Chin. Geogra. Sci. 2017 Vol. 27 No. 4 pp. 516–528
doi: 10.1007/s11769-017-0869-7

Springer Science Press
www.springerlink.com/content/1002-0063

Mapping Soil Organic Carbon Stocks of Northeastern China Using Expert Knowledge and GIS-based Methods

SONG Xiaodong, LIU Feng, JU Bing, ZHI Junjun, LI Decheng, ZHAO Yuguo, ZHANG Ganlin

(*State Key Laboratory of Soil and Sustainable Agriculture*, *Institute of Soil Science*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China*)

Abstract: The main aim of this paper was to calculate soil organic carbon stock (SOCS) with consideration of the pedogenetic horizons using expert knowledge and GIS-based methods in northeastern China. A novel prediction process was presented and was referred to as model-then-calculate with respect to the variable thicknesses of soil horizons (MCV). The model-then-calculate with fixed-thickness (MCF), soil profile statistics (SPS), pedological professional knowledge-based (PKB) and vegetation type-based (Veg) methods were carried out for comparison. With respect to the similar pedological information, nine common layers from topsoil to bedrock were grouped in the MCV. Validation results suggested that the MCV method generated better performance than the other methods considered. For the comparison of polygon based approaches, the Veg method generated better accuracy than both SPS and PKB, as limited soil data were incorporated. Additional prediction of the pedogenetic horizons within MCV benefitted the regional SOCS estimation and provided information for future soil classification and understanding of soil functions. The intermediate product, that is, horizon thickness maps were fluctuant enough and reflected many details in space. The linear mixed model indicated that mean annual air temperature (MAAT) was the most important predictor for the SOCS simulation. The minimal residual of the linear mixed models was achieved in the vegetation type-based model, whereas the maximal residual was fitted in the soil type-based model. About 95% of SOCS could be found in Argosols, Cambosols and Isohumosols. The largest SOCS was found in the croplands with vegetation of *Triticum aestivum* L., *Sorghum bicolor* (L.) Moench, *Glycine max* (L.) Merr., *Zea mays* L. and *Setaria italica* (L.) P. Beauv.

Keywords: soil organic carbon stock; model-then-calculate; random forest; linear mixed model; northeastern China

Citation: Song Xiaodong, Liu Feng, Ju Bing, Zhi Junjun, Li Decheng, Zhao Yuguo, Zhang Ganlin, 2017. Mapping soil organic carbon stocks of northeastern China using expert knowledge and GIS-based methods. *Chinese Geographical Science*, 27(4): 516–528. doi: 10.1007/s11769-017-0869-7

1 Introduction

Small changes in soil organic carbon stocks (SOCS) might have considerable potential to increase atmospheric carbon dioxide (CO_2) concentrations and then influence the global climate. SOCS can be influenced by various abiotic and biotic environmental variables, such as topography, soil types and vegetation types (Jenny, 1941), and hence it varies spatially because of the changing environmental parameters (Kosmas *et al*., 2000). Consequently, the main challenge in quantifying the temporal change in soil C pools at regional scales is how to accurately predict the spatial pattern of SOCS.

Numerous methods for prediction of SOCS have been presented focusing on different pedological and statistical assumptions. There are two kinds of SOCS mapping techniques concerning the pedotransfer rules and pedogenic information, such as soil types, land cover and

Received date: 2016-07-21; accepted date: 2016-11-05
Foundation item: Under the auspices of Basic Project of State Commission of Science Technology of China (No. 2008FY110600), National Natural Science Foundation of China (No. 91325301, 41401237, 41571212, 41371224), Field Frontier Program of Institute of Soil Science, Chinese Academy of Sciences (No. ISSASIP1624)
Corresponding author: ZHANG Ganlin. E-mail: glzhang@issas.ac.cn

other easily accessible environmental variables. One method is based on the assumption that soil variation is homogeneous within soil types or landscape units, such as the professional knowledge-based method (PKB) (Zhang *et al*., 2008), which depends on the SOC density of soil type maps as well as the land unit method (Ottoy *et al*., 2015). Because of the large variability in soil, the landscape-based approaches would be efficient only if detailed soil maps and soil profiles were collected. The other method is a regular grid-based method using discretized real-world soil variation, in which the value of each cell represents the soil properties. This spatial interpolation paradigm can involve different spatial prediction methods, such as multivariate regression models, data mining and machine learning techniques. This method is usually used in practice, as the soil maps in terms of raster grids can reflect the continuous variation in detail and are easily integrated with other terrestrial models.

From the pedogenesis and soil classification perspective, a detailed soil survey should be conducted regarding the soil genetic horizons because the chemical and physical attributes might differ from the layers above and below the horizon of interest (Parras-Alcántara *et al*., 2015). This point has been widely accepted by most pedologists (Ahrens *et al*., 2003). Most soil taxonomy criteria involve soil morphology and laboratory tests to group similar soil pedons. As the horizons are generally parallel to the soil crust, the vertical sequence of soil could well be represented by several horizons. However, soil classification based on the diagnostic horizons could mask the soil continuum, particularly on the border of soil type polygons (Ließ *et al*., 2012). While spatial soil information is difficult to obtain because of the variable thicknesses of soil horizons, several spline algorithms have been proposed to fit a curve for a soil profile, so that a mean soil property value can be derived for the same depth increment. Emerging evidence has indicated that SOCS could be better computed by pedogenetic horizon-based methods than fixed-thickness approaches at the plot scale (Parras-Alcántara *et al*., 2015). Several attempts have been made to predict soil horizons (Chaplot *et al*., 2010; Ließ *et al*., 2012). Little attention has been paid to the quantitative comparison of SOCS estimation based on the entire soil profile considering pedogenetic horizons and fixed-thicknesses with preset depth increments.

Whether we need the depth function for each location, which might have a similar vertical distribution pattern within one soil map unit, depends on the heterogeneity of the soil-scape resulting from various anthropogenic and natural soil-forming processes. The discrete calculation of SOCS based on soil horizons can be easily implemented, when compared with sophisticated 3D modeling techniques (Kempen *et al*., 2011; Liu *et al*., 2013; Lacoste *et al*., 2014) and time-consuming prediction with small intervals (Aitkenhead and Coull, 2016). In general, the model-then-calculate (MC) method has been widely used. Here, the ′model′ represents the spatial prediction of soil properties for different layers, i.e., soil organic carbon (SOC) concentration, bulk density (BD) and the proportion of rock fragments (Dorji *et al*., 2014). The ′calculate′ indicates the calculation of SOCS at each point by summing the SOC density of individual soil layers (Were *et al*., 2015). To date, only the fixed-thickness method (MCF) with the same depth interval for each layer has been widely used within this framework.

The objectives of this study were: i) to design a feasible SOCS estimation workflow based on pedogenetic horizons with variable thicknesses within the model-then-calculate framework (MCV); ii) to evaluate the mapping performance of MCF, MCV and other GIS-based methods under the assumption of homogeneous soil variation, including soil profile statistics (SPS), pedological professional knowledge-based (PKB) and vegetation type-based (Veg) methods (Zhi *et al*., 2014); and iii) to investigate the random effects associated with categorical predictors.

2 Materials and Methods

2.1 Study area and soil sampling

The study area is located in the northeastern part of China and consists of Liaoning, Jilin and Heilongjiang provinces from south to north, extending from 38°43′N to 53°33′N and 118°50′E to 135°05′E (Fig. 1). The landscape is composed of plains and mountains. The altitude of this area varies between 0 and 2615 m a.s.l. The study area is controlled by a temperate, semi-humid continental monsoon climate. The annual precipitation ranges from 372 to 1504 mm, and the mean annual temperature ranges from −8℃ to 11℃. The study area covers about 8.08×10^5 km^2, 43.42% and 38.06% of

which are forests and croplands, respectively.

The field survey of Chinese soil series was conducted from July 2009 to September 2012; 453 locations were sampled, 18 of which (red rectangles in Fig. 1) were based on transect sampling and 435 of which were based on stratified random sampling (de Gruijter *et al*., 2006) considering the pedogenic predictors (i.e., geology, land cover and soil type). Soil profiles were dug to the depth of 1 m or to bedrock; composite samples were collected for each soil horizon.

The core ring sampler was used for bulk density (BD) determination. The SOC concentration and BD were analyzed by using the Walkley-Black wet oxidation method (Nelson and Sommers, 1982) and the core method (Blake, 1965), respectively. The SOC stock (SOCS, kg/m^2) was calculated by summing the SOC density of different soil layers:

$$SOCS = \sum_{i=1}^{n} SOC_i \times BD_i \times \left(1 - \frac{Gr_i}{100}\right) \times T_i \quad (1)$$

where n is the number of soil horizons, and SOC_i, BD_i, Gr_i and T_i denote the SOC concentration (g/kg), BD (g/cm^3), volume percentage of gravel (%) and thickness (m) of horizon i, respectively. A legacy soil texture map was used for the volume percentage of gravel. A log-transformation was carried out if skewness of the targeted variable was higher than 1 according to the rule of thumb presented in Webster and Oliver (2001).

2.2 Environmental variable

Nineteen environmental variables were chosen that were related to soil-forming factors (Dorji *et al*., 2014; Lacoste *et al*., 2014; Aitkenhead and Coull, 2016). Terrain attributes were derived from Shuttle Radar Topographical Mission Digital Elevation Model (SRTM DEM), including elevation, slope, aspect, solar radiation, plan curvature (PlaCur), profile curvature (ProCur), vertical distance to channel network (DistoChann), relative slope position (SlpPosit), modified topographic wetness index (TWI) (Boehner *et al*., 2002), multi-resolution index of valley bottom flatness (MRVBF) and multi-resolution index of the ridge top flatness (MRRTF) (Gallant and Dowling, 2003). The cosine function was performed to transform aspect to the range from −1 to 1

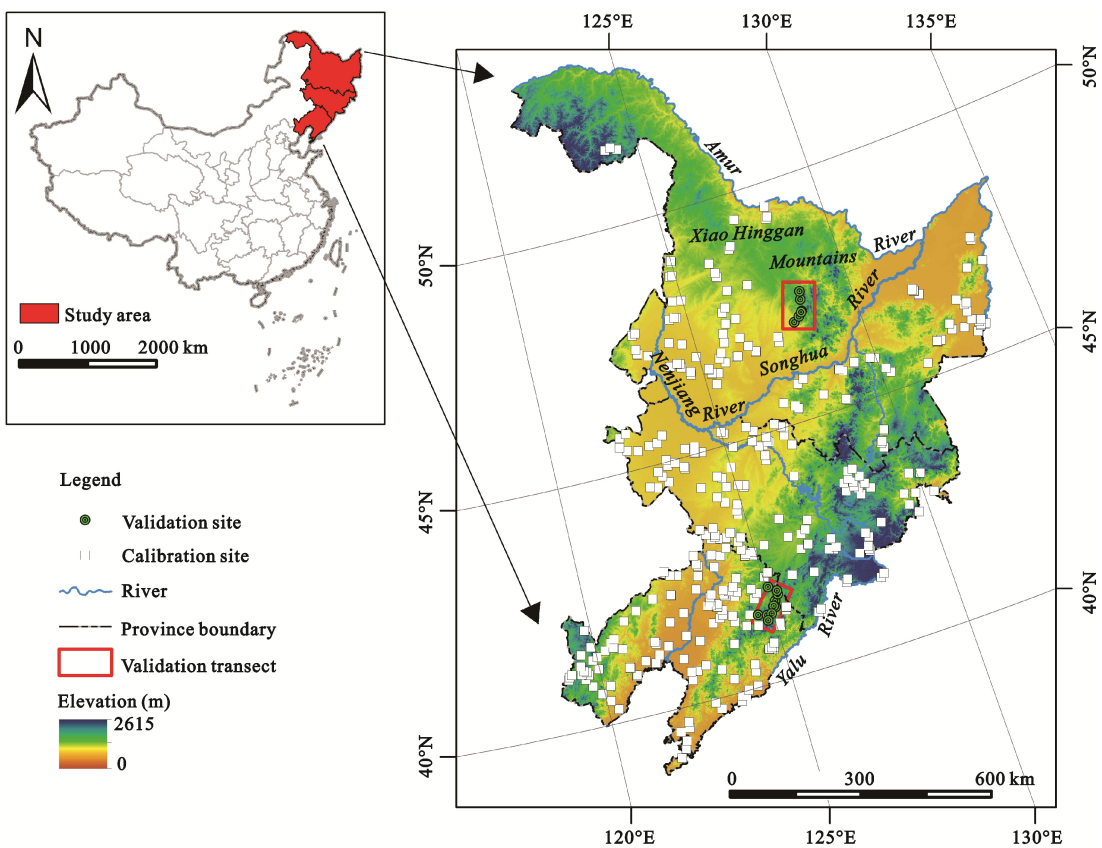

Fig. 1 Location of the study area and sampling sites draped over a hillshade

indicating the degrees of north (Cos(Aspect) for short). Topographic wetness index (TWI) was derived from the modified catchment area calculation algorithm considering the flow width.

Geological, land cover and vegetation maps were provided by the National Science & Technology Infrastructure of China, National Earth System Science Data Sharing Infrastructure (http://www.geodata.cn). Soil type maps based on Chinese Soil Taxonomy (CST) were obtained from the Institute of Soil Science Chinese Academy of Sciences. Climate data included mean annual air temperature (MAAT), mean annual precipitation (MAP) and accumulated annual air temperature higher than 0℃(AAT0) (CMA, 2011). The normalized difference vegetation index (NDVI) was calculated by Landsat 7 at 30-m spatial resolution in 2010 (http://landsat.gsfc.nasa.gov).

Predictors were resampled to 90-m spatial resolution and normalized. Stepwise regression based on the Akaike Information Criterion (AIC) was adopted to select the best predictor sets for different prediction models. Since the predictors used for each model (SOC and BD prediction for each depth) were different, we did not list them all. Interested readers can refer to Section 3.4 for frequently used predictors.

2.3 MCV method

Regarding the variable thicknesses of soil horizons, the model-then-calculate algorithm was improved for the SOCS estimation. Four main steps were as follows:

(1) The horizon thickness data were prepared. It was assumed that soil horizons vary continuously in space (Fig. 2). Based on the description of soil layers in the field using the Chinese Soil Taxonomy specification, the soil continuum was categorized for specific pedons (Table 1). Nine common layers from topsoil to bedrock were grouped according to similar pedological information obtained in this study. To model the spatial distribution of the soil continuum, the missing layer thickness of a site, that is, a potential layer to adjacent sites, was set to a very small value (0.00001), and related soil properties were set to the minimum observed values. However, inconsistent with the classification of horizons (Table 1), some complex horizons might be found within the same location, e.g., the cambic horizon (Bw) and the illuvic horizon (Bt1). In this case, the latter layer should be moved to the next category, i.e., Bt1 should be transferred to the B-II category. A detailed description of the soil horizons was referred to the Chinese Soil Taxonomy (Cooperative Research Group on Chinese Soil Taxonomy, 2001).

Table 1 Classification of soil horizons with similar hierarchy in vertical sequence

Layer	Soil horizons	Definition
A-I	A1, Ap1, Ap11, O1	A horizon coupled with plow disturbance, humic epipedons, and epipedons contain organic soil materials, *etc*.
A-II	2ABhr, 2AC, 2Ah, 2Ahb, AB1, AC1, AE, Ah1, Ap12, Ap2, Au, O2	A horizon, heterogeneous A horizon, organic epipedons, humic epipedons, anthorpic epipedons, hydragric horizon, cumulic epipedons, AC transition horizon, epipedons containing additional horizon characteristics or features, by suffixes p, h, b or u, and transition horizons containing dominantly one horizon characteristics but also containing another horizon attributes
A-III	2A, 2AB, 2ABhw1, 2Ah, 2Ahb, AB2, AC2, Ah2, O3	A horizon, heterogeneous A horizon, organic epipedons, humic epipedons, anthorpic epipedons, hydragric horizon, cumulic epipedons, AB transition horizon, and lower layers of epipedons which may contain additional features of B or C horizons
B-I	B1, 2Ah1, 2B1, 2Bt, 3Ahb, Bb, Be1, Bhn, Bk1, Bn, Br1, Bs1, Bt1, Bw1	B horizon, heterogeneous B horizon, cambic horizon, plinthic horizon, carbonate horizon, claying horizon, salic horizon, subsurface accumulation of clay, Fe, Al, Si, humus, $CaCO_3$, and so on. In this section, more transition attributes transferred from upper horizons can be observed
B-II	2B, 2Bt, 3AB, 2Ah2, 3Bt, B2, B2C, Ba, Bab, BC1, Be2, Bk2, Br2, Bs2, Bt2, Bw2	B horizon, cambic horizon, plinthic horizon, carbonate horizon, claying horizon, salic horizon, BC transition horizon, and obvious accumulation of clay, Fe, Al, Si, humus, $CaCO_3$, and so on
B-III	2BC, BC2, Bgr1, Bq1, Br3, Bt3, Bw3	B horizon, cambic horizon, plinthic horizon, carbonate horizon, claying horizon, salic horizon, BC transition horizon, and more attributes of transition horizons that will appear i.e., like BC
C-I	C1, Cg1, Cr1	C horizon, gleyic features, and hydragric horizon
C-II	2C, C2, Cg2, Cr2	Paralithic contact, C horizon, gleyic features, and hydragric horizon
R	R	Bedrock

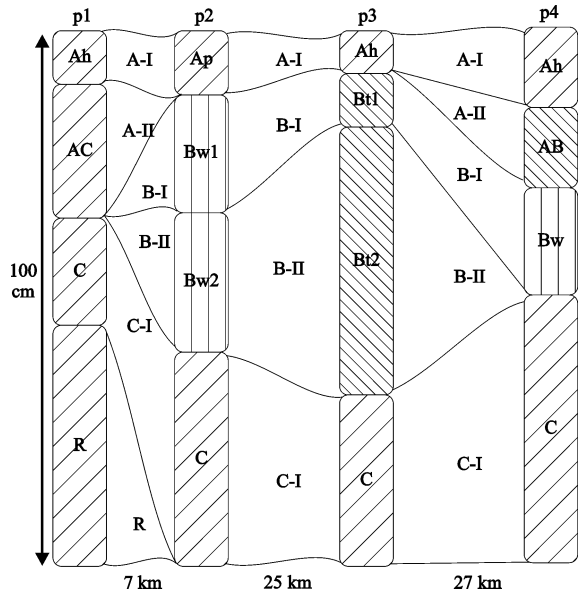

Fig. 2 Illustration of spatial variation of pedogenetic horizons with respect to soil continuum. Four soil profiles (p1, p2, p3 and p4) were used; the horizons were described according to the Chinese Soil Taxonomy

(2) The spatial distribution of horizon thicknesses and soil properties were predicted by the random forest (RF) approach. This is an up-to-date ensemble approach that can explore the relationships between predictors and the soil property of interest (Breiman *et al*., 1984). Note that the values of soil properties were field observed rather than modeled by the spline function used in the MCF method.

(3) The predicted results were back-transformed on the log-scale by exponentiation. The back-transformed prediction obtained was the median of the distribution, as a log-normal distribution for the prediction was assumed.

(4) The SOCS for each point was aggregated by summing the SOC density for each layer. Because uncertainty will be introduced by random errors and the prediction process, the total thickness might be different from 1 m. Hence a restricted coefficient (*RC*) of total horizon thickness was proposed as follows:

$$RC = 1 \Big/ \sum_{i=1}^{n} \hat{T}_i \tag{2}$$

where n is the number of soil horizons, and $\hat{T}_i$ denotes the predicted thickness (m) of horizon i. The value of *RC* for one soil is a constant. Then the SOCS calculation is then updated as follows:

$$SOCS = \sum_{i=1}^{n} SOC_i \times BD_i \times \left(1 - \frac{Gr_i}{100}\right) \times RC \times \hat{T}_i \tag{3}$$

A linear mixed model was used to investigate the fixed effects related to the categorical predictors (i.e., geology, land cover, vegetation and soil types) and the random effects related to the correlated variation of the properties (Bapat, 2012). The observed SOC density up to 100 cm depth was used as the targeted variable and eight predictors were selected. As all the predictors were scaled, the more extreme a regression coefficient of variables is, either positive or negative, the stronger it will be affected by the random effect.

2.4 Other estimation methods

MCF: The vertical distribution of SOC and BD was fitted by the equal-area quadratic splines (Bishop *et al*., 1999). SOC and BD were spatially modeled using fixed depth (Dorji *et al*., 2014). Six fixed depths were employed using GlobalSoilMap specifications (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm and 60–100 cm) (http://www.globalsoilmap.net).

Soil profile statistics (SPS) method: The SOC density was calculated for each sampling plot and then the mean value was set to the overlapping polygons of soil type and county boundary. The SOCS in each unit was calculated by multiplying the SOC density by area. This means that the SOCS of all soil types was the aggregation of all SOCS in different units (Zhi *et al*., 2014).

Pedological professional knowledge-based (PKB) method: Similar to the SPS method, mean SOC density within one unit was set to overlapping polygons of soil type, county boundary and soil parent materials (Zhi *et al*., 2014).

Vegetation type-based (Veg) method: Mean SOC density values were summarized within one vegetation type. The SOCS was calculated by multiplying the mean values by the corresponding areas, and then the SOCS within each polygon was aggregated as the total C stock.

2.5 Cross validation

Leave one out cross validation (LOOCV) and hold-out validation (2-fold cross validation) were performed using three indices: mean error (*ME*), root mean squared errors (*RMSE*) and Lin′s concordance correlation coefficient (ρ_c) (Lin, 1989). As the prediction procedures of the five methods were significantly disparate, LOOCV

was performed for MCF and MCV using 453 stratified sampling points (Section 2.2), and hold-out validation was conducted for all five methods using 18 transect sampling points. Lin's concordance correlation can be calculated as follows:

$$\rho_c = \frac{2S_{xy}}{S_x^2 + S_y^2 + (\bar{x} - \bar{y})^2} \quad (4)$$

where $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, $S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. Here $\bar{x}$ and $\bar{y}$ are the means for values x (observed) and y (estimated), respectively, n denotes the number of validation points, and x_i and y_i are the observed and estimated values at location i, respectively. Lin′s concordance correlation coefficient varies from –1 to 1, in which the complete agreement between all paired sites is represented by 1.

3 Results

3.1 Validation of SOCS

The validation and distribution of the total SOCS estimations are given in Table 2 and Fig. 3, respectively. Both the cross validation and hold-out validation indicated that MCV generally outperformed the other methods. The exception was the ρ_c of MCF (0.14) that was larger than for MCV based on cross validation. The smaller *RMSE* value of MCV compared with MCF indicated that MCV might produce a similar or better accuracy than that of MCF. For the comparison of polygon-based approaches, the Veg method was more accurate than both the SPS and PKB methods. It was suggested that the use of soil type-based information was not advantageous, when limited soil data were incorporated. However, it is noted that the mean errors of MCV illustrated an obvious underestimation of SOCS.

The difference between maps generated by the two assumptions was apparent (Figs. 3a and 3b versus 3c, 3d and 3e), where the maps generated with MCF and MCV showed more consistent spatial patterns than those generated with SPS, PKB and Veg. As expected, the two maps estimated by MCF and MCV presented a reasonable prediction with respect to the terrain topography, which in turn affected the land cover and climate (Fig. 3). The variation in SOCS generated by MCF was very similar to that of MCV, especially in the northern and eastern parts of the study area. When comparing the differences between the predictions (Fig. 3f), it was shown that 61% of SOCS values estimated by MCV were smaller than the estimates by MCF. About 88% of the absolute difference between SPS and PKB was smaller than 2 and 80% was with the same values (Fig. 3g).

Table 2 Performance comparison of soil organic carbon stock (SOCS) estimation based on cross validation and hold-out validation

Method	Validation	*ME* (kg/m^2)	*RMSE* (kg/m^2)	ρ_c
MCF	Cross validation	–4.62	17.28	0.14
MCV	Cross validation	–2.95	15.25	0.11
MCF	Independent validation	–5.43	17.80	0.09
MCV	Independent validation	–6.31	16.14	0.29
SPS	Independent validation	–0.49	23.14	–0.04
PKB	Independent validation	4.54	20.88	–0.01
Veg	Independent validation	–0.82	19.51	–0.13

Notes: MCF, model-then-calculate using fixed-thickness; MCV, model-then-calculate using variable-thicknesses; SPS, soil profile statistics; PKB, pedological professional knowledge-based; Veg, vegetation type-based method; *ME*, mean error; *RMSE*, root mean squared error; ρ_c, Lin's concordance correlation coefficient

3.2 Prediction of soil properties

Generally, the values of soil properties generated by the MCV method had the same or better accuracy than those generated by MCF (Table 3 and Table 4). The *ME* values of SOC and BD were moderately smaller than zero indicating an underestimation. It was suggested that the prediction of SOC and BD using the machine learning technique (i.e., random forest) was not always equivalent to those generated by an unbiased estimator (i.e., ordinary least square). The highest values of ρ_c for SOC were obtained with the MCV method (0.45), and MCF produced better values for BD than MCV (0.39 versus 0.37). The prediction performance of both MCF and MCV methods decreased with increasing soil depth.

Horizon thickness maps generated by random forest models were fluctuant and reflected many details in space (Fig. 4). Note that there was somewhat thickness randomness described during field survey. As can be seen from Fig. 4 and Table 3, the horizon thicknesses generated were moderately accurate, concurring with previous studies (Vanwalleghem *et al*., 2010; Crouvi *et al*., 2013). The complex spatial variation was ascribed to the distinctive geographical features of the study area, including high-relief mountains, different ecosystems and geographic locations. From a qualitative viewpoint,

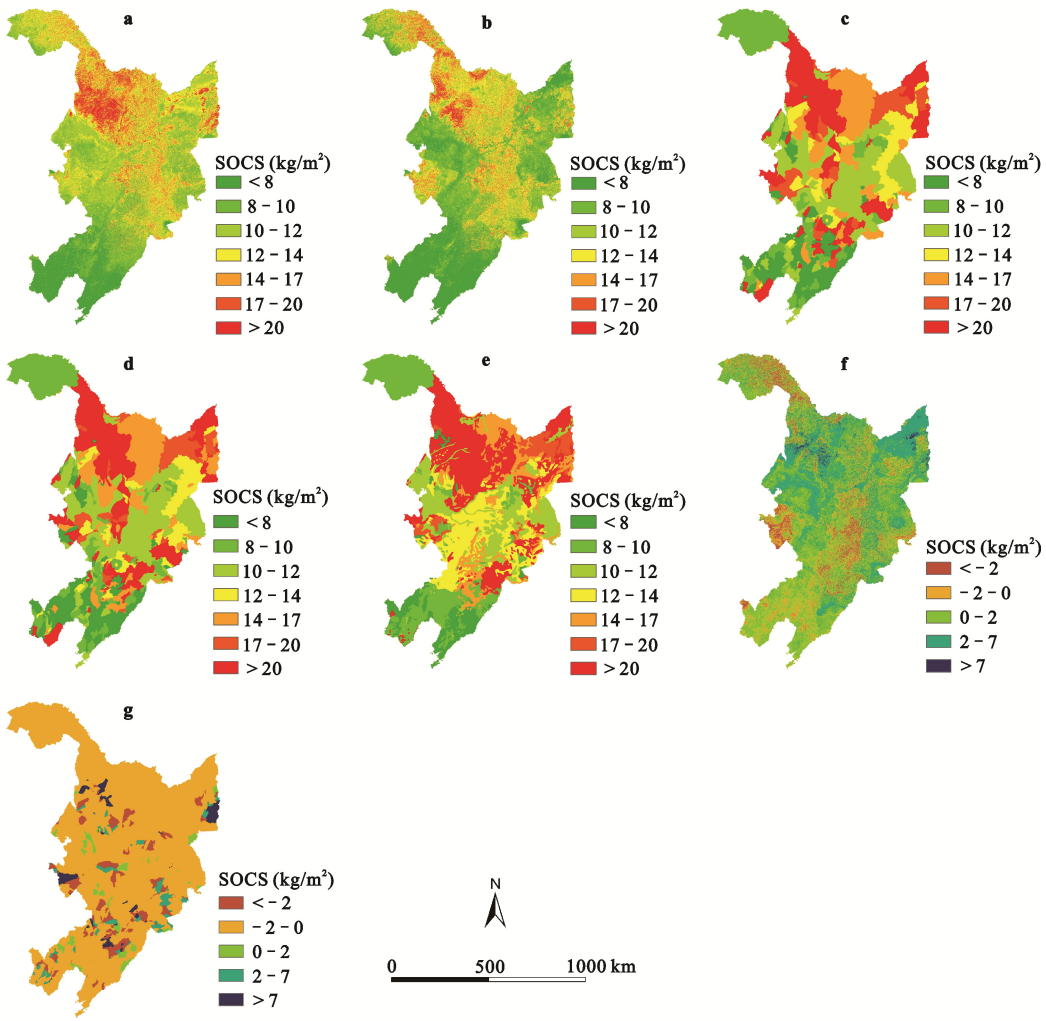

Fig. 3 Spatial variation in the soil organic carbon stock (SOCS) generated by MCF (a), MCV (b), SPS (c), PKB (d) and Veg (e), and the differences between predictions obtained with MCF and MCV (MCF–MCV) (f) and SPS and PKB (SPS–PKB) (g). MCF, model-then-calculate using fixed-thickness; MCV, model-then-calculate using variable-thicknesses; SPS, soil profile statistics; PKB, pedological professional knowledge-based; Veg, vegetation type-based method

Table 3 Cross validation for soil organic carbon (SOC) concentration and bulk density (BD) prediction using MCV method

Horizon	Thickness (m)			SOC (g/kg)			BD (g/cm^3)		
	ME	*RMSE*	ρ_c	*ME*	*RMSE*	ρ_c	*ME*	*RMSE*	ρ_c
A–I	0.0024	0.09	0.36	–6.110	30.73	0.45	–0.044	0.29	0.37
A–II	0.0018	0.14	0.16	–5.410	37.67	0.42	–0.036	0.27	0.31
A–III	0.0012	0.11	0.17	–5.310	26.18	0.12	–0.079	0.37	0.22
B–I	0.0017	0.15	0.18	–4.140	21.37	0.24	–0.260	0.69	0.17
B–II	–0.0012	0.16	0.09	–0.710	3.78	0.01	–0.088	0.45	0.11
B–III	0.0009	0.08	0.16	–0.035	0.51	0.14	–0.006	0.12	–0.15
C–I	0.0045	0.16	0.29	–0.051	0.91	0.09	0.350	0.26	0.24
C–II	0.0014	0.09	0.14	0.063	0.84	0.14	1.990	0.60	0.07
R	0.0027	0.11	0.29	–	–	–	–	–	–

Notes: the definition of horizons is given in Table 1

Table 4 Cross validation for SOC concentration and BD prediction using the MCF method

Depth (cm)	SOC (g/kg)			BD (g/cm^3)		
	ME	*RMSE*	ρ_c	*ME*	*RMSE*	ρ_c
0–5	–4.67	29.36	0.41	–0.0002	0.29	0.39
5–15	–5.99	31.61	0.38	–0.0016	0.31	0.25
15–30	–5.48	30.74	0.32	–0.0012	0.27	0.27
30–60	–5.55	41.73	0.016	–0.0062	0.26	0.37
60–100	–4.77	25.59	0.036	–0.0051	0.42	0.28

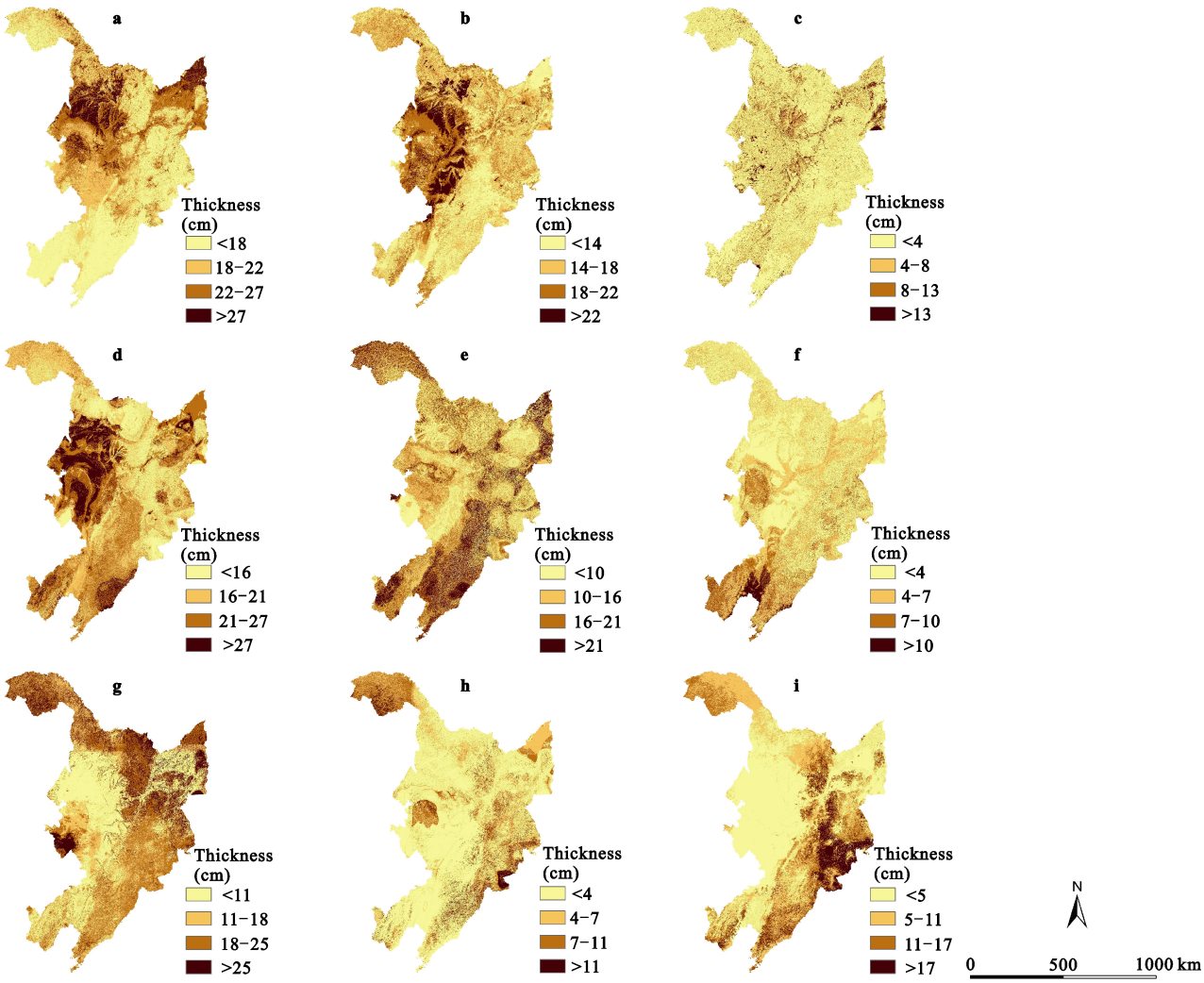

Fig. 4 Distribution of the predicted thickness of each pedogenetic horizon: A–I (a), A–II (b), A–III (c), B–I (d), B–II (e), B–III (f), C–I (g), C–II (h) and R (i)

the maps showed diverse macro-patterns even over the same soil type. The surface and subsurface horizons were highly variable on the plains, and the bedrock thicknesses were varied on the mountainous areas.

The spatial distribution of the restricted coefficient (*RC*) of each soil profile with respect to nine types of horizons is given in Fig. 5. The aforementioned *RC* can be used to represent the overall performance of horizons estimation, in which a value of 1 denotes complete agreement with the prediction over 1 m prediction. The *RC* map yielded a mean of 0.96 and standard deviation of 0.10.

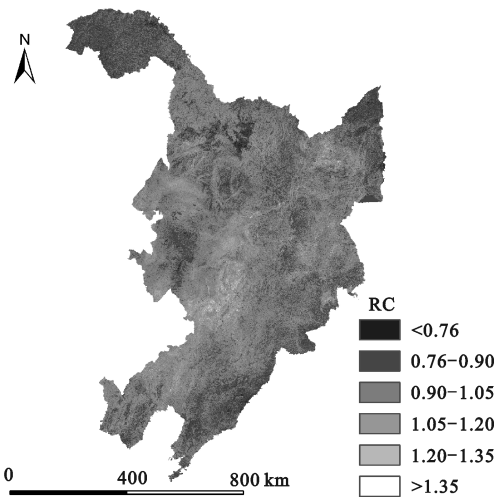

Fig. 5 Spatial variation of restricted coefficient (*RC*). *RC* is dimensionless

3.3 Comparison of SOCS produced by different methods

The SOCS values produced by the five different methods are summarized in Table 5 because of the approximate spatial variation (i.e., MCF versus MCV, and SPS versus PKB in Fig. 3). The lowest values of mean, median and maximum were obtained for the MCV method. As MCV also generated the lowest *RMSEs* of the SOCS estimation, analysis of the SOCS map produced by MCV was carried out, in which the soil type and vegetation type were compared. About 95% of SOCS could be found in Argosols, Cambosols and Isohumosols, with about 58% of SOCS found in Argosols (Table 6). The highest relative SOC densities (percent of SOCD/ percent of area) were found in areas of Udic Cambosols (1.34) and Udic Isohumosols (1.22). These results did not agree with the findings of Liu *et al*. (2012) because of the soil degradation and the inappropriate reference of legacy soil maps. The main vegetation types were focused on the forest and cropland. The largest SOCS was found in the croplands mainly consisting of *Triticum aestivum* L., *Sorghum bicolor* (L.) Moench, *Glycine max* (L.) Merr., *Zea mays* L. and *Setaria italica* (L.) P.

Table 5 Statistics of the predicted soil organic carbon density

Method	Minimum (kg/m^2)	Median (kg/m^2)	Mean (kg/m^2)	Maximum (kg/m^2)	Standard deviation (kg/m^2)	Total SOC stock (10^{12} kg)
MCF	2.31	11.54	11.71	37.33	3.83	9.16
MCV	1.88	10.23	10.62	34.89	3.62	8.32
SPS	1.30	13.08	15.83	121.23	10.99	12.38
PKB	1.30	13.08	15.78	121.39	10.15	12.32
Veg	0.52	13.57	16.62	121.39	10.76	12.99

Notes: MCF, model-then-calculate using fixed-thickness; MCV, model-then- calculate using variable-thicknesses; SPS, soil profile statistics; PKB, pedological professional knowledge-based; Veg, vegetation type-based method

Table 6 Statistics of the SOCS map generated by MCV concerning the soil type (subgroup) and vegetation type (species)

Soil type (suborder)	Area (%)	SOCS (%)	Vegetation type (species)	Area (%)	SOCS (%)
Boric Argosols	48.61	53.79	*Triticum aestivum* L., *Sorghum bicolor* (L.) Moench, *Glycine max* (L.) Merr., *Zea mays* L., and *Setaria italica* (L.) P. Beauv.	33.89	33.70
Udic Isohumosols	10.02	12.20	*Corylus heterophylla* Fisch., *Lespedeza bicolor* Turcz., *Quercus mongolica* Fisch. ex Ledeb.	18.09	19.08
Aquic Cambosols	14.54	11.79	*Larix gmelinii* Rupr. and *Larix olgensis* A. Henry	6.85	7.94
Ustic Isohumosols	6.45	6.51	*Quercus acutissima* Carruth.	6.68	6.64
Gelic Cambosols	5.00	5.26	*Ocimum basilicum* L. and grass-forb	6.37	6.50
Ustic Argosols	4.57	2.26	*Tilia chinensis* Maxim., *Ulmus pumila* L., *Betula platyphylla* Suk. and *Populus adenopoda* Maxim.	5.09	5.53
Udic Argosols	4.46	2.17	*Pinus koraiensis* Siebold & Zucc. and deciduous broad leaved tree	3.90	4.31
Udic Cambosols	0.76	1.02	*Cyperus rotundus* L. and other small grass	3.34	3.46
Ustic Cambosols	0.18	0.17	*Leymus chinensis* (Trin.) Tzvelev	3.21	3.14
Others	5.41	4.82	Others	12.57	9.69

Note: SOCS, soil organic carbon stock

Beauv. The highest relative SOC density was found in areas of *Pinus koraiensis* Siebold & Zucc. and deciduous broad leaved trees (1.11), and *Larix gmelinii* Rupr. and *Larix olgensis* A. Henry (1.16).

3.4 Effects of categorical variables

Specifically, the effect of predictors was explained within each level of categorical variables (Fig. 6). Most predictors were fitted with varied regression coefficients illustrating the significant influence of the random effects. Some predictors, that is, NDVI and Elevation, had different signs in different linear mixed models. The MAAT was the most important predictor for the SOCS simulation, concurring with a recent study (Qin *et al*., 2016). It was worth noting that linear mixed models based on soil type and vegetation type resulted in similar regressions. The minimal residual of the linear mixed models was achieved in the vegetation type-based model, whereas the maximal residual was fitted in the soil type-based model.

4 Discussion

4.1 Comparison of modeling approaches

As an extension of the data driven method (MCF), the MCV method could be referred to as a pedological knowledge driven model, sharing similar model advantages, such as detailed spatial distribution of SOCS and incorporation of more soil-forming factors. The MCV method generated the same or better accuracy when compared with the MCF and other polygon-based techniques. Even if more prediction procedures were involved, MCV produced less prediction errors than MCF. This could be explained by the precise calculation of SOCS based on pedogenetic horizons. According to a recent study (Parras-Alcántara *et al*., 2015), about 59.8% and 28.2% bias would be introduced for the topsoil and total SOCS estimation respectively in the MCF model.

Concerning the soil continuum in space, MCV additionally predicted the spatial variation of pedogenetic horizon thickness, that is, the thickness of soil horizons as it varies in space. The SOCS was then aggregated based on the predicted soil layer thickness. Within MCV, the types of horizons were merged. Numerous SOCS mapping studies focusing on model comparisons have been performed to seek the best prediction model in practice. It is difficult to straightforwardly compare those results, in that pedological conditions often vary in different landscapes. Similar to other studies (Dorji *et al*., 2014; Lacoste *et al*., 2014), our cross validations illustrated that there are still major challenges in simulating the SOC variation in subsoil (Table 3). Our results agreed with the consensus view that the use of machine learning techniques might improve model performance within a specific context (Grunwald, 2009; Martin *et al*.,

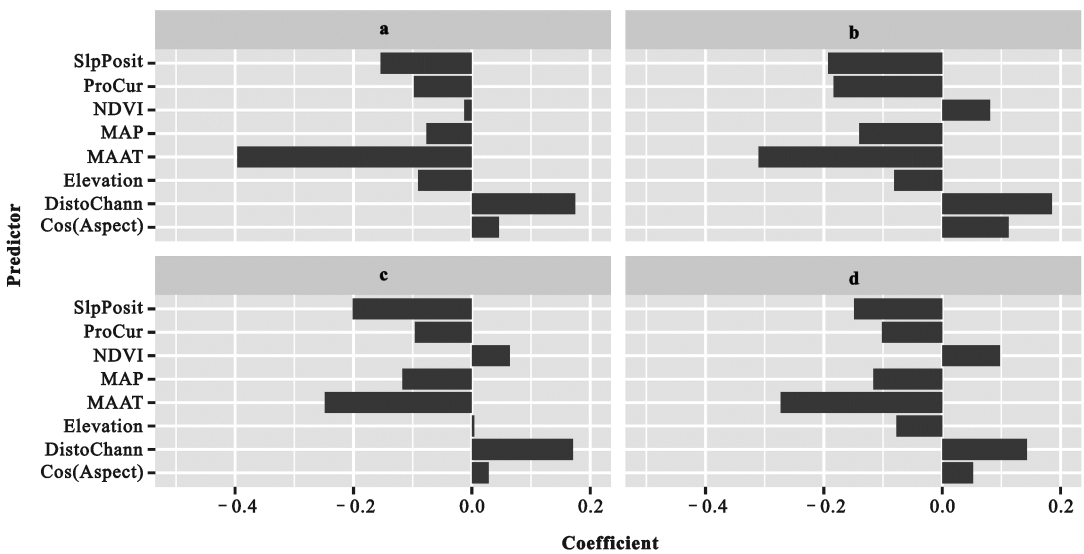

Fig. 6 Regression coefficients of the predictors within linear mixed models considering the random effects of geology (a), land use (b), soil type (c) and vegetation type (d). SlpPosit: relative slope position; ProCur: profile curvature; NDVI: normalized difference vegetation index; MAP: mean annual precipitation; MAAT: mean annual air temperature; DistoChann: vertical distance to channel network; Cos(Aspect): cosine-transform of aspect

2014; Were *et al*., 2015). SOCS estimation can benefit from the use of the MCV framework based on other predictive techniques as well as the random forest method employed in this study. Therefore, we recommend the pedogenetic horizons-based approach because of its characteristics of reliability and portability. Incorporating the MCV and depth functions (Kempen *et al*., 2011), the vertical soil variation can be realistically represented for anthropogenic influences and complex morphology profiles.

4.2 Uncertainty assessment

The five techniques considered generated moderate accuracy because of the relatively low sampling density. SPS achieved a similar level of performance to PKB in terms of validation results and spatial pattern (Table 2 and Fig. 3), especially when the soil data were limited and the high-level classification was adopted. Compared with the study in the Sanjiang Plain (Mao *et al*., 2015), both the MCF and MCV methods underestimated the SOCS. It was possible that the low sampling density prohibited an effective learning process (Bourennane *et al*., 2014). Another reason for the underestimation of MCV was likely because of the use of minimum values of soil properties in the potential pedogenetic layers. Additional predictions concerning the horizon estimation also generated more prediction errors for the calculation of SOCS. For example, the restricted coefficient (*RC*) in Equation (2) might underestimate the SOCS when the topsoil had high SOC concentration and high thickness.

Detailed information about vegetation would further benefit the SOCS estimation (Yu *et al*., 2013; Qi *et al*., 2016). Furthermore, the change in land use also involved a loss of SOCS in this area (Ding *et al.*, 2013; Wei *et al*., 2014), suggesting that the dynamically changing information on land use should be further used. As most of our samples were not collected in urban areas, urbanization was not fully considered (Vasenev *et al*., 2014; Zhang *et al*., 2015).

4.3 Influence of random effects on SOCS

The random effects were relatively significant in this study, showing a potential issue with the use of hybrid geostatistical techniques (Martin *et al*., 2014). Coming from the fitting relationships between the predictors and soil property of interest, these random effects can bring additional soil-landscape knowledge, for example, ranking the SOCS driving factors (Martin *et al*., 2011), analyzing the spatial dependence of model residuals (Song *et al*., 2016), spatial generalization of soil-land units (Ottoy *et al*., 2015), and quantifying SOCS under urbanized and cropping regions (Vasenev *et al*., 2014; Cardinael *et al*., 2015). Increasing the complexity of mixed effect models led to a useful suggestion for the selection of predictors. This was in particular the case for regional soil mapping at different generalization levels.

4.4 Soil taxonomy and soil continuum description

It is well known that the Chinese genetic soil classification (Xiong, 1987) of the legacy soil maps has been unable to keep up with developments in modern soil science because of the lack of quantitative criteria. Therefore, a soil taxonomy system (Cooperative Research Group on Chinese Soil Taxonomy, 2001) was developed more recently in line with international diagnosis-based soil classification systems. To date, maps of diagnostic horizons were not yet available. Numerous diagnostic horizons were usually varied especially across the taxonomic borders (Bockheim, 2014), hampering the ability of pedologists to make classification decisions. Therefore, as an intermediate of the MCV method, another potential application of the predicted horizon thicknesses is an update of the Chinese soil maps. Taking the transect validation as an example, the predicted horizons well described the thicknesses for the soil genesis (not given). The horizon thicknesses could be directly used for the soil classification at high level, which was applicative for the CST, Australian Soil Classification and USDA soil taxonomy. If the soil diagnostic characteristics could be achieved or inferred from legacy data or other information, the regional soil classification would be substantially improved.

5 Conclusions

The proposed MCV method was guided by the soil continuum descriptions in terms of pedogenetic horizons and generated acceptable SOCS estimations. The MCV method generated higher accuracy than the MCF method and other polygon-based approaches (SPS, PKB and Veg). The intermediate product of MCV method, that is, horizon thickness maps were fluctuant enough and reflected many details in space. Difference between maps

maps generated by the two assumptions was apparent. The maps generated with MCF and MCV showed more consistent spatial patterns than others. About 61% of SOCS values estimated by MCV were smaller than the estimates by MCF. Furthermore 88% of the absolute difference between SPS and PKB was smaller than 2 and 80% was with the same values. Random effects of typical soil-forming factors (lithological type, land cover, vegetation and soil type) were investigated by linear mixed model. It was suggested that mean annual air temperature (MAAT) was the most important predictor for the SOCS simulation. Furthermore the minimal residual of the linear mixed models was achieved in the vegetation type-based model.

The prediction of pedogenetic horizons took advantage of the soil continuum and could benefit the soil classification and understanding of soil functions. Thus the statistical analysis of the SOCS map produced by MCV was performed regarding the soil type and vegetation type. About 95% of SOCS could be found in Argosols, Cambosols and Isohumosols, and 58% was found in Argosols. Different with published literatures, high SOC densities were found in areas of Udic Cambosols and Udic Isohumosols, which could be explained by the soil degradation and the inappropriate soil type reference of legacy soil maps. The largest SOCS was found in the croplands mainly consisting of *Triticum aestivum* L., *Sorghum bicolor* (L.) Moench, *Glycine max* (L.) Merr., *Zea mays* L. and *Setaria italica* (L.) P. Beauv. The use of up-to-date information on land use change and vegetation type would benefit the estimation of SOCS, which will be our future research focus.

References

Ahrens R J, Eswaran H, Rice T J, 2003. Soil classification: past and present. In: Eswaran H *et al*. (eds.). *Soil Classification*: *A Global Desk Reference*. Boca Raton: CRC Press.

Aitkenhead M J, Coull M C, 2016. Mapping soil carbon stocks across Scotland using a neural network model. *Geoderma*, 262: 187–198. doi: 10.1016/j.geoderma.2015.08.034

Bapat R B, 2012. *Linear Mixed Models*. Heidelberg: Springer. doi: 10.1007/978-1-4471-2739-0

Bishop T F A, McBratney A B, Laslett G M, 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91(1–2): 27–45. doi: 10.1016/S0016-7061(99)00003-8

Blake G R, 1965. Bulk density. In: Black, C A (eds.). *Methods of Soil Analysis*, *Part1*. *Physical and Mineralogical Properties*, *including Statistics of Measurement and Sampling*. Madison: American Society of Agronomy, Soil Science Society of America.

Bockheim J G, 2014. *Soil Geography of the USA*: *A Diagnostic-Horizon Approach*. Heidelberg: Springer.

Boehner J, Koethe R, Conrad O *et al*., 2002. Soil regionalisation by means of terrain analysis and process parameterisation. In: Micheli E *et al*. (eds.). *Soil Classification 2001*. Luxembourg: European Soil Bureau, 213–222.

Bourennane H, Salvador-Blanes S, Couturier A *et al*., 2014. Geostatistical approach for identifying scale-specific correlations between soil thickness and topographic attributes. *Geomorphology*, 220: 58–67. doi: 10.1016/j.geomorph.2014.05.026

Breiman L, Friedman, J H, Olshen R A *et al*., 1984. *Classification and Regression Trees*. New York: Chapman and Hall.

Cardinael R, Chevallier T, Barthès B G *et al*., 2015. Impact of alley cropping agroforestry on stocks, forms and spatial distribution of soil organic carbon: a case study in a Mediterranean context. *Geoderma*, 259–260: 288–299. doi: 10.1016/j.geoderma.2015.06.015

Chaplot V, Lorentz S, Podwojewski P *et al*., 2010. Digital mapping of A-horizon thickness using the correlation between various soil properties and soil apparent electrical resistivity. *Geoderma*, 157(3–4): 154–164. doi: 10.1016/j.geoderma.2010.04.006

CMA (China Meteorological Administration), 2011. *China Meteorological Data Daily Value*. Beijing: China Meteorological Data Sharing Service System.

Cooperative Research Group on Chinese Soil Taxonomy, 2001. *Chinese Soil Taxonomy*. Beijing: Science Press.

Crouvi O, Pelletier J D, Rasmussen C, 2013. Predicting the thickness and aeolian fraction of soils in upland watersheds of the Mojave Desert. *Geoderma*, 195–196: 94–110. doi: 10.1016/j.geoderma.2012.11.015

de Gruijter J, Brus D J, Bierkens M F P *et al*., 2006. *Sampling for Natural Resource Monitoring*. Berlin: Springer. doi: 10.1007/3-540-33161-1

Ding F, Hu Y L, Li L J *et al*., 2013. Changes in soil organic carbon and total nitrogen stocks after conversion of meadow to cropland in Northeast China. *Plant & Soil*, 373(1–2): 659–672. doi: 10.1007/s11104-013-1827-5

Dorji T, Odeh I O A, Field D J *et al*., 2014. Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. *Forest Ecology and Management*, 318: 91–102. doi: 10.1016/j.foreco.2014.01.003

Editorial board of Series of Chinese Soil Taxonomy Classification, 1993. *Progress of the Chinese Soil Taxonomy Classification*. Beijing: Science Press.

Gallant J C, Dowling T I, 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research*, 39(12): 1347–1359. doi: 10.1029/2002WR001426

Grunwald S, 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, 152(3–4): 195–207. doi: 10.1016/j.geoderma.2009.06.003

Jenny H, 1941. *Factors of Soil Formation*: *A System of Quantitative Pedology*. New York: McGraw Hill.

Kempen B, Brus D J, Stoorvogel J J, 2011. Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. *Geoderma*, 162(1–2): 107–123. doi: 10.1016/j.geoderma.2011.01.010

Kosmas C, Gerontidis S, Marathianou M, 2000. The effect of land use change on soils and vegetation over various lithological formations on Lesvos (Greece). *Catena*, 40(1): 51–68. doi: 10.1016/S0341-8162(99)00064-8

Lacoste M, Minasny B, McBratney A *et al*., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. *Geoderma*, 213: 296–311. doi: 10.1016/j.geoderma.2013.07.002

Ließ M, Glaser B, Huwe B, 2012. Making use of the World Reference Base diagnostic horizons for the systematic description of the soil continuum: application to the tropical mountain soil-landscape of southern Ecuador. *Catena*, 97: 20–30. doi: 10.1016/j.catena.2012.05.002

Lin L I K, 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1): 255–268. doi: 10.2307/2532051

Liu F, Zhang G L, Sun Y J *et al*., 2013. Mapping the three- dimensional distribution of soil organic matter across a subtropical hilly landscape. *Soil Science Society of America Journal*, 77(4): 1241–1253. doi: 10.2136/sssaj2012.0317

Liu X, Burras L, Kravchenko Y S *et al*., 2012. Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science*, 92(3): 383–402. doi: 10.4141/cjss2010-058

Mao D H, Wang Z M, Li L *et al*., 2015. Soil organic carbon in the Sanjiang Plain of China: storage, distribution and controlling factors. *Biogeosciences*, 12(6): 1635–1645. doi: 10.5194/bg-12-1635-2015

Martin M P, Orton T G, Lacarce E *et al*., 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. *Geoderma*, 223–225: 97–107. doi: 10.1016/j.geoderma.2014.01.005

Martin M P, Wattenbach M, Smith P *et al*., 2011. Spatial distribution of soil organic carbon stocks in France. *Biogeosciences*, 8: 1053–1065. doi: 10.5194/bg-8-1053-2011

Nelson D W, Sommers L E, 1982. Total carbon, organic carbon and organic matter. In: Page A L *et al*. (eds.). *Methods of Soil Analysis, Part 2*. *Chemical and Microbiological Properties*. Madison: Agronomy Monograph, 539–579.

Ottoy S, Beckers V, Jacxsens P *et al*., 2015. Multi-level statistical soil profiles for assessing regional soil organic carbon stocks. *Geoderma*, 253–254: 12–20. doi: 10.1016/j.geoderma.2015.04.001

Parras-Alcántara L, Lozano-García B, Brevik E C *et al*., 2015. Soil organic carbon stocks assessment in Mediterranean natural areas: a comparison of entire soil profiles and soil control sections. *Journal of Environmental Management*, 155: 219–228. doi: 10.1016/j.jenvman.2015.03.039

Qi Guang, Chen Hua, Zhou Li *et al*., 2016. Carbon stock of larch plantations and its comparison with an old-growth forest in northeast China. *Chinese Geographical Science*, 26(1): 10–21. doi: 10.1007/s11769-015-0772-z

Qin Falyu, Shi Xuezheng, Xu Shengxiang *et al*., 2016. Zonal differences in correlation patterns between soil organic carbon and climate factors at multi-extent. *Chinese Geographical Science*, 26(5): 670–678. doi: 10.1007/s11769-015-0736-3

Song X D, Brus D J, Liu F *et al*., 2016. Mapping soil organic carbon content by geographically weighted regression: a case study in the Heihe River Basin, China. *Geoderma*, 261: 11–22. doi: 10.1016/j.geoderma.2015.06.024

Vanwalleghem T, Poesen J, McBratney A *et al*., 2010. Spatial variability of soil horizon depth in natural loess-derived soils. *Geoderma*, 157(1–2): 37–45. doi: 10.1016/j.geoderma.2010.03.013

Vasenev V I, Stoorvogel J J, Vasenev I I *et al*., 2014. How to map soil organic carbon stocks in highly urbanized regions? *Geoderma*, 226–227: 103–115. doi: 10.1016/j.geoderma.2014.03.007

Webster R, Oliver M A, 2001. *Geostatistics for Environmental Scientists*. Chichester: John Wiley & Sons.

Wei Yawei, Yu Dapao, Lewis Bernard Joseph *et al*., 2014. Forest carbon storage and tree carbon pool dynamics under natural forest protection program in northeastern China. *Chinese Geographical Science*, 24(4): 397–405. doi: 10.1007/s11769-014-0703-4

Were K, Bui D T, Dick Ø B *et al*., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, 52: 394–403. doi: 10.1016/j.ecolind.2014.12.028

Xiong Yi, 1987. *Chinese Soils* (*Second Edition*). Beijing: Science Press, 20–38. (in Chinese)

Yu P, Li Q, Jia H *et al*., 2013. Carbon stocks and storage potential as affected by vegetation in the Songnen grassland of northeast China. *Quaternary International*, 306(450): 114–120. doi: 10.1016/j.quaint.2013.05.053

Zhang Dan, Zheng Haifeng, Ren Zhibin *et al*., 2015. Effects of forest type and urbanization on carbon storage of urban forests in Changchun, Northeast China. *Chinese Geographical Science*, 25(2): 147–158. doi: 10.1007/s11769-015-0743-4

Zhang Y, Zhao Y C, Shi X Z *et al*., 2008. Variation of soil organic carbon estimates in mountain regions: a case study from Southwest China. *Geoderma*, 146(3–4): 449–456. doi: 10.1016/j.geoderma.2008.06.015

Zhi J, Jing C, Lin S *et al*., 2014. Estimating soil organic carbon stocks and spatial patterns with statistical and GIS-based methods. *Plos One*, 9(5): e97757. doi: 10.1371/journal.pone.0097757