Article

Improving the Prediction of Soil Organic Matter in Arable Land Using Human Activity Factors

Lixin Ning [1,2,3,4], Changxiu Cheng [3,4,*], Xu Lu [1,2], Shi Shen [3,4], Liang Zhang [1,2,*], Shaomin Mu [1,2] and Yunsheng Song [1,2]

[1] College of Information Science and Engineering, Shandong Agricultural University, Taian 271018, China; ninglixin123@163.com (L.N.); lxuu306@sdau.edu.cn (X.L.); msm@sdau.edu.cn (S.M.); sys_sd@126.com (Y.S.)

[2] Agricultural Big Data Research Center, Shandong Agricultural University, Taian 271018, China

[3] Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; shens@bnu.edu.cn

[4] State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China

[*] Correspondence: chengcx@bnu.edu.cn (C.C.); zliang@sdau.edu.cn (L.Z.)

Abstract: Detailed spatial distribution of soil organic matter (SOM) in arable land is essential for agricultural management and decision making. Based on digital soil mapping (DSM) theory, much attention has been focused on the selection of environmental covariates. However, the importance of human activity factors in SOM prediction has not received enough attention, especially in arable soil. Moreover, due to the insufficient amount of soil sampling data used to train and validate the DSM model, the prediction results may be questionable, and some even contradictory. This paper explores the effectiveness of the human footprint, amount of fertilizer application, agronomic management level, crop planting type, and irrigation guarantee degree in SOM mapping of arable land in Heilongjiang Province. The results show that the model only including environmental covariates accounts for 41% of the variation in SOM distribution. The model combining the five human activity factors increases the SOM spatial prediction by 39% in terms of R^2 (coefficient of determination), 12% in terms of RMSE (root mean square error), 15% in terms of MAE (mean absolute error), and 11% in terms of LCCC (Lin's concordance correlation coefficient), showing better prediction accuracy and performance. This indicates that human activity factors play a crucial role in determining SOM distribution in arable land. In the SOM prediction, soil moisture is the most important environmental covariate, and the amount of fertilizer application with a relative importance of 11.36% (ranking 3rd) is the most important human activity factor, higher than the annual average precipitation and elevation. From a spatial point of view, the Sanjiang Plain is a difficult area for prediction.

Keywords: soil organic matter prediction; human activity factor; large amount of data; arable land

1. Introduction

Soil organic matter (SOM) is an important soil property [1,2]. Carbon contained in soil is the largest pool in terrestrial ecosystems, containing three times as much carbon as in the atmosphere [3,4], whose small dynamic changes could affect the overall emissions of greenhouse gases [5,6]. Furthermore, SOM, providing nutrients to plants, is a crucial soil property that affects soil quality and soil fertility [7]. SOM of arable land is particularly important because agricultural soil provides most of the food needed for human survival. Unfortunately, due to intensive human activities, the depletion of SOM has been observed in arable land around the world [8–10]. The change will inevitably influence the normal global climate and agricultural production. Thus, adequate information on the spatial distribution of SOM in arable land is essential for quantifying the carbon budget, modeling the ecosystem and climate change, and evaluating soil quality to improve agricultural management and policymaking.

The well-known state factor equation of soil, referred to as CLORPT, was proposed by Jenny to estimate SOM spatial information [11]. Subsequently, McBratney proposed SCORPAN, which considers the soil to be a product of the interaction of environmental co-variates [12]. In addition, "Digital Soil Mapping (DSM)" was described. As an inexpensive and efficient method, DSM technologies have received increasing attention. Numerous studies have been conducted to obtain the spatial distribution of soil information over the last few decades, including at the regional scale, national scale, and global scale [13–15].

Based on the DSM theory, much attention has been focused on the selection of covariates for predicting SOM. At early times, most variables used in DSM are natural environmental conditions. With the development of the human cognition level, the importance of anthropogenic factors in soil prediction has been gradually recognized, especially in arable soils with strong human activities. An enhanced conceptual soil equation, STEP-AWBH, has been proposed to add human activity elements as explicit soil forming factors [16]. Recently, many studies have been conducted to investigate the importance of human activities to SOM modeling, which are receiving increasing attention. In Anhui Province (China), Yang built four pools of different variables, containing environmental or agricultural management indicators, to test the effect of agricultural parameters in SOM mapping [6]. In the same area, crop rotation information generated using a Fourier transform was used to explore the effectiveness of such information in SOM prediction based on four pools of variables with different categories [4]. Another study in northeastern China used the same method to quantify the influence of cultivation history with only two prediction models with different variables [7]. Although related studies have been carried out gradually, research on the impact of human activities on SOM is insufficient. Through various productive and living activities, humans have a profound impact on soil formation, especially in arable land with strong human influence. Unfortunately, information or data related to human activities are difficult to acquire compared to environmental conditions. Much of the information can only be acquired through field surveys and personnel interviews, which often consume considerable manpower, resources and time. This further limits related research on the effects of human activities on SOM.

The relationship between soil and environment can only be obtained by collecting many known soil points and their covariates. Because it is difficult to obtain a sufficient amount of SOM sampling data on a wide scale, most DSM studies were based on a certain amount of sampling point data that were collected through specific sampling design, field sampling, and laboratory analysis. Data splitting and cross-validation were used to analyze the accuracy of the DSM [1,4,6,7,17–19]. The amount of soil point data used in DSM studies is obviously different, ranging from dozens [4,6] to thousands [20–22]. Unfortunately, due to the small amount of soil sampling data (compared to the amount of prediction in the study area) used to train the DSM model, many similar studies may draw controversial conclusions [22–26]. The contradictory conclusions may be attributed to the insufficient amount of sampling data. In other words, the sampling density may be too low to capture the actual spatial distribution of SOM [22]. What is worse is that less data were used for model validation, and the accuracy was used to represent the prediction accuracy over the whole area. However, one notable question is how representative the sampled data are and whether the accuracy using sample data can be used to reflect the prediction results of the whole study area. Although this framework has been widely used as a consequence of the time consumption of large-scale soil sampling, the problem associated with the method still requires the attention of researchers. However, this issue can be easily avoided if a large range of soil survey data is available for prediction and accuracy verification.

The Ministry of Natural Resources of the People's Republic of China launched the "the national field-scale evaluation of arable land quality project", which is an important task to realize the trinity protection of the quantity, quality, and ecology of arable land. The project will acquire information on soil properties and soil management, including agronomic management level, irrigation water quality, SOM, etc., in each patch of arable land across the whole nation. To date, soil surveys in some regions have been completed. The project

provides sufficient data to solve problems related to DSM, which are mentioned earlier. On the one hand, the large amount of soil survey data can be used to verify the credibility of DSM technologies and theory, which provides a reference for future related DSM research in other specific fields. On the other hand, survey data related to soil management can be employed to explore the impact of human activities on SOM.

In this study, five human activity factors, including the human footprint, amount of fertilizer application, agronomic management level, crop planting type, and irrigation guarantee degree, were used to improve the prediction of SOM of arable land in Heilongjiang Province. To our knowledge, few studies using these variables with sufficient SOM sampling data and field survey data for SOM mapping have been conducted. Little is known about the effectiveness of these factors in SOM mapping. The purpose of this study is: (1) to test the hypothesis that the inclusion of these factors could improve the accuracy of SOM mapping; (2) to identify important environmental and human activity factors on SOM; and (3) to provide credible conclusions for the above research purposes using the soil survey data of entire research areas.

2. Materials and Methods

2.1. Study Area

The research was conducted in Heilongjiang Province, located in northeast China (46°23′~53°24′ N latitude, 121°13′~135°05′ E longitude), covering approximately 4,730,000 km^2 (Figure 1).
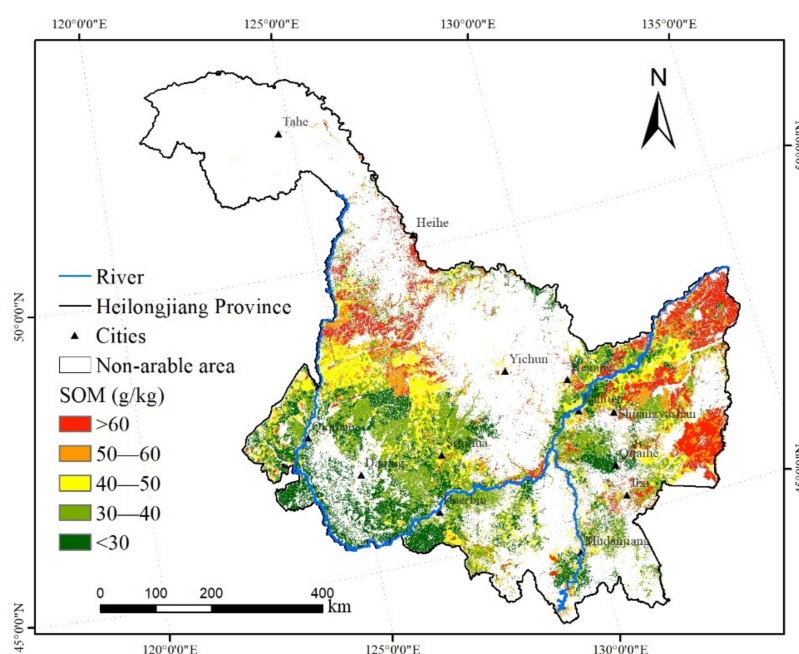

Figure 1. Location of the study area and distribution of observational SOM.

Heilongjiang Province mainly has a continental monsoon climate. In summer, the precipitation is sufficient, and the temperature is high. In winter, the weather is dry and cold for a long time. The annual temperature ranges from −4 to 4 °C. Annual precipitation ranges from 500~600 mm and is mainly concentrated in summer. Mountain areas account for 59% of the province and are concentrated mostly in the northwest, north, and southeast. The Greater Khingan and Lesser Khingan are the two most important mountains in the province. The province has large tracts of land with relatively flat terrain and low elevation, and 80% of the arable land consists of four soil types, Haplic Phaeozem, Haplic Chernozem, Luvic Phaeozem, and Albic Luvisol. The solar resources are relatively abundant, approximately 2300~2800 sunshine hours per year.

The black soil region located in Northeast China is one of the "three major black soil regions" in the world and is mainly concentrated in central and western Heilongjiang Province. The arable land area in the province accounts for 11.75% of the total country's [27], and 83% of the area of the Black Soil region is used as arable land [28]. As a result, Heilongjiang Province is an important commodity grain production base in China, and it became "No.1 Grain Production" in 2016. Total production accounts for 10% of China's total production. Moreover, the province has achieved "12 consecutive increases" in grain production from 2003–2015.

However, the soil resources in the province are facing an over-reclamation problem resulting from the development of the economy and population. Two main factors limiting the utilization of arable land are the thinning of black soil thickness and soil erosion.

2.2. SOM Data

The "Specification of Arable Land Survey, Monitoring and Evaluation" was formulated by the Ministry of Natural Resources of the People's Republic of China to acquire information on soil properties and soil management at each patch of arable land. The SOM data used in this research came from the project that was implemented after 2015.

The project surveyed all arable land in 131 counties in Heilongjiang Province and obtained SOM data for approximately 1.16 million soil patches (Figure 1). Topsoil SOM (g/kg) in each soil patch was determined based on the Method for Determination of Soil Organic Matter in the Agricultural Industry Standard of the People's Republic of China [29] after air-drying, sieving, heating, and titration.

Meanwhile, the project also obtained information on soil management, including the agronomic management level and irrigation water quality.

2.3. Covariates

Fourteen environmental variables for SOM modeling were selected to represent topography, climate, parent materials, vegetation, soil, and others. They are elevation, aspect, slope, landform class, average precipitation, average temperature, lithological unit, average NDVI, sedimentary deposit thickness, average soil moisture, soil type, water table depth, solar radiation, and surface water occurrence (Table 1).

Table 1. Covariates used in the study and their sources.

Forming Factors	Variables	Data Sources	Time Span
Topography	Elevation	GMTED 2010 [30]	-
	Aspect	Processed from Elevation data	-
	Slope	Processed from Elevation data	-
	Landform class	Global Ecological Land Units [31]	2008–2013
Climate	Annual average precipitation	Resource and Environment Data Cloud Platform	2006–2015
	Annual average Temperature	Resource and Environment Data Cloud Platform	2006–2015
Parent material	Lithological unit	Global Ecological Land Units [31]	2008–2013
Vegetation	Annual average NDVI	Resource and Environment Data Cloud Platform	2006–2015
Soil	Sedimentary deposit thickness	ORNL DAAC [32,33]	-
	Average soil moisture	TerraClimate [34]	2009–2017
	Soil type	Resource and Environment Data Cloud Platform	1995
	Water Table Depth	[35]	-
Other	Solar radiation	Global Change Research Data Publishing and Repository	2015
	Surface Water occurrence	[36]	1984–2015
Human activities	Human footprint	Socioeconomic Data and Applications Center [37]	2009
	Amount of fertilizer application		-
	Agronomic management level		-
	Crop planting type		-
	Irrigation guarantee degree		-

- means the data will not change for a long time and is a fixed value.

For the topography, a 7.5 arc-second resolution digital elevation model (DEM) was freely downloaded from the U.S. Geological Survey (USGS) [30]. Three terrain variates (elevation, aspect, and slope) were derived from the DEM. The landform classes were derived based on the USGS's Map of Global Ecological Land Units [31]. The sources and resolutions of the data are shown in Table 1 and Figure 2a–d.
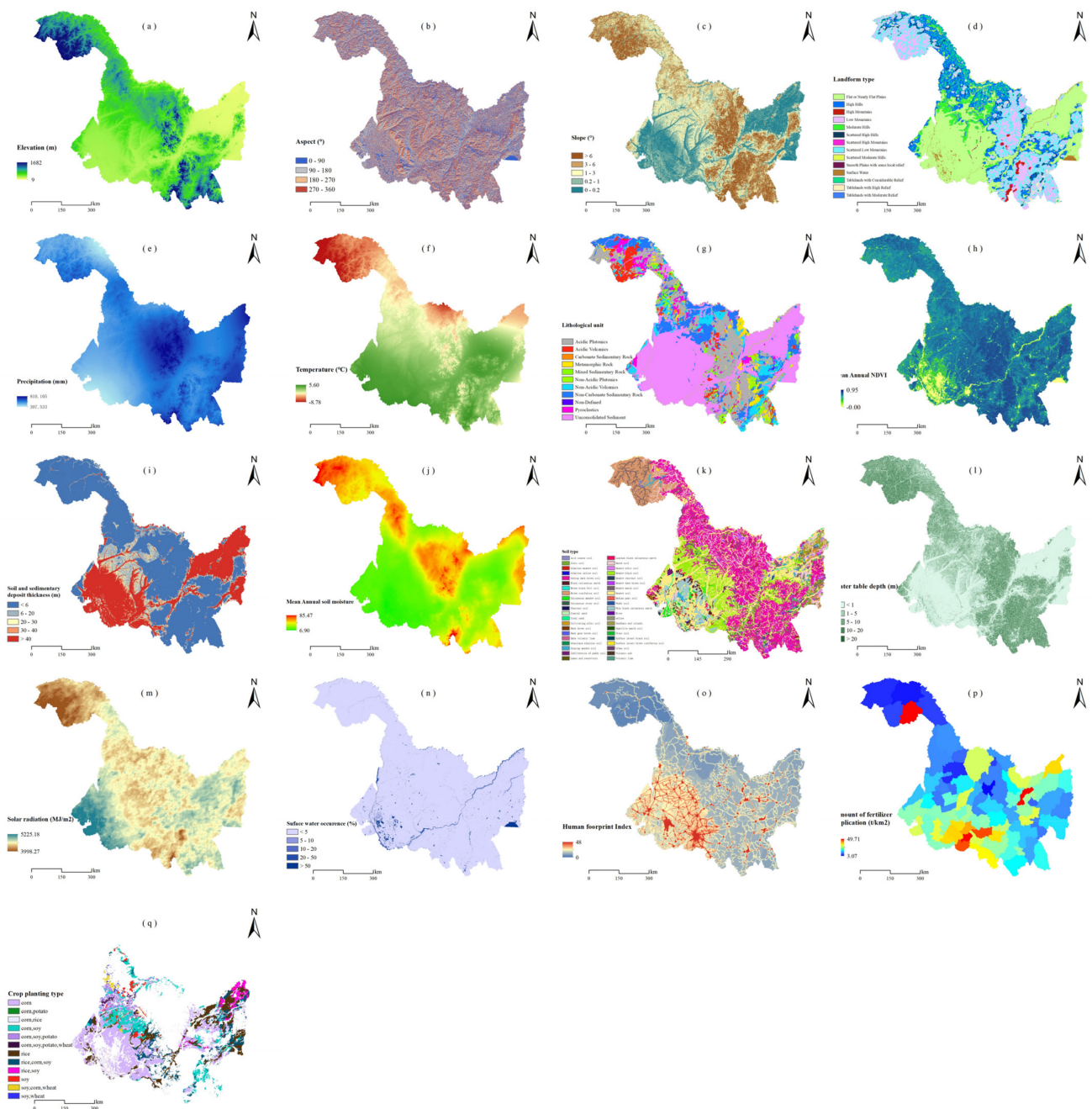

Figure 2. Distribution of covariates used in our study, including environmental covariates and human activities covariates. (**a**): Elevation; (**b**): aspect; (**c**): slope; (**d**): landform class; (**e**): mean annual precipitation; (**f**): mean annual temperature; (**g**): lithological unit; (**h**): mean annual NDVI; (**i**): soil and sedimentary deposit thickness; (**j**): soil moisture; (**k**): soil type; (**l**): water table depth; (**m**): solar radiation; (**n**): surface water occurrence; (**o**): human footprint; (**p**): the amount of fertilizer application; (**q**): agronomic management level.

For climate, two covariates, annual average precipitation and annual average temperature, were selected. Precipitation and temperature data from 2006–2015 were derived from the "Resource and Environment Data Cloud Platform (https://www.resdc.cn/ accessed on 12 April 2022)", which were used to calculate the annual average precipitation and annual average temperature. The source and resolution of the data are shown in Table 1 and Figure 2e,f.

For the parent material, one variate, the lithological unit, was selected. The data were collected from the U.S. Government's open data website. Eleven lithology types exist in the study area, including acidic plutonic, acidic volcanic, carbonate sedimentary rock, metamorphic rock, mixed sedimentary rock, nonacidic plutonic, nonacidic volcanic, noncarbonate sedimentary rock, non-defined, pyroclastic, and unconsolidated sediment [31]. The source and resolution of the data are shown in Table 1 and Figure 2g.

For the vegetation, the annual average NDVI was selected, which exhibits a good correlation with green-leaf density and can be used to estimate aboveground biomass [38]. Mahmoudabadi found that NDVI derived from remote sensing is a very effective parameter for predicting SOC [39,40], although it does not show satisfactory performance compared to other indices, e.g., EVI, at high altitudes [41]. In addition, historical pattern can explain a much larger part of the spatial variability in SOM in comparison to current data [40]. Then, NDVI data of 2007–2016 were obtained from the "Resource and Environment Data Cloud Platform". Then, the annual average NDVI was calculated using these data. The source and resolution of the data are shown in Table 1 and Figure 2h.

For the soil, fore variates, including sedimentary deposit thickness, average soil moisture, soil type, and water table depth, were acquired. Sedimentary deposit thickness was captured from the "Distributed Active Archive Center for Biogeochemical Dynamics" [32,33]. Mean monthly soil moisture data from 2009–2017 were derived from the "Climatology Lab" [34], and the average soil moisture was calculated. Soil type, including 41 types, was obtained from the "Resource and Environment Data Cloud Platform". The water table depth was obtained from the article [35]. The sources and resolutions of the data are shown in Table 1 and Figure 2i–l.

For the others, solar radiation and surface water occurrence were used. Solar radiation was acquired from the "Global Change Research Data Publishing and Repository". The surface water occurrence was derived from the "Global surface Water Explorer"[36]. The sources and resolutions of the data are shown in Table 1 and Figure 2m,n.

For human activity, five factors were used: human footprint, amount of fertilizer application, agronomic management level, crop planting type, and irrigation guarantee degree. Human footprint data were derived from the "Socioeconomic Data and Applications Center (SECAC)" [37]. For the amount of fertilizer application, we reviewed the official statistical yearbooks of cities in 2015 to collect the total amount of fertilizer application and cultivated area in Heilongjiang Province. Then, the total fertilizer application was divided by the cultivated area to obtain the amount of fertilizer application per unit area. The crop planting type in Heilongjiang Province was obtained from the article [42]. In the past decade, the national field-scale evaluation of arable land quality project was implemented to conduct mainland-wide surveys on arable land quality, which invested CNY 0.43 (equivalent to USD 0.067) billion and 1.3 CNY (equivalent to USD 0.2) million by the Ministry of Natural Resources of China [43–45]. The agronomic management level and irrigation guarantee degree were derived from the project, which consists of approximately 1.16 million soil patches over all of Heilongjiang Province. Due to the comprehensive nature, the agronomic management level is mainly graded by local statistical data and the questionnaire. The questionnaire mainly included the selection of good varieties, the planting structure, the popularization of fertilization by soil testing, the cultivation of weeds, water-saving irrig-

ation, and pest control. Level I represents a high level of comprehensive agronomic management. Level III represents a low level of comprehensive agronomic management. The irrigation guarantee degree was obtained by combining field surveys with water map information. Level I indicates that the irrigation requirements for agricultural production are fully met. Level II indicates that the irrigation requirements are met. Level III indicates the general situation, but it is difficult to meet the requirements during a dry year. Level IV indicates no irrigation conditions.

All numeric covariates with a resolution coarser than 250 m were resampled to 250 m using the cubic method. All type covariates with a resolution coarser than 250 m were resampled to 250 m using the nearest method.

2.4. Data Pre-Processing

The original SOM dataset, which is in vector format, was converted into raster format with a resolution of 250 m. The grid dataset was superimposed with 19 covariates, including environmental and human activity covariates after projection transformation and resampling. For type covariates, such as landform class, lithological unit, soil type, agronomic management level, crop planting type, and irrigation guarantee degree, the mean SOM of one type was weighted as the variable value of the type [46].

Collinearity and multicollinearity might exist between series of variables. To avoid this, partial correlation analysis was conducted between different variables to measure the correlation between two variables, removing the effects of other variables [47–49].

The descriptive statistics for SOM and 19 covariates are calculated to show the basic characteristics of the data. Additionally, the box plot method (Tukey's test) was employed to remove outliers of the SOM dataset. Finally, a 2,017,044 grid point dataset with SOM and 19 covariates was derived. A 2-D matrix is created, which is of the size of $n \times c$, where n = 2,017,044 is the number of sampling locations, and c = 20 is the SOM value and covariates values.

Then, n = 100,000 dataset was randomly selected as the training set and performed ten times. All grid points were used as the validation set.

2.5. Modeling and Evaluation

A series of DSM models have been developed to predict SOM spatial distribution, including geostatistical methods, neural networks, and cubists. Random forest (RF) has been widely used in DSM studies. It is an ensemble of regression trees, and each that is built benefits from a random subset of original training data sampling. Only a randomly selected prediction subset is used to generate the best segmentation. RF has a series of advantages compared to other DSM methods, including better error measurement, flexibility with input variable types, and less susceptibility to overfitting [50–52]. In addition, RF has been demonstrated to perform better than other DSM methods in many studies [51,53–56].

To investigate whether adding human activity factors would improve SOM prediction, seven pools of covariates with different categories were derived to investigate whether adding human activity factors would improve SOM spatial prediction (Table 2). Pool 1 only contains the 14 environmental variables, including elevation, aspect, slope, landform class, average precipitation, average temperature, lithological unit, average NDVI, sedimentary deposit thickness, average soil moisture, soil type, water table depth, solar radiation, and surface water occurrence. Pool 2 to Pool 6 were composed of the 14 environmental covariates with the addition of the human footprint, amount of fertilizer application, agronomic management level, crop planting type, and irrigation guarantee degree, respectively. Pool 7 consisted of the 14 environmental covariates and 5 human activity covariates together. For these seven covariate pools, seven prediction models are established, which are Models 1–7.

Table 2. Pools with different covariates. (Environmental covariates consisted of elevation, aspect, slope, landform class, average precipitation, average temperature, lithological unit, average NDVI, sedimentary deposit thickness, average soil moisture, soil type, water table depth, solar radiation, and surface water occurrence).

Pools	Covariates
Pool 1	environmental variates
Pool 2	environmental variates + human footprint
Pool 3	environmental variates + amount of fertilizer application
Pool 4	environmental variates + agronomic management level
Pool 5	environmental variates + crop planting type
Pool 6	environmental variates + irrigation guarantee degree
Pool 7	environmental variates + human footprint, amount of fertilizer application, agronomic management level, crop planting type, irrigation guarantee degree

For evaluation indicators, the mean absolute error (MAE) (Equation (1)), root mean square error (RMSE) (Equation (2)), coefficient of determination (R^2) (Equation (3)), and Lin's concordance correlation coefficient (LCCC) (Equation (4)) were calculated. They are defined as follows:

$$\text{MAE} = \frac{1}{h} \sum_{j=1}^{h} |(P_h - Q_h)| \tag{1}$$

$$\text{RMSE} = \left[\frac{1}{h} \sum_{j=1}^{h} (P_h - Q_h)^2 \right]^{0.5} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{j=1}^{h}(P_h - Q_h)^2}{\sum_{j=1}^{h}(P_h - \overline{Q_h})^2} \tag{3}$$

$$\text{LCCC} = \frac{2 \bullet r \bullet \partial_P \bullet \partial_Q}{\partial_{P^2} + \partial_{Q^2} + (\overline{P} + \overline{Q})^2} \tag{4}$$

where h is the number of predictions, P_h is the observed SOM value at point h, Q_h is the predicted SOM value at point h, $\overline{Q_h}$ is the mean of Q_h, and ∂_P and ∂_Q are variances of P_h and Q_h.

Variable importance was calculated to indicate the predicting power of difference covariates for SOM; it was estimated based on the mean decrease in prediction accuracy of each variable by replacing each covariate in turn by random noise and observing the average increase in the prediction accuracy for all trees [6,19,57–60].

For the RF model, one important parameter, the number of trees, should be set manually before simulation. Here, the GridSearchCV method in the Sklearn package was used to tune the hyperparameters of RF. The tuning result shows that 350 was the best number of trees, which was used in the RF model of the study.

The "RandomForestRegressor" of the sklearn package [61] in Python was used to conduct random forest modeling and calculate variable importance.

3. Results

3.1. Feature Selection

As a result of collinearity examination, Figure 3 shows the partial correlation coefficient between the 19 covariates ranging from −0.48 to 0.56, which are not significant, indicating that there is no collinearity between the covariates. Then, all covariates were used for subsequent modeling and analysis.

3.2. Descriptive Statistics

Table 3 shows the descriptive statistics for SOM and 19 covariates. The SOM content ranged from 2.93 g/kg to 80.13 g/kg, with a mean value of 39.46 g/kg. The SD of SOM was

above the mean value, which indicates a high variability in its distribution. The annual average NDVI has the smallest variation, while the aspect and solar radiation have the largest variation.
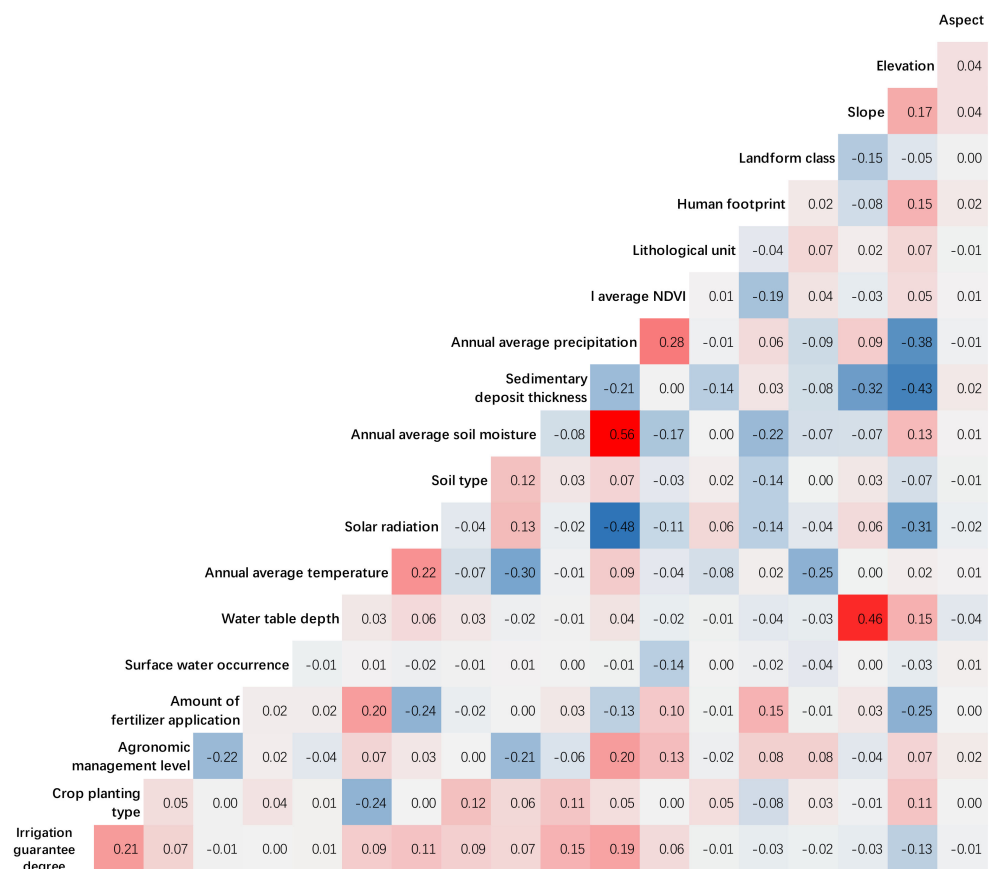

Figure 3. Partial correlation analysis between 19 covariates (the partial correlation coefficients range from −0.48 with blue to 0.56 with red).

Table 3. Summary statistics of SOM and quantitative covariates of all datasets.

Covariates	Max	Mean	Min	SD	Covariates	Max	Mean	Min	SD
SOM (g/kg)	80.13	39.46	2.93	192.75	Annual average soil moisture	56.11	18.39	6.90	59.31
Aspect (°)	359.84	173.08	−1.00	10,414.20	Soil type	111.55	45.63	18.39	205.70
Elevation (m)	799.00	172.88	35.00	7917.64	Solar radiation (MJ/m²)	5151.92	4668.69	4391.01	12,035.60
Slope (°)	25.64	1.01	0.00	2.25	Annual average temperature (°C)	5.60	3.43	−4.87	2.16
Landform class	62.95	46.66	32.91	6.19	Water table depth (m)	159.03	2.01	−16.37	27.53
Human footprint	47.00	12.29	1.00	37.40	Surface Water occurrence	96.42	0.31	0.00	6.55
Lithological unit	53.38	46.66	40.67	5.06	Amount of fertilizer application (t/km²)	37.82	17.50	4.31	51.49
Annual average NDVI	0.95	0.86	0.18	0.00	Agronomic management level	47.28	46.78	45.36	0.71
Annual average precipitation (mm)	690.73	560.67	430.20	1836.79	Crop planting type	80.99	45.92	35.31	116.10
Sedimentary deposit thickness (m)	60.00	32.06	−7.00	399.85	Irrigation guarantee degree	59.98	46.63	39.89	67.59

Max, maximum; Min, Minimum; SD, standard deviation.

As mentioned, ten *n* = 100,000 samples were randomly selected from all *n* = 2,017,044 data as the training set. Figure 4 summarizes the comparison of statistical descriptions

of SOM between all datasets and ten training sets. This shows that the distribution and concentration of SOM in the ten training sets is consistent with the SOM in the original dataset. The training set selected can well represent the SOM distribution of the entire Heilongjiang Province, which provides an important basis for subsequent analysis.
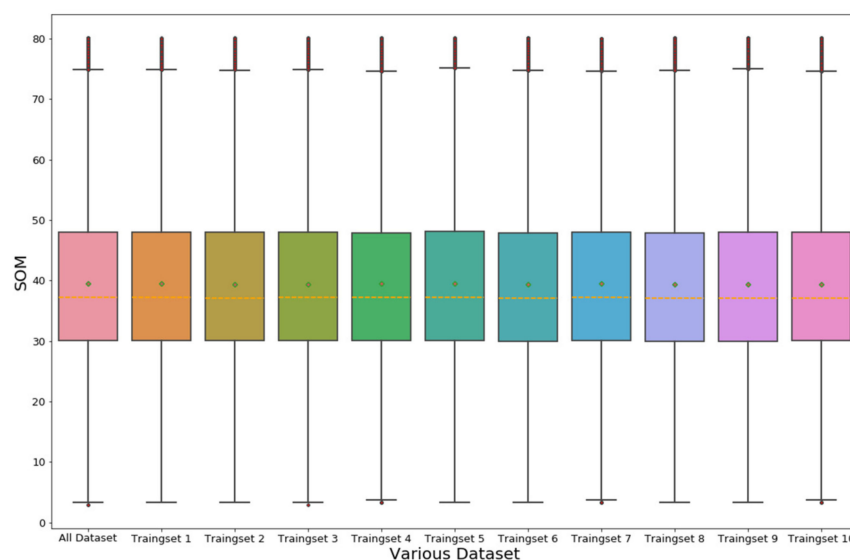

Figure 4. Boxplot of SOM of all datasets and 10 training sets with *n* = 100,000 samples.

3.3. Covariate Importance

The relative importance of each variable for the seven pools is illustrated in Figure 5. Slight differences are shown among the rankings of the same covariates of the seven pools in the figures. This results from the fact that the relative importance of a given variable in a certain model depends on its correlation with other variables [4,59]. For all covariates, soil moisture, with a relative importance of 26.67%, was the most important factor in the prediction of SOM (Figure 5g). It is followed by the annual average temperature. In general, soil moisture, annual average temperature, annual average precipitation, and elevation are among the most important variables of all seven pools. Each of the four variables in the seven pools has a relative importance higher than 11%, which together accounts for more than 70% of the SOM variation. For human activities, the relative importance of the human footprint ranks 8th in Model 2, with a relative importance of 2.66% (Figure 5b). The amount of fertilizer application ranks 3rd in Model 3, with a relative importance of 12.33% (Figure 5c). The agronomic management level, with a relative importance of 3.10%, ranks 7th in Model 4 (Figure 5d). The crop planting type ranks 8th (2.93%) in Model 5 (Figure 5e). The irrigation guarantee degree ranks 10th (2.00%) in Model 6 (Figure 5f). In Model 7 with all 19 covariates (Figure 5g), the amount of fertilizer application was the 3rd most important variable, with a relative importance 11.36% higher than the annual average precipitation and elevation. The human footprint, agronomic management level, crop planting type, and irrigation guarantee degree rank 9th, 11th, 14th, and 16th, with a relative importance of 1.92%, 1.77%, 1.37%, and 1.15%, respectively. The results indicate the important prediction power of the five human activity factors. Remarkably, fertilizer application is the most important factor for SOM prediction. In contrast, surface water occurrence, lithological unit, and landform class are the last three important variables, and each has a relative importance lower than 1% in Model 7.
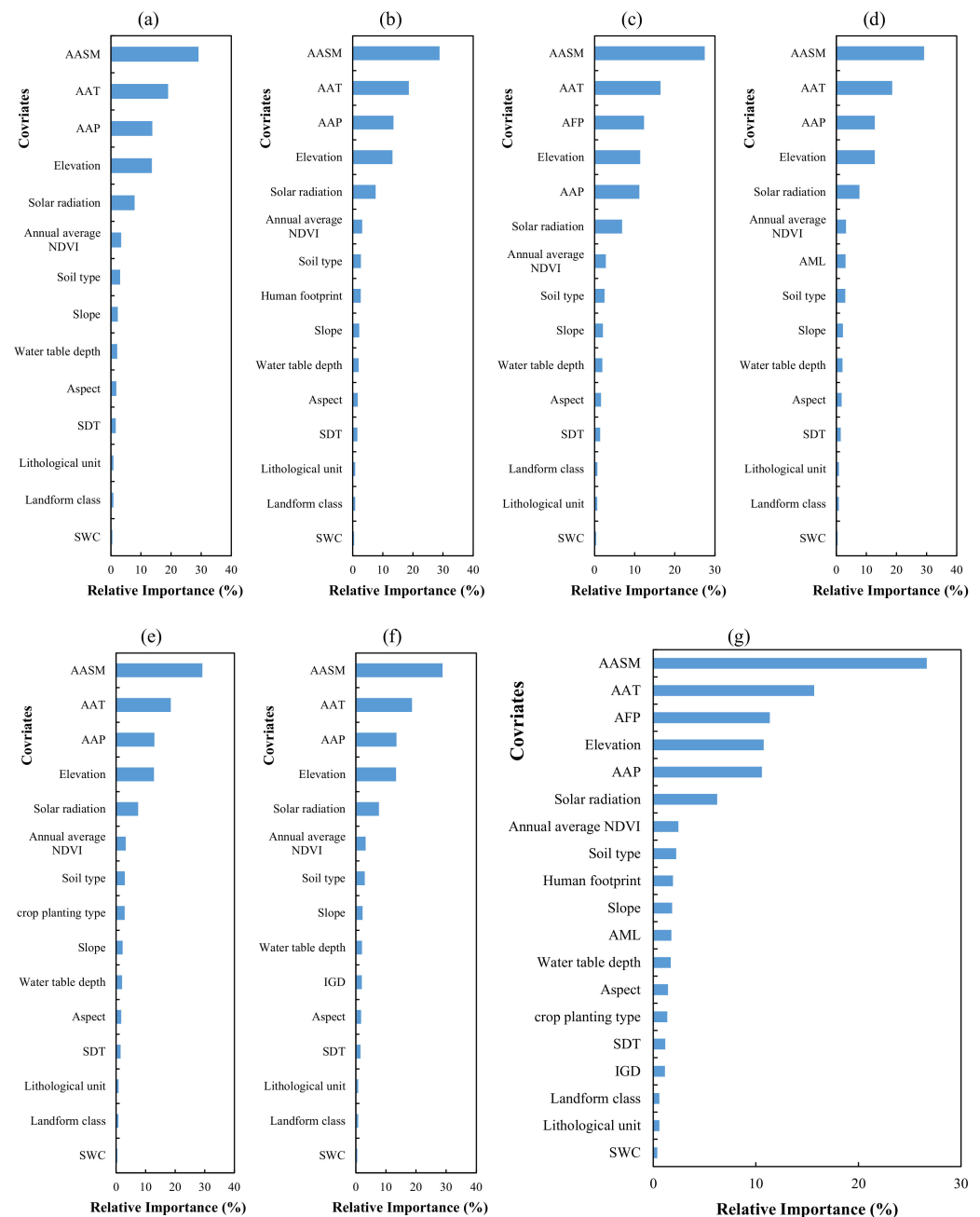
Figure 5. Relative importance of each variable for the seven models calculated by RF (SWC: surface water occurrence; SDT: sedimentary deposit thickness; AAP: annual average precipitation; AAT: annual average temperature; AFP: the amount of fertilizer application; AML: agronomic management level; IGD: irrigation guarantee degree; AASM: annual average soil moisture. (**a**), (**b**), (**c**), (**d**), (**e**), (**f**), (**g**) indicate relative importance of each variable for Model 1, 2, 3, 4, 5, 6, 7, respectively).

3.4. Model Performance and Spatial Difference

The prediction accuracy of SOM with seven combinations of covariates is listed in Table 4. The modeling and validation were performed ten times, and the average value was calculated as the final accuracy value.

Table 4. Prediction performances of SOM for seven models.

No.	Covariates	MAE	RMSE	LCCC	R^2
1	environmental variates	5.87	8.37	0.63	0.41
2	environmental variates + human footprint	5.79	8.27	0.64	0.42
3	environmental variates + amount of fertilizer application	5.29	7.72	0.68	0.53
4	environmental variates + agronomic management level	5.66	8.13	0.65	0.44
5	environmental variates + crop planting type	5.74	8.18	0.65	0.44
6	environmental variates + irrigation guarantee degree	5.73	8.20	0.64	0.43
7	environmental variates + human footprint, amount of fertilizer application, agronomic management level, crop planting type, irrigation guarantee degree	5.02	7.37	0.70	0.57

R^2, coefficient of determination.

The results show that adding the amount of fertilizer application would increase the prediction accuracy by 12% in terms of R^2 compared with when only environmental variables are used. The accuracy is higher than adding the human footprint, agronomic management level, crop planting type, and irrigation guarantee degree only, increasing the prediction accuracy by 1%, 3%, 3%, and 2%, respectively. The MAE and RMSE of the five models show a decreasing trend when only environmental covariates are used. Comparing Model 7 and Model 1, the validation results show that adding the five human activity factors increased the prediction accuracy by 39% in terms of R^2, 12% in terms of RMSE, and 15% in terms of MAE compared to when only environmental covariates were used. Combining the five human activity factors shows more promising prediction power for SOM mapping in arable land. R^2 revealed that the model including the five human activity factors and environmental covariates could explain 57% of the variation in the SOM distribution, whereas including environmental covariates only explained 41% of the variation in the SOM distribution. Additionally, a similar conclusion could be drawn from the higher LCCC of SOM prediction in model 7.

The results prove that the predicted value of Model 7 is closer to the observed value, but the proximity may vary in different SOM intervals. Therefore, this section uses the kernel density method to show the proximity between the predicted and observed values of Model 1 and Model 7. The density distribution of SOM is shown in Figure 6.
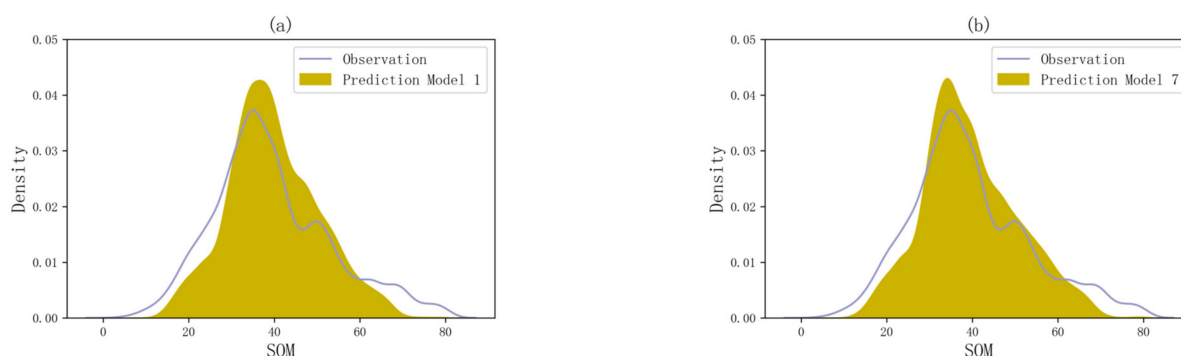
Figure 6. Density distribution of SOM predicted with Model 1 (**a**) and Model 7 (**b**).

A similar conclusion could be drawn from the density distribution. Overall, the prediction of Model 7 is closer to the observed value than Model 1 and better follows the real density distribution. However, the proximity has different performances in different SOM intervals. In the interval of high SOM (>60 g/kg) and low SOM (<30 g/kg), the density distributions of Model 1 and Model 7 were both lower than the observed value,

and the difference was more obvious in the former. However, the predictions of the two models have no significant difference and are almost identical. In the middle SOM interval (30 g/kg~60 g/kg), the density distributions of the two models are higher than the observed values. At the same time, there are obvious differences between the two models. In the SOM interval, the density distribution of Model 1 is higher than that of Model 7, and the difference from the observed value is more obvious. Only in a small interval (approximately 30 g/kg and 60 g/kg), the density of Model 7 is higher than that of Model 1. Moreover, one can find that the peak density of the observation is 35 g/kg, and the peak value of Model 1 and Model 7 are 37 g/kg and 35 g/kg, respectively. This suggested that human activity factors have a profound influence on SOM distribution in arable areas, and the prediction is more consistent with the observation.

The addition of human activity factors can improve the proximity between observations and predictions. To show the spatial distribution of the proximity, this section calculates the difference between the observed value and predicted and of Model 1 and Model 7 and spatializes it to characterize the spatial differences.

Figure 7 shows the spatial distribution of the SOM difference between the observations and predictions of Model 1 (Figure 7a) and Model 7 (Figure 7b). Overall, there is an obvious difference between Model 1's result and the real value. The difference in many areas exceeds 10 g/kg, mainly located in the middle of the Songnen Plain and the northeastern part of the Sanjiang Plain, accounting for 16.4% of the total area. There is no obvious difference (−3~3 g/kg) in most areas (approximately 40.0% of the total area), mainly distributed in the Songnen Plain. Compared with Model 1, the difference between the predictions of Model 7 and the observations is obviously reduced. The area with a difference exceeding 10 g/kg only exists sporadically in the northeastern part of the Sanjiang Plain, accounting for 12.6% of the total area, far less than the 16.4% of Model 1. However, the area with no obvious difference (−3~3 g/kg) increased dramatically, accounting for 47.4% of the region. The area has an increase of 18.5% compared with Model 1, and most areas of Songnen Plain belong to this category.
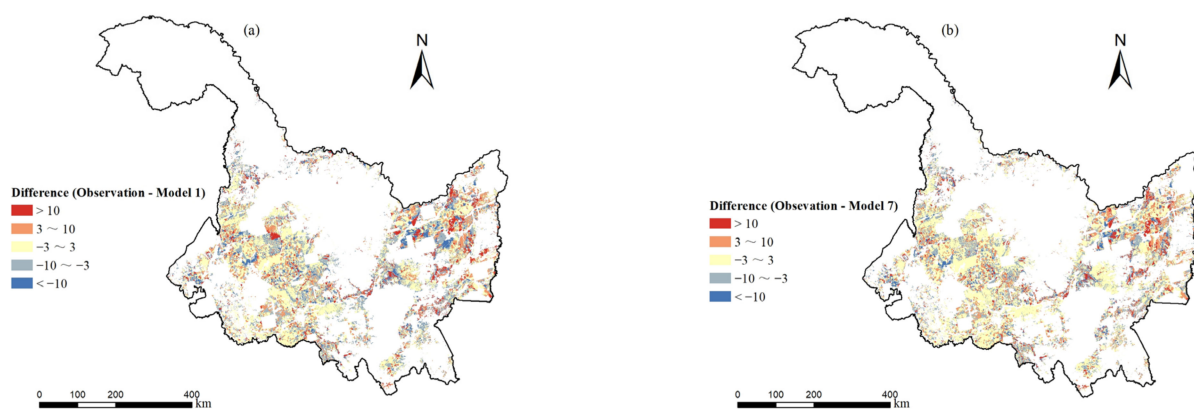

Figure 7. Spatial difference between observation and prediction with Model 1 (**a**) and Model 7 (**b**).

4. Discussion

4.1. Relative Importance of Environmental Covariates

Studies have shown that the importance of explanatory covariates varies with different regions and scales. Generally, due to their high relative importance, topography, climate, and vegetation covariates have been widely used in SOM mapping. A similar conclusion was drawn in this study. The annual average temperature, elevation, annual average precipitation, and annual average NDVI rank 2, 4, 5, and 7, respectively, in Model 7 (Figure 5g).

Topography covariates have been widely employed in SOM prediction [62,63]. Among the series topography covariates, elevation is the most important covariate in this study, followed by slope and aspect. Similarly, previous studies for SOM prediction suggested that

elevation is the most effective topographic covariate [19,64]. Mechanistically, topography could control precipitation, temperature, water flow paths, and discharge and significantly influence erosional processes. Therefore, it plays a crucial role for SOM. Sites with high elevations and gentle slopes favor water accumulation, which controls the SOM input. Furthermore, soil erosion and redistribution determined by water flow paths and local microclimate influenced by topography have profound effects on SOM variation in highly variable terrain, as in our study area [7,65]. In addition to our study area, topographic covariates would also significantly affect SOM in northeastern Iran [39], the Mediterranean region [66], Barro Colorado Island [58], eastern China [67], etc.

Climatic influences the spatial variation of SOM [40,65], which is the "C" factor of soil formation. Precipitation and temperature are two key climatic covariates affecting the spatial variation of SOM, and similar results have been reported in similar regions [1,7,68]. In Model 7, AAT was the second most important covariate, whereas AAP ranked fifth (Figure 5g), indicating that temperature played a greater role in SOM prediction than precipitation, which is consistent with previous studies [4,7,19]. Mechanically, the two covariates could affect both C input and SOM decomposition. Precipitation plays a crucial role in net primary productivity and then the input of C into the soil. A higher humidity favors the weathering of parent material and the formation of a soil carbon stabilizing mineral surface [69,70], which reduces the decomposition of SOM [71]. Temperature plays a crucial role in the microbial decomposition rate of SOM by affecting its complex molecular attributes [72–74].

Vegetation is another important covariate that is frequently used to predict SOM [75,76]. Among many vegetation covariates, the annual average NDVI is frequently employed for SOM mapping [6]. In Model 7, the annual average NDVI ranks 7 of all 19 covariates (Figure 5g). Previous studies also reported a very important effect of NDVI on SOM distribution [1,19,77]. As the main source of SOM, it can enrich SOM by adding organic material, conserving soil moisture, and protecting soil erosion [78,79]. Moreover, vegetation will also affect the decomposition of C [65]. In addition to our study area, vegetation has been confirmed to primarily affect SOM in northern America [80], eastern Australia [81], and northwest Iran [13].

Compared to topography, climate, and vegetation covariates, soil moisture is less used in SOM prediction and mapping. In Model 7, annual average soil moisture has the highest relative importance compared to the other 18 covariates, which indicates that soil moisture could significantly affect the spatial variation of SOM and play the most important role in SOM development (Figure 5g). Soil moisture controls net primary productivity; therefore, SOM input plays a crucial role in soil microbial activity and SOM output. High and low soil moisture could reduce the soil aeration rate, substrate mobility/oxygen availability, and microbial activity and therefore favor SOM accumulation [65]. In addition to our study, soil moisture would also significantly affect SOM variation in an alluvial-diluvial plain in northeastern Ningxia Province [82], Flanders (Belgium) [83], Ohio State of the USA [84], and the Santa Fe River watershed in north-central Florida [85].

4.2. Relative Importance of Human Activity Factors

Human activity factors could also affect soil properties to a certain extent, especially in arable land. In the last decade, many studies have been conducted to investigate the importance of human activities to SOM prediction, which are receiving increasing attention [16]. Yang used phenological parameters extracted from NDVI time-series data to improve the prediction of soil organic carbon (SOC) content and found that the spatial SOC is significantly affected by agricultural management in arable land [6]. Crop rotation can also significantly improve the prediction of SOC information in arable land [4]. Cultivation history is another important human activity factor that can influence the spatial distribution of SOM [7]. In this study, five human activity factors are employed to quantify the importance of SOM prediction in arable land in Heilongjiang Province. The results find that the amount of fertilizer application is the most important human activity factor, ranking 3rd in

our model, which is higher than elevation, precipitation, NDVI, etc. The human footprint, agronomic management level, crop planting type, and irrigation guarantee degree rank 9th, 11th, 14th, and 16th, respectively.

Actually, the amount of fertilizer application is rarely used in SOM prediction. However, the predictive model is considered to be extended using fertilizer application information to improve performance and accuracy [7]. Based on long-term field experiments, many studies have suggested that nitrogen fertilizer application could lead to significant soil acidification worldwide [86–88]. Mechanistically, the chemical could release hydrogen ions (H^+) through nitrification of NH_4^+ and leaching of NO_3^-. The H^+ and base cations leaching with gully erosion runoff may lead to soil acidification [22,89]. Furthermore, the use of chemical fertilizer could increase crop yields by accelerating SOM accumulation [89] and therefore reduce the use of traditional manure application or straw return in similar areas [90–92]. However, some studies have illustrated that massive fertilizer application could increase crop yield over a short period but could not sustain the level of SOM in the long term [93]. Moreover, unreasonable fertilizer input can lead to a decrease in SOM by accelerating soil carbon decomposition in arable land and therefore cannot guarantee sustainable development in the northeastern black soil region in China [22]. Although different studies have controversial conclusions regarding the effect of fertilizer on SOM, our study suggests that adding the amount of fertilizer application could greatly improve the performance and accuracy of SOM prediction. This finding may result from the high acid buffering capacity of black soil. In highly intensive agriculture, biomass production increased with fertilization, leading to enhanced root growth and exudation of organic molecules, increased microorganism biomass and activity, and higher plant residue accumulation. Thus, fertilization up to a certain level may enhance SOM in soil. Additionally, the cumulative number of fertilizers did not reach the threshold that significantly affected SOM in this region. In addition to our study area, the use of fertilizer would also significantly affect SOM in Mediterranean cropping systems [94], northeast China [90,92], south China [88], and Alabama in southeastern USA [95].

The human footprint, agronomic management level, and crop planting type will affect the spatial variation of SOM to varying degrees. The human footprint is measured using eight variables, including built-up environments, population density, electric power infrastructure, crop lands, pasture lands, roads, railways, and navigable waterways, determining human pressure on the soil. Our finding is consistent with previous research [96]. Dong considered that the distance to the river is the most important variable of SOM prediction in alluvial-diluvial plains in China [82,97]. Yang used phenological parameters extracted from NDVI time-series data to improve the prediction of SOC content, and she found that the spatial variation of SOC is significantly affected by agricultural management in arable land [6]. Crop rotation can also significantly affect the amount and spatial variation of SOC in arable land [4]. The use of the irrigation guarantee degree will increase the accuracy of SOM mapping. It can not only meet crop water requirements but also improve soil quality by leaching salts to deep soil horizons [98]. The study indicated that SOM in arable land decreased with the reduction in the degree of irrigation guarantee (Table 5). The average SOM for different irrigation guarantee degrees ranged from 44.82 g/kg to 35.04 g/kg, and the highest mean value was observed in Level I and the lowest in Level III; however, the mean SOM with no irrigation conditions (Level IV) was higher than that in Level III. This may be a result of the low utilization of arable land with no irrigation conditions.

Table 5. Changes in SOM under different irrigation guarantee degrees.

	Number of Samples	Min (g/kg)	Mean (g/kg)	Max (g/kg)	Variance
Level I	536,753	2.93	44.82	80.12	203.93
Level II	517,441	5.90	39.94	80.13	189.80
Level III	609,342	3.30	35.04	80.12	170.48
Level IV	353,508	9.00	38.26	80.13	139.31

Meanwhile, lots of studies have investigated the influence of other human management on SOM in various regions based on field experiments. In Zhejiang Province (China), Wissing analyzed the management-induced organic carbon accumulation in paddy soils and found that the organic carbon concentrations in paddy soils increased from 18 mg/g to 30 mg/g, resulting from the iron oxides strongly interacting with organic matter and playing an important role in the stabilization of SOM [99]. In the same province, Mi reported the effect of four organic materials mulching on the variance of organic soil and found that cattle manure showed the most profound influence, which, combined with NPK fertilizer, resulted in the highest level of SOM [100]. In contrast to the impact of other indicators, such as spent mushroom compost and rice straw residues.

4.3. Model Performance

As mentioned, DSM is conducted based on the relationship between soil and environment. The relationship can only be acquired by collecting soil sampling data and covariate data. Unfortunately, due to the small amount of soil sampling data (compared to the amount of prediction in the study area) used to train the DSM model, many similar studies may draw controversial conclusions [22–26]. Some studies have found that SOM in the black soil region of northeast China has declined rapidly over the past 30 years due to serious soil erosion [23,24]. Other research similarly found that SOM decomposition has exceeded sequestration as a result of the use of chemical fertilization [25]. However, research [26] has found that the SOM in three counties located in the same region did not show a significant difference between 1980 and 2010. The contradictory conclusions may be attributed to the insufficient amount of sampling data [22]. Another notable question is whether the accuracy using a small part of the soil sampling data can be used to reflect the prediction results of the whole study area. The "Specification of Arable Land Survey, Monitoring and Evaluation" project developed by the Ministry of Natural Resources has obtained information on soil properties and soil management at each patch of arable land in Heilongjiang Province. The large amount of soil survey data provided by the project can be used to solve problems related to DSM, which are mentioned earlier, and provide us with more accurate and credible evaluation results.

Table 4 shows that Model 7 has a lower MAE (5.02) and RMSE (7.37) compared to the MAE (5.87) and RMSE (8.37) in Model 1. Additionally, due to the higher R^2 and LCCC, Model 7 exhibited better performance. R^2 indicates that Model 1 with environmental covariates could explain 41% of the variation in SOM distribution. However, Model 7 with environmental covariates and the five human activity factors could explain 57% of the SOM variation.

Previous studies have predicted SOM spatial distribution, and the R^2 from many studies was smaller than 0.5 [63,101–105] (Table 6). In arable land with relatively flat terrain, many widely used environmental variates, such as vegetables and topography, may be too homogenous to represent SOM variation effectively [106,107]. Similarly, in the flat arable land in Heilongjiang Province, Model 1 with only environmental covariates explains 41% of the variation in SOM distribution, which is lower than 0.5. However, Model 7, containing the five human activity factors, finally obtains a higher explanatory power for SOM. This further illustrates the importance of human activity factors as a predictor for SOM distribution.

For SOM prediction in similar areas, our accuracy result ($R^2 = 0.57$) is lower than those of some other studies. Qi used ten covariates to predict SOC in Liaoning Province with the help of the random forest method, obtaining an accuracy result of $R^2 = 0.58$ [1]. Wang combined the cultivation history and environmental variates to predict the SOC of arable land, obtaining an accuracy result of $R^2 = 0.76$, whereas it was 0.65 when considering only environmental variates [7]. This illustrates that if only a small part of the data is used as the validation set, the evaluation result will be overestimated.

Table 6. Comparison of the results of this study with published achievements.

Areas	R^2 (SOM)	Predictive Models	Reference
Brazil	0.33	RF	[104]
Eastern Himalayas	0.36	RF	[102]
Denmark	0.42	Cubist	[63]
Australia	0.25	SVR	[101]
Jiangsu, China	0.53	RK-REML	[103]
China	0.35	XGBoost	[105]
Liaoning, China	0.63	RF	[1]
Northeastern China	0.76	BRT	[7]

SVR: support vector regression; RK-REML: regression kriging with a mixed linear model fitted by residual maximum likelihood; XGBoost: scalable and efficient tree boosting system, eXtreme gradient boost; BRT: boosted regression trees.

4.4. Limitations and Outlook

In relatively flat arable land, many widely used environmental variates may be too homogenous to represent SOM variation effectively. Some environmental variates, such as topography parameters, are not important in predicting SOM. In contrast, due to the excellent production and living conditions in these areas, strong human activity and urbanization could lead to rapid land use change, which will have a significant impact on SOM variation. However, information about human activities with fine spatiotemporal resolution was not available to us. The data obtained from the city statistical yearbook are too coarse for further research. Obtaining data about human activity with fine spatiotemporal resolution is a great challenge.

Due to the unavailability of human activity information, many studies obtained relative data by field surveys and personnel interviews, which is time-consuming and expensive. Therefore, data sharing is a very important choice to promote scientific research. However, the formulation of related policies is still a major challenge.

5. Conclusions

In the study, five human activity factors, including the amount of fertilizer application, human footprint, agronomic management level, crop planting type, and irrigation guarantee degree, were used to explore their effectiveness in predicting SOM in arable land in Heilongjiang Province, China. As a result of the analysis, answers can be drawn about the initial motivation of the study in the Introduction; (1) The model, by combining the five activity factors, increases the SOM spatial prediction by 39% in terms of R^2, 12% in terms of RMSE, 15% in terms of MAE, and 11% in terms of LCCC, showing better prediction accuracy and performance, whereas only environmental covariates account for 41% of the variation in SOM distribution. (2) In the SOM prediction model, soil moisture was the most important environmental covariate, followed by annual average temperature. The amount of fertilizer application, ranking 3rd, is the most important human activity factor. (3) Sufficient SOM sampling data and field survey data were employed for prediction and accuracy verification, finding that the evaluation result will be overestimated when only a small part of the sampling data is used.

However, the relative importance of environmental conditions and human activity covariates may vary in other regions, which requires more analysis and discussion. During the research process, the data about human activity with fine spatiotemporal resolution are still a great challenge in SOM prediction in arable land.

Author Contributions: Conceptualization, L.N. and C.C.; methodology, S.S.; software, L.Z.; validation, S.M.; writing—original draft preparation, L.N.; writing—review and editing, L.N.; visualization, Y.S.; supervision, C.C. and L.Z.; Software, X.L.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

References

1. Qi, L.; Wang, S.; Zhuang, Q.; Yang, Z.; Bai, S.; Jin, X.; Lei, G. Spatial-temporal changes in soil organic carbon and pH in the Liaoning Province of China: A modeling analysis based on observational data. *Sustainability* **2019**, *11*, 3569. [CrossRef]
2. Dick, W. Organic carbon, nitrogen, and phosphorus concentrations and pH in soil profiles as affected by tillage intensity. *Soil Sci. Soc. Am. J.* **1983**, *47*, 102–107. [CrossRef]
3. Post, W.M.; Kwon, K.C. Soil carbon sequestration and land-use change: Processes and potential. *Glob. Chang. Biol.* **2000**, *6*, 317–327. [CrossRef]
4. Yang, L.; Song, M.; Zhu, A.-X.; Qin, C.; Zhou, C.; Qi, F.; Li, X.; Chen, Z.; Gao, B. Predicting soil organic carbon content in croplands using crop rotation and Fourier transform decomposed variables. *Geoderma* **2019**, *340*, 289–302. [CrossRef]
5. Hoffmann, M.; Pohl, M.; Jurisch, N.; Prescher, A.-K.; Campa, E.M.; Hagemann, U.; Remus, R.; Verch, G.; Sommer, M.; Augustin, J. Maize carbon dynamics are driven by soil erosion state and plant phenology rather than nitrogen fertilization form. *Soil Tillage Res.* **2018**, *175*, 255–266. [CrossRef]
6. Yang, L.; He, X.; Shen, F.; Zhou, C.; Zhu, A.-X.; Gao, B.; Chen, Z.; Li, M. Improving prediction of soil organic carbon content in croplands using phenological parameters extracted from NDVI time series data. *Soil Tillage Res.* **2020**, *196*, 104465. [CrossRef]
7. Wang, Y.; Wang, S.; Adhikari, K.; Wang, Q.; Sui, Y.; Xin, G. Effect of cultivation history on soil organic carbon status of arable land in northeastern China. *Geoderma* **2019**, *342*, 55–64. [CrossRef]
8. Paustian, K.; Andren, O.; Janzen, H.; Lal, R.; Smith, P.; Tian, G.; Tiessen, H.; Van Noordwijk, M.; Woomer, P. Agricultural soils as a sink to mitigate CO_2 emissions. *Soil Use Manag.* **1997**, *13*, 230–244. [CrossRef]
9. Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **2004**, *304*, 1623–1627. [CrossRef]
10. Smith, P. Carbon sequestration in croplands: The potential in Europe and the global context. *Eur. J. Agron.* **2004**, *20*, 229–236. [CrossRef]
11. Hans, J. *Factors of Soil Formation: A System of Quantitative Pedology*; Dover Publication: Mineola, NY, USA, 1941.
12. McBratney, A.B.; Santos, M.M.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
13. Hamzehpour, N.; Shafizadeh-Moghadam, H.; Valavi, R. Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. *Catena* **2019**, *182*, 104141. [CrossRef]
14. Ramcharan, A.; Hengl, T.; Nauman, T.; Brungard, C.; Waltman, S.; Wills, S.; Thompson, J. Soil property and class maps of the conterminous United States at 100-meter spatial resolution. *Soil Sci. Soc. Am. J.* **2018**, *82*, 186–201. [CrossRef]
15. Hengl, T.; de Jesus, J.M.; Heuvelink, G.B.; Gonzalez, M.R.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]
16. Grunwald, S.; Thompson, J.; Boettinger, J. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Sci. Soc. Am. J.* **2011**, *75*, 1201–1213. [CrossRef]
17. Liang, Z.; Chen, S.; Yang, Y.; Zhao, R.; Shi, Z.; Rossel, R.A.V. National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. *Geoderma* **2019**, *335*, 47–56. [CrossRef]
18. Hengl, T.; Leenaars, J.G.; Shepherd, K.D.; Walsh, M.G.; Heuvelink, G.B.; Mamo, T.; Tilahun, H.; Berkhout, E.; Cooper, M.; Fegraus, E. Soil nutrient maps of Sub-Saharan Africa: Assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutr. Cycl. Agroecosyst.* **2017**, *109*, 77–102. [CrossRef]
19. Zhou, Y.; Hartemink, A.E.; Shi, Z.; Liang, Z.; Lu, Y. Land use and climate change effects on soil organic carbon in North and Northeast China. *Sci. Total Environ.* **2019**, *647*, 1230–1238. [CrossRef]
20. Wadoux, A.M.-C. Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma* **2019**, *351*, 59–70. [CrossRef]
21. Zhao, R.; Biswas, A.; Zhou, Y.; Zhou, Y.; Shi, Z.; Li, H. Identifying localized and scale-specific multivariate controls of soil organic matter variations using multiple wavelet coherence. *Sci. Total Environ.* **2018**, *643*, 548–558. [CrossRef]
22. Ou, Y.; Rousseau, A.N.; Wang, L.; Yan, B. 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]
23. Zhang, Y.; Wu, Y.; Liu, B.; Zheng, Q.; Yin, J. Characteristics and factors controlling the development of ephemeral gullies in cultivated catchments of black soil region, Northeast China. *Soil Tillage Res.* **2007**, *96*, 28–41. [CrossRef]

24. Wu, Y.; Zheng, Q.; Zhang, Y.; Liu, B.; Cheng, H.; Wang, Y. Development of gullies and sediment production in the black soil region of northeastern China. *Geomorphology* **2008**, *101*, 683–691. [CrossRef]

25. Jiao, X.; Gao, C.; Sui, Y.; Lü, G.; Wei, D. Effects of long-term fertilization on soil carbon and nitrogen in Chinese Mollisols. *Agron. J.* **2014**, *106*, 1018–1024. [CrossRef]

26. Chun-hua, Z.; Zong-ming, W.; Chun-ying, R.; Bai, Z.; Kai-shan, S.; Dian-wei, L. Temporal and spatial variations of soil organic and total nitrogen in the Songnen Plain maize belt. *Geogr. Reserach* **2011**, *30*, 256–268.

27. Zhao, Y.; Jiang, Q.; Wang, Z. The System Evaluation of Grain Production Efficiency and Analysis of Driving Factors in Heilongjiang Province. *Water* **2019**, *11*, 1073. [CrossRef]

28. Xu, S. Temporal and Spatial Characteristics of the Change of Cultivated Land Resources in the Black Soil Region of Heilongjiang Province (China). *Sustainability* **2019**, *11*, 38. [CrossRef]

29. *NY/T1121.6-2006*; Soil Testing-Part 6: Method for Determination of Soil Organic Matter. Ministry of Agriculture: Beijing, China, 2006.

30. Danielson, J.J.; Gesch, D.B. *Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010)*; US Department of the Interior, US Geological Survey: Washington, DC, USA, 2011.

31. Sayre, R.; Dangermond, J.; Frye, C.; Vaughan, R.; Aniello, P.; Breyer, S.; Cribbs, D.; Hopkins, D.; Nauman, R.; Derrenbacher, W. *A New Map of Global Ecological Land Units—An Ecophysiographic Stratification Approach*; Association of American Geographers: Washington, DC, USA, 2014.

32. Pelletier, J.D.; Broxton, P.D.; Hazenberg, P.; Zeng, X.; Troch, P.A.; Niu, G.Y.; Williams, Z.; Brunke, M.A.; Gochis, D. A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. *J. Adv. Model. Earth Syst.* **2016**, *8*, 41–65. [CrossRef]

33. Pelletier, J.; Broxton, P.; Hazenberg, P.; Zeng, X.; Troch, P.; Niu, G.; Williams, Z.; Brunke, M.; Gochis, D. *Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers*; ORNL DAAC: Oak Ridge, TN, USA, 2016.

34. Abatzoglou, J.T.; Dobrowski, S.Z.; Parks, S.A.; Hegewisch, K.C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* **2018**, *5*, 170191. [CrossRef]

35. Fan, Y.; Li, H.; Miguez-Macho, G. Global patterns of groundwater table depth. *Science* **2013**, *339*, 940–943. [CrossRef]

36. Pekel, J.-F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422. [CrossRef] [PubMed]

37. Venter, O.; Sanderson, E.W.; Magrach, A.; Allan, J.R.; Beher, J.; Jones, K.R.; Possingham, H.P.; Laurance, W.F.; Wood, P.; Fekete, B.M. Global terrestrial Human Footprint maps for 1993 and 2009. *Sci. Data* **2016**, *3*, 160067. [CrossRef] [PubMed]

38. Mallick, J.; AlMesfer, M.K.; Singh, V.P.; Falqi, I.I.; Singh, C.K.; Alsubih, M.; Kahla, N.B. Evaluating the NDVI–Rainfall Relationship in Bisha Watershed, Saudi Arabia Using Non-Stationary Modeling Technique. *Atmosphere* **2021**, *12*, 593. [CrossRef]

39. Mahmoudabadi, E.; Karimi, A.; Haghnia, G.H.; Sepehr, A. Digital soil mapping using remote sensing indices, terrain attributes, and vegetation features in the rangelands of northeastern Iran. *Environ. Monit. Assess.* **2017**, *189*, 500. [CrossRef]

40. Lamichhane, S.; Kumar, L.; Wilson, B.J. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma* **2019**, *352*, 395–413. [CrossRef]

41. Kumari, N.; Srivastava, A.; Dumka, U.C.J.C. A long-term spatiotemporal analysis of vegetation greenness over the Himalayan Region using Google Earth Engine. *Climate* **2021**, *9*, 109. [CrossRef]

42. Liu, H.; Yan, Y.; Zhang, X.; Qiu, Z.; Wang, N.; Yu, W. Remote sensing extraction of crop planting structure oriented to agricultural regionalizaiton. *Chin. J. Agric. Resour. Reg. Plan.* **2017**, *38*, 43–54.

43. Yao, X.; Zhu, D.; Ye, S.; Yun, W.; Zhang, N.; Li, L.J.C.; Agriculture, E.i. A field survey system for land consolidation based on 3S and speech recognition technology. *Comput. Electron. Agric.* **2016**, *127*, 659–668. [CrossRef]

44. Ye, S.; Song, C.; Shen, S.; Gao, P.; Cheng, C.; Cheng, F.; Wan, C.; Zhu, D. Spatial pattern of arable land-use intensity in China. *Land Use Policy* **2020**, *99*, 104845. [CrossRef]

45. Wan, C.; Kuzyakov, Y.; Cheng, C.; Ye, S.; Gao, B.; Gao, P.; Ren, S.; Yun, W. A soil sampling design for arable land quality observation by using SPCOSA–CLHS hybrid approach. *Land Degrad. Dev.* **2021**, *32*, 4889–4906. [CrossRef]

46. Liao, Y.; Wang, J.; Meng, B.; Li, X. Integration of GP and GA for mapping population distribution. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 47–67. [CrossRef]

47. Kenett, D.Y.; Tumminello, M.; Madi, A.; Gur-Gershgoren, G.; Mantegna, R.N.; Ben-Jacob, E. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS ONE* **2010**, *5*, e15032. [CrossRef]

48. Eichler, M.; Dahlhaus, R.; Sandkühler, J. Partial correlation analysis for the identification of synaptic connections. *Biol. Cybern.* **2003**, *89*, 289–302. [CrossRef]

49. Peng, S.; Piao, S.; Ciais, P.; Myneni, R.B.; Chen, A.; Chevallier, F.; Dolman, A.J.; Janssens, I.A.; Penuelas, J.; Zhang, G. Asymmetric effects of daytime and night-time warming on Northern Hemisphere vegetation. *Nature* **2013**, *501*, 88–92. [CrossRef]

50. Heung, B.; Bulmer, C.E.; Schmidt, M.G. Predictive soil parent material mapping at a regional-scale: A random forest approach. *Geoderma* **2014**, *214*, 141–154. [CrossRef]

51. Zhi, J.; Zhang, G.; Yang, F.; Yang, R.; Liu, F.; Song, X.; Zhao, Y.; Li, D. Predicting mattic epipedons in the northeastern Qinghai-Tibetan Plateau using Random Forest. *Geoderma Reg.* **2017**, *10*, 1–10. [CrossRef]

52. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

53. Forkuor, G.; Hounkpatin, O.K.; Welp, G.; Thiel, M. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PLoS ONE* **2017**, *12*, e0170478. [CrossRef]

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

55. Deng, X.; Chen, X.; Ma, W.; Ren, Z.; Zhang, M.; Grieneisen, M.L.; Long, W.; Ni, Z.; Zhan, Y.; Lv, X. Baseline map of organic carbon stock in farmland topsoil in East China. *Agric. Ecosyst. Environ.* **2018**, *254*, 213–223. [CrossRef]

56. Jeong, G.; Oeverdieck, H.; Park, S.J.; Huwe, B.; Ließ, M. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. *Catena* **2017**, *154*, 73–84. [CrossRef]

57. Behrens, T.; Schmidt, K.; Ramirez-Lopez, L.; Gallant, J.; Zhu, A.-X.; Scholten, T. Hyper-scale digital soil mapping and soil formation analysis. *Geoderma* **2014**, *213*, 578–588. [CrossRef]

58. Grimm, R.; Behrens, T.; Märker, M.; Elsenbeer, H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma* **2008**, *146*, 102–113. [CrossRef]

59. Shi, J.; Yang, L.; Zhu, A.; Qin, C.; Liang, P.; Zeng, C.; Pei, T. Machine-learning variables at different scales vs. Knowledge-based variables for mapping multiple soil properties. *Soil Sci. Soc. Am. J.* **2018**, *82*, 645–656. [CrossRef]

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

61. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

62. Sumfleth, K.; Duttmann, R. Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. *Ecol. Indic.* **2008**, *8*, 485–501. [CrossRef]

63. Adhikari, K.; Hartemink, A.E.; Minasny, B.; Kheir, R.B.; Greve, M.B.; Greve, M.H. Digital mapping of soil organic carbon contents and stocks in Denmark. *PLoS ONE* **2014**, *9*, e105519. [CrossRef]

64. Yang, R.-M.; Zhang, G.-L.; Liu, F.; Lu, Y.-Y.; Yang, F.; Yang, F.; Yang, M.; Zhao, Y.-G.; Li, D.-C. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecol. Indic.* **2016**, *60*, 870–878. [CrossRef]

65. Wiesmeier, M.; Urbanski, L.; Hobley, E.; Lang, B.; von Lützow, M.; Marin-Spiotta, E.; van Wesemael, B.; Rabot, E.; Ließ, M.; Garcia-Franco, N. Soil organic carbon storage as a key function of soils—A review of drivers and indicators at various scales. *Geoderma* **2019**, *333*, 149–162. [CrossRef]

66. Schillaci, C.; Acutis, M.; Lombardo, L.; Lipani, A.; Fantappie, M.; Märker, M.; Saia, S. Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. *Sci. Total Environ.* **2017**, *601*, 821–832. [CrossRef] [PubMed]

67. Ma, Y.; Minasny, B.; Wu, C. Mapping key soil properties to support agricultural production in Eastern China. *Geoderma Reg.* **2017**, *10*, 144–153. [CrossRef]

68. Wang, S.; Zhuang, Q.; Wang, Q.; Jin, X.; Han, C. Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China. *Geoderma* **2017**, *305*, 250–263. [CrossRef]

69. Chaplot, V.; Bouahom, B.; Valentin, C. Soil organic carbon stocks in Laos: Spatial variations and controlling factors. *Glob. Change Biol.* **2010**, *16*, 1380–1393. [CrossRef]

70. Doetterl, S.; Stevens, A.; Six, J.; Merckx, R.; Van Oost, K.; Pinto, M.C.; Casanova-Katny, A.; Muñoz, C.; Boudin, M.; Venegas, E.Z. Soil carbon storage controlled by interactions between geochemistry and climate. *Nat. Geosci.* **2015**, *8*, 780–783. [CrossRef]

71. Meier, I.C.; Leuschner, C. Variation of soil and biomass carbon pools in beech forests across a precipitation gradient. *Glob. Change Biol.* **2010**, *16*, 1035–1045. [CrossRef]

72. Conant, R.T.; Ryan, M.G.; Ågren, G.I.; Birge, H.E.; Davidson, E.A.; Eliasson, P.E.; Evans, S.E.; Frey, S.D.; Giardina, C.P.; Hopkins, F.M. Temperature and soil organic matter decomposition rates–synthesis of current knowledge and a way forward. *Glob. Chang. Biol.* **2011**, *17*, 3392–3404. [CrossRef]

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

74. Von Lützow, M.; Kögel-Knabner, I. Temperature sensitivity of soil organic matter decomposition—What do we know? *Biol. Fertil. Soils* **2009**, *46*, 1–15. [CrossRef]

75. Stumpf, F.; Keller, A.; Schmidt, K.; Mayr, A.; Gubler, A.; Schaepman, M. Spatio-temporal land use dynamics and soil organic carbon in Swiss agroecosystems. *Agric. Ecosyst. Environ.* **2018**, *258*, 129–142. [CrossRef]

76. Song, X.-D.; Yang, F.; Ju, B.; Li, D.-C.; Zhao, Y.-G.; Yang, J.-L.; Zhang, G.-L. The influence of the conversion of grassland to cropland on changes in soil organic carbon and total nitrogen stocks in the Songnen Plain of Northeast China. *Catena* **2018**, *171*, 588–601. [CrossRef]

77. Peng, Y.; Xiong, X.; Adhikari, K.; Knadel, M.; Grunwald, S.; Greve, M.H. Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra. *PLoS ONE* **2015**, *10*, e0142295. [CrossRef]

78. Paul, E.A. The nature and dynamics of soil organic matter: Plant inputs, microbial transformations, and organic matter stabilization. *Soil Biol. Biochem.* **2016**, *98*, 109–126. [CrossRef]

79. Brady, N.C.; Weil, R.R.; Weil, R.R. *The Nature and Properties of Soils*; Prentice Hall: Hoboken, NJ, USA, 2008; Volume 13.

80. Frank, D.A.; Pontes, A.W.; McFarlane, K.J. Controls on soil organic carbon stocks and turnover among North American ecosystems. *Ecosystems* **2012**, *15*, 604–615. [CrossRef]

81. Gray, J.M.; Bishop, T.F.; Wilson, B.R. Factors controlling soil organic carbon stocks with depth in eastern Australia. *Soil Sci. Soc. Am. J.* **2015**, *79*, 1741–1751. [CrossRef]

82. Dong, W.; Wu, T.; Luo, J.; Sun, Y.; Xia, L. Land parcel-based digital soil mapping of soil nutrient properties in an alluvial-diluvia plain agricultural area in China. *Geoderma* **2019**, *340*, 234–248. [CrossRef]

83. 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]

84. Tan, Z.; Lal, R.; Smeck, N.; Calhoun, F. Relationships between surface soil organic carbon pool and site variables. *Geoderma* **2004**, *121*, 187–195. [CrossRef]

85. Vasques, G.; Grunwald, S.; Comerford, N.; Sickman, J. Regional modelling of soil carbon at multiple depths within a subtropical watershed. *Geoderma* **2010**, *156*, 326–336. [CrossRef]

86. Russell, A.E.; Laird, D.; Parkin, T.B.; Mallarino, A.P. Impact of nitrogen fertilization and cropping system on carbon sequestration in Midwestern Mollisols. *Soil Sci. Soc. Am. J.* **2005**, *69*, 413–422. [CrossRef]

87. Vieira, F.; Bayer, C.; Mielniczuk, J.; Zanatta, J.; Bissani, C. Long-term acidification of a Brazilian Acrisol as affected by no till cropping systems and nitrogen fertiliser. *Soil Res.* **2008**, *46*, 17–26. [CrossRef]

88. Zhou, J.; Xia, F.; Liu, X.; He, Y.; Xu, J.; Brookes, P.C. Effects of nitrogen fertilizer on the acidification of two typical acid soils in South China. *J. Soils Sed.* **2014**, *14*, 415–422. [CrossRef]

89. Haynes, R.J.; Naidu, R. Influence of lime, fertilizer and manure applications on soil organic matter content and soil physical conditions: A review. *Nutr. Cycl. Agroecosyst.* **1998**, *51*, 123–137. [CrossRef]

90. Yang, X.; Zhang, X.; Fang, H.; Zhu, P.; Ren, J.; Wang, L. Long-term effects of fertilization on soil organic carbon changes in continuous corn of northeast China: RothC model simulations. *Environ. Manag.* **2003**, *32*, 459–465.

91. Yang, X.; Zhang, X.; Deng, W.; Fang, H. Black soil degradation by rainfall erosion in Jilin, China. *Land Degrad. Dev.* **2003**, *14*, 409–420. [CrossRef]

92. Liu, Z.; Yang, X.; Hubbard, K.G.; Lin, X. Maize potential yields and yield gaps in the changing climate of northeast China. *Glob. Change Biol.* **2012**, *18*, 3441–3454. [CrossRef]

93. 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]

94. Aguilera, E.; Lassaletta, L.; Gattinger, A.; Gimeno, B.S. Managing soil carbon for climate change mitigation and adaptation in Mediterranean cropping systems: A meta-analysis. *Agric. Ecosyst. Environ.* **2013**, *168*, 25–36. [CrossRef]

95. Sainju, U.M.; Senwo, Z.N.; Nyakatawa, E.Z.; Tazisong, I.A.; Reddy, K.C. Soil carbon and nitrogen sequestration as affected by long-term tillage, cropping systems, and nitrogen fertilizer sources. *Agric. Ecosyst. Environ.* **2008**, *127*, 234–240. [CrossRef]

96. Liu, X.; Herbert, S.; Hashemi, A.; Zhang, X.; Ding, G. Effects of agricultural management on soil organic matter and carbon transformation-a review. *Plant Soil Environ.* **2006**, *52*, 531. [CrossRef]

97. Syswerda, S.; Corbin, A.; Mokma, D.; Kravchenko, A.; Robertson, G. Agricultural management and soil carbon storage in surface vs. deep layers. *Soil Sci. Soc. Am. J.* **2011**, *75*, 92–101. [CrossRef]

98. Yang, J.; Zhang, S.; Li, Y.; Bu, K.; Zhang, Y.; Chang, L.; Zhang, Y. Dynamics of saline-alkali land and its ecological regionalization in western Songnen Plain, China. *Chin. Geogr. Sci.* **2010**, *20*, 159–166. [CrossRef]

99. Wissing, L.; Kölbl, A.; Häusler, W.; Schad, P.; Cao, Z.-H.; Kögel-Knabner, I.J.S.; Research, T. Management-induced organic carbon accumulation in paddy soils: The role of organo-mineral associations. *Soil Tillage Res.* **2013**, *126*, 60–71. [CrossRef]

100. Mi, W.; Wu, L.; Brookes, P.C.; Liu, Y.; Zhang, X.; Yang, X.J.S.; Research, T. Changes in soil organic carbon fractions under integrated management systems in a low-productivity paddy soil given different organic amendments and chemical fertilizers. *Soil Tillage Res.* **2016**, *163*, 64–70. [CrossRef]

101. Somarathna, P.; Malone, B.; Minasny, B. Mapping soil organic carbon content over New South Wales, Australia using local regression kriging. *Geoderma Reg.* **2016**, *7*, 38–48. [CrossRef]

102. Dorji, T.; Odeh, I.O.; Field, D.J.; Baillie, I.C. Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. *For. Ecol. Manag.* **2014**, *318*, 91–102. [CrossRef]

103. Zhao, M.-S.; Rossiter, D.G.; Li, D.-C.; Zhao, Y.-G.; Liu, F.; Zhang, G.-L. Mapping soil organic matter in low-relief areas based on land surface diurnal temperature difference and a vegetation index. *Ecol. Indic.* **2014**, *39*, 120–133. [CrossRef]

104. Gomes, L.C.; Faria, R.M.; de Souza, E.; Veloso, G.V.; Schaefer, C.E.G.; Fernandes Filho, E.I. Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma* **2019**, *340*, 337–350. [CrossRef]

105. Liang, Z.; Chen, S.; Yang, Y.; Zhou, Y.; Shi, Z. High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling. *Sci. Total Environ.* **2019**, *685*, 480–489. [CrossRef]

106. Zhu, A.; Liu, F.; Li, B.; Pei, T.; Qin, C.; Liu, G.; Wang, Y.; Chen, Y.; Ma, X.; Qi, F. Differentiation of soil conditions over low relief areas using feedback dynamic patterns. *Soil Sci. Soc. Am. J.* **2010**, *74*, 861–869. [CrossRef]

107. Zeng, C.; Zhu, A.-X.; Liu, F.; Yang, L.; Rossiter, D.G.; Liu, J.; Wang, D. The impact of rainfall magnitude on the performance of digital soil mapping over low-relief areas using a land surface dynamic feedback method. *Ecol. Indic.* **2017**, *72*, 297–309. [CrossRef]