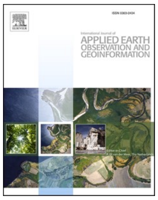
International Journal of Applied Earth Observations and Geoinformation 102 (2021) 102428

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables

Lin Yang [a], Yanyan Cai [a], Lei Zhang [a], Mao Guo [a], Anqi Li [a], Chenghu Zhou [a,b,*]

[a] School of Geography and Ocean Science, Nanjing University, Nanjing, 210023 China

[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, CAS, Beijing 100101, China

ARTICLE INFO

Keywords:
Deep learning
Soil organic carbon
Convolutional neural network (CNN)
Land surface phenology

ABSTRACT

Obtaining the spatial distribution information of soil organic carbon (SOC) is significant to quantify the carbon budget and guide land management for migrating carbon emissions. Digital soil mapping of SOC at a regional scale is challenging due to the complex SOC-environment relationships. Vegetation phenology that directly indicates a long time vegetation growth characteristics can be potential environmental covariates for SOC prediction. Deep learning has been developed for soil mapping recently due to its ability of constructing high-level features from the raw data. However, only dozens of predictors were used in most of those studies. It is not clear that how deep learning with long term land surface phenology product performs for SOC prediction at a regional scale. This paper explored the effectiveness of ten-years MODIS MCD12Q2 phenology variables for SOC prediction with a convolutional neural network (CNN) model in Anhui province, China. Random forest (RF) was applied to compare with CNN using three groups of environmental variables. The results showed that adding the land surface phenology variables into the pool of the natural environmental variables improved the prediction accuracy of CNN by 5.57% of RMSE and 31.29% of R^2. Adding phenology variables obtained a higher accuracy improvement than adding Normalized Differences Vegetation Indices. The CNN obtained a higher prediction accuracy than RF regardless of using which group of variables. This study proved that land surface phenology metrics were effective predictors and CNN was a promising method for soil mapping at a regional scale.

1. Introduction

Soil organic carbon (SOC), as one of the most important soil property, is crucial for many soil functions and ecosystem services (Grinand et al., 2017). It also plays an important role in global carbon cycle (Lal, 2004; Hamzehpour et al., 2019), and drew wide attentions in climate change studies (Bradford et al., 2016). Predicting the spatial distribution of soil organic carbon is significant to guide land management for soil health and migrating carbon emissions (Vaudour et al., 2016; Angelopoulou et al., 2019).

The methods for SOC spatial prediction have been rapidly developed in the past decades. One common approach is interpolation (such as, Kriging) based on sample data (Dou et al., 2010; Elbasiouny et al., 2014). However, a high interpolation accuracy requires a large amount of samples which is labor-extensive and costly to collect. Since SOC at a location are regulated by interactions of climate, terrain, and vegetation, etc., environmental covariates are increasingly used for SOC prediction to increase prediction accuracy with limited samples (Grinand et al., 2017; Lamichhane et al., 2019). Those approaches to predict SOC based on the soil-environment relationships are so called digital soil mapping (DSM) (McBratney et al., 2003; An et al., 2018). Both the environmental covariates and the prediction methods determine the DSM accuracy. It is, therefore, meaningful to develop influential environmental covariates and effective prediction methods for producing accurate SOC maps.

Numerous variables have been used for mapping SOC, such as climatic variables, terrain attributes, soil parent materials, and so on (McBratney et al., 2003; Grinand et al., 2017). Remote sensing (RS) data has been increasingly used in DSM because of its easy accessibility, large spatial coverage and long time series. Those commonly-used RS variables include surface reflectance, band ratios representing soil or vegetation, and vegetation indices (McBratney et al., 2003; Zhou et al., 2021). The surface reflectance or band ratios indicating soil is mainly effective for bare soil (Peng et al., 2003), and vegetation indices such as

* Corresponding author.
E-mail addresses: yanglin@nju.edu.cn (L. Yang), MG1927050@smail.nju.edu.cn (Y. Cai), zhanglei@smail.nju.edu.cn (L. Zhang), zhouch@lreis.ac.cn (C. Zhou).

https://doi.org/10.1016/j.jag.2021.102428
Received 21 February 2021; Received in revised form 3 June 2021; Accepted 29 June 2021
Available online 12 July 2021

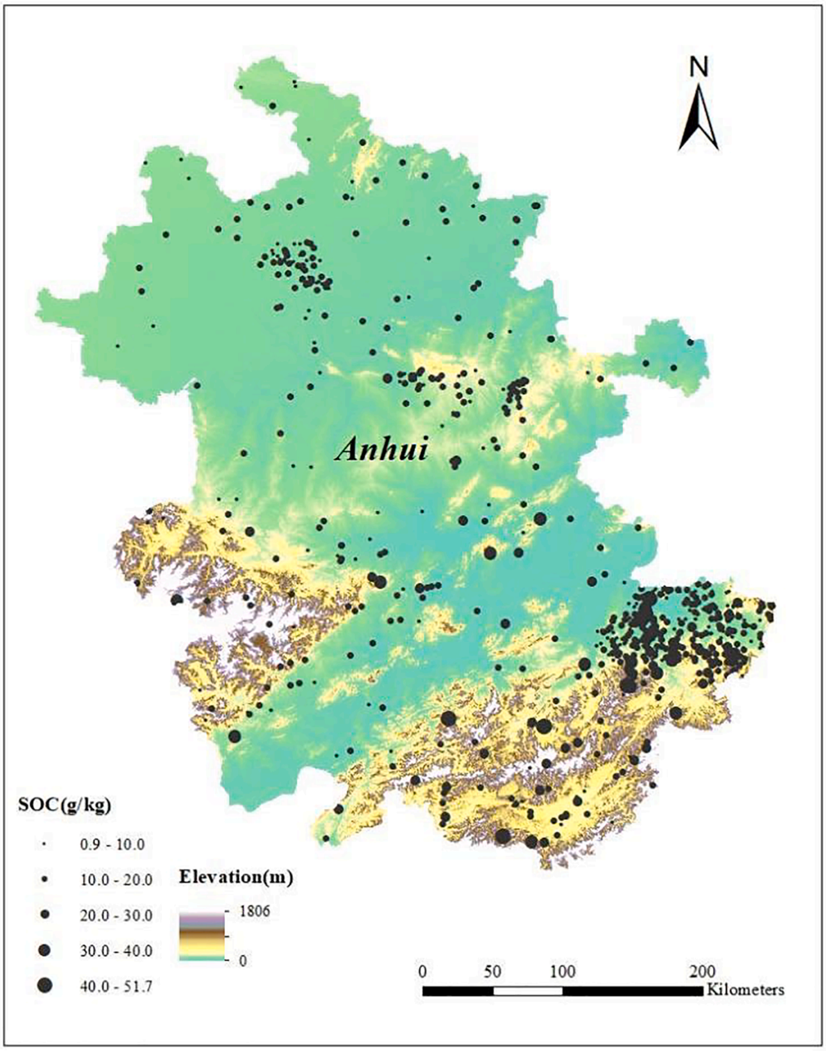


Fig. 1. The study area and the sample locations.

Normalized Differences Vegetation Index (NDVI) are more widely used (Mora-Vallejo et al., 2008). However, the use of vegetation indices would not always improve the mapping accuracy much (Wang et al. 2018; Schillaci et al., 2017). In case of interactions between SOC and vegetation over a long period, variables based on the long term RS data better indicating the vegetation growth status instead of several single vegetation indices can be developed for SOC mapping.

Vegetation phenology is promising variables directly indicating vegetation growth characteristics of a long period. Vegetation phenology is reported to be closely related to soil functions (Araya et al., 2016) and is largely linked to SOC dynamics (Hoffmann et al., 2018). The study of Yang et al. (2020) verified the effectiveness of the phenological parameters extracted based on NDVI profiles in a cropland SOC prediction study. However, the extraction of phenological parameters from NDVI profiles needs field observations to train the extraction model, which is costly especially for a large area. Alternatively, land surface products such as Moderate-resolution Imaging Spectroradiometer (MODIS) Land Cover Dynamics product, which provide global annual land surface phenology measurements over years, may serve as facilitate and powerful predictors for SOC prediction at a regional scale.

Many prediction methods have been employed for digital soil mapping over the past decades (Grunwald, 2009; Wadoux, 2019; Zhang et al., 2021). Machine learning is one of the most accurate methods (Grinand et al., 2017; Hamzehpour et al., 2019). Currently, deep learning has become a promising direction in machine learning (LeCun et al., 2015). The major advantage of deep learning models is their capability of extracting high-level features from raw data through a series of processing layers (LeCun et al., 2015; Tien Bui et al., 2020). Deep learning have been successfully employed in time series analysis (LeCun and Bengio, 1995), computer vision (Krizhevsky et al., 2012), nature language processing (Lee et al., 2018), landslide susceptibility assessment (Tien Bui et al., 2020), and so on. As for soil mapping, several studies have been conducted with deep learning recently. Behrens et al. (2018) developed a deep learning method for soil texture and zinc mapping based on multi-scale terrain attributes. Convolutional neural network (CNN) has been used used to simultaneously predict multiple soil properties (Ng et al., 2019; Wadoux, 2019; Padarian et al., 2019), and classified soil aggregates based on images (Azizi et al., 2020). The above studies verify the out-performance of deep learning in soil mapping.

In most of the above soil mapping studies using deep learning, only dozens of predictors were used. The number of inter-annual phenological parameters are usually hundreds, which is advantageous but not used for SOC prediction with deep learning. It is not clear that how effective deep learning with long term land surface phenology product for SOC prediction is at a regional scale. Thus, the objectives of this paper are, 1) to develop a deep learning model, specifically convolutional neural network (CNN) for SOC prediction in Anhui province with time series satellite-based land surface phenological variables, 2) to evaluate whether adding phenologcial variables would improved the

Table 1
The environment variables of the study area.

Factors	Variables	Data Sources	Original resolution
Climate	Annual mean temperature Annual precipitation Annual mean evaporation Annual accumulated temperature above 10 °C (Acc10) Arid index Moisture index	Chinese Academy of Agricultural Sciences (An et al., 2018)	1, 000 *m*
Terrain	Elevation	The Spaceshuttle Radar Topographical Mission (SRTM) Digital Elevation Model (DEM) (https://search.earthdata.nasa.gov/).	90 *m*
	Slope gradient Planform curvature Profile curvature	Calculated with the terrain analysis software 3DMapper (www.terriananalytics.com) based on the SRTM DEM.	
	Topographic wetness index (TWI)	Calculated with the multiple flow direction strategy (MFD-md, Qin et al., 2007) using a software SimDTA V1.0.	
Parent materials	Parent materials	Extracted from the 1:500, 000 geological map database of China (http://www.ngac.org.cn/).	–
Vegetation	Annual normalized differences vegetation index (NDVI) from 2007 to 2016	Downloaded from Data Center for Resources and Environmental Sciences Chinese Academy of Sciences (RESDC), http://www.resdc.cn.	1, 000 *m*

SOC prediction accuracy compared with common-used variables including climate, terrain with/without vegetation index, and 3) to compare the developed CNN with an accurate machine learning method, random forest, for SOC prediction.

2. Study area and data

2.1. Study area

Anhui province is located in central-eastern China (29°23′44″N-34°39′5″N, 114°52′35″E-119°39′37″E), which covers an area of 1.40 × 10^5 km^2. Elevation varies from 0 to 1806 m with large flat plains in northern areas, low hills in middle areas and mountains in southwestern and southern areas (Fig. 1). The average annual temperature of Anhui is 14–16 °C, and the annual precipitation is 750–2000 mm. The parent materials in Anhui are variable, including granite, basalt, schist, perknite, diorite, sandstone, shale, and others. According to the national soil genetic classification system of China, five soil orders occur in Anhui, including Semi-hydromorphic soils in the northern area, Primitive soils, Anthropogenic soils and Eluvial soils in the central area, and Ferro-allitic soils in the southern areas (Data Center for Resources and Environmental Sciences Chinese Academy of Sciences (RESDC)). Anhui has experienced intensive human activities. There are several land use types, i.e. cropland, forest, shrub and grass, and construction land.

2.2. Sample data

In the study area, 733 samples were collected in 2011, 2015, and 2016 from three projects with different sampling strategies (Fig. 1). Two hundred eighty nine samples were collected across the whole province using a stratified random sampling strategy with the parent materials as strata, 183 samples collected based on experts' knowledge in three counties (Dingyuan, Mengcheng and Xuanzhou), and 261 samples collected in Xuancheng (the dense samples mainly in southern Anhui in Fig. 1). The Xuancheng samples included 60 samples collected based on a stratified random sampling strategy, 57 sample points collected with 10 km by 10 km grid arrangement based on a systematic sampling strategy, 57 samples collected to cover environmental feature space with an integrative hierarchical stepwise sampling (Yang et al., 2013), 57 samples using a heuristic uncertainty directed sampling (Zhang et al., 2016) and 30 samples based on environmental similarity (Ma et al, 2020). Although the samples were not uniformly distributed over the geographic space, the samples represented the distribution of the environmental variables well. The representation on the environmental feature space ensured a good chance of fitting soil-environment relationships through deep learning or machine learning.

Soils were sampled at a depth of 0–20 cm at each sample location. A dichromate oxidation method (external heat applied) (Nelson et al., 1996) was used to measure soil organic carbon concentration (g/kg).

Table 2
The description of the phenology metrics in MODIS MCD12Q2 product.

Phenology metrics	Description	Unit
NumCycle	The total number of vegetation cycles with peak in product year	–
Greenup	The date of EVI2 first crossing 15% of the segment EVI2 amplitude	Days
MidGreenup	The date of EVI2 first crossing 50% of the segment EVI2 amplitude	Days
Maturity	The date of EVI2 first crossing 90% of the segment EVI2 amplitude	Days
Peak	The date of EVI2 get to the top of segment EVI2 amplitude	Days
Senescence	The date of EVI2 last crossing 90% of the segment EVI2 amplitude	Days
MidGreendown	The date of EVI2 last crossing 50% of the segment EVI2 amplitude	Days
Dormancy	The date of EVI2 last crossing 15% of the segment EVI2 amplitude	Days
EVI_Minimum	The minimum value of segment EVI2 (the 2-band Enhanced Vegetation Index) amplitude	NBAR (Nadir Bidirectional Reflectance Distribution Function-Adjusted Reflectance)-EVI2
EVI_Amplitude	The value of segment EVI2 amplitude maximum minus segment EVI2 amplitude minimum	NBAR-EVI2
EVI_Area	The value of sum of daily interpolation of EVI2 from Greenup to Dormancy	NBAR-EVI2

2.3. The conventional environmental variables

In this study, we selected 22 environment variables to represent climate, terrain, parent materials and vegetation index, as shown in Table 1. The annual NDVI data from 2007 to 2016 were downloaded from RESDC (http://www.resdc.cn), which were produced using a maximum value composite (MVC) method based on the ten-day SPOT/VEGTATION NDVI data. The annual NDVI data is able to represent the vegetation growth for each year. To be consistent with the resolution of DEM, the original parent material data was rasterized into a raster layer with a 90 m cell size, and the climate and vegetation data were resampled to 90 m using a nearest neighbor assignment algorithm. All

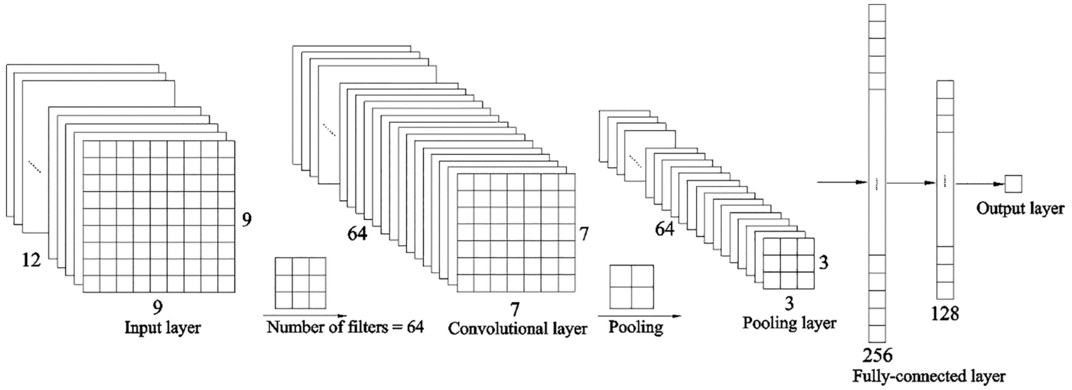


Fig. 2. The structure of CNN for SOC prediction.

variables were re-projected to the WGS1984_UTM_50N projection system.

2.4. The MODIS phenology product

Satellite remote sensing has proven a powerful tool to monitor large-scale land surface phenology (Moon et al., 2019; Hmimina et al., 2013). The MODIS Land Cover Dynamics (LCD) product (MCD12Q2) serves as annual measurements of global land surface phenology at a 500 m spatial resolution since 2001 to 2018. It has been well assessed by several scholars (Ganguly et al., 2010; Moon et al., 2019). It provides 11 phenology metrics for each vegetation cycle detected per year, and at most two vegetation cycles can be detected. Due to its long time series record, we used the MODIS LCD Product in this study.

We selected the 11 phenology metrics of the first vegetation cycle (Table 2) from the Collection 6 MODIS LCD Product since not all pixels have the second vegetation cycle. To cover all the sampling time and take consideration of the impact of historical cropping, we selected the annual MCD12Q2 phenologcy metrics from 2007 to 2016, resulting a total of 110 phenology variables. To be consistent with the above environment variables, we resampled the phenology variables to a resolution of 90 m.

3. Methodology

We developed a CNN model for predicting the spatial distribution of SOC content, and used a widely-used and accurate machine learning method, random forest (RF) compared with CNN.

3.1. Development of different environmental variable sets

The environmental variables were grouped into three pools to evaluate the performance of phenological parameters on soil prediction with CNN and RF. The first group was climatic, topographic variables and parent materials listed in Table 1, which we called the natural environmental variables. The second group was the first group (the natural variables) and the vegetation indices (annual NDVI from 2007 to 2016). The third group was the first group (the natural variables) and the phenology variables listed in Table 2 from 2007 to 2016.

3.2. Convolutional neural network

CNN is one mainstream deep learning model and includes one or more convolutional layers. It has been proven to be successful in image and video processing (Krizhevsky et al., 2012; Azizi et al., 2020), soil prediction (Wadoux, 2019; Ng et al., 2019) and other geographical element prediction (Tien Bui et al., 2020). A CNN model is typically established with an input layer, several hidden layers (including convolution layers, pooling layers and fully-connected layers), and an output layer. The CNN architecture of our study is shown in Fig. 2.

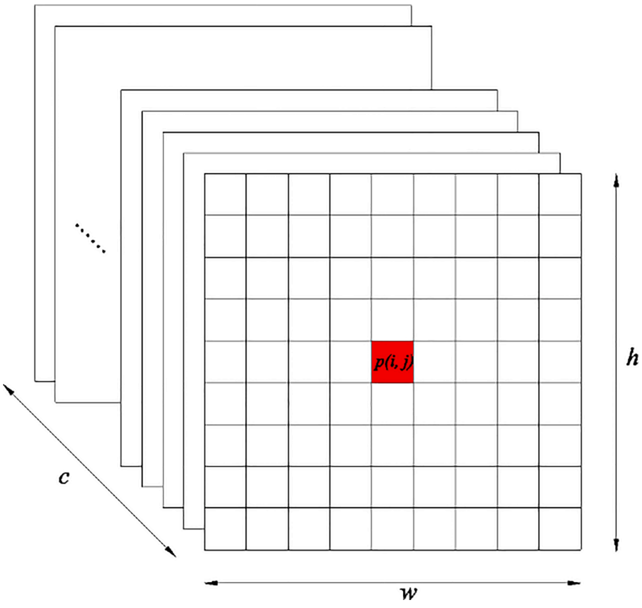


Fig. 3. The input data structure, $p(i, j)$ indicates the training point, w is width, h is height, c is the number of channels (hereafter the number of environment variables).

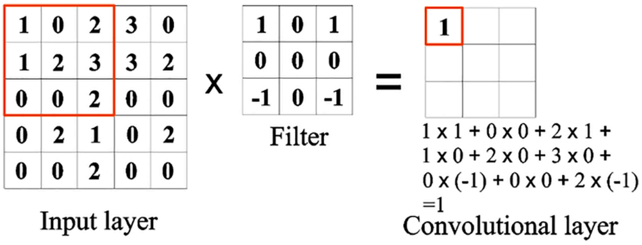


Fig. 4. Illustration of the convolution operation with a 3*3 filter over a 5*5 array.

The input layer is training points with environmental data. It can be specified with width (w), height (h) and several channels (c). Due to that SOC of a pixel is impacted by its neighborhood pixels, the n environmental variables of every training point (pixel p(i, j)) and its neighborhood pixels was organized as the input data (Fig. 3.). In Fig. 3., w*h is the neighborhood size of pixel p(i, j), for example $w = h = 9$ in Fig. 2. And the number of environmental variables was the number of channels.

The convolution layer is used to extract features from the input data

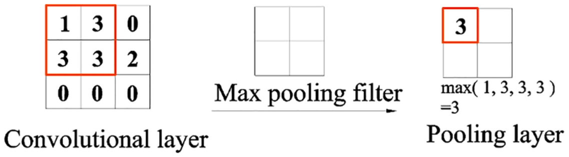


Fig. 5. Illustration of the first step of the max pooling operation with a 2*2 filter.

Table 3
Specifications of layers used in CNN for SOC prediction.

Layers	Filter size	Number of filters/neurons	Activation
Convolutional	3x3	32	ReLU
Max-Pooling	2x2	–	–
Convolutional	3x3	64	ReLU
Max-Pooling	2x2	–	–
Fully-connected	–	512	ReLU
Dropout	–	–	–
Fully-connected	–	1	Linear

through the convolution operation, which uses a filter sliding over the input data (Fig. 4). A single or multiple convolutional layers can be used for identifying features. The output of this layer is used as an input in the next layer after passing through an activation function. The activation function is used to realize nonlinear transformation for achieving a rapid convergence of the network. There are some activation functions, including sigmoid, tanh, and Rectified Linear Units (ReLU) (Ng et al., 2019). ReLU is the most common activation function, and is capable to converge faster than using sigmoid or tanh (Krizhevsky et al., 2012).

The pooling layer is employed to reduce the dimensions of the output data from the last step by providing a more abstract representation, which reduces computational costs and prevents over-fitting of the network. The common pooling operation is max and average pooling (Zuo et al., 2019). In this study, the max-pooling with a filter of 2 × 2 was used (Fig. 5).

The last output data from pooling is flattened into a one-dimension vector, and connected to fully-connected layers. The fully connected layer contains numerous neurons connecting to the output, i.e. SOC content in this study. A dropout layer is used to reduce the over-fitting risk in fully-connected layers (Krizhevsky et al., 2012).

Due to that we have 733 training points, constructing a too complex network could increase the risk of over-fitting. We constructed a network including two convolutional layers, two max-pool layers, three full-connected layers, and one dropout layer (Table 3). The layers in CNN were connected by weights and trained to reduce the difference between the predicted and the observed output with an initial learning rate of 0.01.

Different neighborhood size for the input layers would contain different ranges of spatial information. In this paper, we selected multiple neighborhood sizes, including 3*3, 5*5, 7*7, 9*9, 13*13, 17*17, 21*21, 25*25, and 29*29 to examine the impact of spatial neighborhood size on soil prediction. And an optimal size with the highest prediction accuracy was chosen for CNN with each environmental variable set. The evaluation of the predictions were demonstrated in Section 3.4.

The CNN was implemented in Python (v3.6.1, Anaconda 4.4.0) using Tensorflow backend.

3.3. Random forest as a comparison model

Random forest has been proven as an accurate machine learning method in digital soil mapping (Grinand et al., 2017; Wadoux, 2019). RF integrates multiple decision trees based on the idea of ensemble learning (Breiman, 2001; Cutler et al., 2012). It utilizes the boostrap strategy to randomly select two thirds of the training data to construct each decision tree, and leave the rest of samples as validation set. Every decision tree randomly selects some predictors to find the best node splits. The final prediction is determined based on all decision trees.

Two important parameters, the number of randomly selected predictors for each tree building (*mtry*) and the number of trees to be learned in forest (*ntree*), need to be set. As for *mtry*, the rounded down square root of the total number of environmental variables was taken as its value by default (Breiman, 2001). We set *ntree* as 15,000 because it meets the requirement to acquire stable results in our experiments.

Random forest was conducted using the 'randomForest' package (Breiman and Cutler, 2012) in the R language.

3.4. Evaluation of the prediction accuracy

A ten-fold cross validation was conducted to validate the performance of CNN and RF for SOC prediction. All the samples were partitioned equally in ten sub-sets of sample points that were stratified on parent materials. Each sub set was taken as to validate the prediction results based on the calibration set compiled from the remaining sample points. Two indices were adopted to evaluate the generated SOC maps based on the validation sample set, including the root-mean-square error (RMSE) and coefficient of determination (R^2). The two indices were established as follows:

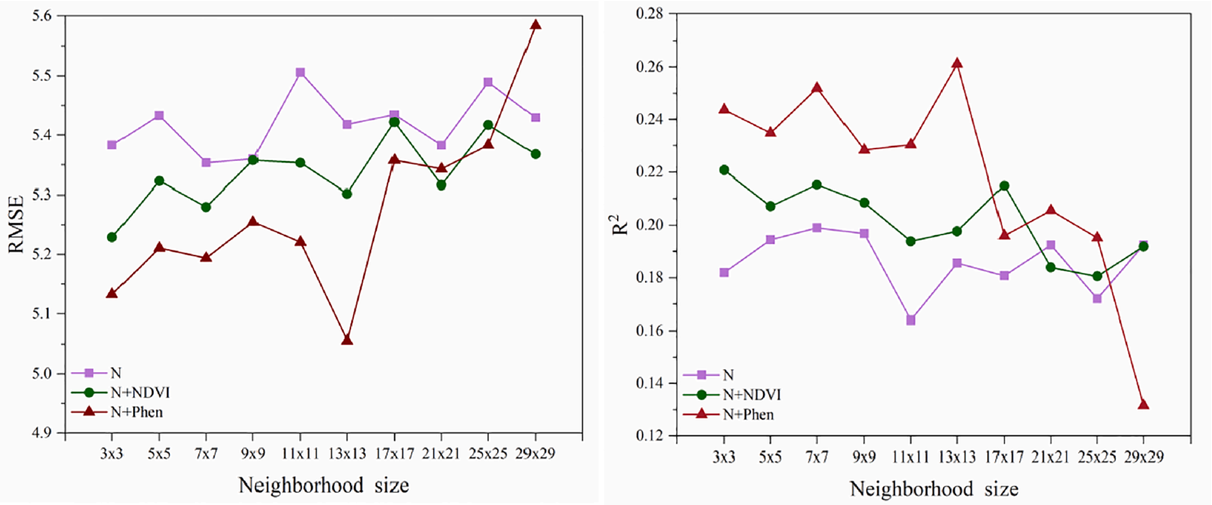
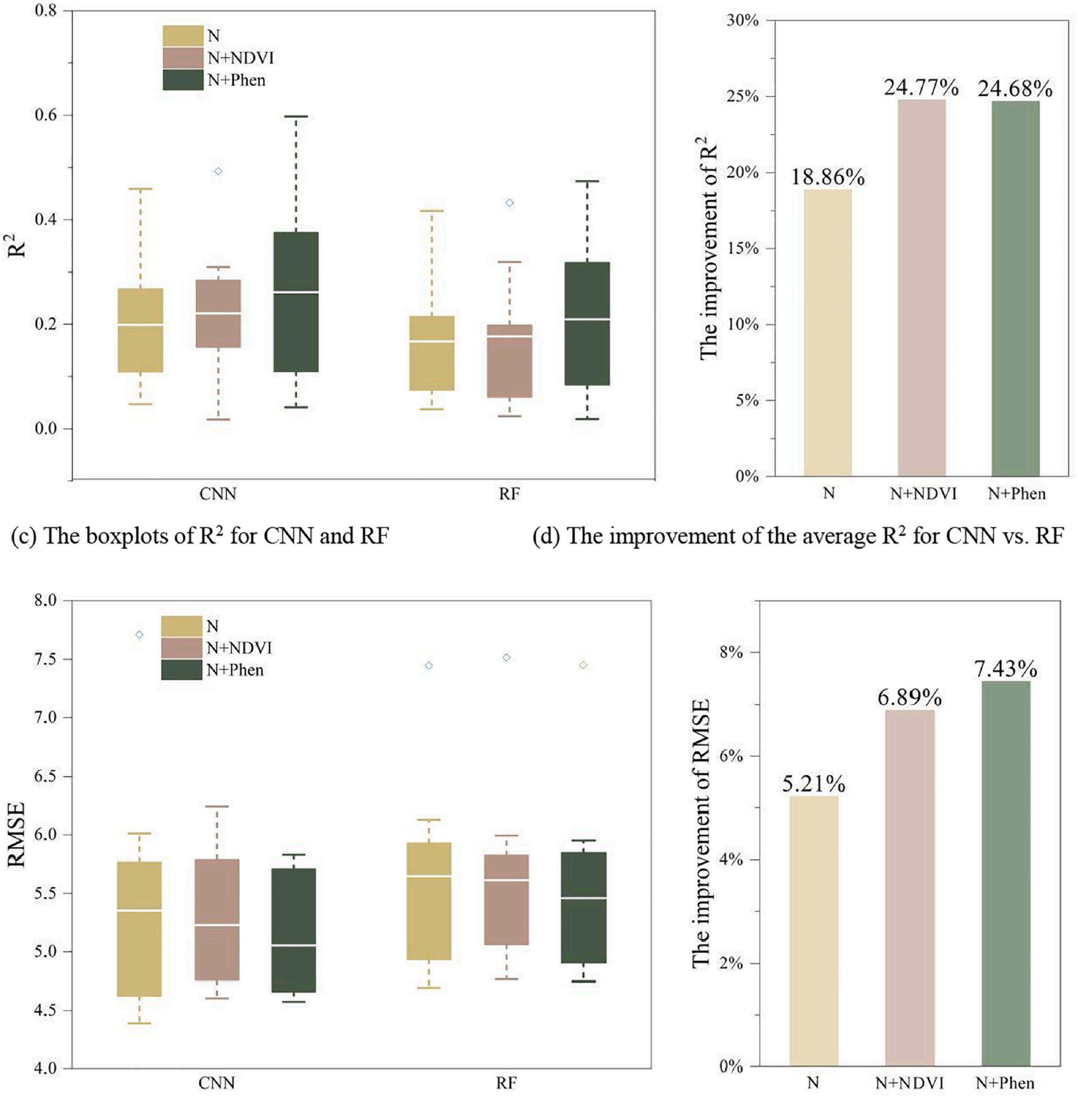


Fig. 6. The validation results (RMSE and R^2) of the SOC prediction for Anhui province with different neighborhood sizes for CNN, N: the natural environment variables, N + NDVI: the natural variables and NDVI, N + Phen: the natural variables and phenology variables.



(c) The boxplots of R^2 for CNN and RF

(d) The improvement of the average R^2 for CNN vs. RF

(a) The boxplots of RMSE for CNN and RF

(b) The improvement of the average RMSE for CNN vs. RF

Fig. 7. The boxplots of RMSE (a) and R^2 (c) for CNN and RF, and the improvement of the average RMSE (b) and R^2 (d) for CNN compared with RF. N: the natural environment variables, N + NDVI: the natural variables and NDVI, N + Phen: the natural variables and phenology variables.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - o_i)^2}{n}} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(o_i - \overline{o})^2} \tag{2}$$

where p_i and o_i are the predicted and observed SOC content of the i_{th} validation sample point, respectively; n is the sample size of validation points; $\bar{o}$ is the average observed SOC content of validation samples.

The above indices were calculated for 10 times based on each validation sample set, and the average of the 10 results were taken as the final evaluation results.

4. Results

4.1. The descriptive characteristics of SOC

Based on the 733 samples in Anhui province (with a sample density of 5.24 10^{-3} per/km^2), Anhui province has a large range of SOC content (0.91 ~ 89.12 g/kg) with a mean of 13.11 g/kg and a middle coefficient of variation (46.76%), indicating a heterogeneity of SOC in Anhui. Generally, the SOC in northern Anhui and its variation is smaller than those is middle and southern Anhui (as seen in Fig. 1). And southern Anhui has the largest variance of SOC mainly due to its varying topography.

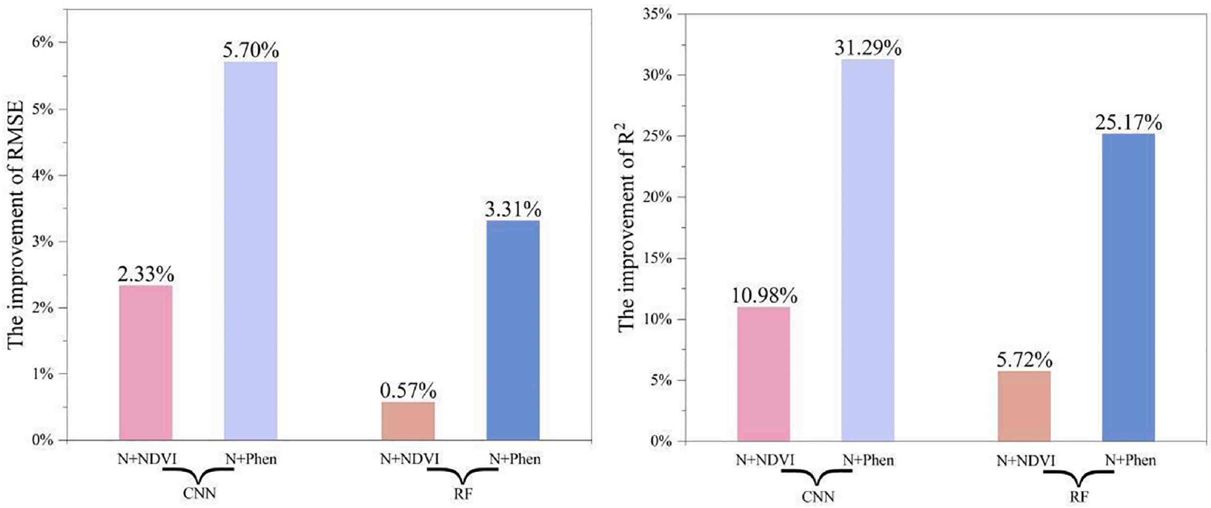


Fig. 8. The improvement of the average prediction accuracy (%) for CNN and RF using the two groups of environmental variables compared with using only the natural environmental variables. N + NDVI: the natural variables and NDVI, N + Phen: the natural and phenology variables.

4.2. The optimal neighborhood sizes for CNN

The evaluation results of the SOC prediction for Anhui province with different neighborhood sizes are shown in Fig. 6. It shows that the changing of RMSE and R^2 with different environmental variable sets have different trends. The RMSE and R^2 when using the natural environmental variables show a fluctuation with the increasing of neighborhood size, and the smallest RMSE with the highest R^2 occurs in neighborhood size of 7 and 9. There is a large decrease in the RMSE or a large increase in R^2 when the neighborhood size equals to 13 with the natural environmental and phenological variables. For the natural environmental and NDVIs, the RMSE increases slowly with the increasing of neighborhood size, the neighborhood size with the smallest RMSE or largest R^2 is 3*3. Thus, we chose the neighborhood size of 7*7, 3*3, 13*13 for the natural environmental variables, the natural environmental variables and NDVI, the natural environmental and phenological variables, respectively.

4.3. Comparisons of the evaluation results for CNN and RF

The boxplots of RMSE and R^2 of SOC prediction using CNN and RF, and the improvement of the average RMSE and R^2 for CNN compared with RF using each group of environmental variables are displayed in Fig. 7. It suggests that the increase of the average prediction accuracy with CNN vs. RF is larger with more predictors. The decrease of the average RMSE for CNN when using both the natural environmental and phenology variables is 7.43% compared with RF, and the increase of the average R^2 is 24.68%.

The improvement of the average prediction accuracy (%) for CNN and RF using the other two groups of environmental variables compared with using only the natural environmental variables is shown in Fig. 8. It shows that adding either the phenology variables or NDVI into the natural environmental variables improves the prediction performance no matter using CNN or RF. The improvement of the average accuracy using the natural environmental and phenology variables is larger than that using the natural environmental variables and NDVIs. Furthermore, the improvement with CNN is larger than that with RF.

5. Discussions

5.1. Applicability and limitations of CNN

CNN has a good learning ability when processing complex and large-volume input data due to that it is capable of modeling nonlinear relations (Tien Bui et al., 2020). Another advantage is that it leverages the spatial contextual information of predictors surrounding the sampling points (LeCun et al., 2015). This makes it suitable for modeling the soil-environment relationships because soil at a location is influenced by both the environmental conditions of this location and the surrounding locations. With more and more data available in a big data era, such as the land surface phenology data in our study, deep learning has a potential to be an effective soil prediction method.

Our results suggested that the neighborhood size of input data exerted a strong influence on prediction accuracy, which is consistent with previous studies (Padarian et al., 2019; Wadoux, 2019). The neighborhood size was closely related to the amount of input contextual information. A suitable neighborhood size may relate to the scale that the environmental covariates impact the soil development. The scale effect of environmental covariates on soil prediction have been examined (Smith et al., 2006; Shi et al., 2018). Their results indicated that appropriate scales matching the scales of soil property–landscape process would improve the soil mapping accuracy, and the appropriate scale for soil properties is different and very likely case-dependent.

CNN also has some limitations. First, a reliable CNN model usually require a large amount of data for training. Correspondingly, high demand on the computing power is needed. Yet in case of insufficient input data, data augmentation is an alternative way to improve the model accuracy (Saleh and Hamoud, 2021). Second, several parameters are needed to be calibrated, and preventing overfitting should be paid attentions (Darwish et al., 2020). Also, it is not easy to interpret the neural network results (Lee et al., 2018).

5.2. Model performance

Although the phenological metrics improved the prediction accuracy, the accuracy of our study is not high. A possible reason is the driving mechanism of environmental variables on SOC in this large area is complicated and varying over space. The sampling density of our study area is relatively small ($5.24*10^{-3}$ per point/km^2). Furthermore, most of the Anhui province is generally flat with croplands, where the variables such as terrain variables are probably not effective in DSM (Zhu et al., 2010).

It is often that R^2 of soil organic carbon/matter prediction in many previous studies was not larger than 0.5 because of the complicated interactions between soil and environmental variables (Wiesmeier et al., 2013; Liang et al., 2019; Funes et al., 2019). Our prediction accuracy is consistent with the relevant studies, such as Wiesmeier et al. (2013) with a R^2 of 0.21, and Funes et al. (2019) with R^2 of 0.20–0.35. In the study of

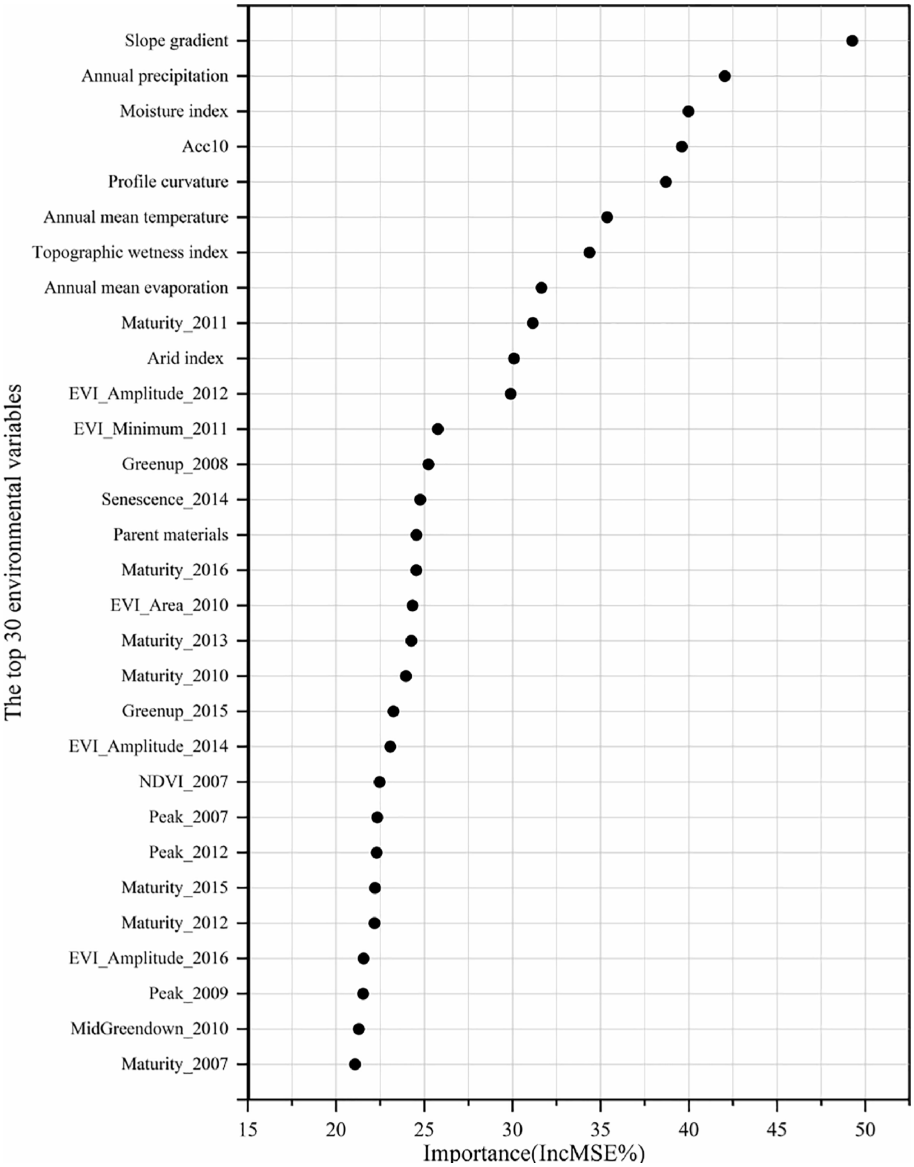


Fig. 9. The variable importance of the environmental variables for SOC variation based on random forest, only the top 30 variables are shown. Acc10: Annual accumulated temperature above 10 °C, The number in the phenology or NDVI indicates its observation year, for example, Maturity_2007 means the maturity in year 2007.

Wadoux (2019), the R^2 for SOC prediction with CNN was only 0.15. Their sample density was 0.36 per point/km^2, much larger than ours.

5.3. The phenology variables used for SOC prediction

Our results showed that using land surface phenology parameters from the MODIS MCD12Q2 product improved the SOC prediction ability with CNN or RF. Long term phenological variables reflect a long time vegetation growth characteristics, resulting from interactions of soil impacted by land management over time. In turn, vegetation is a main source of carbon input for soil. This may explain the effectiveness of the phenology parameters in SOC prediction.

The easy availability and long time series record of MODIS phenology metric product makes it a valuable environmental covariate for soil mapping at a regional or global scale. A similar product is the Suomi National Polar-Orbiting Partnership NASA Visible Infrared Imaging Radiometer Suite (VIIRS) Land Cover Dynamics data product which provides global yearly phenological metrics from 2013. The VIIRS sensor is a long-term continuity of the MODIS land surface phenology data (Román et al., 2011; Zhang et al. 2018). The comparison of the MODIS and VIIRS products (Moon et al., 2019) suggests that phenometrics from the two products show only minor differences, but merging the two products should exploit their overlap period. Therefore, the future work using the two products together for SOC prediction should consider this.

The 500 m resolution of this product may limit its applicability in soil mapping at a detailed scale. A possible way is to downscale the phenology metric product with coarse spatial resolution to detailed resolution data with data fusion methods (Zeng et al., 2020).

5.4. The influential environmental variables for SOC variation

In order to examine the influential environmental variables for SOC prediction, we generated the variable importance of the environmental variables for SOC variation based on random forest in Fig. 9, only the top 30 variables are listed. It shows that the topographic and climatic variables are the most important variables impacting the spatial distribution of SOC in Anhui, followed by the phenological variables. Only one annual mean NDVI occurs in the top 30 important variables. Slope ranks the first in Fig. 9, and annual precipitation, moisture index, and annual accumulated temperature above 10 °C are the three most important variables among the influential climatic variables. Maturity_2011, EVI-Amplitude_2012, EVI-Minimum_2011, greenup_2008, senescene_2014, maturity_2016, and EVI_area_2010, indicating different characteristics of vegetation growth, are the seven most important phenological variables. It seems that there is no obvious rule that which variable or year plays a more important role among the phenological variables. This may indicate that variables representing different characteristics of the vegetation growth together are related to the SOC variation in Anhui.

6. Conclusions

A CNN model was developed for SOC prediction in Anhui province with three groups of environmental variables. Adding the long term MODIS MCD12Q2 land surface phenology parameters or annual NDVIs to the natural environmental variables improved the prediction accuracy. The phenology variables obtained a lager accuracy improvement compared with NDVIs. In addition, the CNN obtained a higher prediction accuracy than RF regardless of using either group of variables. It suggests that the land surface phenology metrics indicating the long term vegetation growth characteristics could be effective or even better predictors for SOC prediction at a regional scale. Meanwhile, CNN could be a promising method for soil mapping in case studies with large data.

CRediT authorship contribution statement

Lin Yang: Conceptualization, Investigation, Writing - original draft. **Yanyan Cai:** Methodology, Investigation, Visualization, Writing - review & editing. **Lei Zhang:** Software, Investigation. **Mao Guo:** Validation. **Anqi Li:** Data curation, Validation. **Chenghu Zhou:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (Project No. 41971054), and the Leading Funds for the First Class Universities (020914912203020 9-14380079).

References

An, Y., Yang, L., Zhu, A.X., Qin, C., Shi, J.J., 2018. Identification of representative samples from existing samples for digital soil mapping. Geoderma 311, 109–119. https://doi.org/10.1016/j.geoderma.2017.03.014.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: A review. Remote Sens. 11 (6), 676. https://doi.org/10.3390/rs11060676.

Araya, S., Lyle, G., Lewis, M., Ostendorf, B., 2016. Phenologic metrics derived from MODIS NDVI as indicators for Plant Available Water-holding Capacity. Ecol. Indic. 60 (60), 1263–1272. https://doi.org/10.1016/j.ecolind.2015.09.012.

Azizi, A., Gilandeh, Y.A., Mesri-Gundoshmian, T., Saleh-Bigdeli, A.A., Moghaddam, H.A., 2020. Classification of soil aggregates: A novel approach based on deep learning. Soil Tillage Res. 199, 104586 https://doi.org/10.1016/j.still.2020.104586.

Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018. Multi-scale digital soil mapping with deep learning. Sci. Rep. 8, 15244. https://doi.org/10.1038/s41598-018-33516-6.

Bradford, M.A., Wieder, W.R., Bonan, G.B., Fierer, N., Raymond, P.A., Crowther, T.W., 2016. Managing uncertainty in soil carbon feedbacks to climate change. Nat. Clim. Chang. 6, 751–758. https://doi.org/10.1038/nclimate3071.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Cutler, A., 2012. Breiman and Cutler's random forests for classification and regression. Packag. "randomForest." 29.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random forests. Ensemble Mach. Learn. Methods Appl. 157–175 https://doi.org/10.1007/9781441993267_5.

Darwish, A., Ezzat, D., Hassanien, A.E., 2020. An optimized model based on convolutional neural networks and orthogonal learning particle swarm optimization algorithm for plant diseases diagnosis. Swarm Evol. Comput. 52, 100616 https://doi.org/10.1016/j.swevo.2019.100616.

Dou, F.G., Yu, X., Ping, C.L., Michaelson, G., Guo, L.D., Jorgenson, T., 2010. Spatial variation of tundra soil organic carbon along the coastline of northern Alaska. Geoderma 154 (3–4), 328–335.

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

Funes, I., Savé, R., Rovira, P., Molowny-Horas, R., Alcañiz, J.M., Ascaso, E., Herms, I., Herrero, C., Boixadera, J., Vayreda, J., 2019. Agricultural soil organic carbon stocks in the north-eastern Iberian Peninsula: Drivers and spatial variability. Sci. Total Environ. 668, 283–294. https://doi.org/10.1016/j.scitotenv.2019.02.317.

Ganguly, S., Friedl, M.A., Tan, B., Zhang, X., Verma, M., 2010. Land surface phenology from MODIS: Characterization of the Collection 5 global land cover dynamics product. Remote Sens. Environ. 114, 1805–1816. https://doi.org/10.1016/j.rse.2010.04.005.

Grinand, C., Maire, G. Le, Vieilledent, G., Razakamanarivo, H., Razafimbelo, T., Bernoux, M., 2017. Estimating temporal changes in soil carbon stocks at ecoregional scale in Madagascar using remote-sensing. Int. J. Appl. Earth Obs. Geoinf. 54, 1–14. https://doi.org/10.1016/j.jag.2016.09.002.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207. https://doi.org/10.1016/j.geoderma.2009.06.003.

Hamzehpour, N., Shafizadeh-Moghadam, H., Valavi, R., 2019. Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. CATENA 182, 104141. https://doi.org/10.1016/j.catena.2019.104141.

Hmimina, G., Dufrêne, E., Pontailler, J.Y., Delpierre, N., Aubinet, M., Caquet, B., de Grandcourt, A., Burban, B., Flechard, C., Granier, A., Gross, P., Heinesch, B., Longdoz, B., Moureaux, C., Ourcival, J.M., Rambal, S., Saint André, L., Soudani, K., 2013. Evaluation of the potential of MODIS satellite data to predict vegetation phenology in different biomes: An investigation using ground-based NDVI measurements. Remote Sens. Environ. 132, 145–158. https://doi.org/10.1016/j.rse.2013.01.010.

Hoffmann, M., Pohl, M., Jurisch, N., Prescher, A.K., Mendez Campa, E., Hagemann, U., Remus, R., Verch, G., Sommer, M., Augustin, J., 2018. Maize carbon dynamics are driven by soil erosion state and plant phenology rather than nitrogen fertilization form. Soil Tillage Res. 175, 255–266. https://doi.org/10.1016/j.still.2017.09.004.

Krizhevsky, A., Sutskever, L., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304, 1623–1627. https://doi.org/10.1126/science.1097396.

Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352, 395–413. https://doi.org/10.1016/j.geoderma.2019.05.031.

LeCun, Y., Bengio, Y., 1995. Convolutional Networks for Images, Speech, and Time Series, in: Arbib, M.A. (Ed.), Handbook of Brain Theory and Neural Networks. MIT Press, p. 3361.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Lee, W.Y., Park, S.M., Sim, K.B., 2018. Optimal hyperparameter tuning of convolutional neural networks based on the parameter-setting-free harmony search algorithm. Optik (Stuttg). 172, 359–367. https://doi.org/10.1016/j.ijleo.2018.07.044.

Liang, Z., Chen, S., Yang, Y., Zhou, Y., Shi, Z., 2019. High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling. Sci. Total Environ. 685, 480–489. https://doi.org/10.1016/j.scitotenv.2019.05.332.

Ma, T., Wei, T., Qin, C.Z., Zhu, A.X., Qi, F., Liu, J., Zhao, F., Pan, H., 2020. In-situ recommendation of alternative soil samples during field sampling based on environmental similarity. Earth Sci. Informatics. 13, 39–53. https://doi.org/10.1007/s12145-019-00407-x.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Moon, M., Zhang, X., Henebry, G.M., Liu, L., Gray, J.M., Melaas, E.K., Friedl, M.A., 2019. Long-term continuity in land surface phenology measurements: A comparative assessment of the MODIS land cover dynamics and VIIRS land surface phenology products. Remote Sens. Environ. 226, 74–92. https://doi.org/10.1016/j.rse.2019.03.034.

Mora-Vallejo, A., Claessens, L., Stoorvogel, J., Heuvelink, G.B.M., 2008. Small scale digital soil mapping in Southeastern Kenya. CATENA 76 (1), 44–53. https://doi.org/10.1016/j.catena.2008.09.008.

Nelson, D.W., Sommers, L.E., Sparks, D.L., Page, A.L., Helmke, P.A., Loeppert, R.H., Sumner, M.E., 1996. Total carbon, organic carbon, and organic matter. Methods Soil Anal. 9, 961–1010.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251–267. https://doi.org/10.1016/j.geoderma.2019.06.016.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. Geoderma Reg. 16, e00198 https://doi.org/10.1016/j.geodrs.2018.e00198.

Peng, W., Wheeler, D.B., Bell, J.C., Krusemark, M.G., 2003. Delineating patterns of soil drainage class on bare soils using remote sensing analyses. Geoderma 115, 261–279. https://doi.org/10.1016/S0016-7061(03)00066-1.

Qin, C., Zhu, A.X., Pei, T., Li, B., Zhou, C., Yang, L., 2007. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. Int. J. Geogr. Inf. Sci. 21, 443–458. https://doi.org/10.1080/13658810601073240.

Román, M.O., Gatebe, C.K., Schaaf, C.B., Poudyal, R., Wang, Z., King, M.D., 2011. Variability in surface BRDF at different spatial scales (30m–500m) over a mixed agricultural landscape as retrieved from airborne and satellite spectral measurements. Remote Sens. Environ. 115, 2184–2203. https://doi.org/10.1016/j.rse.2011.04.012.

Saleh, A.M., Hamoud, T., 2021. Analysis and best parameters selection for person recognition based on gait model using CNN algorithm and image augmentation. J. Big Data. 8, 1. https://doi.org/10.1186/s40537-020-00387-6.

Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., Acutis, M., 2017. Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. Geoderma 286, 35–45. https://doi.org/10.1016/j.geoderma.2016.10.019.

Shi, J.J., Yang, L., Zhu, A.-X., Qin, C.Z., Liang, P., Zeng, C.Y., Pei, T., 2018. Machine-Learning Variables at Different Scales vs. Knowledge-based Variables for Mapping Multiple Soil Properties. Soil Sci. Soc. Am. J. 82, 645–656. https://doi.org/10.2136/sssaj2017.11.0392.

Smith, M.P., Zhu, A.X., Burt, J.E., Stiles, C., 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma 137, 58–69. https://doi.org/10.1016/j.geoderma.2006.07.002.

Tien Bui, D., Hoang, N.D., Martínez-Álvarez, F., Ngo, P.T.T., Hoa, P.V., Pham, T.D., Samui, P., Costache, R., 2020. A novel deep learning neural network approach for predicting flash flood susceptibility: A case study at a high frequency tropical storm area. Sci. Total Environ. 701, 134413 https://doi.org/10.1016/j.scitotenv.2019.134413.

Vaudour, E., Gilliot, J.M., Bel, L., Lefevre, J., Chehdi, K., 2016. Regional prediction of soil organic carbon content over temperate croplands using visible near-infrared airborne hyperspectral imagery and synchronous field spectra. Int. J. Appl. Earth Obs. Geoinf. 49, 24–38. https://doi.org/10.1016/j.jag.2016.01.005.

Wadoux, A.M.J.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70. https://doi.org/10.1016/j.geoderma.2019.05.012.

Wang, S., Adhikari, K., Wang, Q., Jin, X., Li, H., 2018. Role of environmental variables in the spatial distribution of soil carbon (C), nitrogen (N), and C: N ratio from the northeastern coastal agroecosystems in China. Ecol. Indic. 84, 263–272. https://doi.org/10.1016/j.ecolind.2017.08.046.

Wiesmeier, M., Hübner, R., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling, B., von Lützow, M., Kögel-Knabner, I., 2013. Amount, distribution and driving factors of soil organic carbon and nitrogen in cropland and grassland soils of southeast Germany (Bavaria). Agric. Ecosyst. Environ. 176, 39–52. https://doi.org/10.1016/j.agee.2013.05.012.

Yang, L., He, X., Shen, F., Zhou, C., Zhu, A.X., Gao, B., Chen, Z., Li, M., 2020. Improving prediction of soil organic carbon content in croplands using phenological parameters extracted from NDVI time series data. Soil Tillage Res. 196, 104465 https://doi.org/10.1016/j.still.2019.104465.

Yang, L., Zhu, A.X., Qi, F., Qin, C.Z., Li, B., Pei, T., 2013. An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping. Int. J. Geogr. Inf. Sci. 27, 1–23. https://doi.org/10.1080/13658816.2012.658053.

Zeng, C., Yang, L., Zhu, A.X., 2020. The generation of soil spectral dynamic feedback using landsat 8 data for digital soil mapping. Remote Sens. 12 (10) https://doi.org/10.3390/rs12101691.

Zhang, L., Yang, L., Ma, T., Shen, F., Cai, Y., Zhou, C., 2021. A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data. Geoderma 384. https://doi.org/10.1016/j.geoderma.2020.114809.

Zhang, S.J., Zhu, A.X., Liu, J., Yang, L., Qin, C.Z., An, Y.M., 2016. An heuristic uncertainty directed field sampling design for digital soil mapping. Geoderma 267, 123–136. https://doi.org/10.1016/j.geoderma.2015.12.009.

Zhang, X., Jayavelu, S., Liu, L., Friedl, M.A., Henebry, G.M., Liu, Y., Schaaf, C.B., Richardson, A.D., Gray, J., 2018. Evaluation of land surface phenology from VIIRS data using time series of PhenoCam imagery. 256,137–149. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2018.03.003.

Zhou, Y., Chen, S., Zhu, A.-X., Hu, B., Shi, Z., Li, Y., 2021. Revealing the scale- and location-specific controlling factors of soil organic carbon in Tibet. Geoderma 382, 114713.

Zhu, A.-X., Liu, F., Li, B., Pei, T., Qin, C., Liu, G., Wang, Y., Chen, Y., Ma, X., Qi, F., Zhou, C., 2010. Differentiation of Soil Conditions over Low Relief Areas Using Feedback Dynamic Patterns. Soil Sci. Soc. Am. J. 74, 861–869. https://doi.org/10.2136/sssaj2008.0411.

Zuo, R., Xiong, Y., Wang, J., Carranza, E.J.M., 2019. Deep learning and its application in geochemical mapping. Earth-Science Rev. 192, 1–14. https://doi.org/10.1016/j.earscirev.2019.02.023.