Geoderma 335 (2019) 47–56

Contents lists available at ScienceDirect

Geoderma

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

National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China

Zongzheng Liang[a,b], Songchao Chen[c,d], Yuanyuan Yang[a], Ruiying Zhao[a], Zhou Shi[a,*], Raphael A. Viscarra Rossel[b]

[a] Institute of Agricultural Remote Sensing and Information Technology Application, Zhejiang University, 310058 Hangzhou, China
[b] CSIRO Land & Water, Bruce E. Butler Laboratory, PO Box 1700, Canberra, ACT 2601, Australia
[c] INRA Unité InfoSol, 45075 Orléans, France
[d] UMR SAS, INRA, Agrocampus Ouest, 35000 Rennes, France

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Soil organic matter
Spatial modeling
Cubist machine learning algorithm
Soil map
Uncertainty assessment

ABSTRACT

Accurate digital soil maps of soil organic matter (SOM) are needed to evaluate soil fertility, to estimate stocks, and for ecological and environment modeling. We used 5982 soil profiles collected during the second national soil survey of China, along with 19 environment predictors, to derive a spatial model of SOM concentration in the topsoil (0–20 cm layer). The environmental predictors relate to the soil forming factors, climate, vegetation, relief and parent material. We developed the model using the Cubist machine-learning algorithm combined with a non-parametric bootstrap to derive estimates of model uncertainty. We optimized the Cubist model using a 10-fold cross-validation and the best model used 17 rules. The correlation coefficient between the observed and predicted values was 0.65, and the root mean squared error was 0.28 g/kg. We then applied the model over China and mapped the SOM distribution at a resolution of 90 × 90 m. Our predictions show that there is more SOM in the eastern Tibetan Plateau, northern Heilongjiang province, northeast Mongolia, and a small area of Tianshan Mountain in Xinjiang. There is less SOM in the Loess Plateau and most of the desert areas in northwest China. The average topsoil SOM content is 24.82 g/kg. The study provides a map that can be used for decision-making and contribute towards a baseline assessment for inventory and monitoring. The map could also aid the design of future soil surveys and help with the development of a SOM monitoring network in China.

1. Introduction

Soil organic matter (SOM) is an important component of soil that helps to determine crop yield and carbon sequestration (Manlay et al., 2007). It is a key property that affects soil quality and the assessment of soil resources. The amount of carbon stored in soil is three times that in the atmosphere (Post and Kwon, 2000), and thus, small losses of soil carbon to the atmosphere can have a significant impact on the overall emissions of greenhouse gases and the greenhouse effect (Raich and Potter, 1995).

Soil in China, like elsewhere, is subject to complex soil forming environments, with persistent soil erosion and degradation, and long-term intensive farming. As a consequence, the spatial distribution of soil properties is very heterogeneous and existing soil property maps have considerable uncertainty. There is a growing demand for fine-resolution soil property maps for applications in environmental modeling and monitoring. Traditional polygon-based soil maps are less useful for these purposes because they do not adequately characterize the spatial variation of continuous soil properties. For instance, there is a need for precise spatially explicit estimates of SOM at the national scale for providing baselines for monitoring and to inform national greenhouse gas inventories (Viscarra Rossel et al., 2014).

Dokuchaev firstly developed a scientific classification of soils, methods for soil mapping and established the foundation for the study of both soil genesis and soil geography (Buol et al., 2011). Later Jenny proposed the well-known State Factor Equation of soil, where soil is described as a function of CLimate, Organisms, Relif, Parent material and Time, referred to as *CLORPT* (Jenny, 1941). McBratney et al. (2003) and Scull et al. (2003) reviewed methods for soil mapping which they defined as a spatial soil information system using field and laboratory observational methods, coupled with spatial and non-spatial soil inference systems. Lagacherie and McBratney (2007) formalized the "*SCORPAN*" framework of McBratney et al. (2003) and in a collection of manuscripts described "digital soil mapping". Since then, much of the work on soil mapping with linear regression, geostatistical

* Corresponding author at: Environmental & Resource Sciences College, Zhejiang University, Hangzhou 310058, China.
E-mail address: shizhou@zju.edu.cn (Z. Shi).

https://doi.org/10.1016/j.geoderma.2018.08.011
Received 18 October 2017; Received in revised form 4 August 2018; Accepted 7 August 2018
Available online 10 August 2018
0016-7061/

methods, and data mining methods have fallen under the "digital soil mapping" umbrella (Adhikari et al., 2014; Chen et al., 2018; Grunwald, 2009; McBratney et al., 2003; Sun et al., 2012; Viscarra Rossel and Chen, 2011; Zhou et al., 2016).

In the last decade, much attention has been focused on soil carbon storage at the national scale. In China, most researchers have used classification statistics and interpolation methods to obtain average soil organic carbon content, soil depth, and other information (Pan, 1999; Wang et al., 2000; Wu et al., 2003; Xie, 2004; Yu et al., 2005). However, due to the inaccuracies of the input point data, the studies have produced diverse results. Other studies have used statistical models to map the spatial distribution of soil organic carbon in China, such as multiple regression combined with high accuracy surface modeling (HASM) and neural networks (Li et al., 2010; Q.Q. Li et al., 2013), and land surface modeling (Shangguan et al., 2013). But the spatial resolution of these studies is relatively coarse ($> 1 \times 1$ km). Large area, country, continental and global scale mapping at a fine resolution is now a major research emphasis that will allow for a better understanding of the soil resource and our environment (Arrouays et al., 2014).

Our aim here was to use a machine-learning model with data from the second national soil survey of China and covariates that represent the environmental factors, to map the spatial distribution of topsoil (0–20 cm) organic matter in China and its uncertainty at 90×90 m spatial resolution.

2. Materials and methods

2.1. Collection and processing of soil profile data

The study used a dataset of 5982 soil profiles derived from the second national soil survey of China (SNSSC), which was undertaken in the 1980s and is mainly recorded in the Soil Series of China (National Soil Survey Office, 1993, 1994a, 1994b, 1995a, 1995b, 1996) and the Soil Series of Provinces (National Soil Survey Office, 1998). The carbon determination was carried out by rational wet combustion (Pan et al., 2004). The soil data covers most geographical regions of China and is the most detailed soil survey available at the national scale.

The soil profiles were sampled and analyzed by genetic horizons, and thus the depth intervals for each soil profile are inconsistent. As variation in SOM down a profile is usually continuous, we used equal-area splines (Ponce-Hernandez et al., 1986; Bishop et al., 1999; Malone et al., 2009) to harmonize the SOM content of the topsoil, which we defined as the 0–20 cm depth layer. To fit the splines to the SOM values in the profiles we tested different tuning parameter values, λ: 10, 1, 0.1, 0.01, 0.001, 0.0001, and 0.00001. We found that $\lambda = 0.01$ produced the best fits with the smallest root mean square error (RMSE). The splines were fitted to a maximum depth of 1 m, and we aggregated the spline predictions of SOM over the 0–20 cm to represent topsoil. Fig. 1 displays SOM depth function curves for three random soil profiles under different land uses.

The soil profiles from the SNSSC lack precise geographical registration and have no data on latitude and longitude. However, they do have detailed sampling location information that can be accurate to the villages, fields. To verify the spatial location accuracy of the digitized soil profiles, we compared elevation, mean annual precipitation, and mean annual temperature (below) recorded in each soil profile with data extracted from a high-resolution digital elevation model (DEM) and digital climate map using linear regression analysis. Coefficients of determination (R^2) for both elevation and temperature were larger than 0.90, and it was 0.80 for precipitation (Fig. 2). These results confirmed that the spatial accuracy of the digitized soil profiles was adequate for our study. The spatial distribution of the profiles is shown in Fig. 3.

Statistical analysis of the dataset showed that the distribution of SOM was skewed, with a mean of 24.82 g/kg, maximum of 560.1 g/kg for a peat soil type, found in Ganzi in Sichuan Province, minimum of 0.6 g/kg for a clay soil type found in Gaolan in Gansu Province. SOM of topsoil showed strong spatial variation, with a coefficient of variation (CV) of 140%, which is mainly attributed to diverse soil types, land uses, ecosystems, etc., at the national scale. To ensure the data is normally distributed, the SOM data was log-transformed prior to modeling with logs to the base 10.

2.2. Environmental covariates

Following soil formation theory, a number of environmental covariates were chosen for our modeling. They include covariates that represent terrain, climate, biota, geology, and human activities (Table 1). Terrain information was derived from the 90-m shuttle radar topographic mission (STRM) DEM. All terrain attributes, including elevation, slope, aspect, curvature, slope length (LS), slope steepness, mass balance index (MBI), terrain ruggedness index (TRI), topographic wetness index (TWI), and multiresolution index of valley bottom flatness (MrVBF) were derived from the DEM with the System for Automated Geoscientific Analyses (SAGA) geographic information system (GIS) (http://www.saga-gis.org).

Data on daytime land surface temperature (LST_D), nighttime land surface temperature (LST_N), normalized difference vegetation index (NDVI), evapotranspiration (ET), and net primary productivity (NPP) were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) (Justice et al., 1998). The resolution of different data is shown in Table 1.

Monthly precipitation products were obtained from the Tropical Rainfall Measuring Mission (TRMM), which measures tropical and subtropical precipitation (http://trmm.gsfc.nasa.gov/data_dir/data.html).These data were at a coarse resolution (0.25° resolution) (Huffman et al., 2007) and so we downscaled it using a geographically weighted regression (Ma et al., 2017) to derive 90-m resolution mean annual precipitation that we could use here.

Daily air temperature data for the period 1951–2014 from 754 base meteorological observation stations distributed throughout mainland China were used to calculate annual mean temperature (http://cdc.nmic.cn/home.do). Mean annual temperature maps were produced at 90-m resolution using a regression-kriging approach with elevation, latitude, and longitude as the auxiliary variables.

Annual mean solar radiation data (1950–1980) were derived from the National Earth System Science Data Sharing Infrastructure (http://www.geodata.cn). Land use and land cover data were obtained from the Resource and Environment Data Center of the Chinese Academy of Sciences (http://www.resdc.cn/).

These environmental covariates (LST_D, LST_N, ET, NDVI, NPP and Solar Radiation) with resolution coarser than 90 m were resampled to 90-m using bilinear method in ArcGIS 10.0.

2.3. Digital soil mapping

2.3.1. Modeling

The sites recorded covariates (precipitation, temperature, and elevation) as well as other environmental covariates at the sampling locations are easy to obtain by overlaying the sampling locations with the covariates maps and can be used as predictors from which to predict soil properties such as SOM. We took advantage of this by setting up a model in the form of a decision tree at the sites for which we had data and then using the model to predict SOM elsewhere. Decision trees have become one of the most commonly used data mining algorithms and are ideally suited for dealing with complex nonlinear relationships and missing values (Quinlan, 1992).

We used the algorithm that is implemented in the 'cubist' library (Kuhn et al., 2014) in the R software (R Core Team, 2013). Cubist uses conditional functions to build rules that partition the data into regions that are similarly defined by the characteristics of the predictor variables. If the condition is true, then an ordinary least-squares linear model predicts the response. If the condition is false, then the rule

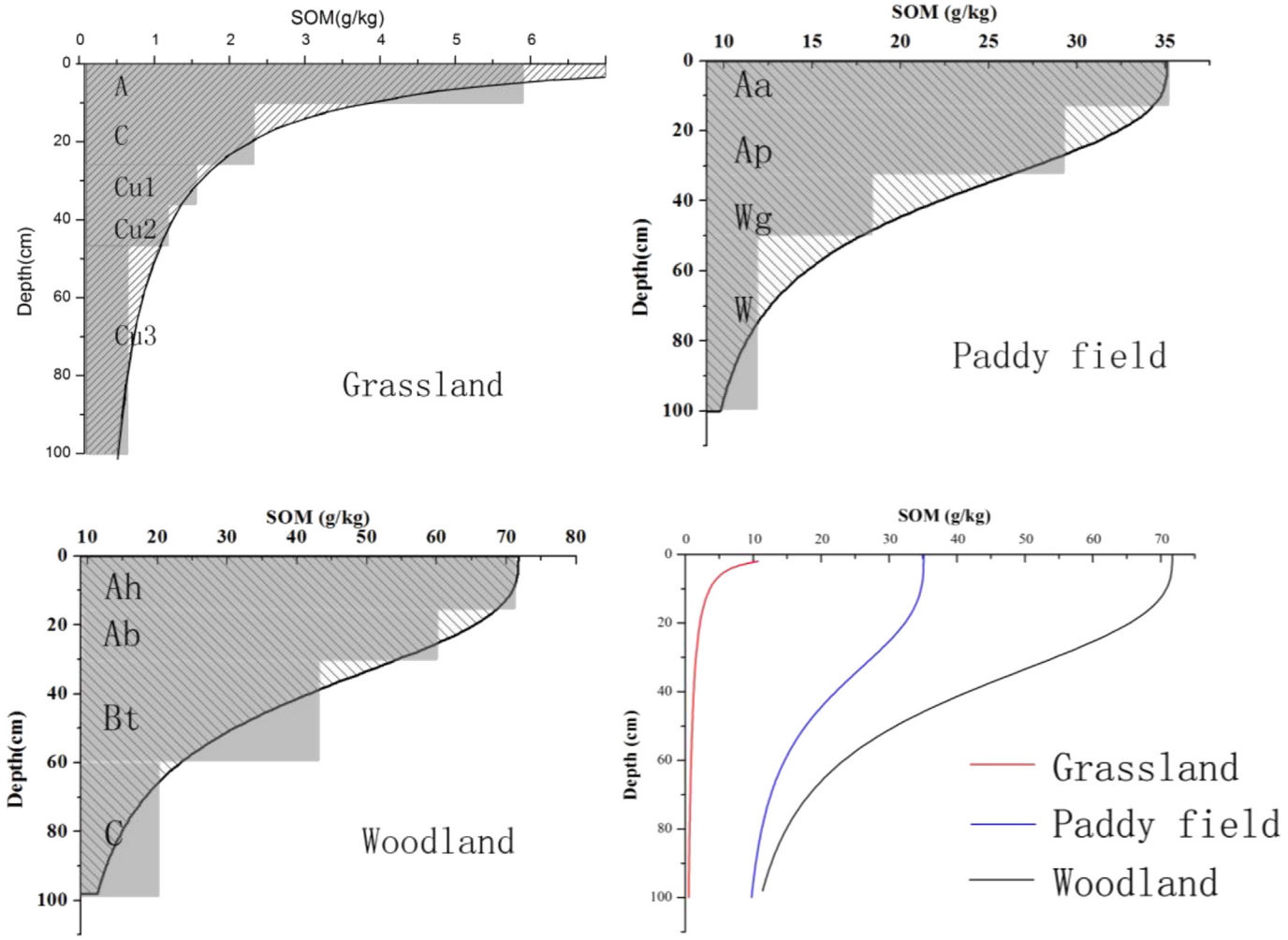

Fig. 1. Example of soil organic matter (SOM) depth function curves for three different land uses. Horizontal bars represent measured soil organic matter at different soil horizons, continuous line through horizons represents a fitted spline.

defines the next node in the tree. The sequence if, then, else is repeated. The result is that the regression equations, although general in form, are local to the partitions and have smaller errors smaller than other methods. The advantage of conditional rules is that they enable different linear models to capture the local linearity in different parts of the landscape, as represented by the predictor variable space, leading to smaller, more interpretable trees and better prediction accuracy than regression trees. Cubist has been extensively used in soil science to model and map soil properties (e.g. Bui et al., 2009; Henderson et al., 2005; Viscarra Rossel, 2011), to downscale remote sensing data (Ma et al., 2017), and to model soil sensor data and soil spectra (Viscarra Rossel and Webster, 2012).

We used Cubist models to make predictions at the notes of the 90 m grid. But, there are errors in the predictions from the Cubist model, and we quantified them by analyzing the model residuals. These residuals were autocorrelated. To account for them, and to improve prediction,

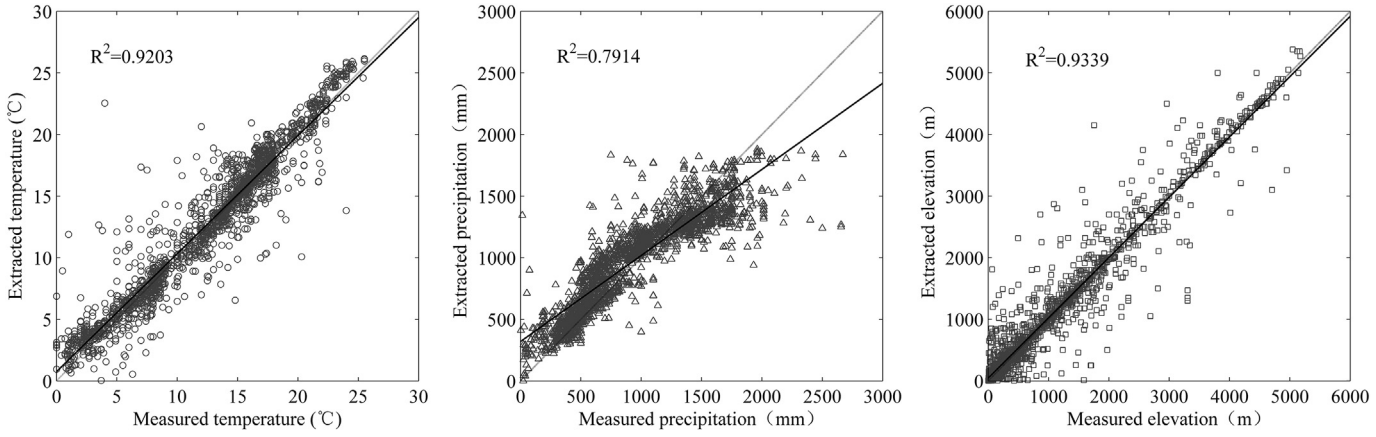

Fig. 2. Comparison of temperature, precipitation, and elevation measured in the in original record and against that extracted from remote sensing data.

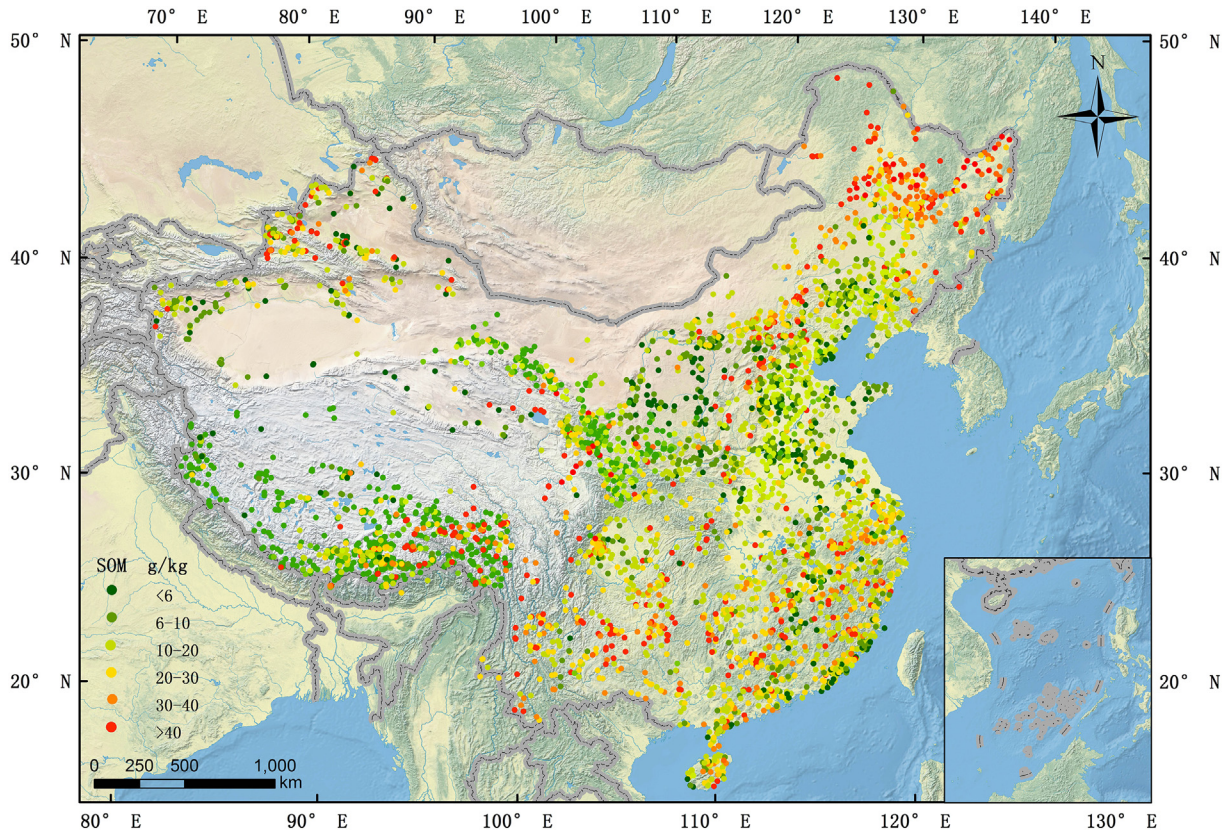

Fig. 3. Soil sample distribution and topsoil (0–20 cm) organic matter content, different colors representing different concentrations of SOM in topsoil. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Environmental covariates and their resolution.

Theme	Surrogate variable[a]	Resolution	Theme	Surrogate variable	Resolution
Climate	Annual mean temp, °C	90 m	Terrain	DEM, m	90 m
	LST_D, °C	5 km		Slope	90 m
	LST_N, °C	5 km		Aspect	90 m
	Potential evapotranspiration, mm	1 km		Curvature	90 m
	Rainfall, mm	90 m		MBI	90 m
	Radiation	1 km		TRI	90 m
Biota and vegetation	Net primary productivity, kg C·m^{-2}	1 km		TWI	90 m
	NDVI	250 m		LS	90 m
Soil	Soil land use types			MrVBF	90 m

[a] Key to terms: LST_D, daytime land surface temperature; LST_N, nighttime land surface temperature; NDVI, normalized difference vegetation index; DEM, digital elevation model; MBI, mass balance index; TRI, terrain ruggedness index; TWI, topographic wetness index; LS, slope length; MrVBF, multiresolution index of valley bottom flatness.

we used ordinary kriging to also predict the residuals at the nodes of the 90 m grid. The final Cubist-kriging SOM predictions summed the predictions from Cubist and the predicted residuals.

2.3.2. Bootstrapping

The non-parametric bootstrap (Efron and Tibshirani, 1993) is used to assess the uncertainties of the Cubist modeling. We bootstrapped the modeling, as described above, 50 times ($B = 50$) so that at each point of the 90 m grid, we generated 50 realizations of each of the final Cubist-kriging predictions of SOM. We averaged the 50 Cubist-kriging predictions to derive the mean estimates of SOM and we computed the uncertainty by adding the bootstrap variance of the Cubist predictions and the average kriging variances of the residuals (Viscarra Rossel et al., 2015). For clarity, a summary of the spatial modeling is presented in Fig. 4.

Because we modelled the data as logs, we then back-transformed the Cubist-kriging predictions of SOM by accounting for the variances in the data as follows:

$$\widehat{SOM} = \exp\{\ln(10) \times log\widehat{SOM} + \ln(10) \times 0.5\mathrm{Var}[log\widehat{SOM}]\} \quad (1)$$

where $log\widehat{SOM}$ is the average Cubist-kriging prediction from the 50 bootstraps on the log scale and its variance is expressed as $\mathrm{var}[log\widehat{SOM}]$. Cox's method proposed by Zhou and Gao (1997) was used to calculate and back-transform 95% confidence intervals. It is expressed as:

$$\exp\left\{\ln(10) \times (log\widehat{SOM}) + \ln(10) \times \frac{\widehat{V}}{2} \pm Z_{1-\alpha/2}\sqrt{\frac{\widehat{V}}{B} + \frac{\widehat{V}^2}{2(B-1)}}\right\} \quad (2)$$

where $\widehat{V}$ is the summation of the variance in the Cubist-kriging predictions and the average kriging variances of the residuals ($\overline{\sigma}^2$), $Z_{1-\alpha/2}$ is the standard normal deviate for the chosen probability $\alpha = 0.05$.

$$\widehat{V} = \mathrm{Var}(\widehat{\log SOM_{CK}}) + \overline{\sigma}^2 \quad (3)$$

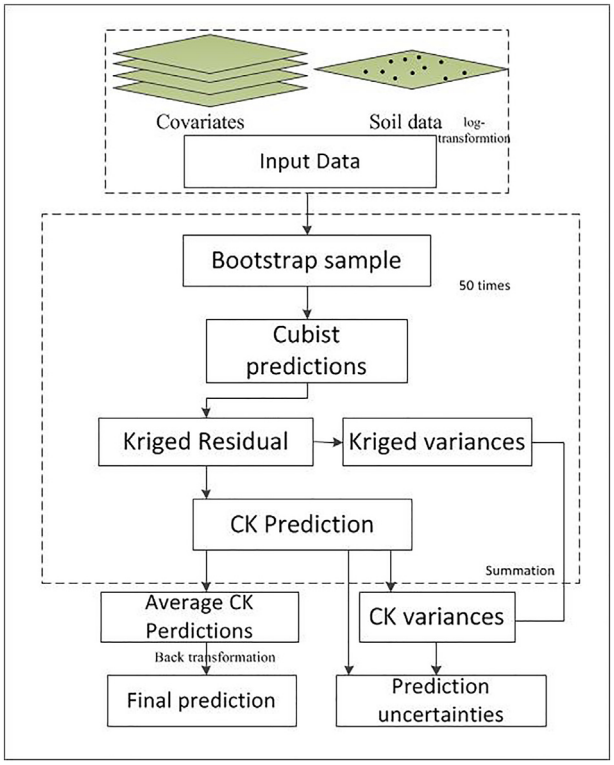

Fig. 4. Workflow of soil organic matter modeling and uncertainties assessment.

We assessed the uncertainties of the CK prediction, which can be defined as follows:

$$SOM_{uncertainty} = (T1 - T2)/\widehat{SOM} \quad (4)$$

where T1 and T2 are the upper and lower limits of 95% confidence intervals calculated in Eq. (2), respectively. We calculated the uncertainty of our estimates as the range of the 95% confidence intervals divided by their mean.

2.3.3. Model selection and validation

Models with a 2:1 training to test data split were implemented. Data from the 5982 soil profiles were randomly split into training (3982) and independent validation dataset (2000). We used the training set to perform 10-fold cross validation to test different models with up to 20 rules in each. We selected the model that produced the highest accuracy and independently assess the performance of the model with the validation data set. Finally, we refitted the best model with all 5982 observations to predict the "unknown" SOM at the nodes of the 90 m grid.

We used a range of statistics to assess the quality of the predictions. The Pearson correlation coefficient (r) was used to assess variation and correspondence between the predictions and original data, the root mean squared error (RMSE) to quantify the inaccuracy of the predictions, the mean error (ME) to assess bias, and finally the standard deviation of the error (SDE) to assess the precision of the predictions.

3. Results and discussion

3.1. Cubist spatial modeling

Climatic, ecological, and soil properties vary enormously at the national scale in China, resulting in high spatial variability of SOM and a complex correlation between soil organic matter and environmental covariates. In this study, we consider environmental covariates in each rule as variables that capture local variations of SOM in the different landscapes. Therefore, the relationship between environmental covariates and SOM was used to establish rule sets and independent models based on correlation to predict the SOM of corresponding soil landscape units. We tried a different number of rules and found that for this study, the prediction accuracy of the SOM model increases with the number of model rules. As shown in Fig. 5, the correlation coefficient reaches 0.62

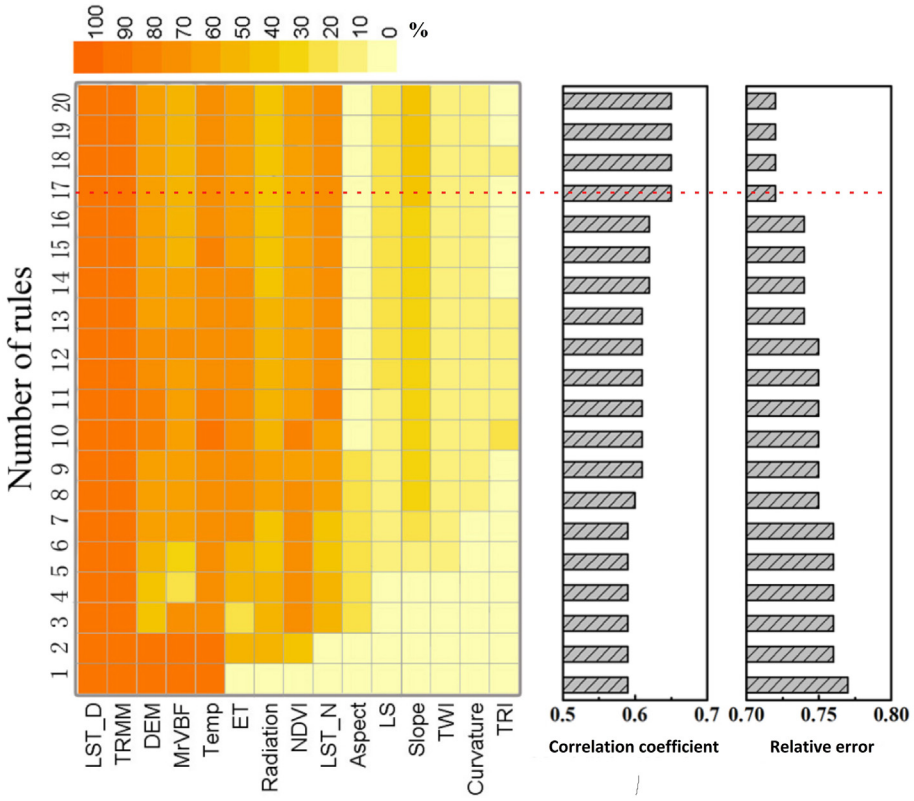

Fig. 5. Contribution rate of each environmental factor in different rule sets and corresponding prediction error associated with Cubist modeling. Key to variables: LST_D, daytime land surface temperature; TRMM, monthly precipitation from the tropical rainfall monitoring mission; DEM, digital elevation model; MrVBF, multiresolution index of valley bottom flatness; Temp, temperature; ET, evapotranspiration; NDVI, normalized difference vegetation index; LST_N, nighttime, land surface temperature; LS, slope length; TWI, topographic wetness index; TRI, terrain ruggedness index.

when the number of rules is 17, and the RE reaches 0.72. Model accuracy appears to level off above 17. To ensure both model accuracy and rule simplification, we set the number of rules in the Cubist model at 17.

Of the 17 environmental covariates, 15 were used in SOM modeling. Fig. 5 shows that in the Cubist SOM modeling process, the contribution rates of LST-D and TRMM are 96% and 90%, respectively, followed by LST-N at 75% and temperature at 72%. Contribution rates of DEM, MrVBF, ET, and NDVI are 60%, and those of the remaining factors (Radiation, Slope, LS, TWI, Curvature, TRI, and Aspect) are 47, 41, 12, 12, 12, 8, and 4%, respectively.

3.2. Construction of the Cubist rules

The Cubist model can reclassify and model the study area data by establishing different rule sets based on the characteristics of environment variables. Viscarra Rossel et al. (2014) showed how the rulesets of the Cubist model could be mapped to provide a better interpretation of the local drivers (determined form the linear models within each of the mapped rulesets) of SOC at the continental (Australian) scale. Viscarra Rossel and Bui (2016) used a similar mapping of the Cubist rulesets to better interpret the drivers for soil phosphorus in Australia. The dominant environmental factors influencing SOM vary in different regions. The Cubist rule set contained 17 individual linear models. The rule regions established by the Cubist algorithm (Fig. 6) are consistent with climatic zones in the Tianshan, Altai, and Kunlun Mountains in Xinjiang province. The rules are relatively consistent within similar regions, such as northeastern China and the Tibetan Plateau.

Soil is complex and SOM is affected by the interactions of numerous variables. Huang et al. (2018) pointed that the correlation between SOC and soil temperature can be positive or negative across the world. Large areas in China are characterized by a heterogeneous mix of multiple rules, which reflects the real-world situation. This is especially clear in the humid region of south China, due to the changeable terrain and variable coastal rainfall. The NDVI and the topographical characteristic were the two most important factors in the semi-humid region and the semi-arid region, which are mainly covered by vegetation (especially Tibet plateau and northeast China region). And there are complex terrain changes in two climate regions. The most important factor was NDVI in the desert arid region, where the poor vegetation coverage.

3.3. Digital SOM map

The spherical model is used for ordinary kriging to interpolate the residuals. The fitted range of spatial correlation of residual was about 12.3 km. The nugget value in the study is about 0.0072, whereas the sill is 0.07 (nugget to sill ratio was 10.3%). Due to the small range, thus the accuracy improvement is not significant in the western region, but in the eastern intensive sampling region, the accuracy has improved. Chien et al. (1997) pointed that the ratio of nugget to sill can be used as a criterion to classify the spatial dependence of properties. If this ratio is < 25%, the variable has strong spatial dependence, for 25%–75%, it has moderate spatial dependence, and for > 75%, only weak spatial dependence. So the residual variation showed a strong spatial structure.

The model predictive value and the kriging prediction of residuals were added to give our final predictions of the SOM, as in Fig. 4. The prediction accuracy in terms of r increased from 0.62 to 0.65 when considering the residual spatial variation in CK. The spatial distribution of SOM in top soil (0–20 cm depth) predicted by the model (Fig. 7) shows the higher SOM in southwest and northeast of China than in the northwest. Specifically, most of the higher values are located in eastern Tibet, the Three Rivers Watershed, western Sichuan, northern Heilongjiang, northeastern Inner Mongolia, and a small area of the Tianshan Mountains in Xinjiang province. Commonly, these areas are characterized by a cool climate and high percentage of forest cover. Lower SOM is mainly found in hilly areas of the Sichuan basin, the Loess Plateau, and most desert

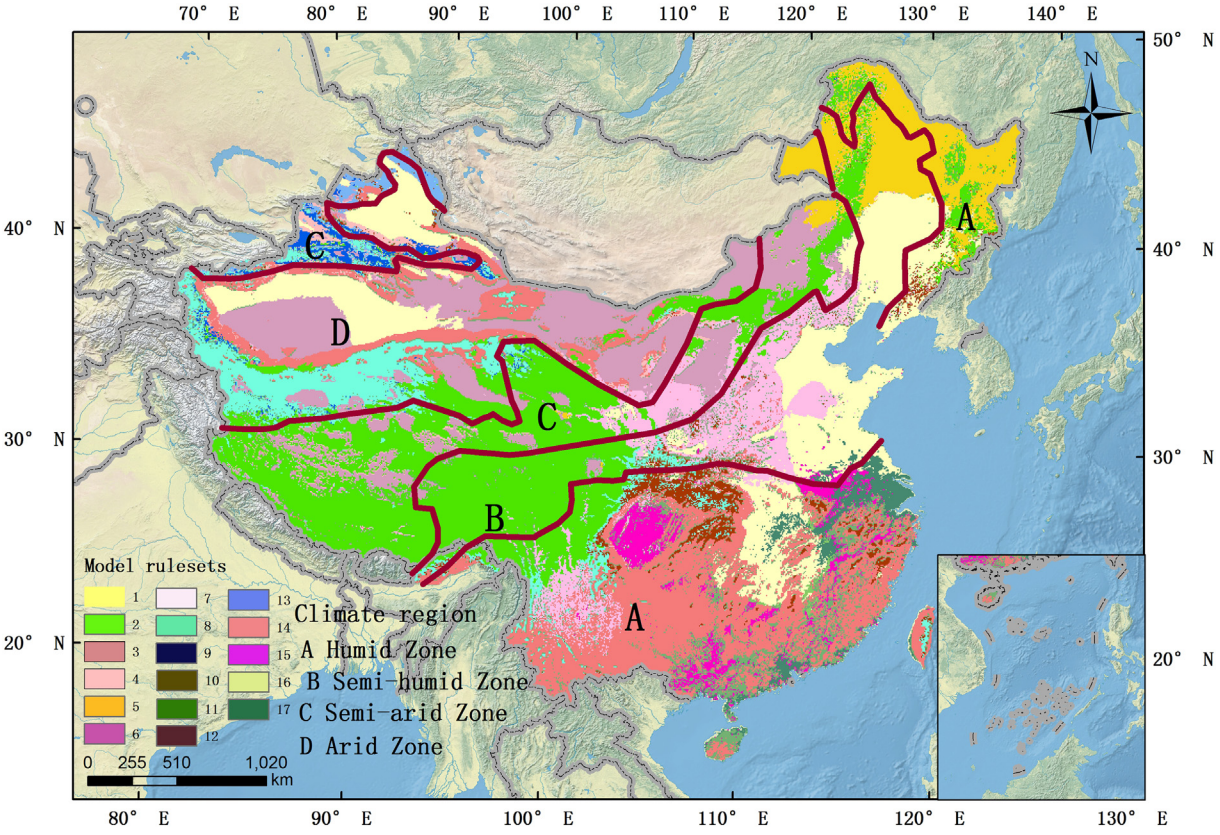

Fig. 6. Rule regions derived from the Cubist model of SOM.

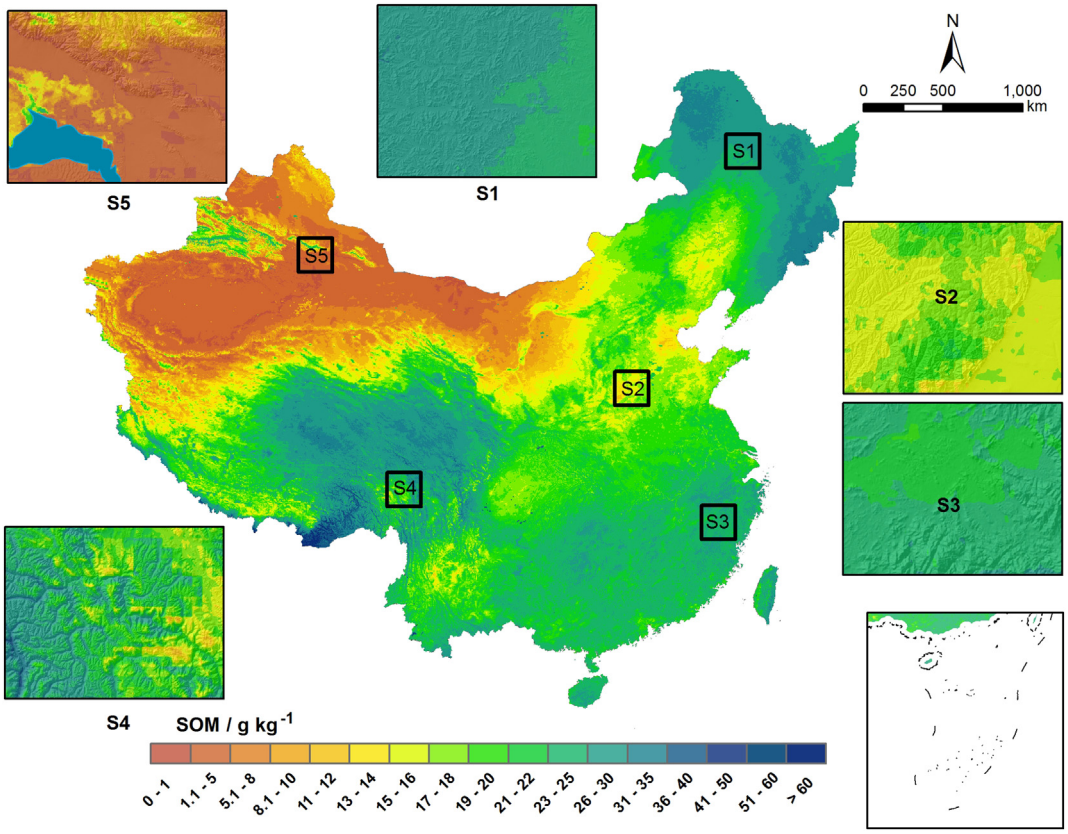

Fig. 7. Prediction map of SOM in topsoil (0–20 cm depth).

areas in northwest China, which are characterized by relatively high temperature and poor vegetation cover. SOM content gradually decreases from southeast to northwest in Tibet and Sichuan, which is consistent with the transition from humid, to sub-humid, semi-arid, and arid climates. The phenomenon indicates that the pattern of SOM content results from the distinct hydrothermal condition of the Tibet region. The Songliao and Sanjiang plains in northeast China are one of the world's three major areas of black soil (chernozem). The cold and humid climate contributes to the rich SOM content of black soil. Under the dense forest cover in the zone of high SOM in Tibet and northeast China, SOM is promoted by abundant standing litter, while the cold environment means that the organic matter decomposition rate is low, which contributes to SOM accumulation. According to the formula of carbon density, we know that both bulk density and SOM are the key attributes for calculating soil organic carbon storage, and the bulk density values generally have a significant correlation with the SOM. However, in most of the soil databases, the bulk density values are extremely scarce, so the map of SOM content plays a very important role in the later calculation of soil organic carbon stocks.

The spatial distribution of SOM in our study is largely consistent with the distribution of SOCD as estimated by Xie (2004) and Yu et al. (2005) using statistical methods. Our results are also broadly consistent with Q.Q. Li et al. (2013) estimate of SOC in topsoil in China using an artificial neural network, and the Harmonized World Soil Database (HWSD) of the Food and Agriculture Organization of the United Nations. Results of the latter two studies show the same patterns of high and low SOM in Eastern Tibet, northern Heilongjiang, northeastern Inner Mongolia, and a small area of the Tianshan Mountains. Our estimates of SOM in China using the Cubist algorithm fall in between the other two studies, with those of Q.Q. Li et al. (2013) higher and the HWSD lower. Two previous national maps of SOC in China, produced by Q.Q. Li et al. (2013) and Shangguan et al. (2013), show a low spatial resolution of 1 km.

3.4. Verification of model accuracy

The original SOM data presented a skewed distribution, therefore, before construction of the Cubist model, the data was log-transformed to satisfy the normal distribution. The statistics of model accuracy in terms of r, RMSE, ME and SDE were shown in Fig. 8. China has a larger land area and greater spatial climatic and topography variability, including the Tibet Plateau which has been termed the 'roof of the world'. China also has more latitudinal variation in climate. Different soil

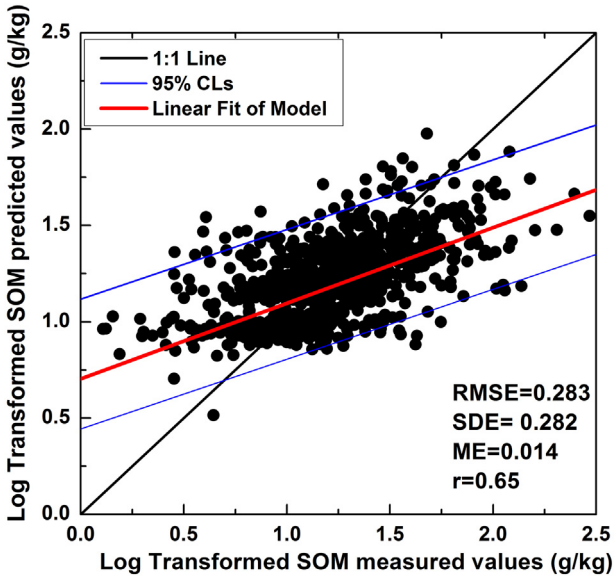

Fig. 8. Scatter plot of predicted against measured SOM values. Key to terms: CLs, confidence intervals; RMSE, root mean squared error.

properties analysis methods and the lack of standardized sampling methods are sources of error. These sophisticated factors make it difficult to build a very accurate prediction model of SOM in China. With the limited data available, our estimates suggest the model has a level of predictability that has a practical significance for large scale digital soil mapping.

3.5. Spatial distribution of uncertainty

The analysis of accuracy and quality control of spatial data are frontline problems in the earth sciences, but research on uncertainty in mapping is relatively rare in China. The uncertainty problem is unavoidable and widespread in the process of digital soil mapping and decision-making based on spatial data. Analyzing and evaluating uncertainty helps data users to understand its existence, and it also can help to improve decision quality and, thus, improve the accuracy and reliability of the decision. Our study is the first to assess the spatial distribution of uncertainty in the estimation of SOM at the national scale in China, and, as such, it is expected to provide a baseline reference for future estimates of the spatial distribution of soil critical attributes.

The uncertainty of our estimates using Cubist-kriging method is calculated by adding the bootstrap variance of the Cubist regression predictions and the average kriging variances of the residuals. We do not claim that the method we used is optimal, and the kriging with external drift (KED) approach (Viscarra Rossel et al., 2016) could directly provide the estimation of the prediction uncertainty from both of regression prediction and residuals. This method has been successfully applied in digital mapping and uncertainty assessment for soil organic carbon at regional scale (Viscarra Rossel et al., 2016). However, we could not use the KED method in our study across China because it is computationally difficult to map at the scale of China. Thus, we preferred the Cubist-kriging approach for our study as we could use more easily-developed parallel computing framework for the mapping.

As the spatial variability in SOM at a large scale is intense, we quantified the spatial distribution of uncertainty in top soil (Fig. 9). The uncertainties are large in Xinjiang and the desert areas of Inner Mongolia, the Altyn Tagh Mountains, and Qilian mountain range where data were lacking or environmental prediction conditions were poor. In future national carbon accounting, these unpopulated regions of desert and rangeland need to be sampled more densely to improve the certainty of predictions. This would allow the spatial distribution and reserve of soil organic matter in China to be objectively and comprehensively evaluated. All soil profiles used in this study derived from the second national soil survey of China. Most of relevant research in China used same dataset with us, which is public available and most exhaustive one. However, it has its limitation that the error occurred in matching sampling location information and the actual spatial coordinate, which is also the source of uncertainty.

3.6. SOM under different land uses and soil types

The magnitude and distribution of SOM content through the profile appears to vary with land use (Fig. 1); this is probably related to vegetation roots that regulate SOM distribution through the soil profile, similar findings have been reported in previous studies (M. Li et al., 2013). In the three soil profiles we randomly selected, SOM of grassland sharply decreased at a depth of 10 cm, where there is also a significant decrease in roots. In contrast, SOM in the woodland profile decreased only slightly until a depth of 50 cm as forest roots are abundant at greater depths. In the farmland (paddy field) profile, SOM decreased below the plowing layer, around 20 cm depth, which is the consequence of long term tillage and natural soil formation.

The China land use map, which was interpolated from Landsat TM imagery from 1990 by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn), was used as the land use reference in our study. The map consists of seven major land cover types: paddy field, dry field, forest, grassland, construction area, vacant land, and water body. Table 2 lists the estimated SOM for the different land uses (excluding water body). The highest average SOM is shown by forest (24 g/kg), followed by paddy field, dry field and grassland, with vacant land having the lowest SOM (< 10 g/kg). Vacant land also shows the highest coefficient of variance (90.37%), which was much higher than other land uses. According to the *China Soil Fertility (1998)* report, the average SOM for paddy field and dry field is 25.6 and 18.5 g/kg, and nearly 60% of paddy field and dry field had a SON content of 15–25 g/kg and 8–15 g/kg, respectively. Our predictions are consistent with the reported data ranges.

We constructed a map of SOM in topsoil (0–20 cm) in China at 1-km resolution using a depth weighted algorithm based on three SOM maps from SoilGrids (1-km resolution) (www.soilgrids.org), for 0–5, 5–15, and 15–30 cm depth, respectively. We extracted SOM statistics from our results and the SoilGrids product based on the China soil type map, which is available from RESDC (http://www.resdc.cn). Using the report of *China Soil Fertility* as a reference (Shen, 1998), we compared the prediction accuracy of our results and SoilGrids (Fig. 10). Our product using Cubist more similar to the China Soil Fertility than SoilGrids: SoilGrids overestimates SOM in most soil types while our results are closer to the data in the *China Soil Fertility* report. The differences between SoilGrids and our results are mainly due to the data source and modeling approach used. SoilGrids uses a global model for soil data, while our Cubist model is more suitable for mapping at national scale. As the series of covariates used in the Cubist model were obtained at 90 m resolution, they provide more detail about soil spatial variance, especially for those regions with large heterogeneity. Consequently, our produced predictions of SOM would be more accurate than SoilGrids in China national scale.

4. Conclusions

Our results show that the Cubist model can analyze the relationships between SOM and environmental covariates under different landscapes and complex topography at a national scale. The approach allows efficient calibration of models for each sub-region based on characteristics of environmental covariates. Thus, we can confirm the effectiveness of Cubist for high-resolution digital soil mapping at a large scale.

The spatial distribution of SOM is controlled by the collective effect of the environmental covariates, and therefore, theoretically, all available environmental covariates should be added into the model for more accurate prediction. However, we found that using 19 environmental covariates in modeling did not improve prediction accuracy, and resulted in higher computing time, than using 15 environmental covariates. Therefore, when prediction accuracy is comparable, use of a simplified subset of environmental covariates is preferable for producing large scale, high-resolution soil maps.

Our results also indicate that large spatial heterogeneity, lack of representative soil samples, and shortcomings of available soil samples affect the prediction accuracy and uncertainty of digital soil mapping. A more comprehensive sample distribution and sampling strategy would improve model accuracy, and the uncertainty map provides a guide for decision-making in additional sampling. In the current model, we did not take artificial interventions, such as tillage system, and population density into consideration, but it would be worthwhile to add these factors into future modeling research.

The 1980s Chinese national SOM baseline map provides a reference for monitoring and evaluating how land cover change, soil management practices, and climate change affect the distribution of SOM. Our 90-m grid high-resolution SOM map offers more detail, especially for those areas with high spatial heterogeneity. The map could contribute to national decision-making on agriculture and related studies, provide data support for research on terrestrial carbon circulation in China and

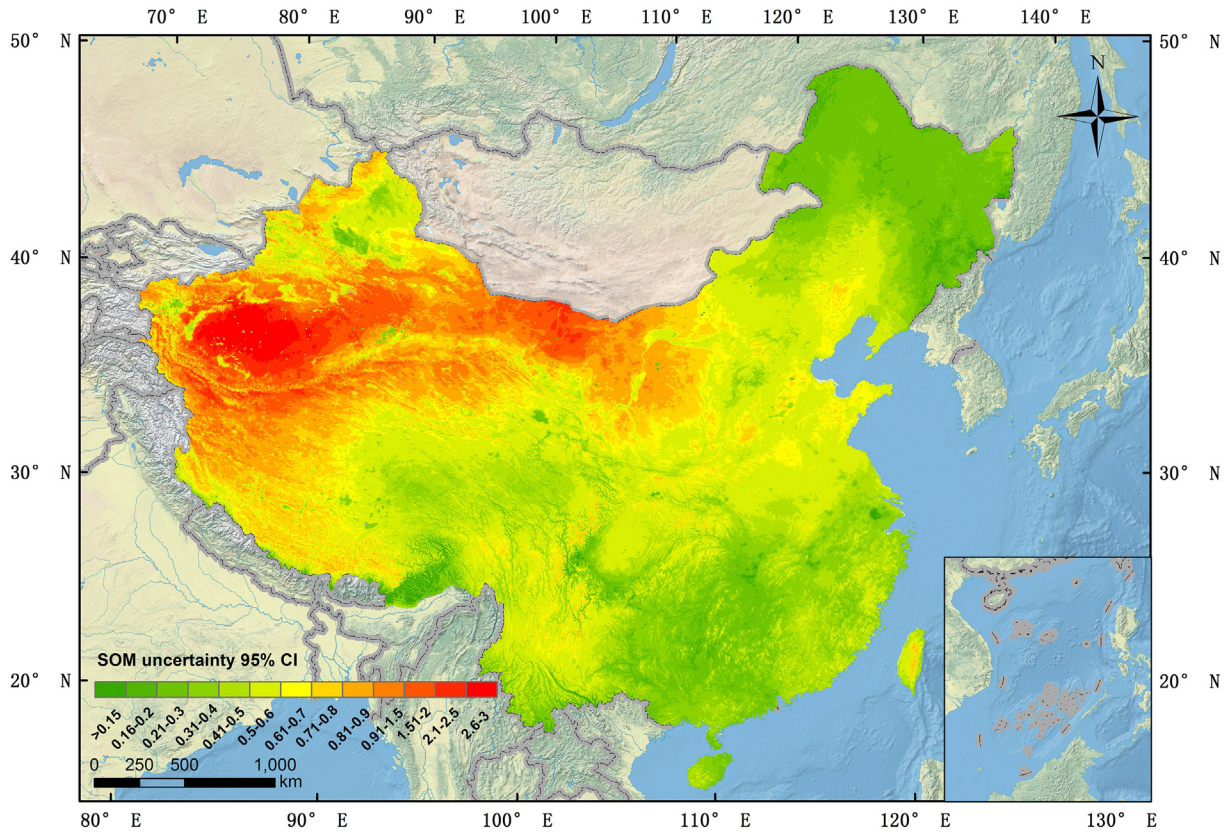

Fig. 9. Uncertainty of SOM expressed as the range of the 95% confidence intervals (CI) divided by their mean.

Table 2
Statistical characteristics of soil organic matter (SOM) under different land uses.

Type	Mean (g/kg)	SD[a]	CV[b]	Skewness	Kurtosis	95% upper	95% lower
Paddy	21.38	2.5632	11.99%	0.339	2.505	21.45	21.30
Dry field	18.98	4.4584	23.49%	−0.56	1.224	19.06	18.90
Forest	23.97	4.9413	20.61%	0.072	2.13	24.03	23.90
Grassland	18.02	7.4719	41.46%	−0.295	−0.403	18.10	17.93
Construction area	18.65	3.7301	20.00%	−0.314	1.182	18.84	18.47
Vacant land	9.29	8.3950	90.37%	−0.205	0.035	9.41	9.18

[a] SD standard deviation.
[b] CV coefficient of variance.

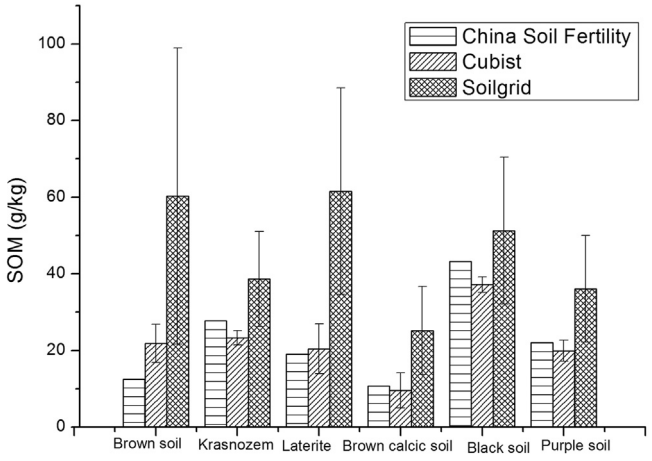

Fig. 10. Comparison of SOM predictions for different soil types using the Cubist and SoilGrids models with data from Soil Fertility in China.

global carbon stock estimation, and make a contribution to global change research and global soil mapping.

Acknowledgement

This study was supported by the National Key Research and Development Program (2017YFD0700501), and the Research Fund of State Key Laboratory of Soil and Sustainable Agriculture, Nanjing Institute of Soil Science, Chinese Academy of Sciences (No. Y412201430). We appreciate the editor and anonymous reviewers for their valuable comments to improve this manuscript.

References

Adhikari, K., Hartemink, A.E., Minasny, B., Kheir, R.B., Greve, M.B., Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS One 9 (8), e105519.
Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., 2014. Chapter Three-GlobalSoilMap: toward a fine-resolution global grid of soil properties. Adv. Agron. 125, 93–134.
Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45.

Bui, E., Henderson, B., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Glob. Biogeochem. Cycles 23, GB4033.
Buol, S.W., Southard, R.J., Graham, R.C., McDaniel, P.A., 2011. Soil Genesis and Classification. (J.).
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.
Chien, Y.J., Lee, D.Y., Guo, H.Y., Houng, K.H., 1997. Geostatistical analysis of soil properties of mid-west Taiwan soils. Soil Sci. 162 (4), 291–298.
Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. London, Chapman and Hall.
Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207.
Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124 (3), 383–398.
Huang, J., Minasny, B., McBratney, A.B., Padarian, J., Triantafilis, J., 2018. The location- and scale-specific correlation between temperature and soil carbon sequestration across the globe. Sci. Total Environ. 615, 540–548.
Huffman, G.J., Adler, R.F., Bolvin, D.T., Gu, G.J., Nelkin, E., Bowman, K.P., Hong, Y., Stocker, E.F., Wolff, D.B., 2007. The TRMM multisatellite precipitation analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. Hydrometeorology 8, 38–55.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York, pp. 1–270.
Justice, C.O., Vermote, E., Townshend, J.R., 1998. The moderate resolution imaging spectroradiometer (MODIS): land remote sensing for global change research. IEEE Trans. Geosci. Remote Sens. 36, 1228–1249.
Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2014. C Code for Cubist by Ross Quinlan. Cubist: Rule- and Instance-Based Regression Modeling. R Package Version 0.0.18.
Lagacherie, P., McBratney, A.B., 2007. Spatial soil information systems and spatial soil inference systems: perspectives for Digital Soil Mapping. Dev. Soil Sci. 31, 3–22.
Li, Q.Q., Yue, T.X., Fan, Z.M., 2010. Study on method for spatial simulation of topsoil SOM at national scale in China. J. Nat. Res. 25, 1385–1399.
Li, Q.Q., Yue, T.X., Wang, C.Q., Zhang, W.J., Yu, Y., Li, B., Bai, G.C., 2013a. Spatially distributed modeling of soil organic matter across China: an application of artificial neural network approach. Catena 104, 210–218.
Li, M., Zhang, X., Pang, G., Han, F., 2013b. The estimation of soil organic carbon distribution and storage in a small catchment area of the loess plateau. Catena 101 (2), 11–16.
Ma, Z., Shi, Z., Zhou, Y., Xu, J., Yu, W., Yang, Y., 2017. A spatial data mining algorithm for downscaling tmpa 3b43 v7 data over the Qinghai-Tibet plateau with the effects of systematic anomalies removed. Remote Sens. Environ. 200.
Malone, B.P., Mcbratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152.
Manlay, R.J., Feller, C., Swift, M.J., 2007. Historical evolution of soil organic matter concepts and their relationships with the fertility and sustainability of cropping systems. Agric. Ecosyst. Environ. 119 (3), 217–233.
Mcbratney, A.B., Mendinca, S.M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
National Soil Survey Office, 1993. Chinese Soil Genus Records. vol. 1 China Agriculture Press, Beijing (in Chinese).
National Soil Survey Office, 1994a. Chinese Soil Genus Records. vol. 2 China Agriculture Press, Beijing (in Chinese).
National Soil Survey Office, 1994b. Chinese Soil Genus Records. vol. 3 China Agriculture Press, Beijing (in Chinese).
National Soil Survey Office, 1995a. Chinese Soil Genus Records. vol. 4 China Agriculture Press, Beijing (in Chinese).
National Soil Survey Office, 1995b. Chinese Soil Genus Records. vol. 5 China Agriculture Press, Beijing (in Chinese).
National Soil Survey Office, 1996. Chinese Soil Genus Records. vol. 6 China Agriculture Press, Beijing (in Chinese).
National Soil Survey Office, 1998. Soils of China. China Agricultural Press, Beijing (in Chinese).
Pan, G.X., 1999. Study on carbon reservoir in soils of China. Bull. Sci. Technol. 15 (5), 330–332.
Pan, G., Li, L., Wu, L., Zhang, X., 2004. Storage and sequestration potential of topsoil organic carbon in china's paddy soils. Glob. Chang. Biol. 10 (1), 79–92.
Ponce-Hernandez, R., Marriott, F.H.C., Beckett, P.H.T., 1986. An improved method for reconstructing a soil-profile from analysis of a small number of samples. J. Soil Sci. 37, 455–467.
Post, W.M., Kwon, K.C., 2000. Soil carbon sequestration and land use change: processes and potential. Glob. Chang. Biol. 6 (3), 317–327.
Quinlan, J.R., 1992. Learning with continuous classes. In: Adams, S. (Ed.), Proceedings of the 5th Australian Joint Conference on Artificial Intelligence. World Scientific, Singapore, pp. 343–348.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria.
Raich, J.W., Potter, C.S., 1995. Global patterns of carbon dioxide emissions from soils. Glob. Biogeochem. Cycles 9 (1), 23–36.
Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27, 171–197.
Shangguan, W., Dai, Y.J., Liu, B.Y., Zhu, A.X., Duan, Q.Y., Wu, L.Z., Ji, D.Y., Ye, A.Z., Yuan, H., Zhang, Q., Chen, D.D., Chen, M., Chu, J.T., Duo, Y.J., Guo, J.X., Li, H.Q., Li, J.J., Liang, L., Liang, X., Liu, H.P., Liu, S.Y., Miao, C.Y., Zhang, Y.Z., 2013. A China data set of soil properties for land surface modeling. J. Adv. Model. Earth Syst. 5, 212–224.
Shen, S.M., 1998. China Soil Fertility. China Agriculture Press, Beijing (in Chinese).
Sun, X.L., Zhao, Y.G., Wu, Y.J., 2012. Spatio-temporal change of soil organic matter content of Jiangsu Province, China, based on digital soil maps. Soil Use Manag. 28 (3), 318–328.
Viscarra Rossel, R.A., 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. J. Geophys. Res. F Earth Surf. 116, F04023.
Viscarra Rossel, R.A., Bui, E.N., 2016. A new detailed map of total phosphorus stocks in Australian soil. Sci. Total Environ. 542, 1040–1049.
Viscarra Rossel, R.A., Chen, C., 2011. Digitally mapping the information content of visible-near infrared spectra of surficial Australian soils. Remote Sens. Environ. 115, 1443–1455.
Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63 (6).
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).
Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res. 53 (8), 845–864.
Viscarra Rossel, R.A., Brus, D.J., Lobsey, C., Shi, Z., McLachlan, G., 2016. Baseline estimates of soil organic carbon by proximal sensing: comparing design-based, model-assisted and model-based inference. Geoderma 265.
Wang, S.Q., Zhou, C.H., Li, K.R., 2000. Analysis on spatial distribution characteristics of soil organic carbon reservoir in China. Acta Geograph. Sin. 55, 533–544.
Wu, H.B., Guo, Z.T., Peng, C.H., 2003. Distribution and storage of soil organic carbon in China. Glob. Biogeochem. Cycles 17, 67–80.
Xie, X., 2004. Organic carbon density and storage in soils of China and spatial analysis. Acta Pedol. Sin. 41 (1), 35–43.
Yu, D., Shi, X., Sun, W., Wang, H., Liu, Q., Zhao, Y., 2005. Estimation of China soil organic carbon storage and density based on 1: 1,000,000 soil database. J. Appl. Ecol. 16 (12), 2279–2283.
Zhou, X., Gao, S., 1997. Confidence intervals for the log-normal mean. Stat. Med. 16 (7), 783–790.
Zhou, Y., Biswas, A., Ma, Z., Lu, Y., Chen, Q., Shi, Z., 2016. Revealing the scale-specific controls of soil organic matter at large scale in Northeast and North China Plain. Geoderma 271, 71–79.