic matter decomposition and
460 erosion and tillage losses (Burke et al., 1989; Lettens et al., 2004; Adhikari and Hartemink,
461 2015).
462   We observed that with increasing elevation, this is true for forest pixels (Figure 8). The
463 effect of elevation on SOC and STN has been examined in a number of studies (e.g., Kieft 1998;

464 Kunkel et al. 2011; Minasny et al., 2013; Adhikari et al. 2013; Chen et al. 2015; Minasny et al.,
465 2017; Liang et al., 2019). Wang et al. (2017) reported that SOC and STN significantly increased
466 with elevation. Differences in elevation gradients may have affected the input and loss of soil
467 carbon and nitrogen mainly through indirect factors along the gradient, such as precipitation and
468 temperature (Minasny et al., 2013; Elbasiouny et al. 2014; Were et al. 2015; Adhikari &
469 Hartemink 2015).
470   The predicted distribution of SOC and STN showed a discontinuous pattern (Figure 7 c and
471 7d). We found that the abrupt change in SOC and STN contents were located at the junction of
472 landscapes. Such a pattern was due to the method of landscape division leading to an abrupt
473 change of different landscape boundaries (Jaeger 2000; Peilin et al. 2010). The apparent
474 partitioning between landscapes led to spatially discontinuous distribution of SOC and STN,
475 resulting in may lead to certain unreasonable predictions (Rosenbloom et al. 2006). This pattern
476 may be attributed to the evident transition between the two landscapes, but the transition zone
477 was smaller than the pixel resolution (Belshe et al. 2012).
478 **Limitation of the study**
479 Although the ISA model constructed in this study had a good performance in predicting SOC
480 and STN content, there were still some limitations. Firstly, we used a soil-landscape model to
481 collect soil samples, which is idiosyncratically suitable for the ISA model. An independent
482 probability sampling method would be required in our subsequent research. Secondly, this study
483 area was divided into several typical climate-vegetation landscape units and then applied the
484 model to simulate their SOC and STN. This treatment ignored the gradual transition between
485 landscape units with respect to SOC and STN, which might have biased our regional prediction.
486 In the future, the model should be revised to avoid this bias. Thirdly, because the study area was
487 dominated by mountains, which were easily affected by the terrain and clouds, the high-altitude
488 areas were prone to produce shadows in the process of image segmentation, resulting in large
489 reflectivity errors of satellite image data. This might have also introduced prediction errors.
490 Finally, this research was limited to topsoil (0-20 cm) SOC and STN content, which might have
491 led to underestimating SOC and STN due to a large amount of SOC and STN typically stored in
492 deeper soil layers in the region.
493 **Conclusions**
494 This study advanced an ISA model to predict the spatial distribution of SOC and STN in the
495 Northeast coastal area of China. The ISA model incorporated a GMMC method to divide the
496 study area into six climatic-vegetation landscapes first. The six landscapes were then modeled
497 using a t theory of environmental similarity. Comparing the prediction of GWR, RK, and BRT
498 models, we found that the ISA model was the most robust and effective method in mapping SOC
499 and STN, which explained 78% and 83% of the spatial variation of SOC and STN in the study
500 area, respectively. Of the six climate-vegetation landscape units, the soils under low-elevation
501 forest had the highest level of SOC and STN content than soils under the rest of the landscape
502 units including coastal plain, interior plain, medium-coverage grasslands, low-mountain shrub,
503 and high-coverage grasslands. Topography is the main driving force of SOC and STN

504 distribution in the Northeast coastal area. Therefore, we suggested that terrain variables should
505 be included in future SOC and TN mapping studies, especially in coastal and hilly areas of China.
506 Overall, the improved ISA method better predicted SOC and STN in the region. Our predicted
507 SOC and STN distribution shall provide important information for soil and water conservation,
508 ecological restoration, environmental management and agricultural production planning in the
509 Northeast coastal agricultural region.
510

511 **Acknowledgements**
512

513 **References**

514 **Adhikari K, Hartemink AE. 2015.** Digital mapping of topsoil carbon content and changes in
515 the Driftless Area of Wisconsin, USA. *Soil Science Society of America Journal* 79(1):155-164.
516 **Adhikari K, Hartemink AE, Minasny B, Kheir RB, Greve MB, Greve MH. 2014.** Digital
517 mapping of soil organic carbon contents and stocks in Denmark. *PloS one* 9(8): e105519.
518 **Adhikari K, Kheir RB, Greve MB, Greve MH. 2013.** Comparing kriging and regression
519 approaches for mapping soil clay content in a diverse Danish landscape. *Soil Science* 178(9):
520 505-517.
521 **Adhikari K, Owens PR, Ashworth AJ, Sauer TJ, Libohova Z, Richter JL, Miller DM. 2018.**
522 Topographic controls on soil nutrient variations in a silvopasture system. *Agrosystems,*
523 *Geosciences & Environment* 1(1):1-15.
524 **An YM, Yang L, Zhu AX, Qin CZ, Shi JJ. 2018.** Identification of representative samples from
525 existing samples for digital soil mapping. *Geoderma* 311:109-119.
526 **Banfield JD, Banfield AE. 1993.** Model-based Gaussian and non-Gaussian clustering.
527 *Biometrics* 803-821.
528 **Batjes NH. 1996.** Total carbon and nitrogen in the soils of the world. *European journal of soil*
529 *science* 47:151-163.
530 **Belshe EF, Schuur EAG, Bolker BM, Bracho R. 2012.** Incorporating spatial heterogeneity
531 created by permafrost thaw into a landscape carbon estimate. *Journal of Geophysical*
532 *Research: Biogeosciences* 117: (G1).
533 **Bhunia GS, Kumar Shi P, Pourghasemi HR. 2019.** Soil organic carbon mapping using remote
534 sensing techniques and multivariate regression model. *Geocarto International* 34(2): 215-226.
535 **Brunsdon C, Fotheringham S, Chariton M. 1998.** Geographically weighted regression-
536 modeling spatial non-stationarity. *The Statistician* 47:431-443.
537 **Burke IC, Yonker CM, Parton WJ, Cole CV, Schimel DS, Flach K. 1989.** Texture, climate,
538 and cultivation effects on soil organic matter content in US grassland soils. *Soil Science*
539 *Society of America Journal* 53(3):800-805.
540 **Causarano, HJ, Franzluebbers AJ, Shaw JN, Reeves DW, Raper RL, Wood C. 2008.** Soil
541 organic carbon fractions and aggregation in the Southern Piedmont and Coastal Plain. *Soil*
542 *Science Society of America Journal* 72(1): 221-230.

543 **Chen C, Hu K, Li H, Yun A, Li B. 2015.** Three-dimensional mapping of soil organic carbon by
544   combining kriging method with profile depth function. *PloS one* 10(6):e0129038.
545 **Clement F, Orange D, Williams M, Mulley C, Epprecht M. 2009.** Drivers of afforestation in
546   Northern Vietnam: assessing local variations using geographically weighted regression.
547   *Applied Geography* 29: 561-576.
548 **Conrad O., Bechtel B., Bock M., Dietrich H., Fischer E., Gerlitz L., Wehberg J., Wichmann**
549   **V., Böhner J., 2015.** System for automated geoscientific analyses (SAGA) v. 2.1.4.
550   *Geoscientific Model Development* 8 (7), 1991–2007.
551 **Danielsson PE. 1980.** Euclidean distance mapping. *Computer Graphics and image processing*
552   14(3):227-248.
553 **Davidson EA, Janssens IA. 2006.** Temperature sensitivity of soil carbon decomposition and
554   feedbacks to climate change. *Nature* 440:165-173
555 **Dempster AP, Laird NM, Rubin DB. 1977.** Maximum Likelihood from Incomplete Data Via
556   Em Algorithm. *Journal of the Royal Statistical Society* 39:1-38.
557 **Ding J., Yang A., Wang J., Sagan V., Yu D. 2018.** Machine-learning-based quantitative
558   estimation of soil organic carbon content by VIS/NIR spectroscopy. *PeerJ* 6, e5714.
559 **Elbasiouny H, Abowaly M, Abu_Alkheir A, Gad A. 2014.** Spatial variation of soil carbon and
560   nitrogen pools by using ordinary Kriging method in an area of north Nile Delta, Egypt. *Catena*
561   113: 70-78.
562 **Farr TG, Kobrick M. 2000.** Shuttle Radar Topography Mission produces a wealth of data. *Eos,*
563   *Transactions American Geophysical Union* 81(48): 583-585.
564 **Fraley C, Raftery AE. 2002.** Model-based clustering, discriminant analysis, and density
565   estimation. *Journal of the American statistical Association* 97:611-631
566 **Friedman J, Hastie T, Tibshirani R. 2000.** Additive logistic regression: a statistical view of
567   boosting. *The Annals of Statistics* 28(2):337-407.
568 **Fuchs H, Magdon P, Kleinn C, Flessa H. 2009.** Estimating aboveground carbon in a catchment
569   of the Siberian forest tundra: Combining satellite imagery and field inventory. *Remote Sensing*
570   *of Environment* 113(3):518-531.
571 **Garten Jr CT, Ashwood TL. 2002.** Landscape level differences in soil carbon and nitrogen:
572   implications for soil carbon sequestration. *Global Biogeochemical Cycles* 16(4): 61-1.
573 **Grimm R, Behrens T, Märker M, Elsenbeer H. 2008.** Soil organic carbon concentrations and
574   stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis.
575   *Geoderma* 146(1):102-113.
576 **Hengl T, Heuvelink G, Stein A. 2004.** A Generic Framework for Spatial Prediction of Soil
577   Variables Based on Regression Kriging. *Geoderma* 122(1-2):75-93.
578 **Huang H, Cheng X, Li B, Liu Y. 2007.** Application of High-Resolution Remote Sensing Image
579   Using ENVI. *Geospatial Information* 3.
580 **Hudson BD. 1992.** The soil survey as paradigm-based science. *Soil Science Society of America*
581   *Journal* 56: 836-841.

582 **Lettens S, Van Orshoven J, van Wesemael B, Muys B. 2004.** Soil organic and inorganic
583   carbon contents of landscape units in Belgium derived using data from 1950 to 1970. *Soil Use*
584   *and Management* 20(1): 40-47.
585 **Jaeger JAG. 2000.** Landscape division, splitting index, and effective mesh size: new measures
586   of landscape fragmentation. *Landscape ecology* 15(2):115-130.
587 **Jenny H. 1941.** Factors of Soil Formation. McGraw-Hill, New York.
588 **Jobbagy EG, Jackson RB. 2000.** The vertical distribution of soil organic carbon and its relation
589   to climate and vegetation. *Ecological applications* 10: 423-436.
590 **Kieft TL, White CS, Loftin SR, Aguilar R, Craig J, Skaar DA. 1998.** Temporal dynamics in
591   soil carbon and nitrogen resources at a grassland–shrubland ecotone. *Ecology* 79(2): 671-683.
592 **Kovačević M, Bajat B, Gajić B. 2010.** Soil type classification and estimation of soil properties
593   using support vector machines. *Geoderma* 154(3):340-347.
594 **Kumar S., Lal R., Liu, D. 2012.** A geographically weighted regression kriging approach for
595   mapping soil organic carbon stock. *Geoderma*, 189, 627-634.
596 **Kunkel ML, Flores AN, Smith TJ, McNamara JP, Benner SG. 2011.** A simplified approach
597   for estimating soil carbon and nitrogen stocks in semi-arid complex terrain. *Geoderma*
598   165(1):1-11.
599 **Lal R. 2004.** Soil carbon sequestration impacts on global climate change and food security.
600   *Science* 304 (5677):1623-1627 (2004).
601 **Liu F, Rossiter D-G, Song XD, Zhang GL, Yang RM, Zhao YG, Ju B. 2016.** A similarity-
602   based method for three-dimensional prediction of soil organic matter concentration. *Geoderma*
603   263: 254-263.
604 **Liu J. 2010.** Mapping Soil Properties Using Individual Representativeness of Samples over
605   Large Area. Beijing, Beijing Normal University. (in Chinese).
606 **Matejovic I. 1993.** Determination of carbon, hydrogen, and nitrogen in soils by automated
607   elemental analysis (dry combustion method). *Communications in Soil Science and Plant*
608   *Analysis* 24(17-18): 2213-2222.
609 **McBratney AB, Santos MLM, Minasny B. 2003.** On digital soil mapping. *Geoderma* 17: 3-52.
610 **Minasny B., Malone B.P., McBratney A.B., Angers D.A., Arrouays D., Chambers A., Field**
611   **D.J. 2017.** Soil carbon 4 per mille. *Geoderma* 292, 59-86.
612 **Minasny B, McBratney AB. 2008.** Regression rules as a tool for predicting soil properties from
613   infrared reflectance spectroscopy. *Chemometrics and intelligent laboratory systems* 94(1): 72-
614   79.
615 **Minasny B., McBratney A.B., Malone B.P., Wheeler, I. 2013.** Digital mapping of soil carbon.
616   In Advances in Agronomy. *Academic Press* 118, 1-47.
617 **Mondal A, Khare D, Kundu S, Mondal S, Mukherjee S, Mukhopadhyay A. 2017.** Spatial
618   soil organic carbon (SOC) prediction by regression kriging using remote sensing data. *The*
619   *Egyptian Journal of Remote Sensing and Space Science* 20(1), 61-70.
620 **Moore ID, Gessler PE, Nielsen GAE, Peterson GA. 1993.** Soil attribute prediction using
621   terrain analysis. *Soil science society of america journal* 57(2): 443-452.

622 **Mulder VL, Lacoste M, Richer-de-Forges AC. 2016.** National versus global modelling the 3D
623   distribution of soil organic carbon in mainland France. *Geoderma* 263: 16-34.
624 **Odeh IOA, McBratney AB, Chittleborough DJ. 1995.** Further Results on Prediction of Soil
625   Properties from Terrain Attributes: Heterotopic Cokriging and Regression-Kriging. *Geoderma*
626   67:215-226.
627 **Peilin L, Chunla L, Yunyuan D, Xiuying SHEN, Bohua LI, Zui HU. 2010.** Landscape
628   division of traditional settlement and effect elements of landscape gene in China. *Acta*
629   *Geographica Sinica* 65(12): 1496-1506.
630 **Post WM, Kwon KC. 2000.** Soil carbon sequestration and land-use change: processes and
631   potential. Global change biology 6:317-327.
632 **R Development Core Team. 2013.** R: A language and environment for statistical computing.
633   Vienna, Austria; R Foundation for Statistical Computing.
634 **Rial M, Cortizas AM, Taboada T, Rodríguez-Lado L. 2017.** Soil organic carbon stocks in
635   Santa Cruz Island, Galapagos, under different climate change scenarios. *Catena* 156, 74-81.
636 **Rosenbloom NA, Harden JW, Neff JC, Schimel DS. 2006.** Geomorphic control of landscape
637   carbon accumulation. *Journal of Geophysical Research: Biogeosciences* 111: (G1).
638 **Scull P., Franklin J., Chadwick O.A., McArthur D. 2003.** Predictive soil mapping: a review.
639   *Progress in Physical Geography* 27(2): 171-197.
640 **Schad P, Van Huyssteen C, Micheli E. 2014.** International Soil Classification System for
641   Naming Soils and Creating Legends for Soil Maps; World Reference Base for Soil Resources:
642   Rome, Italy. 106:246-385.
643 **Shi X, Zhu AX, Burt JE, Qi F, Simonson D. 2004.** A case-based reasoning approach to fuzzy
644   soil mapping. *Soil Science Society of America Journal* 68(3):885-894.
645 **Song XD, Brus DJ, Liu F, Li DC, Zhao YG, Yang JL, Zhang GL. 2016.** Mapping soil
646   organic carbon content by geographically weighted regression: A case study in the Heihe
647   River Basin, China. *Geoderma* 261: 11-22.
648 **Stevens A, Udelhoven T, Denis A, Tychon B, Lioy R, Hoffmann L, Van Wesemael B. 2010.**
649   Measuring soil organic carbon in croplands at regional scale using airborne imaging
650   spectroscopy. *Geoderma* 158(1):32-45.
651 **Taghizadeh-Mehrjardi R, Minasny B, Sarmadian F, Malone BP. 2014.** Digital mapping of
652   soil salinity in Ardakan region, central Iran. *Geoderma* 213:15–28.
653 **Tesfaye MA, Bravo F, Ruiz-Peinado R, Pando V, Bravo-Oviedo A. 2016.** Impact of changes
654   in land use, species and elevation on soil organic carbon and total nitrogen in Ethiopian
655   Central Highlands. *Geoderma* 261:70-79
656 **Wang K, Zhang C, Li W. 2013.** Predictive mapping of soil total nitrogen at a regional scale: a
657   comparison between geographically weighted regression and cokriging. *Applied Geography*
658   42:73-85.
659 **Wang S, Adhikari K, Wang Q, Jin X, Li, H. 2018a.** Role of environmental variables in the
660   spatial distribution of soil carbon (C), nitrogen (N), and C: N ratio from the northeastern
661   coastal agroecosystems in China. *Ecological indicators* 84: 263-272.

662 **Wang S, Wang Q, Adhikari K, Jia S, Jin X, Liu H. 2016.** Spatial-Temporal Changes of Soil
663   Organic Carbon Content in Wafangdian, China. *Sustainability*. 8(11):1154.

664 **Wang S, Zhuang Q, Wang Q, Jin X, Han C. 2017.** Mapping stocks of soil organic carbon and
665   soil total nitrogen in Liaoning Province of China. *Geoderma* 305: 250-263.

666 **Wang S, Zhuang Q, Jia S, Jin X, Wang Q. 2018b.** Spatial variations of soil organic carbon
667   stocks in a coastal hilly area of China. Geoderma, 314, 8-19.

668 **Were K, Bui DT, Dick ØB, Singh BR. 2015.** A comparative assessment of support vector
669   regression, artificial neural networks, and random forests for predicting and mapping soil
670   organic carbon stocks across an Afromontane landscape. *Ecological indicators* 52: 394-403.

671 **Xu Y, Smith SE, Grunwald S, Abd-Elrahman A, Wani SP, Nair VD. 2018.** Estimating soil
672   total nitrogen in smallholder farm settings using remote sensing spectral indices and regression
673   kriging. *Catena* 163: 111-122.

674 **Yang L, Huang C, Liu G, Liu J, Zhu A-X. 2015.** Mapping soil salinity using a similarity-based
675   prediction approach: a case study in Huanghe River Delta, China. *Chinese geographical*
676   *science* 25(3):283-294.

677 **Yang L, Qi F, Zhu AX, Shi JJ, An YM. 2016.** Evaluation of Integrative Hierarchical Stepwise
678   Sampling for Digital Soil Mapping. *Soil Science Society of America Journal* 80 (3):637-651.

679 **Yang L, Zhu AX, Qi F, Qin CZ, Li BL, Pei T. 2013.** An integrative hierarchical stepwise
680   sampling strategy and its application in digital soil mapping. *International Journal of*
681   *Geographical Information Science* 27(1): 1-23.

682 **Yang RM, Rossiter DG, Liu F, Lu, Y, Yang F, Yang F. 2015.** Predictive mapping of topsoil
683   organic carbon in an Alpine environment aided by Landsat TM. *PloS one* 10(10):e0139042.

684 **Yang RM, Zhang GL, Liu F, Lu YY, Yang F, Yang F, Li DC. 2016.** Comparison of boosted
685   regression tree and random forest models for mapping topsoil organic carbon concentration in
686   an alpine ecosystem. *Ecological indicators* 60: 870-878.

687 **Zhao MS, Zhang GL, Wu YJ, Li DC, Zhao Y-G. 2015.** Driving forces of soil organic matter
688   change in Jiangsu Province of China. *Soil use and management* 31(4):440-449.

689 **Liang Z, Chen S, Yang Y, Zhou Y, Shi Z. 2019**. High-resolution three-dimensional mapping of
690   soil organic carbon in China: Effects of SoilGrids products on national modeling. *Science of*
691   *The Total Environment* 685: 480-489.

692 **Zhu AX, Liu J, Qin CZ, Zhang SJ**, **Chen YN, Ma XW. 2010.** Soil property mapping over
693   large areas using sparse ad-hoc samples. *19th World Congress of Soil Science, Soil Solutions*
694   *for a ChangingWorld. the International Union of Soil Science, Brisbane, Australia* (1–6
695   August 2010).

696 **Zhu AX, Yang L, Li B, Qin C, English E, Burt JE, Zhou C. 2008.** Purposive sampling for
697   digital soil mapping for areas with limited data. In: Hartemink, A.E. (Ed.), Digital Soil
698   Mapping With Limited Data. Springer 33-245.

699 **Zhu AX. 1997.** A similarity model for representing soil spatial information. *Geoderma* 77:217-
700   242.

Table 1

Table 1. Summary statistics of SOC, STN, and environmental variables at sampling points.

1   **Tables**

2   **Table 1.** Summary statistics of SOC, STN, and environmental variables at sampling points

Property	Description	Units	Minimum	Mean	Maximum	SD	Skewness	Kurtosis
SOC	Level of organic carbon content in soil	g kg^{-1}	8.50	12.95	22.02	3.78	1.07	0.06
STN	Sum of various forms of nitrogen in soil	g kg^{-1}	0.57	0.85	1.49	0.23	0.81	-0.31
NDVI	The difference between near infrared and red band reflectance is divided by the sum of their	index	0.15	0.40	0.56	0.08	-0.25	0.15
MAT	Mean annual temperature from 1980 to 2010	Celsius degree	9.5	10.39	10.91	0.26	-1.29	1.80
MAP	Mean annual precipitation from 1980 to 2010	mm	605.8	609.2	622.6	3.97	1.54	1.95
Elevation	Absolute vertical distance to geoid	m	1	66.53	309	61.48	1.67	2.87
Slope gradient	Maximum rate of change between cells and neighbors	degree	0	9.46	41.90	9.78	1.33	1.29
TWI	Calculates slope and specific catchment area based topographic wetness index	index	2.33	5.04	10.13	1.87	0.89	0.52

3   Note: SOC, soil organic carbon; STN, soil total nitrogen; TWI, topographic wetness index; MAP, mean annual precipitation; MAT, mean annual temperature; NDVI: Normalized

4   Difference Vegetation Index; and SD, Standard deviation.

Table 2

Table 2. Comparison of the performances of ISA, GWR, RK, and BRT models using MAE, RMSE, and R^2 with validation Data in predicting SOC and STN.

1   **Table 2.** Comparison of the performances of ISA, GWR, RK, and BRT models using MAE, RMSE, and

2   R^2 with validation Data in predicting SOC and STN.

Property	Category	Item	MAE	RMSE	R^2
SOC (g kg $^{-1}$)	Classified	ISA	0.74	1.01	0.76
		GWR	0.96	2.35	0.57
		RK	1.12	2.87	0.48
		BRT	0.87	1.68	0.64
	Unclassified	ISA	0.98	1.34	0.54
		GWR	1.17	2.67	0.46
		RK	1.27	2.97	0.41
		BRT	1.05	2.13	0.51
STN (g kg $^{-1}$)	Classified	ISA	0.03	0.04	0.83
		GWR	0.08	0.09	0.54
		RK	0.11	0.15	0.43
		BRT	0.06	0.07	0.67
	Unclassified	ISA	0.12	0.09	0.57
		GWR	0.13	0.16	0.41
		RK	0.15	0.19	0.38
		BRT	0.11	0.13	0.51

3   Note: SOC, soil organic carbon; STN, soil total nitrogen; MAE, mean absolute prediction error; RMSE, root mean square error;

4   R^2, coefficient of determination; GWR, geographically weighted regression; and RK, regression kriging (RK); BRT, boosted

5   regression trees.

Table 3

Table 3. Statistics of prediction performance and mean prediction results of each climate-vegetation landscapes by ISA model.

1 **Table 3.** Statistics of prediction performance and mean prediction results of each climate-vegetation landscapes by ISA model

Property	Index	Unit	Climate-vegetation landscapes					
			Coastal plain	Interior plain	Medium-coverage grasslands	High-coverage grasslands	Low-elevation forest	Low-mountain shrub
SOC	Observed	(g kg^{-1})	9.11	11.04	11.29	12.38	14.41	20.02
	Predicted	(g kg^{-1})	8.80	10.26	11.06	12.19	14.78	20.86
	RD	(%)	3.40	7.07	2.04	1.53	-2.57	-4.20
	MAE	(g kg^{-1})	0.31	0.81	0.28	0.58	1.44	0.84
	RMSE	(g kg-1)	0.14	1.56	0.13	0.46	2.53	0.82
	R^2		0.64	0.77	0.6	0.55	0.49	0.37
STN	Observed	(g kg^{-1})	0.61	0.72	0.76	0.83	0.99	1.22
	Predicted	(g kg^{-1})	0.59	0.69	0.76	0.81	1	1.27
	RD	(%)	3.28	4.17	0.00	2.41	-1.01	-4.10
	MAE	(g kg^{-1})	0.03	0.04	0.01	0.03	0.03	0.05
	RMSE	(g kg^{-1})	0.01	0.04	0.01	0.02	0.06	0.02
	R^2		0.76	0.82	0.89	0.5	0.89	0.39

2 Note: SOC, soil organic carbon; STN, soil total nitrogen; RD, relative difference; MAE, mean absolute prediction error; RMSE, root mean square error; R^2, coefficient of
3 determination.

Table 4

Table 4. Summary statistics of soil organic carbon (SOC) and soil total nitrogen (STN) prediction in different climate-vegetation landscapes using the ISA method based on 38 independent validation points.

1   **Table 4.** Summary statistics of soil organic carbon (SOC) and soil total nitrogen (STN) prediction in different

2   climate-vegetation landscapes using the ISA method based on 38 independent validation points.

Climate-vegetation landscapes	Number of independent validation points	SOC (g kg $^{-1}$)		STN(g kg $^{-1}$)	
		Range of change	Mean ± SD	Range of change	Mean ± SD
Coastal plain	5	0.85-17.64	6.97±5.47	0.34-1.13	0.63±0.62
Interior plain	7	2.38-33.98	9.68±6.69	0.43-1.32	0.8±0.69
Medium-coverage grasslands	6	7.61-29.74	11.12±6.32	0.74-2.07	0.88±0.67
High-coverage grasslands	7	8.04-33.25	12.14±6.91	0.77-2.28	0.94±0.7
Low-elevation forest	8	10.24-33.8	16.84±6.14	0.9-2.31	1.22±0.66
Low-mountain shrub	5	7.32-26.33	11.49±5.34	0.73-1.86	0.91±0.61

3   Note: SD, Standard deviation.

Figure 1

Figure 1

Fig 1. Location of the study area and 126 sampling sites (c) in Liaoning Province (b), China (a), which are shown superimposed on a 30-m resolution digital elevation model.
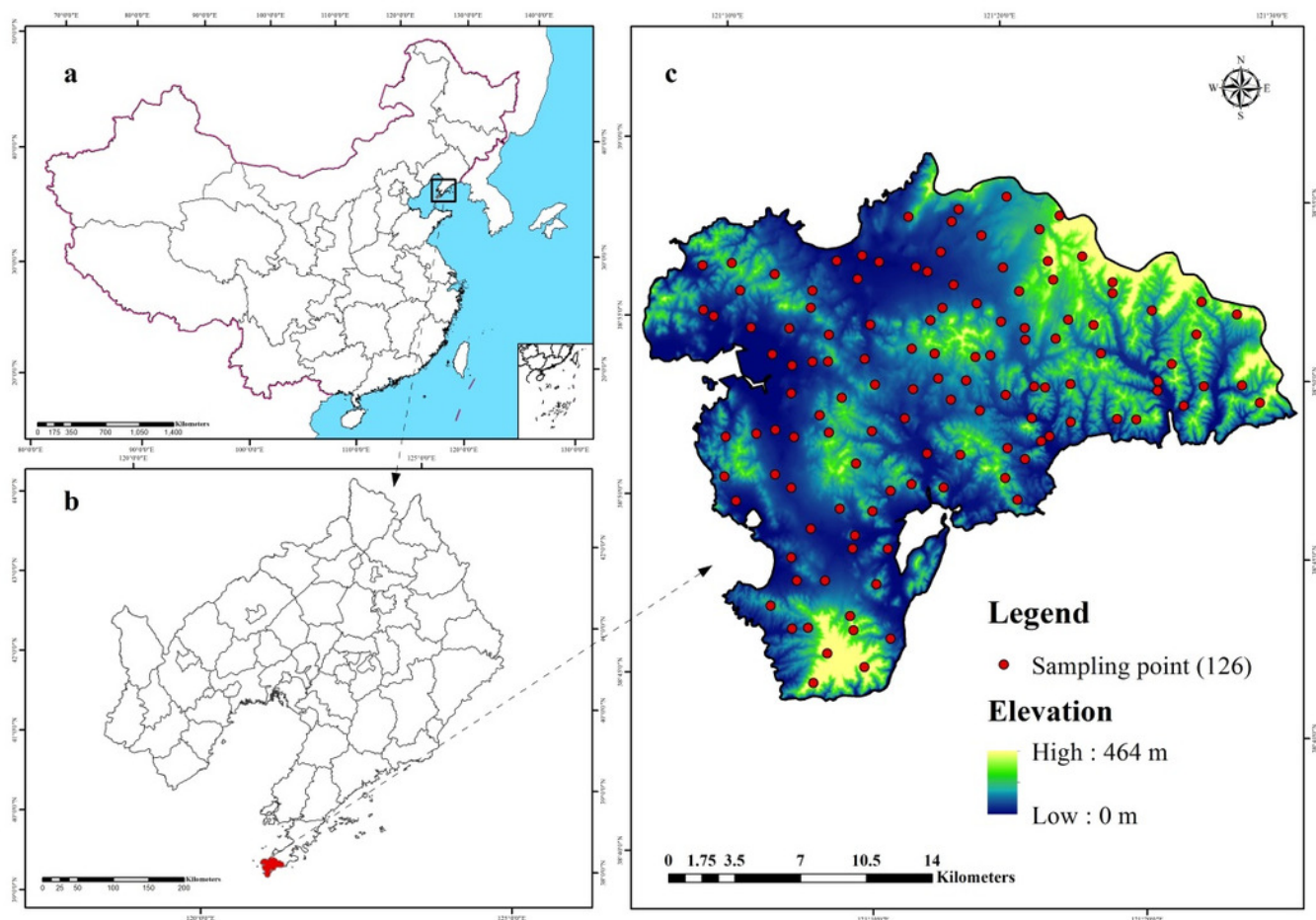
Figure 2

Figure 2

Fig 2. Climate-vegetation landscapes in the study area: Landscape 1, coastal plain;
Landscape 2, interior plain; Landscape 3, medium-coverage grasslands; Landscape 4, high-
coverage grasslands; Landscape 5, low-elevation forest; Landscape 6, low-mountain shrub.
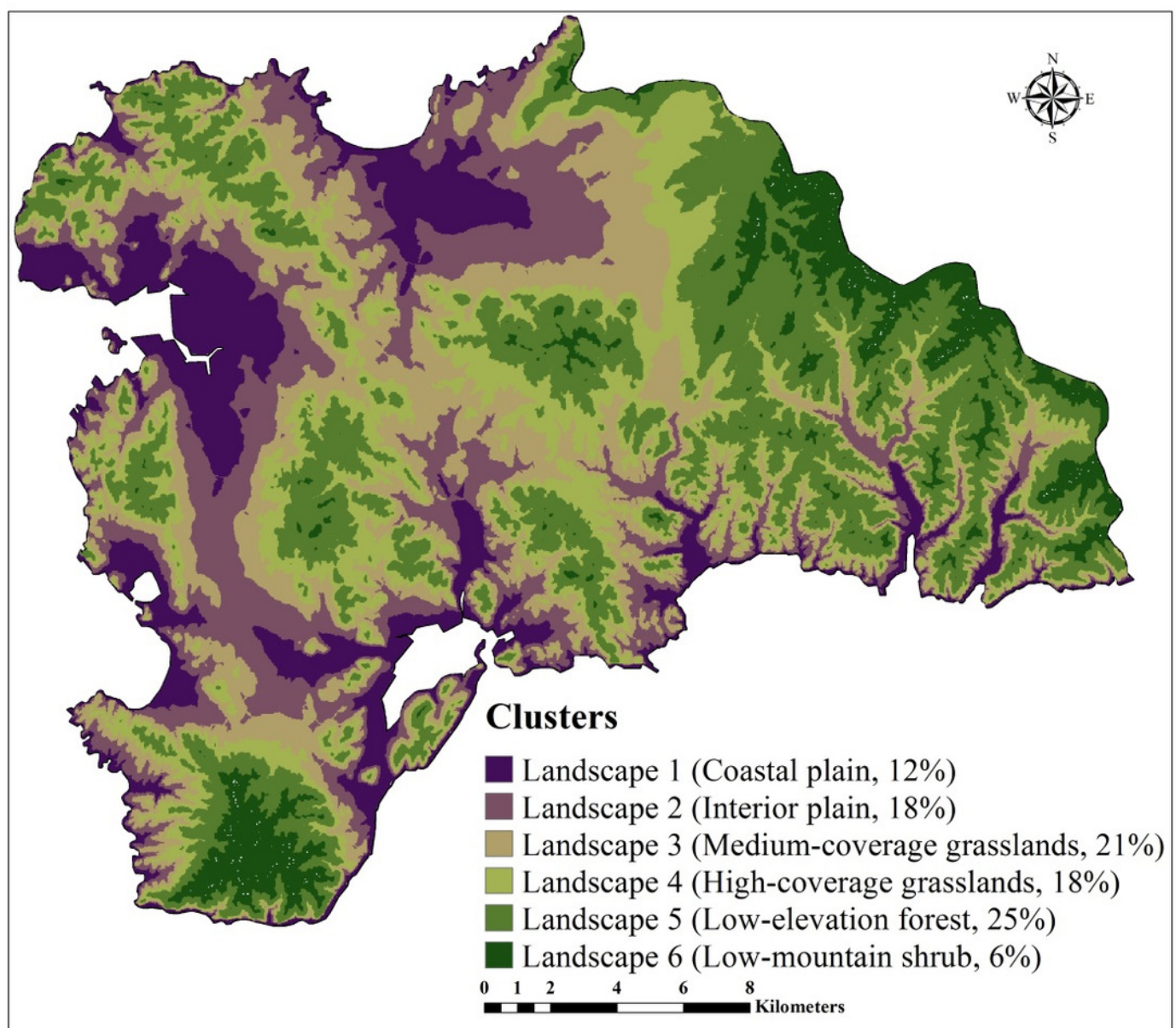
Figure 3

Figure 3

Fig 3. Flowchart of the methodology adopted in the study.
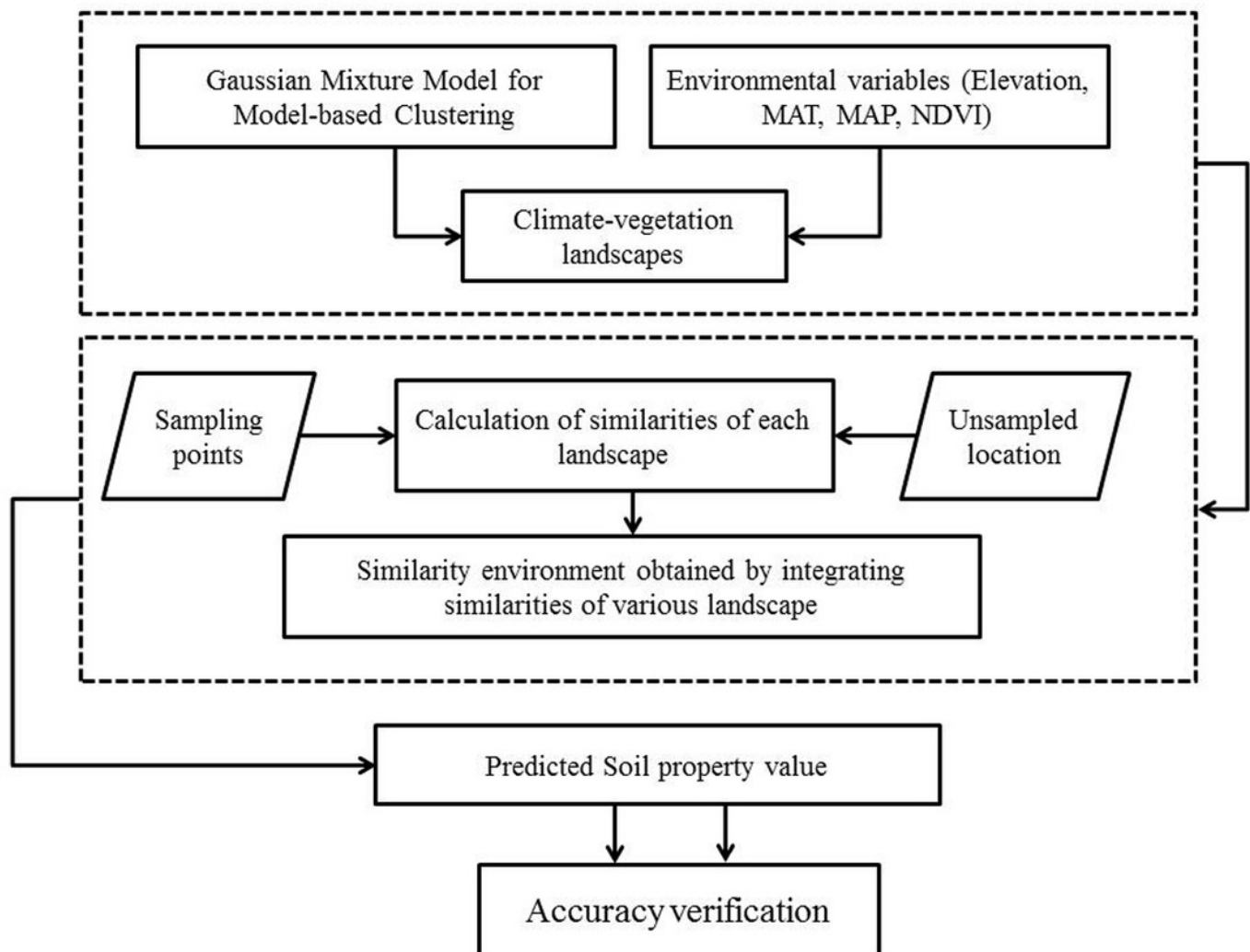
Figure 4

Figure 4

Fig 4. Pearson correlation coefficients of SOC and STN with environmental variables: SOC, soil organic carbon; STN, soil total nitrogen; TWI, topographic wetness index; SG, slope gradient; MAP, mean annual precipitation; MAT, mean annual temperature; and NDVI: Normalized Difference Vegetation Index.
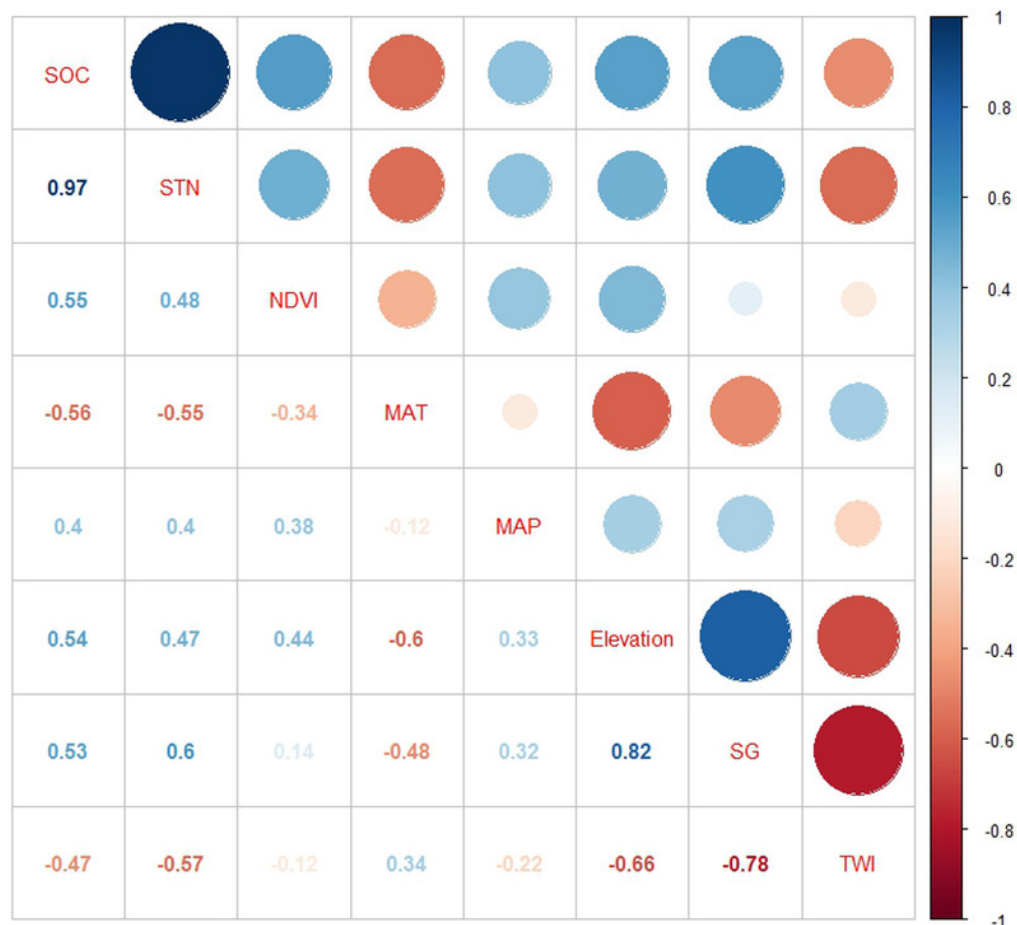
Figure 5

Figure 5

Fig 5. Boxplots of main environmental features related (elevation (a), mean annual precipitation (MAP) (b), mean annual temperature (MAT) (c), and normalized difference vegetation index (NDVI)) to climate-vegetation landscapes including CP: coastal plain; IP: interior plain; MCG: medium-coverage grasslands; HCG: high-coverage grasslands; LEF: low-elevation forest; and LMS: low-mountain shrub. Red is the training dataset (n=88) and green is the independent verification dataset (n=38).
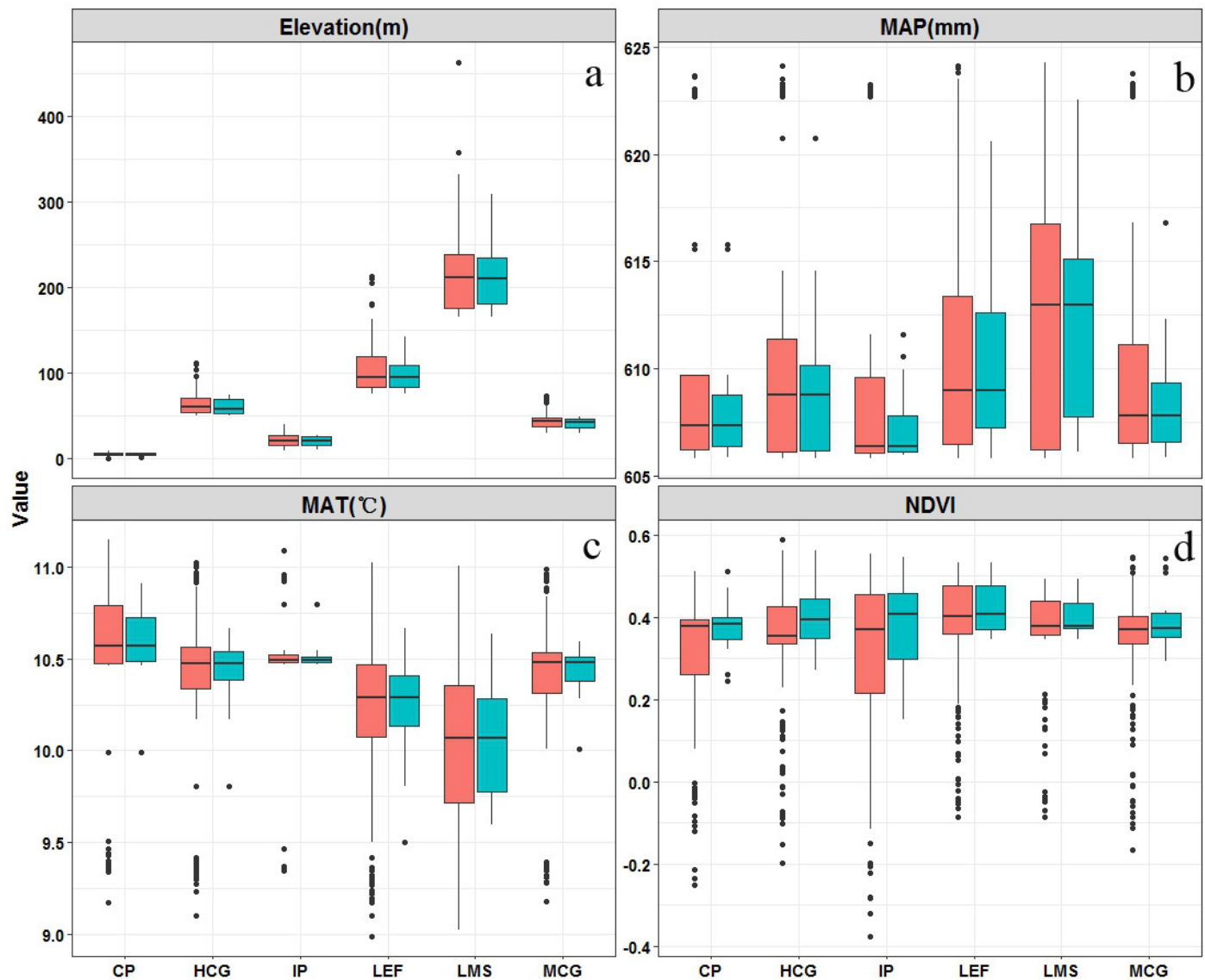
Figure 6

Figure 6

Fig 6. Scatter plot between the observed soil organic carbon (SOC) (a) and soil total nitrogen (STN) (b) content with its predicted values using the ISA model based on independent validation samples.
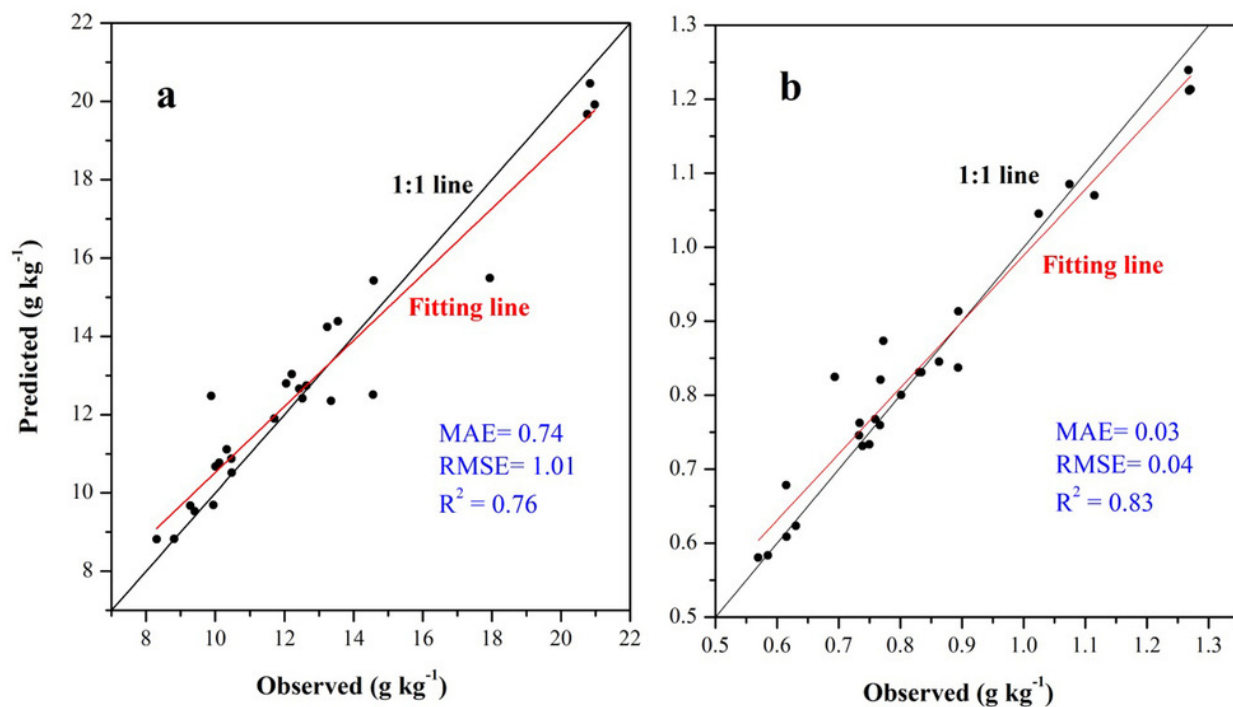
Figure 7

Figure 7

Fig 7. Prediction consistency map of SOC (a) and STN (b) and spatial distribution of SOC (g kg^{-1}) (a) and STN (g kg^{-1}) (b) predicted using an improved similarity-based approach (ISA).
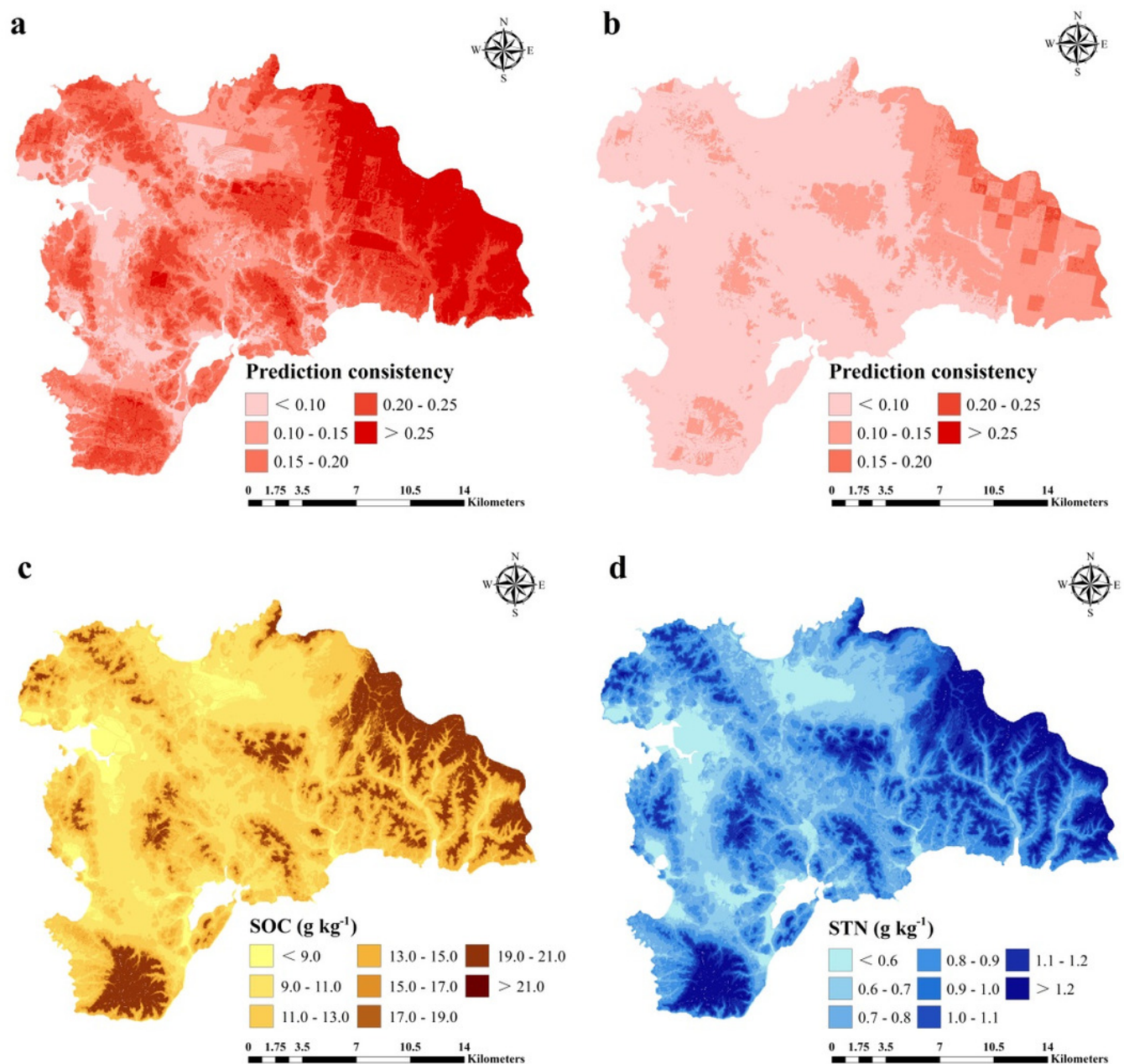
Figure 8

Figure 8

Figure 8. Relative importance of each variable as determined from the boosted regression trees (BRT) model in predicting soil organic carbon (SOC) (a) and soil total nitrogen (STN) (b).