Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping.

Xianglin Zhang, Songchao Chen, Jie Xue, Nan Wang, Yi Xiao, Qianqian Chen, Yongsheng Hong, Yin Zhou, Hongfen Teng, Bifeng Hu, et al.

Geoderma 432 (2023) 116383

Contents lists available at ScienceDirect

Geoderma

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

Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping

Xianglin Zhang [a,b], Songchao Chen [a,b,*], Jie Xue [c], Nan Wang [b], Yi Xiao [b], Qianqian Chen [b], Yongsheng Hong [b], Yin Zhou [d], Hongfen Teng [e], Bifeng Hu [f], Zhiqing Zhuo [g], Wenjun Ji [h], Yuanfang Huang [h], Yuxuan Gou [h], Anne C. Richer-de-Forges [i], Dominique Arrouays [i], Zhou Shi [b]

[a] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou 311200, China
[b] Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[c] Department of Land Management, Zhejiang University, Hangzhou 310058, China
[d] Institute of Land and Urban-Rural Development, Zhejiang University of Finance and Economics, Hangzhou 310018, China
[e] School of Environmental Ecology and Biological Engineering, Wuhan Institute of Technology, Wuhan 430205, China
[f] Department of Land Resource Management, School of Tourism and Urban Management, Jiangxi University of Finance and Economics, Nanchang 330013, China
[g] Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China
[h] College of Land Science and Technology, China Agricultural University, Beijing, 100193, China
[i] INRAE, Unité InfoSol, Orléans 45075, France

ARTICLE INFO

Handling Editor: Morgan Cristine L.S.

Keywords:
Digital soil mapping
Variable selection
Quantile regression forests
Computation efficiency
Northeast and North China

ABSTRACT

In the context of increasing soil degradation worldwide, spatially explicit soil information is urgently needed to support decision-making for sustaining limited soil resources. Digital soil mapping (DSM) has been proven as an efficient way to deliver soil information from local to global scales. The number of environmental covariates used for DSM has rapidly increased due to the growing volume of remote sensing data, therefore variable selection is necessary to deal with multicollinearity and improve model parsimony. Compared with Boruta, recursive feature elimination (RFE), and variance inflation factor (VIF) analysis, we proposed the use of modified greedy feature selection (MGFS), for DSM regression. For this purpose, using quantile regression forest, 402 soil samples and 392 environmental covariates were used to map the spatial distribution of soil organic carbon density (SOCD) in Northeast and North China. The result showed that MGFS selected the most parsimonious model with only 9 covariates (e.g., brightness index, mean annual temperature), much lower than RFE (22 covariates), VIF (30 covariates), and Boruta (76 covariates). The repeated validation (50 times) showed that the MGFS derived model performed better (R^2 of 0.60, LCCC of 0.74, RMSE of 13.80 t ha^{-1}) than these using full covariates, Boruta, RFE and VIF (R^2 of 0.48–0.57, LCCC of 0.64–0.72, RMSE of 14.24–15.79 t ha^{-1}). Despite the similar performance of the uncertainty estimate (PICP), the model using MGFS and RFE had the lowest global uncertainty (0.86) as indicated by the uncertainty index. In addition, MGFS had the best computation efficiency when considering the steps of variable selection and map prediction. Given these advantages over Boruta, RFE and VIF, MGFS has a high potential in fine-resolution soil mapping practices, especially for these studies at a broad scale involving heavy computation on millions or billions of pixels.

1. Introduction

Soil is one of the Earth's most essential and finite resources. It enables life on Earth by delivering crucial ecosystem services, including the provision of food, fibre and fuel, water purification, contaminant reduction, nutrient cycling, carbon sequestration, climate and flood regulation, and biodiversity conservation (McBratney et al., 2014; Adhikari and Hartemink, 2016; Baveye et al., 2016; Pereira et al., 2018). Under the tremendous pressure of population growth, economic development and climate change, global soils are continuously degraded (e.g., erosion, salinization, fertility decline). FAO (2015) reported that 33 % of the Earth's soils are already degraded, and over 90 % could become degraded by 2050. For sustaining soil resources for the next generation, there is an urgent demand to improve management practices which

* Corresponding author at: ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou 311200, China.
E-mail address: chensongchao@zju.edu.cn (S. Chen).

https://doi.org/10.1016/j.geoderma.2023.116383
Received 27 October 2022; Received in revised form 9 January 2023; Accepted 12 February 2023
Available online 24 February 2023

require spatially explicit soil information to support decision-making. However, many conventional soil maps do not provide relevant information for the current global challenges, and most of the data for creating the soil maps is outdated (Sanchez et al., 2009; Arrouays et al., 2014, Arrouays et al., 2017). A digital revolution has taken place in soil mapping in the last 20 years, which resulted in the concept of digital soil mapping (DSM, McBratney et al., 2003).

DSM has emerged as a fast-growing sub-discipline of soil science by integrating soil surveys, geostatistics, geographic information system, remote sensing and machine learning (Minasny and McBratney, 2016). DSM has been widely used in delivering fine-resolution maps of soil information across scales with a focus switching from "primary" soil properties (e.g., soil organic carbon, particle size fractions, pH, soil depth) to "secondary" soil properties (e.g., available water capacity, carbon sequestration potential, carbon vulnerability) in the last decade (Padarian et al., 2014; Hengl et al., 2017; Chen et al., 2018; Román Dobarco et al., 2019; Viscarra Rossel et al., 2019; Liu et al., 2022).

The number of environmental covariates used for DSM has rapidly increased due to the growing volume of remote sensing data (Chen et al., 2022a). Therefore, variable selection is increasingly used before fitting the final predictive model. It has several advantages, such as (1) calibrating the predictive model faster; (2) reducing model complexity; (3) increasing model performance; (4) avoiding multicollinearity; (5) producing the map faster (Wadoux et al., 2020). Currently, there are mainly-two strategies for variable selection in DSM (Wadoux et al., 2020): (1) covariate reduction as a pre-processing step (before calibrating a machine learning model), such as selecting the most relevant covariates by Pearson's correlation coefficient between soil properties and covariates, discarding the covariates which are highly correlated with other covariates, or keeping the first several components using principle component analysis (Mosleh et al., 2016; Hamzehpour et al., 2019; Poggio et al., 2021; Taghizadeh-Mehrjardi et al., 2021); (2) wrapper methods which rely on inference made by calibrating a machine learning model to assess covariate importance (Xiong et al., 2014; Brungard et al., 2015; Nussbaum et al., 2018; Amiri et al., 2019; Keskin et al., 2019; Poggio et al., 2021). In addition to the limitation of not being applicable to categorical variables, most of the first strategy only accounts for the linear relationship between soil properties and covariates or even omits their correlations, so it can potentially neglect these covariates non-linearly correlated to the soil properties and thus decrease model performance (Camera et al., 2017; Zeraatpisheh et al., 2019). In addition, though machine learning models can be affected by multicollinearity (Strobl et al., 2008; Drobnič et al., 2020), they are still more robust than traditional multiple linear regression (Dormann et al., 2013); therefore, the second strategy is more appropriate for machine learning based DSM studies. Among these wrapper methods for variable selection, the most popular ones are Boruta (Xiong et al., 2014; Amiri et al., 2019; Keskin et al., 2019; Rasaei et al., 2020; Xu et al., 2022, Zeraatpisheh et al., 2022) and recursive feature elimination (RFE, Brungard et al., 2015; Nussbaum et al., 2018; Gomes et al., 2019; Chen et al., 2021; Poggio et al., 2021; Yang et al., 2022). Previous studies commonly chose either Boruta or RFE in variable selection, and only a few studies compared RFE and Boruta in DSM practices (Chen et al., 2022b; Luo et al., 2022). In addition, all these relevant studies only focused on the impact of variable selection on the model performance while ignoring its effect on the uncertainty estimate. Therefore, it remains unclear which variable selection method is superior in model performance and uncertainty estimates. From our previous experience with Boruta and RFE, we also found that they can potentially ignore some useful variables so as to decrease model performance (Xiao et al., 2022). Therefore, based on the recently proposed greedy feature selection (GFS, Drobnič et al., 2020), a modified variable selection algorithm is expected to solve this problem for mapping soil properties.

Addressing the current knowledge gap and limitation, the objectives of this study are twofold: (1) apply a modified GFS algorithm to DSM regression problem; (2) compare the modified GFS algorithm to commonly used variable selection methods on model performance, uncertainty estimate and computation efficiency in soil organic carbon density (SOCD) mapping.

2. Materials and methods

2.1. Soil data

The study area is located in Northeast and North China, covering a total area of 56.30×10^4 km^2 in which 37.67×10^4 km^2 are used as cropland (Fig. 1). It has a temperate continental monsoon climate with mean annual precipitation (MAP) between 400 and 1200 mm and mean annual temperature (MAT) between −1.1 and 16 °C. According to the Genetic Soil Classification of China, the main soil types are Black (Phaeozems), Chernozem and Brown Earth (Luvisols) in Northeast Plain, and Fluvo-aquic (Eutric Cambisols), Shajiang Black (Vertisols), and Cinnamon (Eutric Luvisols) in Huai-Hai Plain (Zhuo et al., 2022). The single cropping system (i.e., spring corn, spring wheat, soybean) dominates in the Northeast Plain, and the double cropping system (winter wheat and summer corn) is the main farming system in the Huai-Hai Plain. Accounting for around 30 % of the national grain production, the study area is a crucial agricultural region in China.

A total of 402 sampling sites were selected in the cropland by stratified random sampling representing the major soil types, clay content and agricultural system. For each sampling site, three soil cores were collected at four depth intervals (0–10, 10–20, 20–30, and 30–40 cm) using an undisturbed soil sampler (diameter of 5.1 cm and height of 10 cm) between April and May in 2017. All the composite soil samples were air-dried and sieved to < 2 mm, and SOC content (g kg^{-1}) was determined by the dichromate oxidation–external heating method (Bao, 2007). Soil bulk density (BD, g cm^{-3}) was determined by the mean of three replicates of undisturbed soil cores (100 cm^3 in volume). More details about sampling design and laboratory analysis can be found in Zhuo et al. (2022). In this study, we focused on the topsoil (0–30 cm) by integrating the soil information from the first three depth intervals (0–10, 10–20, 20–30 cm) using the weighted average method. SOCD (t ha^{-1}) was calculated by the equation below:

$$\mathrm{SOCD} = \mathrm{SOC} \times \mathrm{BD} \times \mathrm{Depth} \times 10 \quad (1)$$

where Depth is the soil depth (cm) which is 30 cm. In this study, coarse fragments did not present in the topsoil (0–30 cm) of cropland, so it was not considered in this equation.

2.2. Environmental covariates

Following the Scorpan framework (McBratney et al., 2003), 392 environmental covariates relevant to SOCD were investigated and listed in Table 1. The environmental covariates were derived from multiple sources: Landsat 5 Thematic Mapper (TM), Landsat 7 Enhanced Thematic Mapper (ETM +), Landsat 8 Operational Land Imager (OLI), Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM), products from Moderate Resolution Imaging Spectroradiometer (MODIS), and legacy digital soil maps.

All available Landsat 5, 7 and 8 surface reflectance images at 30 m resolution covering the entire study area in the Collection 2 Tier1 Level 2 database were collected from 2003 to 2017 (https://www.usgs.gov/). These images received radiometric, geometric and atmospheric corrections with Landsat Ecosystem Disturbance Adaptive Processing System (Landsat 5 TM and Landsat 7 ETM+) and Land Surface Reflectance Code (Landsat 8 OLI). To ensure the image quality, we filtered the image collection with the criteria of cloud cover < 20 %, root mean square error (RMSE) of geometric residuals measured on the ground control points < 10 m, and the best image quality level. Afterwards, the CFmask algorithm was used to identify and remove the dilated cloud, cirrus, cloud and cloud shadow (Zhu et al., 2015; Foga et al., 2017).

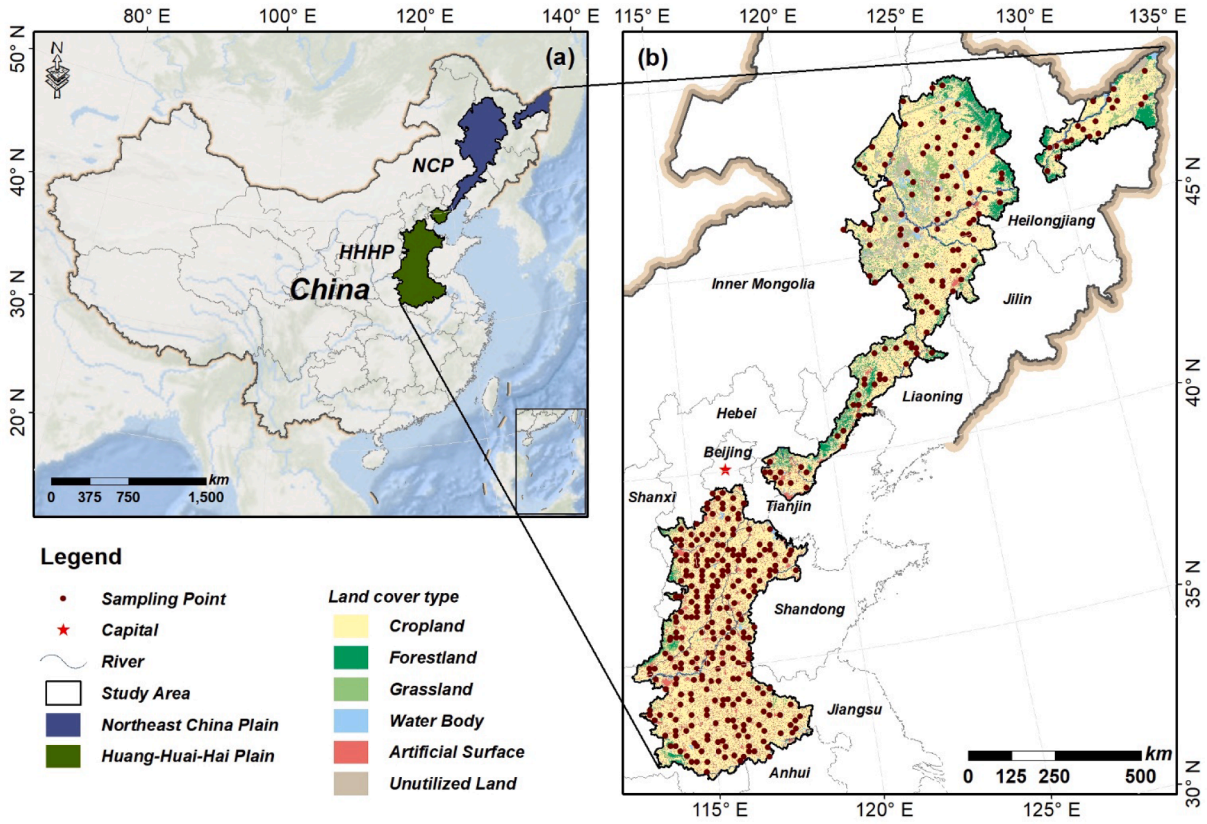

Fig. 1. The location of study area (a) and the distribution of soil sampling sites (b).

Considering the difference in the spectral characteristics of Landsat TM, ETM+ and OLI, we harmonized blue, green, red, near-infrared, shortwave infrared 1, and shortwave infrared 2 in Landsat TM, ETM+ to OLI with the ordinary least square regression coefficient (Roy et al., 2016). The products of MODIS from 2003 to 2017 were also used as environmental covariates.

The Geospatial Soil Sensing System (GEOS3) algorithm was used to detect the bare soil pixels in all available Landsat images from 2003 to 2017 (Demattê et al., 2018), and then a median composite image of all bare soil pixels was used to calculate 11 soil indices such as Brightness Index (BI), Hue Index (HI) and Saturation Index (SI). Additionally, legacy digital soil maps such as soil erosion (SE) and soil particle size fractions were included (Hengl et al., 2017; Teng et al., 2019).

The climate covariates mainly originated from WorldClim2, including 7 climatological and 16 bioclimatic (BIO) variables at 1 km resolution for the period of 1970–2000 (Fick and Hijmans, 2017). Moreover, the average of Day-time and Night-time Land Surface Temperature (LSTD, LSTN) at 1 km resolution were calculated using MOD11A1 (Wan et al., 2021). As for Normalized Difference Snow Index (NDSI), we used Landsat (2003–2017) to calculate the intra-annual and intra-quarterly mean and standard deviation (10 variables), mean of the intra-annual and intra-quarterly maximum and minimum (10 variables), and averaged intra-quarterly standard deviation (1 variable), resulting in 21 variables to better represent its temporal change (Zhou et al., 2019a). As for Potential Evapotranspiration (PET) and Evapotranspiration (ET), we used MOD16A2 at 500 m resolution to calculate the mean and standard deviation of the intra-annual and intra-quarterly aggregation (10 variables), and averaged standard deviation of intra- quarterly aggregation in 2003–2017 (1 variable) (Running et al., 2017).

A total of 253 covariates related to organisms were calculated by Landsat and MODIS products (Loveland and Dwyer, 2012; Myneni et al., 2021; Running et al., 2015). Leaf Area Index (LAI), Enhanced Vegetation Index (EVI), Normalized Difference Vegetation Index (NDVI) and eight other indices were calculated in the same way as NDSI (21 variables for each). The calculation of Gross Primary Production (GPP) and Net Primary Production (NPP) from MOD17A2H was consistent with PET and ET (11 variables for each) (Running et al., 2015).

The DEM at 90 m resolution was derived from SRTM of National Aeronautics and Space Administration (NASA) (Andy et al., 2008). Based on the elevation, we used SAGA GIS to derive 45 relief derivatives (Conrad et al., 2015), such as Analytical Hillshading (AH), Slope (SLO), Channel Network Base Level (CNBL), Plan Curvature (PLC), Multi-resolution Index of Valley Bottom Flatness (MRVBF) and Topographic Wetness Index (TWI).

Based on the latitude and longitude, oblique geographic coordinates (OGC) were calculated at 15°, 30°, 45°, 60°, 75°, 105°, 120°, 135°, 150° and 165° angles, with the following equation (Møller et al., 2020):

$$OGC = \sqrt{Lat^2 + Lon^2} \times \cos\left(\alpha - \tan^{-1}(Lat/Lon)\right) \quad (2)$$

where Lon and Lat are the longitude and latitude, α is the angle (in degrees) of the titled axis relative to the x axis.

All environmental covariates were resampled to 250 m resolution using the bilinear method for spatial modelling and map prediction. The collection and pre-processing of remote sensing data were performed in Google Earth Engine (Gorelick et al., 2017). All covariates were unified into the CGCS WGS 1984 geographic coordinate system, and predicted maps were projected into the Albers Conic Equal Area coordinate system.

2.3. Variable selection methods

2.3.1. Boruta

The Boruta algorithm is one of the widely used variable selection methods in DSM (Xiong et al., 2014; Xu et al., 2017; Keskin et al., 2019; Rasaei et al., 2020). Proposed by Kursa and Rudnicki (2010), Boruta first

Table 1
List of environmental covariates used in this study. The number in the columns of Scorpan factor and Covariate indicates the number of variables within each group.

Scorpan factor	Covariate	Abbreviation	Scale	Reference
Soil (15)	Soil Erosion (1)	SE	1000 m	Teng et al. (2019)
	Clay, Silt and Sand Content (3)	Clay, Silt, Sand	250 m	Hengl et al. (2017)
	Brightness Index (1)	BI	30 m	Abbas and Khan (2007)
	Carbonate Index (1)	CarI	30 m	Boettinger et al. (2008)
	Coloration Index (1)	Coll	30 m	Escadafal (1994)
	Ferrous Minerals (1)	FM	30 m	Imbroane et al. (2007)
	Gypsum index (1)	GI	30 m	Boettinger et al. (2008)
	Hue Index, Saturation Index (1)	HI, SI	30 m	Mandal (2016)
	Iron Oxide (1)	IO	30 m	Hewson et al. (2001)
	Reflectance Absorption Index (1)	RAI	30 m	Ghaemi et al. (2013)
	Redness Index (1)	RI	30 m	Madeira et al. (1997)
	Stress Related (1)	SR	30 m	Foody et al. (2001)
Climate (68)	Bioclimatic Variables (Mean Diurnal Range, Isothermality, Temperature Seasonality, Temperature Annual Range, Mean Temperature Of Wettest Quarter, Mean Temperature of Driest Quarter, Mean Temperature of Warmest Quarter, Mean Temperature of Coldest Quarter, Annual Precipitation, Precipitation of Wettest Month, Precipitation of Driest Month, Precipitation Seasonality, Precipitation of Wettest Quarter, Precipitation of Driest Quarter, Precipitation of Warmest Quarter, Precipitation of Coldest Quarter) (16)	BIO02, 03, 04, 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19	1000 m	Fick and Hijmans (2017)
	Wind Speed, Vapor Pressure, Solar Radiation, Average Precipitation (4)	Wind, VP, Sol, Prec	1000 m	Fick and Hijmans (2017)
	Temperature (Average, Maximum, Minimum) (3)	Tavg, Tmax, Tmin	1000 m	Fick and Hijmans (2017)
	Day-time and Night-time Land Surface Temperature (2)	LSTD, LSTN	1000 m	Wan et al. (2021)
	Potential Evapotranspiration (11)[a] and Evapotranspiration (11)[a]	PET, ET	500 m	Running et al. (2017)
	Normalized Difference Snow Index (21)[b]	NDSI	30 m	Riggs et al. (1994)
Organisms (253)	Canopy Index (21)[b]	CANI	30 m	Vescovo and Gianelle (2008)
	Differenced Vegetation Index (21)[b]	DVI	30 m	Richardson and Wiegand (1977)
	Enhanced Vegetation Index (21)[b]	EVI	30 m	Huete et al. (2002)
	Green Atmospherically Resistant Vegetation Index (21)[b]	GARI	30 m	Gitelson et al. (1996)
	Normalized Difference Red/Green Redness Index (21)[b], Ratio Vegetation Index (21)[b]	NDRI, RVI	30 m	Bannari et al. (1995)
	Normalized Difference Vegetation Index (21)[b]	NDVI	30 m	Tucker (1979)
	Normalized Difference Water Index (21)[b]	NDWI	30 m	Gao (1996)
	Soil Adjusted Vegetation Index (21)[b]	SAVI	30 m	Huete (1988)
	Gross (11)[a] and Net Primary Production (11)[a]	GPP, NPP	500 m	Running et al. (2015)
	Fraction of Photosynthetically Active Radiation (21)[b], Leaf Area Index (21)[b]	FPAR, LAI	500 m	Myneni et al. (2021)
Relief (46)	Elevation, Analytical Hillshading, Aspect, Slope, Catchment Area, Total Catchment Area, Channel Network Base Level, Channel Network Distance, Closed Depressions, Clusters, Landform, Morphometric Features, Convergence Index, Flow Accumulation, Flow Path Length, Generalized Surface, Geomorphons, Gradient, Curvature (Cross-Sectional, Classification, Downslope, Local, Local Downslope, Local Upslope, Longitudinal, Maximum, Minimum, Plan, Profile, Upslope), Slope Length Factor, Slope Length, Relative Slope Position, Melton Ruggedness Number, Multiresolution Index of Ridge Top Flatness, Multiresolution Index of Valley Bottom Flatness, Stream Power Index, Topographic Position Index, Topographic Wetness Index, Upslope Height,, Valley Depth, Vector Terrain Ruggedness, Diurnal Anisotropic	ELE, AH, ASP, SLO, CA, TCA, CNBL, CND, CD, CLU, LAN, MF, CGI, FA, FPL, GS, GEO, GRA, CSC, CVC, DSC, LC, LDC, LUC, LTC, MAC, MIC, PLC, PRC, UC, LSF, SL, RSP, MRN, MRRTF, MRVBF, SPI, TPI, TWI, UH, MAH, VD, VTR, DAH, WE, SI	90 m	Jarvis et al. (2008)

(continued on next page)

Table 1 (*continued*)

Scorpan factor	Covariate	Abbreviation	Scale	Reference
	Heating, Wind Exposition and Shelter Index (46)			
Position (10)	Oblique geographic coordinate at 15°, 30°, 45°, 60°, 75°, 105°, 120°, 135°, 150°, 165° (10)	OGC15, 30, 45, 60, 75, 105, 120, 135, 150, 165	30 m	Møller et al. (2020)

[a] Mean and standard deviation of the intra-annual and intra-quarterly aggregation in 2003–2017 (10 variables), and averaged standard deviation of intra- quarterly aggregation in 2003–2017 (1 variable).

[b] Intra-annual and intra-quarterly mean and standard deviation in 2003–2017 (10 variables), mean of the intra-annual and intra-quarterly maximum and minimum in 2003–2017 (10 variables), and averaged intra-quarterly standard deviation in 2003–2017 (1 variable).

duplicates the data set and then shuffles its predictors in each column (which are called shadow predictors). Afterwards, it trains a Random Forest (RF) model using the original and shuffled data sets combined and evaluates the variable importance (Z score) for each predictor. In each iteration, it checks whether a real predictor has higher importance (RMSE) than the best of its shadow predictors and marks the predictor as either confirmed (important) or rejected (unimportant). Finally, it stops when all the predictors are confirmed or rejected. The Boruta algorithm was performed in R package "Boruta" (Kursa and Rudnicki, 2010).

2.3.2. Recursive feature elimination

The RFE algorithm is commonly used to select the most relevant predictors for machine learning methods (Gomes et al., 2019; Chen et al., 2021; Poggio et al., 2021; Hounkpatin et al., 2022). Based on the backward selection, RFE works as follows: (1) fit a model using all the n predictors, calculate model performance by k-fold cross-validation (RMSE) and the variable importance; (2) remove the least important predictor from the pool, refit the model, assess model performance and remove the least important predictor again; (3) repeat the second procedure down to a pool from n to 1 with a step of 1; (4) determine the optimal number of predictors by taking the model with the best performance (RMSE). The RFE was implemented in R package "caret" using RF as an internal model (Kuhn, 2021).

2.3.3. Variance inflation factor analysis

The VIF analysis is a commonly used method to evaluate the multicollinearity among predictors and select the optimal subset of predictors that are not correlated (Curto and Pinto, 2011). During the VIF process, each predictor is regressed against the remaining predictors by a linear model using ordinary least squares regression. After the determination of the coefficient of determination (R^2) from the fitted linear model, the VIF is calculated by equation 3:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{3}$$

where R_i^2 is the R^2 for predictor i.

According to previous studies, predictors with VIF $>$ 10 were determined to be highly correlated with other predictors and were therefore removed from the final DSM model (Odhiambo et al., 2020; Taghizadeh-Mehrjardi et al., 2021). The R package "car" was used to perform VIF analysis (Fox and Weisberg, 2019).

2.3.4. Modified greedy feature selection

Based on the GFS algorithm (Drobnič et al., 2020), a modified greedy feature selection (MGFS), is proposed for DSM regression problem. Opposite to RFE, MGFS adopts a forward selection strategy which includes steps (Fig. 2): (1) fit a model using all the n predictors, and calculate the variable importance; (2) select the most important predictor (only one) to fit an initial model, and calculate the model performance (RMSE) by k-fold cross-validation (note that there is only one predictor in the pool); (3) fit a list of models using 2 predictors (the combinations of predictor(s) in the pool and one of the remaining predictors), calculate their model performance, and record the model with the best performance; (4) update the pool by taking the predictors from the best model in the previous step; (5) repeat steps 3 and 4 by increasing the number of predictors from 3 to n. The predictors in the model with the best performance (RMSE) are selected for the final model. It is possible to set an early stop when the model performance starts to decrease for a large number of predictors (>50). Since Boruta and RFE both used RF, the same internal model was used in MGFS for a fair comparison in this study. The differences between original GFS and MGFS are in two aspects:

(1) GFS adds an additional constraint (least-trees-used criterion) when adding predictors to the model. This additional constraint likely has little impact since a tie would be required between 2 predictors, which is highly unlikely, and therefore it is not adopted in MGFS;

(2) Inspired by RFE, MGFS started with the most important predictor based the fitted model using all the predictors, while GFS tested all the single predictor in the first iteration. Therefore, MGFS should be slightly

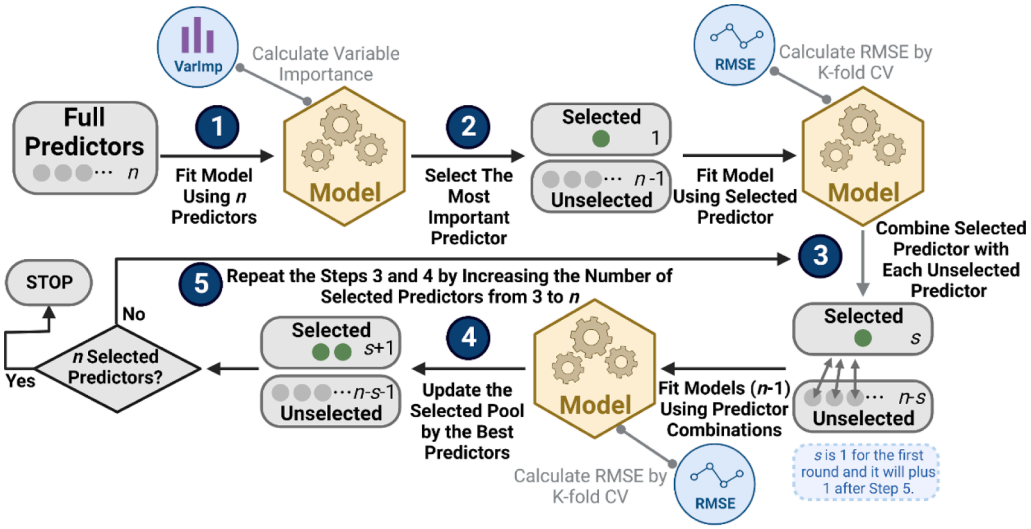

Fig. 2. Diagram of modified greedy feature selection (MGFS).

more efficient than GFS.

In summary, VIF and Boruta are fundamentally different from RFE and MGFS. VIF operates independently from the dependent predictor and deals primarily with multicollinearity, and Boruta is based on individual predictor importance with regards to shadow predictors. Solely based on the variable importance, RFE and MGFS are equivalent.

2.4. Calibration model

Quantile regression forest (QRF) has seen increased use in DSM studies, especially at a broad scale, because it provides calculation of uncertainty and good model performance (Meinshausen and Ridgeway, 2006; Vaysse and Lagacherie, 2017; Loiseau et al., 2019; Kasraei et al., 2021; Poggio et al., 2021).

We define X and Y as the predictor variables and target variables, QRF generates a large number of trees (b) using bootstrapping (random sampling with replacement) from p training samples (X_i,Y_i), $i = 1,\ldots,p$. A random subset of the predictor variables is then used to select split-point for each node of the bootstrap tree. For a new sample $N = X_n$, its prediction for each bootstrap tree is the conditional mean estimate of Y, which can be formulated as below:

$$\widehat{X} = \sum_{i=1}^{p} w_i Y_i \tag{4}$$

where w_i is the weight for the sample (X_i,Y_i) in the same leaf within the same bootstrap tree.

The mean prediction of b bootstrap trees is used to represent the final prediction of the new sample N. Using the weighted samples, QRF can also derive a conditional distribution from which the probability of Y being lower than a given percentile can be determined and thus calculate the prediction intervals. We refer to Meinshausen and Ridgeway (2006) for more details relevant to the calculation of conditional distribution. The variables to possibly split at each node (mtry) were optimized by 5-fold cross-validation with grid searching, and other parameters were set to default values as suggested by (Kuhn, 2021). The "caret" (Kuhn, 2021) and "quantregForest" (Meinshausen, 2017) packages were used for optimizing and running QRF in R (R Core Team, 2021). Predictions at mean, 5th and 95th percentiles can be derived from the fitted QRF model directly (Vaysse and Lagacherie, 2017; Poggio et al., 2021).

2.5. Evaluation of model performance and computation efficiency

The whole dataset (402 samples) was randomly split into calibration (281 samples) and validation (121 samples) sets at a ratio of 70 % to 30 % with 50 repeats. The R^2, Lin's concordance correlation coefficient (LCCC), and RMSE were used to evaluate model accuracy on the validation set. Higher R^2 and LCCC close to 1 and a lower RMSE close to 0 mean better model accuracy.

$$R^2 = 1 - \frac{\sqrt{\sum_i^n (\widehat{y}_i - y_i)^2}}{\sqrt{\sum_i^n (y_i - \overline{y})^2}} \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i^n (\widehat{y}_i - y_i)^2}{n}} \tag{6}$$

$$\mathrm{LCCC} = \frac{2\rho\sigma_y\sigma_{\widehat{y}}}{(\overline{y} - \overline{\widehat{y}})^2 + \sigma_y^2 + \sigma_{\widehat{y}}^2} \tag{7}$$

where y_i and $\widehat{y}_i$ are observation and prediction for sample i, $\overline{y}$ and $\overline{\widehat{y}}$ are the mean of all the observations and predictions, n is the number of samples, ρ is the correlation coefficient, σ_y and $\sigma_{\widehat{y}}$ are the variances of all the observations and predictions.

The 90th prediction interval coverage probability (PICP90), calculated by the proportion of observations that fall within the 90 % prediction intervals (PIs), was used to determine the performance of the uncertainty estimate on the validation set (Malone et al., 2016). A PICP90 close to 0.9 (or 90 %) means the uncertainty estimate is efficacious. To determine the magnitude of uncertainty, we calculated the uncertainty index (UI) by equation 8 (Viscarra Rossel et al., 2014; Zhou et al., 2019b). A greater UI indicates higher model uncertainty.

$$\mathrm{UI} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{Q95}_i - \mathrm{Q5}_i}{Mean}\right) \tag{8}$$

where Mean, $\mathrm{Q95}_i$ and $\mathrm{Q5}_i$ are the prediction at mean, 95th and 5th percentiles for sample i in the validation set.

The computation efficiency was determined by the computation time using a DELL Precision 3650 Tower Workstation (Intel i9-10900 K CPU with 10 cores and 64 GB RAM). Here we recorded the computation time in two steps, variable selection and map prediction, which are the most time-consuming. The total number of pixels to be predicted was about 8.7 million for the cropland in the study area, and parallel computation was not adopted in map prediction.

3. Results

3.1. Statistical summary of soil properties

Table 2 shows the statistical summary of SOCD in all, calibration and validation datasets. For all dataset, SOCD ranged from 10.41 to 149.04 t ha^{-1}. The median and mean of SOCD were 41.29 and 46.26 t ha^{-1} at 0–30 cm. According to the coefficient of variation (47.65 %), SOCD had high heterogeneity. The skewness of 1.61 and kurtosis of 6.36 indicated that SOCD had a slightly positive and leptokurtic distribution in all dataset. The calibration and validation for SOCD had similar distributions to all dataset, showing that the repeated random data split (50 times) was reasonable for proper model calibration and validation.

3.2. Number of covariates selected by variable selection methods

Table 3 indicates the number of final covariates selected by the four variable selection methods. The optimal covariates were determined by Z score for Boruta, VIF score (<10) for VIF, and best model performance (RMSE) for RFE and MGFS (Figures S1, S2, S3 and S4). Among 392 environmental covariates, Boruta retained 76 covariates, within which 3, 26, 36, 6 and 5 covariates were related to soil, climate, organisms, relief and position. VIF kept 30 covariates linked to soil (3), climate (4), organisms (21), and relief (2). RFE selected 22 covariates, including 3, 5, 10, and 4 covariates relevant to soil, climate, organisms, and position, respectively. Selecting the minimum number of covariates, MGFS only kept 9 covariates for the final modelling, including BI (Brightness Index, soil), Tavg (Mean annual temperature, climate), UH (Upslope height, relief), OGC105 (Oblique geographic coordinate at 105°, position), NDVI3Min (Minimum NDVI at 3rd quarter, organisms), LAI2Min (Minimum LAI at 2nd quarter, organisms), NDRI1Avg (Average NDRI at 1st quarter, organisms), GARI3Max (Maximum GARI at 3rd quarter, organisms) and GPP3Avg (Average GPP at 3rd quarter, organisms) ranked descending by importance.

3.3. Model performance of four variable selection methods

The model performance using all 392 covariates and selected covariates by the four variable selection methods is presented in Fig. 3. VIF had the lowest model performance with R^2 of 0.48, LCCC of 0.64, and RMSE of 15.79 t ha^{-1}. The R^2, LCCC and RMSE for the model using all the covariates (named Full hereafter) were 0.54, 0.70 and 14.79 t ha^{-1}

Table 2
Descriptive statistics for SOCD (t/ha) in the all, calibration and validation datasets. The statistics for calibration and validation datasets are the mean of 50 replicates.

Dataset	No	Min	Q1	Median	Mean	Q3	Max	SD	%CV	Skew	Kurt
All	402	10.41	31.89	41.29	46.26	53.66	149.04	22.04	47.65	1.61	6.36
Calibration	281	10.82	31.83	40.97	46.08	53.07	144.69	21.82	47.34	1.60	6.33
Validation	121	11.98	32.36	41.56	46.67	53.99	136.01	22.28	47.64	1.52	6.00

Number of samples (No), minimum (Min), first quantile (Q1), third quantile (Q3), maximum (Max), standard deviation (SD), coefficient of variation (%CV), skewness (Skew) and kurtosis (Kurt).

Table 3
Number of covariates after variable selection.

Category	Number of covariates			
	RFE	Boruta	VIF	MGFS
Whole	22	76	30	9
Soil	3	3	3	1
Climate	5	26	4	1
Organisms	10	36	21	5
Relief	0	6	2	1
Position	4	5	0	1

in SOCD prediction. Compared to the Full model, RFE performed slightly better with R^2 of 0.57, LCCC of 0.72, and RMSE of 14.23 t ha^{-1}, while Boruta had a lower model performance with lower R^2 (0.55), LCCC (0.71) and greater RMSE (14.66 t ha^{-1}). Among all the models, MGFS performed best with R^2 of 0.60, LCCC of 0.74 and RMSE of 13.80 t ha^{-1}. Regarding the PICP, all the models were around 0.90 (0.88–0.92), indicating a similar quantification ability of 90 % PIs. The UI indicated that MGFS had the lowest global uncertainty (0.86) while the VIF model had the highest global uncertainty (0.98).

3.4. Computation efficiency of three variable selection methods

Table 4 shows the computation efficiency of variable selection and map prediction using the Full model and four variable selection methods. Since the Full model did not involve variable selection, the total computation time was 364 min which was equal to the time of map prediction. In the step of variable selection, Boruta (17 min) was the most efficient method, followed by MGFS (67 min), VIF (100 min) and RFE (150 min). In the step of map prediction, MGFS (34 min) performed much more efficient than RFE (51 min), VIF (59 min) and Boruta (102 min). Summing up two procedures, MGFS was the most efficient model (101 min), followed by Boruta (119 min), VIF (159 min), RFE (201 min), and the Full model (364 min).

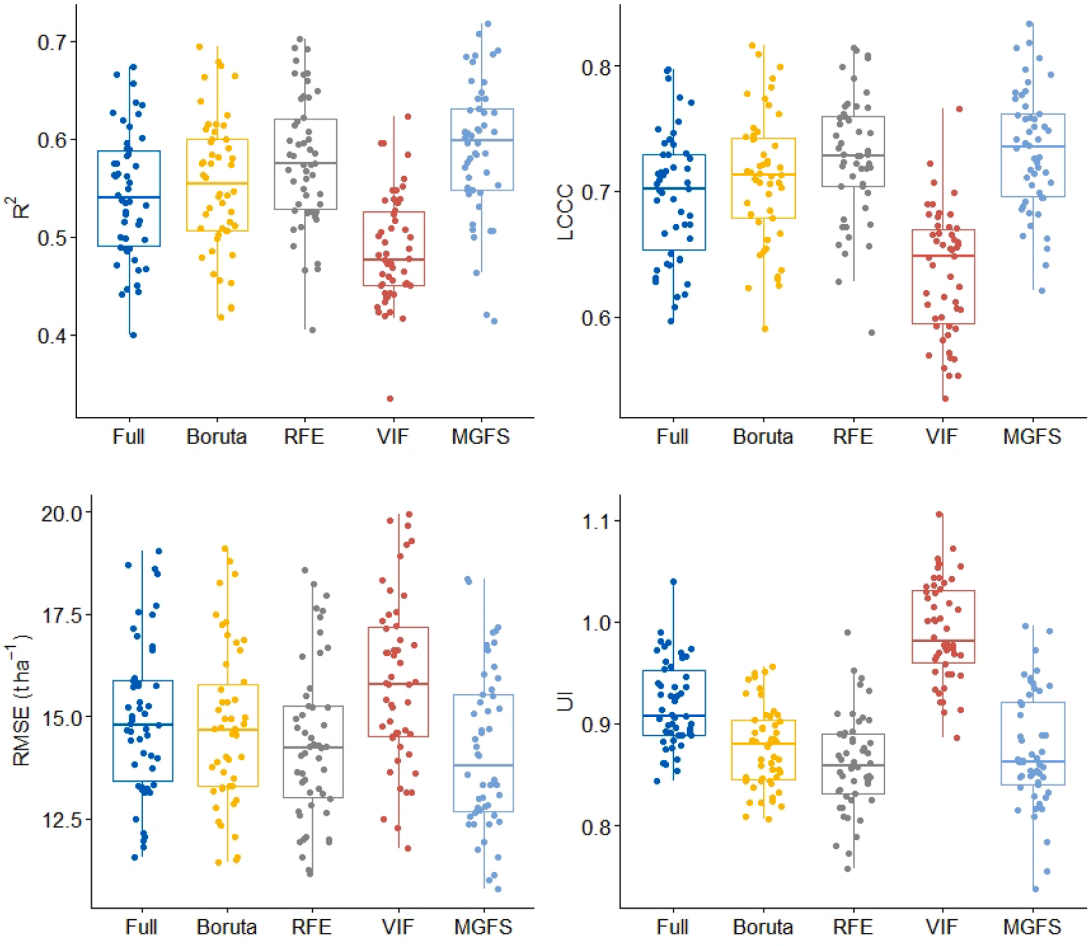

Fig. 3. Boxplots of QRF model performance with 50 repeats using all the predictors (Full), Boruta, recursive feature elimination (RFE), variance inflation factor (VIF), and modified greedy feature selection (MGFS).

Table 4
Computation efficiency of feature selection and map prediction (in minutes). The most efficient method is shown in bold font for each procedure.

Method	Feature selection	Map prediction	All procedures
Full	/	364	364
Boruta	**17**	102	119
RFE	150	51	201
VIF	100	59	159
MGFS	67	**34**	**101**

3.5. Spatial distribution of SOCD and associated uncertainty

The spatial distribution of SOCD and 90 % PIs are shown in Fig. 4 and Figure S5. A general increasing gradient was observed in all the SOCD maps from the south-western part (20–50 t ha^{-1}) to the north-eastern part (>70 t ha^{-1}), and the difference among the four SOCD maps was relatively small. The lower and upper limits (5 % and 95 % PI) of 90 % PIs had a similar spatial pattern, while the lower limits were generally<50 t ha^{-1} and upper limits were commonly>57 t ha^{-1} over the whole study area.

Fig. 5 presents the spatial prediction of UI. The general spatial patterns of UI using five models were similar and the area with a lower density of calibration samples had greater UI (Fig. 1). It was also evident from the frequency plots that MGFS had fewer pixels with UI > 1.6 and more pixels with UI < 0.7, indicating a low global uncertainty.

4. Discussion

4.1. Comparison of variable selection methods

Our results showed that MGFS selected the most parsimonious model with only 9 covariates from a total of 392 covariates compared to Boruta (76 variables), VIF (30 variables), and RFE (22 variables) (Table 3). Among 9 covariates selected by MGFS, BI was the most important one sensitive to the brightness of soils. Liu et al. (2020) noted that soil colour is dominated by SOC in Northeast and North China due to the cold continental semi-humid temperature regime. Therefore, being a proxy of soil colour, BI can provide crucial information in SOCD modelling. The good correlation between BI and SOC (or SOCD) was also confirmed by previous studies (Gholizadeh et al., 2018; Ayala Izurieta et al., 2021; Wang et al., 2021). Mean annual temperature was the second important covariate in SOCD modelling. Previous studies have demonstrated the importance of temperature in SOC variation because temperature predominantly affects the microbial decomposition of SOC (Wiesmeier et al., 2019). A low temperature leads to low SOC decomposition, so numerous studies indicated an increase in SOC with decreasing temperature (e.g., Koven et al., 2017; Chen et al., 2019). The latitude of the study area expands from 30°N to 48°N, leading to an apparent temperature gradient from the south-western part (16 °C) to the north-eastern part (-1.1 °C) that influences SOC decomposition. The upslope height (relief factor) is relevant to the potential energy entering the

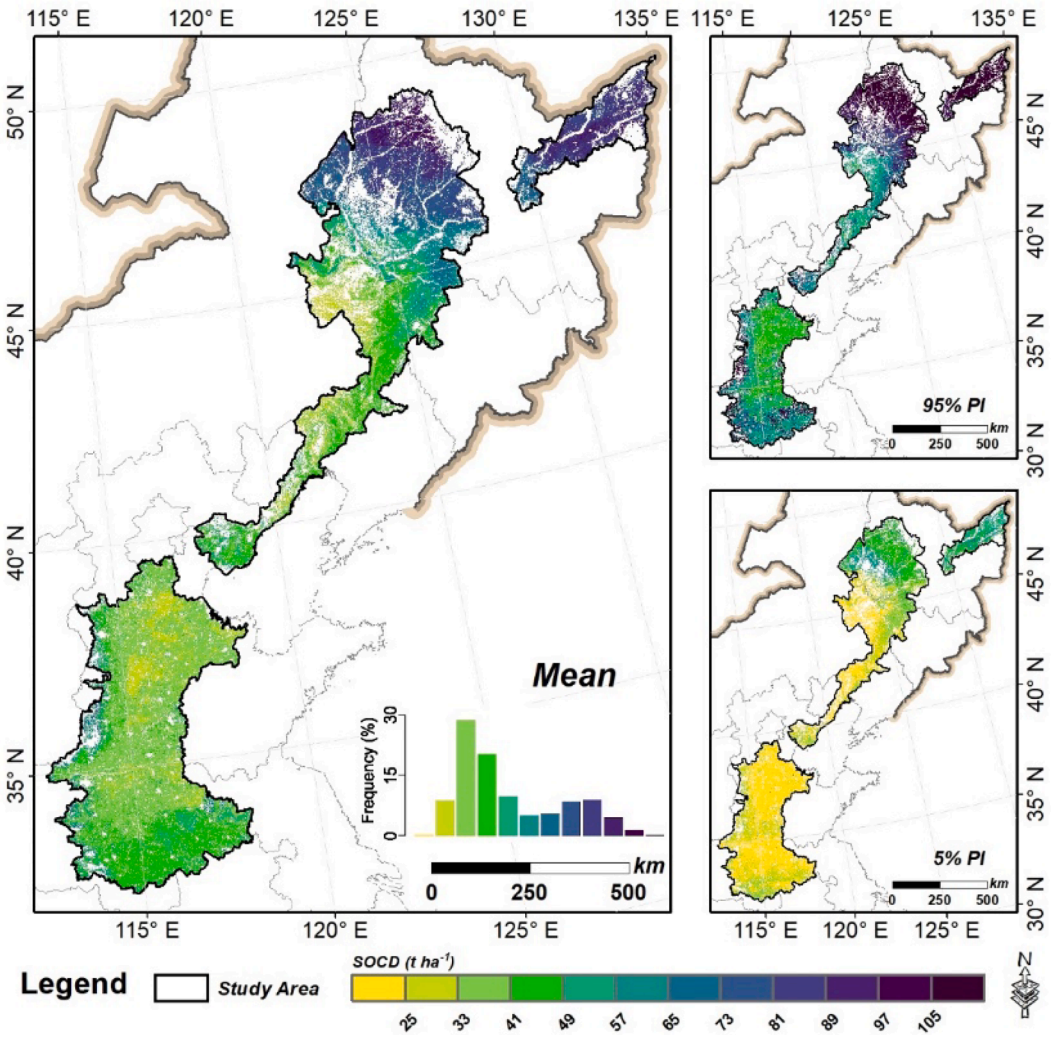

Fig. 4. Spatial distribution of SOCD produced by QRF using modified greedy feature selection (MGFS) and its associated 90% prediction intervals.

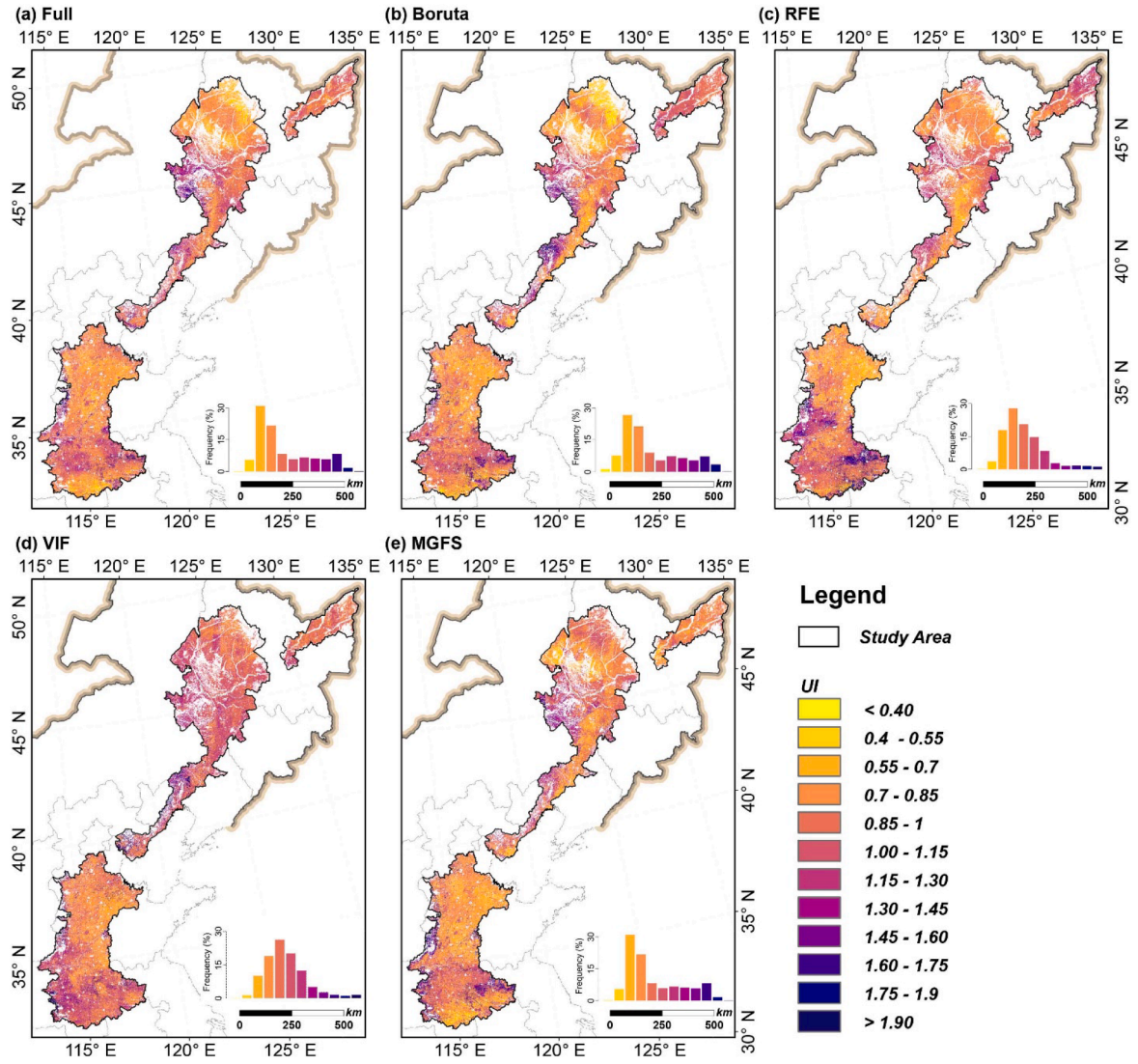

Fig. 5. Uncertainty index of SOCD produced by QRF using all the predictors(Full) (a), Boruta (b), recursive feature elimination (RFE) (c), variance inflation factor (VIF) (d) and modified greedy feature selection (MGFS) (e).

surface and indirectly affects the carbon input into the soil (Thompson et al., 2012). The importance of OGC at 105° indicates a spatial trend of SOCD at 105° relative to the × axis (Møller et al., 2020). The remaining 5 covariates (NDVI3Min, LAI2Min, NDRI1Avg, GARI3Max, GPP3Avg) were related to the organisms factor. It is interesting to note that all these covariates were calculated within a given quarter. This finding suggests that the remote sensing derived indicators calculated within a shorter time window (e.g., month, quarter) can capture more helpful vegetation phenology information related to soil properties than annual ones. This finding is in line with previous results from Hengl et al. (2017), Keskin et al. (2019) and Kempen et al. (2019).

Being the most parsimonious model, MGFS also performed best among four variable selection methods with R^2 of 0.60, LCCC of 0.74, and RMSE of 13.80 t ha^{-1} (Fig. 3, Table 3). Compared to the Full model (R^2 of 0.54, LCCC of 0.70, RMSE of 14.49 t ha^{-1}), MGFS increased 11.1 % and 5.7 % in R^2 and LCCC, and decreased 6.8 % in RMSE. RFE and Boruta performed slightly better than the Full model while VIF had the lowest model performance. The unsatisfactory result of VIF is rooted in the fact that VIF only accounted for the correlation between predictors and neglected their relationships to the target variable. In addition, VIF eliminates multicollinearity but it also removed part of the useful information between correlated variables, leading to a lower model performance. The implication of this result is that highly correlated variables are not mandatory to be removed before fitting a machine learning model (i.e. QRF) because part of their uncorrelated information can provide useful information for modelling. Our result demonstrates that variable selection methods can efficiently decrease the number of covariates, but they do not ensure improving model performance simultaneously (Chen et al., 2022b; Luo et al., 2022). Therefore, digital soil mappers should compare the Full model with the parsimonious model derived from the selected variable selection method and then decide whether the latter is preferable. There was no considerable difference in 90 % prediction intervals since all the PICP were around 0.9 (0.88–0.92), indicating all these uncertainty estimates were efficacious. The UI indicated that MGFS and RFE had the lowest global uncertainty (0.86) when compared to Boruta (0.88), Full model (0.91) and VIF (0.98). These results indicated that we need to include UI for reporting the magnitude of uncertainty in DSM studies since it is beyond the capacity of PICP (e.g., Viscarra Rossel et al., 2014; Zhou et al., 2019b; Poggio et al., 2021).

From the aspect of computation efficiency, MGFS was the most efficient variable selection method (101 min) when considering the variable selection (392 variables, 67 min) and map prediction (8.7 million pixels, 34 min) together (Table 4). When improving the map resolution from 250 m to 30 m, it can be expected that the superiority of MGFS in computation efficiency would be amplified. As noted by Chen

et al. (2022a), benefiting from rapidly growing remote sensing data and new soil observations obtained from multiple platforms, the DSM studies are moving forward to a broader spatial extent with a finer spatial resolution. For fine-resolution DSM at a broad scale, map prediction is the most time-consuming step. Take SoilGrids 2.0 for example, the total computation time (RFE, model training and map prediction) for a single soil property was 1500 CPU hours of which map prediction accounted for about two-thirds (Poggio et al., 2021). MGFS can efficiently promote computation efficiency by constructing a parsimonious model with a fair model performance, so it has a high potential for broad-scale DSM studies at a fine resolution.

4.2. Comparison SOCD product with SoilGrids 2.0

We compared our SOCD product with SoilGrids 2.0 mapped for the globe at 250 m resolution in Fig. 6 (Poggio et al., 2021). After evaluating the map accuracy using the same validation sets, we found that SoilGrids 2.0 had R^2 of 0.24, LCCC of 0.60 and RMSE of 18.84 t ha^{-1}, which indicated a poor accuracy of SOCD estimation compared to our product (R^2 of 0.60, LCCC of 0.74, RMSE of 13.80 t ha^{-1}). The spatial distribution maps showed that our SOCD estimate was generally higher than SoilGrids 2.0, especially in the northern part and the centre of the southwestern part as illustrated in the difference map between our SOCD product and SoilGrids 2.0. The frequency plot indicated that>88 % of the region was underestimated by SoilGrids 2.0. The estimated SOC stock was 2.89 Pg for the study area in our product while the SOC stock estimation from SoilGrids 2.0 product was 24.2 % lower (2.19 Pg). The poor SOCD estimate of SoilGrids 2.0 resulted from the fact that the soil database (>140,000 locations) used for producing SoilGrids 2.0 only had<80 sampling sites in the study area, leading to a limited weight in the spatial predictive model to represent the unique pedo-climatic condition for the study area. Therefore, when using the global soil maps for regional or national studies, it is necessary to correct them by the regional or national soil observations, especially for these regions with limited soil observations in producing the global soil maps. For this purpose, previous studies have proved that model averaging was a potential solution for producing more accurate "local" digital soil property maps by coupling "global" soil maps and "local" soil observations (Malone et al., 2014; Clifford and Guo, 2015, Caubet et al., 2019; Chen et al., 2020).

4.3. Limitations and perspectives

One limitation of MGFS is that it may take a long time to select the optimal variables for a large number of covariates (e.g., $n > 50$) as it needs to run the model $(n^2 + n)/2$ times. Therefore, the early stop is suggested in this study (see details in section 2.3.3) when using MGFS for a large set of covariates. It should be noted that the use of early stop may result in a local optimum which can be slightly different from the global optimum. From our personal experience, the local optimum was quite close to the global optimum but it still needs to be confirmed by other datasets (Xiao et al., 2022).

From theory, MGFS method can work on all the machine learning methods, such as RF, support vector machine, gradient boosted machine, XGBoost. In this study, we only evaluated its performance on RF, so its applicability remains to be tested for other machine learning methods in future studies.

5. Conclusions

In this study, we proposed the use of a modified greedy feature selection (MGFS), to improve model parsimony and keep model accuracy simutanouly for digital soil mapping (DSM) regression. Taking soil organic carbon density (SOCD) mapping of cropland in Northeast and North China as a case study, we evaluated the model performance, uncertainty estimate and computation efficiency on MGFS together with most currently used variable selection methods Boruta, recursive feature elimination (RFE), and variance inflation factor (VIF). Our results showed that MGFS selected the most parsimonious model with only 9 covariates from a list of 392 covariates while RFE, VIF and Boruta retained 22, 30 and 76 covariates, respectively. The repeated validation (50 times) indicated that the model using MGFS selected covariates performed much better (R^2 of 0.60, LCCC of 0.74, and RMSE of 13.80 t ha^{-1}) in SOCD mapping than VIF, Boruta and RFE. When compared to the model using all 392 covariates (R^2 of 0.54, LCCC of 0.70, RMSE of 14.79 t ha^{-1}), MGFS had an increase of R^2 and LCCC at 11.1 %, 5.7 %

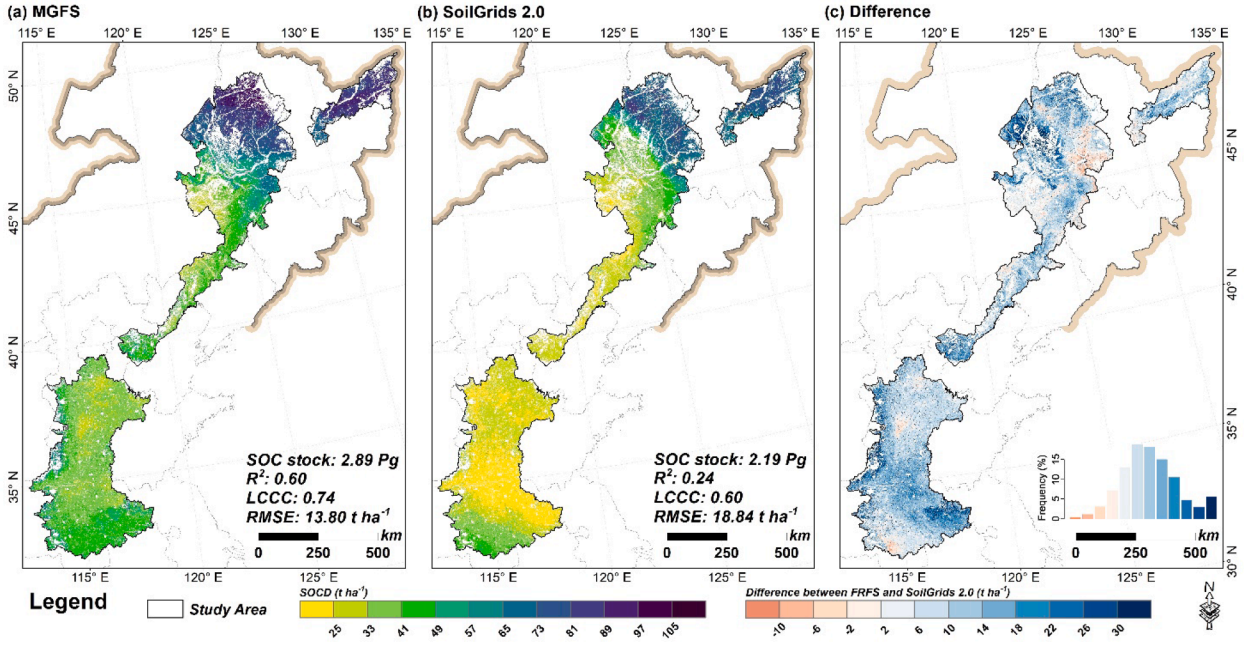

Fig. 6. Spatial distribution of SOCD produced by modified greedy feature selection (MGFS) (a) and SoilGrids 2.0 (b), and difference between MGFS and SoilGrids 2.0 (c).

and decrease of RMSE at 6.8 %. In addition, MGFS had the lowest global uncertainty (UI of 0.86) and the highest computation efficiency among all the models. Accordingly, we concluded that MGFS was superior to RFE, Boruta, VIF, and Full model in terms of model parsimony, model performance, map uncertainty and computation efficiency; therefore, it has a high potential in fine-resolution DSM studies at broad scales. Theoretically, MGFS can work on all machine learning methods; however, further tests on other datasets and machine learning models are still needed to evaluate its robustness.

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.

Data availability

The authors do not have permission to share data.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 41930754, 42001047, 41901055) and Ten-thousand Talents Plan of Zhejiang Province, China (2019R52004). Dominique Arrouays (DA) and Anne C. Richer-de-Forges (ARdF) are members of the research consortium GLADSOILMAP supported by LE STUDIUM Institute of Advanced Research Studies, France. ARdF is coordinator and DA is member of the French "Centre d'Expertise Scientifique Cartographie Numérique des sols" granted by the CNES-TOSCA program. We also would like to acknowledge two anonymous reviewers for their constructive suggestions which improved our manuscript a lot.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2023.116383.

References

Abbas, A., Khan, S., 2007. Using Remote Sensing Techniques for Appraisal of Irrigated Soil Salinity. Modsim 2007: International Congress on Modelling and Simulation (January), 2632-2638.
Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services—A global review. Geoderma 262, 101–111.
Amiri, M., Pourghasemi, H.R., Ghanbarian, G.A., Afzali, S.F., 2019. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. Geoderma 340, 55–69.
Andy, J., Isaak, R.H., Andrew, N., Edward, G., 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database: https://srtm.csi.cgiar.org.
Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.d.L., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. Advances in Agronomy, 125, 93-134.
Arrouays, D., Lagacherie, P., Hartemink, A.E., 2017. Digital soil mapping across the globe. Geoderma Reg. 9, 1–4.
Ayala Izurieta, J.E., Márquez, C.O., García, V.J., Jara Santillán, C.A., Sisti, J.M., Pasqualotto, N., Van Wittenberghe, S., Delegido, J., 2021. Multi-predictor mapping of soil organic carbon in the alpine tundra: a case study for the central Ecuadorian páramo. Carbon Balance Manag. 16 (1), 1–19.
Bannari, A., Morin, D., Bonn, F., Huete, A.R., 1995. A review of vegetation indices. Remote Sens. Rev. 13 (1–2), 95–120.
Bao, S.D., 2007. Soil agro-chemistrical analysis, 3rd ed. China Agriculture Press, Beijing, China.
Baveye, P.C., Baveye, J., Gowdy, J., 2016. Soil "ecosystem" services and natural capital: critical appraisal of research on uncertain ground. Front. Environ. Sci. 4, 41.
Boettinger, J., Ramsey, R., Bodily, J., Cole, N., Kienast-Brown, S., Nield, S., Saunders, A., Stum, A., 2008. Landsat spectral data for digital soil mapping, Digital soil mapping with limited data. Springer 193–202.
Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68–83.
Camera, C., Zomeni, Z., Noller, J.S., Zissimos, A.M., Christoforou, I.C., Bruggeman, A., 2017. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. Geoderma 285, 35–49.
Caubet, M., Román Dobarco, M., Arrouays, D., Minasny, B., Saby, N.P.A., 2019. Merging country, continental and global predictions of soil texture: lessons from ensemble modelling in France. Geoderma 337, 99–110.
Chen, S., Martin, M.P., Saby, N.P., Walter, C., Angers, D.A., Arrouays, D., 2018. Fine resolution map of top-and subsoil carbon sequestration potential in France. Sci. Total Environ. 630, 389–400.
Chen, S., Arrouays, D., Angers, D.A., Chenu, C., Barré, P., Martin, M.P., Saby, N.P., Walter, C., 2019. National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones. Sci. Total Environ. 666, 355–367.
Chen, S., Mulder, V.L., Heuvelink, G.B., Poggio, L., Caubet, M., Román Dobarco, M., Walter, C., Arrouays, D., 2020. Model averaging for mapping topsoil organic carbon in France. Geoderma 366, 114237.
Chen, S., Richer-de-Forges, A.C., Mulder, V.L., Martelet, G., Loiseau, T., Lehmann, S., Arrouays, D., 2021. Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France. Catena 198, 105062.
Chen, S., Arrouays, D., Mulder, V.L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Walter, C., 2022a. Digital mapping of soil properties at a broad scale: A review. Geoderma 409, 115567.
Chen, Y., Ma, L., Yu, D., Zhang, H., Feng, K., Wang, X., Song, J., 2022b. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. Ecol. Ind. 135, 108545.
Clifford, D., Guo, Y., 2015. Combining two soil property rasters using an adaptive gating approach. Soil Res. 53, 907–912.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8 (7), 1991–2007.
Curto, J.D., Pinto, J.C., 2011. The corrected vif (cvif). J. Appl. Stat. 38 (7), 1499–1507.
Demattê, J.A.M., Fongaro, C.T., Rizzo, R., Safanelli, J.L., 2018. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sens. Environ. 212, 161–175.
Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.
Drobnič, F., Kos, A., Pustišek, M., 2020. On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. Electronics 9 (5), 761.
Escadafal, R., 1994. Soil spectral properties and their relationships with environmental parameters-examples from arid regions, Imaging spectrometry—A tool for environmental observations. Springer 71–87.
Fao, 2015. Status of the world's soil resources: main report. Italy, Rome.
Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315.
Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Hughes, M.J., Laue, B., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. Remote Sens. Environ. 194, 379–390.
Foody, G.M., Cutler, M.E., McMorrow, J., Pelz, D., Tangki, H., Boyd, D.S., Douglas, I., 2001. Mapping the biomass of Bornean tropical rain forest from remotely sensed data. Glob. Ecol. Biogeogr. 10 (4), 379–387.
Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression, Third Edition. Sage, Thousand Oaks CA. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
Gao, B.C., 1996. NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58 (3), 257–266.
Ghaemi, M., Astaraei, A.R., Sanaeinejad, S.H., Zare, H., 2013. Using satellite data for soil cation exchange capacity studies. Int. Agrophys. 27 (4), 409–417.
Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sens. Environ. 218, 89–103.
Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58 (3), 289–298.
Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G., Fernandes Filho, E. I., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.
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.
Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.
Hewson, R.D., Cudahy, T., Huntington, J., 2001. Geologic and alteration mapping at Mt Fitton, South Australia, using ASTER satellite-borne data. In: IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat, No. 01CH37217). IEEE, pp. 724–726.

Hounkpatin, K.O., Bossa, A.Y., Yira, Y., Igue, M.A., Sinsin, B.A., 2022. Assessment of the soil fertility status in Benin (West Africa)–Digital soil mapping using machine learning. Geoderma Reg. 28, e00444.
Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25 (3), 295–309.
Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83 (1–2), 195–213.
Imbroane, M.A., Melenti, C., Gorgan, D., 2007. Mineral explorations by Landsat image ratios, Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2007). IEEE 335–340.
Jarvis, A., Reuter., H.I., Nelson., A., Guevara., E., 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database: https://srtm.csi.cgiar.org.
Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., Bethel, W., 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environ. Model. Softw. 144, 105139.
Kempen, B., Dalsgaard, S., Kaaya, A.K., Chamuya, N., Ruipérez-González, M., Pekkarinen, A., Walsh, M.G., 2019. Mapping topsoil organic carbon concentrations and stocks for Tanzania. Geoderma 337, 64–180.
Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58.
Koven, C.D., Hugelius, G., Lawrence, D.M., Wieder, W.R., 2017. Higher climatological temperature sensitivity of soil carbon in cold than warm climates. Nat. Clim. Chang. 7 (11), 817–822.
Kuhn, M., 2021. caret: Classification and Regression Training. R package version 6.0-88. https://CRAN.R-project.org/package=caret.
Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Softw. 36, 1–13.
Liu, F., Rossiter, D.G., Zhang, G.L., Li, D.C., 2020. A soil colour map of China. Geoderma 379, 114556.
Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.-L., Song, X., Shi, Z., Zhu, A.-X., Zhang, G.-L., 2022. Mapping high resolution National Soil Information Grids of China. Science Bulletin 67 (3), 328–340.
Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges, A.C., Lehmann, S., Bourennane, H., Saby, N.P., Martin, M.P., Vaudour, E., Gomez, C., 2019. Satellite data integration for soil clay content modelling at a national scale. Int. J. Appl. Earth Obs. Geoinf. 82, 101905.
Loveland, T.R., Dwyer, J.L., 2012. Landsat: Building a strong future. Remote Sens. Environ. 122, 22–29.
Luo, C., Zhang, X., Wang, Y., Men, Z., Liu, H., 2022. Regional soil organic matter mapping models based on the optimal time window, feature selection algorithm and Google Earth Engine. Soil Tillage Res. 219, 105325.
Madeira, J., Bedidi, A., Cervelle, B., Pouget, M., Flay, N., 1997. Visible spectrometric indices of hematite (Hm) and goethite (Gt) content in lateritic soils: the application of a Thematic Mapper (TM) image for soil-mapping in Brasilia. Brazil. International Journal of Remote Sensing 18 (13), 2835–2852.
Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232, 34–44.
Malone, B.P., Jha, S.K., Minasny, B., McBratney, A.B., 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. Geoderma 262, 243–253.
Mandal, U.K., 2016. Spectral Color Indices Based Geospatial Modeling of Soil Organic Matter in Chitwan District, Nepal. International Archives of the Photogrammetry, Remote Sensing & Spatial. Inf. Sci. 41 (B2), 43–48.
McBratney, A.B., Mendonça Santos, M.d.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117(1-2), 3-52.
McBratney, A.B., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213.
Meinshausen, N., Ridgeway, G., 2006. Quantile regression forests. J. Mach. Learn. Res. 7 (6), 983–999.
Meinshausen, N., 2017. quantregForest: Quantile Regression Forests. R package version 1.3-7. https://CRAN.R-project.org/package=quantregForest.
Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311.
Møller, A.B., Beucher, A.M., Pouladi, N., Greve, M.H., 2020. Oblique geographic coordinates as covariates for digital soil mapping. Soil 6 (2), 269–289.
Mosleh, Z., Salehi, M.H., Jafari, A., Borujeni, I.E., Mehnatkesh, A., 2016. The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environ. Monit. Assess. 188 (3), 1–13.
Myneni, R., Knyazikhin, Y., Park, T., 2021. MODIS/Terra+Aqua Leaf Area Index/FPAR 4-Day L4 Global 500m SIN Grid V061 . In: N.E.L.P. DAAC (Ed.).
Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4 (1), 1–22.
Odhiambo, B.O., Kenduiywo, B.K., Were, K., 2020. Spatial prediction and mapping of soil pH across a tropical afro-montane landscape. Appl. Geogr. 114, 102129.
Padarian, J., Minasny, B., McBratney, A.B., Dalgliesh, N., 2014. Predicting and mapping the soil available water capacity of Australian wheatbelt. Geoderma Reg. 2, 110–118.
Pereira, P., Bogunovic, I., Muñoz-Rojas, M., Brevik, E.C., 2018. Soil ecosystem services, sustainability, valuation and management. Current Opinion in Environmental Science & Health 5, 7–13.
Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240.
R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.
Rasaei, Z., Rossiter, D.G., Farshad, A., 2020. Rescue and renewal of legacy soil resource inventories in Iran as an input to digital soil mapping. Geoderma Reg. 21, e00262.
Richardson, A.J., Wiegand, C., 1977. Distinguishing vegetation from soil background information. Photogramm. Eng. Remote Sens. 43 (12), 1541–1552.
Riggs, G.A., Hall, D.K., Salomonson, V.V., 1994. A snow index for the Landsat thematic mapper and moderate resolution imaging spectroradiometer, Proceedings of IGARSS'94-1994 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 1942-1944.
Román Dobarco, M., Bourennane, H., Arrouays, D., Saby, N.P.A., Cousin, I., Martin, M.P., 2019. Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma 344, 14–30.
Roy, D.P., Kovalskyy, V., Zhang, H.K., Vermote, E.F., Yan, L., Kumar, S.S., Egorov, A., 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens. Environ. 185, 57–70.
Running, S., Mu, Q., Zhao, M., 2015. MOD17A2H MODIS/Terra Gross Primary Productivity 8-Day L4 Global 500m SIN Grid V006 . In: N.E.L.P. DAAC (Ed.).
Running, S., Mu, Q., Zhao, M., 2017. MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006 . In: N.E.L.P. DAAC (Ed.).
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.d.L., Minasny, B., 2009. Digital soil map of the world. Science, 325(5941), 680-681.
Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9 (1), 1–11.
Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Goydaragh, M. G., Schmidt, K., Scholten, T., 2021. Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. Geoderma 399, 115108.
Teng, H.F., Hu, J., Zhou, Y., Zhou, L.Q., Shi, Z., 2019. Modelling and mapping soil erosion potential in China. J. Integr. Agric. 18 (2), 251–264.
Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R., 2012. Digital Soil Mapping: Interactions with and Applications for Hydropedology. Hydropedology 665–709.
Tucker, C.J., 1979. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. Remote Sens. Environ. 8 (2), 127–150.
Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.
Vescovo, L., Gianelle, D., 2008. Using the MIR bands in vegetation indices for the estimation of grassland biophysical parameters from satellite remote sensing in the Alps region of Trentino (Italy). Adv. Space Res. 41 (11), 1764–1772.
Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Glob. Chang. Biol. 20 (9), 2953–2970.
Viscarra Rossel, R.A., Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat. Geosci. 12 (7), 547–552.
Wadoux, A.M.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359.
Wan, Z., Hook, S., Hulley, G., 2021. MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 . In: N.E.L.P. DAAC (Ed.).
Wang, H., Zhang, X., Wu, W., Liu, H., 2021. Prediction of Soil Organic Carbon under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed. Remote Sens. (Basel) 13 (7), 1229.
Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., 2019. Soil organic carbon storage as a key function of soils-A review of drivers and indicators at various scales. Geoderma 333, 149–162.
Xiao, Y., Xue, J., Zhang, X., Wang, N., Hong, Y., Jiang, Y., Zhou, Y., Teng, H., Hu, B., Lugato, E., Richer-de-Forges, A.C., Arrouays, D., Shi, Z., Chen, S., 2022. Improving pedotransfer functions for predicting soil mineral associated organic carbon by ensemble machine learning. Geoderma 428, 116208.
Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Holistic environmental soil-landscape modeling of soil organic carbon. Environ. Model. Softw. 57, 202–215.
Xu, Y., Smith, S.E., Grunwald, S., Abd-Elrahman, A., Wani, S.P., 2017. Incorporation of satellite remote sensing pan-sharpened imagery into digital soil prediction and mapping models to characterize soil property variability in small agricultural fields. ISPRS J. Photogramm. Remote Sens. 123, 1–19.
Xu, Y., Li, B., Shen, X., Li, K., Cao, X., Cui, G., Yao, Z., 2022. Digital soil mapping of soil total nitrogen based on Landsat 8, Sentinel 2, and WorldView-2 images in smallholder farms in Yellow River Basin, China. Environ. Monitor. Assess. 194 (4), 1–15.
Yang, R.M., Liu, L.A., Zhang, X., He, R.X., Zhu, C.M., Zhang, Z.Q., Li, J.G., 2022. The effectiveness of digital soil mapping with temporal variables in modeling soil organic carbon changes. Geoderma 405, 115407.
Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452.
Zeraatpisheh, M., Garosi, Y., Owliaie, H.R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., Xu, M., 2022. Improving the spatial prediction of soil organic carbon

using environmental covariates selection: A comparison of a group of environmental covariates. Catena 208, 105723.
Zhou, Y., Hartemink, A.E., Shi, Z., Liang, Z., Lu, Y., 2019a. Land use and climate change effects on soil organic carbon in North and Northeast China. Sci. Total Environ. 647, 1230–1238.
Zhou, Y., Webster, R., Rossel, R.V., Shi, Z., Chen, S., 2019b. Baseline map of soil organic carbon in Tibet and its uncertainty in the 1980s. Geoderma 334, 124–133.
Zhu, Z., Wang, S.X., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. Remote Sens. Environ. 159, 269–277.
Zhuo, Z., Chen, Q., Zhang, X., Chen, S., Gou, Y., Sun, Z., Huang, Y., Shi, Z., 2022. Soil organic carbon storage, distribution, and influencing factors at different depths in the dryland farming regions of Northeast and North China. Catena 210, 105934.