SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

No Silver Bullet for Digital Soil Mapping: Country-specific Soil Organic Carbon Estimates across Latin America

Mario Guevara[1], Guillermo Federico Olmedo[2,3], Emma Stell[1], Yusuf Yigini[3], Yameli Aguilar Duarte[4], Carlos Arellano Hernández[5], Gloria E Arévalo[6], Carlos Eduardo Arroyo-Cruz[7], Adriana Bolivar[8], Sally Bunning[9], Nelson Bustamante Cañas[10], Carlos Omar Cruz-Gaistardo[5], Fabian Davila[11], Martin Dell Acqua[11], Arnulfo Encina[12], Hernán Figueredo Tacona[13], Fernando Fontes[11], José Antonio Hernández Herrera[14], Alejandro Roberto Ibelles Navarro[5], Veronica Loayza[15], Alexandra M. Manueles[6], Fernando Mendoza Jara[16], Carolina Olivera[17], Rodrigo Osorio Hermosilla[10], Gonzalo Pereira[11], Pablo Prieto[11], Iván Alexis Ramos[18], Juan Carlos Rey Brina[19], Rafael Rivera[20], Javier Rodríguez-Rodríguez[7], Ronald Roopnarine[21,22], Albán Rosales Ibarra[23], Kenset Amaury Rosales Riveiro[24], Guillermo Andrés Schulz[25], Adrian Spence[26], Gustavo M Vasques[27], Ronald R Vargas[3], and Rodrigo Vargas[1]

[1]University of Delaware, Department of Pland and Soil Sciences, Newark DE, USA. 19713
[2]INTA EEA Mendoza, San Martín 3853, Luján de Cuyo, Mendoza, Argentina, M5507EVY
[3]FAO, Vialle de Terme di Caracalla, Rome, Italy
[4]Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Mexico
[5]Instituto Nacional de Estadísitica y Geografía, Aguascalientes, Mexico
[6]Zamorano University of Honduras and Asociación Hondureña de la Ciencia del Suelo
[7]National Commission for the Knowledge and Use of Biodiversity, Mexico City, Mexico
[8]Subdirección Agrología, Instituto Geográfico Agustín Codazzi, Colombia
[9]Oficina Regional de la FAO para América Latina y el Caribe, Chile
[10]Servicio Agrícola y Ganadero, Chile
[11]Direccion General de Recursos Naturales, Ministerio de Ganaderia, Agricultura y Pesca, Uruguay
[12]Facultad de Ciencias Agrarias de la Universidad Nacional de Asunción, Paraguay
[13]Land Viceministry, Ministry of Rural Development and Land, Bolivia
[14]Universidad Autónoma Agraria Antonio Narro Unidad Laguna, Mexico
[15]Ministerio de Agricultura y Ganaderia, Quito, Ecuador
[16]Universidad Nacional Agraria, Nicaragua
[17]Representación de FAO en Colombia
[18]Instituto de Investigación Agropecuaria de Panamá, Panama
[19]Sociedad Venezolana de la Ciencia del Suelo, Venezuela
[20]Ministerio de Medio Ambiente, Republica Dominicana
[21]Department of Natural and Life Sciences, COSTAATT, Port-of Spain, Trinidad and Tobago
[22]University of the West Indies, St Augustine Campus, Trinidad and Tobago
[23]Instituto de Innovación en Transferencia y Tecnología Agropecuaria, Costa Rica
[24]Ministerio de Ambiente y Recursos Naturales de Guatemala
[25]INTA CNIA, Buenos Aires, Argentina
[26]International Centre for Environmental and Nuclear Sciences, University of the West Indies, Jamaica
[27]Embrapa Solos, Rio de Janeiro, Brazil

Correspondence to: Rodrigo Vargas (rvargas@udel.edu)

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

Abstract. Country-specific soil organic carbon (SOC) maps are the baseline for the Global SOC Map of the Global Soil Partnership (GSOCmap-GSP). This endeavor requires harmonizing heterogeneous datasets and building country-specific capacities for digital soil mapping (DSM). We identified country-specific predictors for SOC and tested the performance of five predictive algorithms for mapping SOC across Latin America. The algorithms included: support vector machines, random forest, kernel weighted nearest neighbors, partial least squares regression, and regression-Kriging based on stepwise multiple linear models. Country-specific training data and SOC predictors (5x5km pixel resolution) were obtained from ISRIC-World-Soil-Information-System. In general, temperature, soil type, vegetation indices and topographic constraints were the best predictors for SOC, but country-specific predictors and their respective weights varied across Latin America. We compared a large diversity of country-specific data scenarios and were able to explain 53% of SOC variability (range <1% and 80%) with no universal predictive algorithm among countries. Overall, countries with large area (i.e., Brazil, Bolivia, Mexico, Peru) and large spatial SOC heterogeneity had lower SOC stocks per unit area and larger uncertainty in their predictions. We highlight that setting unreliable (excessive or low) model prediction limits can have important effects (under or overestimating) for predicting SOC; thus expert opinion is needed to set boundary prediction limits. Selection of predictive algorithms should consider density and variability of country-specific available SOC data and country-specific environmental gradients to maximize explained variance while minimizing prediction bias. To progress with country-specific SOC mapping, we call for improvements on quality and quantity of country-specific SOC measurements and associated predictors. This study highlights the large degree of spatial heterogeneity of SOC across Latin America, and provides a reproducible framework that could be used for building DSM capacity to improve country-specific SOC estimates.

1 Introduction

Soils store nearly 1500 Pg of carbon and represent the largest terrestrial carbon pool (Jackson et al., 2017); thus, it is critical to accurately quantify the variability of soil organic carbon (SOC) from local-to-global scales. During the 4th Session of the Global Soil Partnership (GSP) Plenary Assembly held in May 2016 in Rome, it was agreed to develop a Global Soil Organic Carbon Map (GSOCmap) (FAO, 2017). The overarching goal is that a Global SOC Map of the Global Soil Partnership (GSOCmap-GSP) will be developed using a distributed approach relying on country-specific SOC maps. The Food and Agriculture Organization (FAO) recently compiled how different statistical methods (e.g., regression-kriging and machine learning) could be used to generate country-specific SOC maps and uncertainty estimates (Yigini et al., 2017). All these approaches consider the reference framework of the SCORPAN model for digital soil mapping (DSM; McBratney et al. (2003)). In the SCORPAN reference framework a soil attribute (e.g., SOC) can be predicted as a function of the soil forming environment, in correspondence with soil forming factors from the Dokuchaev hypothesis and Jenny's soil forming equation based on climate, organisms, relief, parent material and elapsed time of soil formation (Florinsky, 2012). The SCORPAN (Soils,

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

Climate, Organisms, Parent material, Age and (N) space or spatial position, see McBratney et al. (2003)) reference framework is a empirical approach that can be expressed as in Eq. (1):

$$Sa_{[x;y\ t]} = f(S_{[x;y\ t]}, C_{[x;y\ t]}, O_{[x;y\ t]}, R_{[x;y\ t]}, P_{[x;y\ t]}, A_{[x;y\ t]}) \quad (1)$$

where Sa is the soil attribute of interest at a specific location N (represented by the spatial coordinates of field observations x; y) and representative for a specific time frame (t); S is the soil or other soil properties that are correlated with Sa; C is the climate or climatic properties of the environment; O are the organisms, vegetation, fauna or human activity; R is topography or landscape attributes; P is parent material or lithology; and A is the substrate age or the time factor. To generate predictions of Sa across places where no soil data is available, N should be explicit for the information layers representing the soil forming factors. These predictions will be representative of the time period (t) when soil available data was collected. Therefore, the prediction factors ideally should represent, the conditions of the soil forming environment for the same period of time (as much as possible) when soil available data was collected. In Eq. (1) the left side is usually represented by the available geo-spatial soil observational data (e.g., from legacy soil profile collections) and the right side of the equation is represented by the soil prediction factors. These prediction factors are normally derived from four main sources of information: a) thematic maps (i.e., soil type, rock type, land use type); b) remote sensing (i.e., active and passive); c) climate surfaces and meteorological data; and d) digital terrain analysis or geomorphometry. The SCORPAN reference framework is widely used, but one critical challenge is to quantify the relative importance of the soil forming factors (i.e., prediction factors) that could explain the underlying soil processes controlling the spatial variability of a specific soil attribute (i.e., SOC).

Arguably, there are two cultures for statistical modeling (Breiman, 2001) that influence the predictions of the spatial variability of SOC. One assumes that the variability of observations can be reproduced by a given stochastic data model (e.g., with hypothesis about the spatial structure of the variable). The other uses algorithmic models and treats as unknown, the mechanisms generating the structure of values in available datasets (e.g., with hypothesis about the statistical distribution and moments of the variable). Different mapping approaches use a set of given available predictors in different ways. Thus, comparing different approaches and methods is useful to quantify the relative importance of prediction factors across data configurations and distributional properties. We argue that a systematic analysis of predictive algorithms and consequently selection of predictors (by each one of the algorithms) could provide insights about the underlying factors that control the spatial variability of SOC.

The last decade has seen an increasing diversity of approaches for DSM. Data mining techniques have been successfully used to model and predict the spatial variability of soil properties (Rossel and Behrens, 2010; Hengl et al., 2017; Shangguan et al., 2017) and generate country-specific SOC maps (Viscarra Rossel et al., 2014; Adhikari et al., 2014). The combination of regression modeling approaches with geostatistics of model residuals (i.e., regression Kriging) is a combined strategy that has been widely used to map SOC (Hengl et al., 2004; Mishra et al., 2009; Marchetti et al., 2012; Kumar et al., 2012; Peng et al., 2013; Adhikari et al., 2014; Yigini and Panagos, 2016; Nussbaum et al., 2014; Mondal et al., 2017). Machine learning algorithms such as random forests or support vector machines have also been used to increase statistical accuracy of soil

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

carbon models (Martin et al., 2011; Hashimoto et al., 2017; Hengl et al., 2017) including applications for SOC mapping (Grimm et al., 2008; Sreenivas et al., 2016; Yang et al., 2016; Hengl et al., 2017; Delgado-Baquerizo et al., 2017; Ließ et al., 2016; Viscarra Rossel et al., 2014).Machine learning methods do not necessarily allow to extract information about the main effects of prediction factors in the response variable (e.g., SOC); consequently, a selection strategy is always useful to increase the interpretability of machine learning algorithms. With this diversity of approaches one constant question is if there is a method that systematically improve the prediction capacity of the others aiming to predict SOC across large geographic areas (e.g., Latin America). We postulate that probably there is no universal method (i.e., silver bullet) for DSM, and country-specific efforts are needed to test a variety of predictive algorithms to maximize explained variance while minimizing prediction bias.

The overarching goal of this study is to compare different predictive algorithms across 19 data/country scenarios with publicly available information to support the development of country-specific SOC maps to be included in the GSOCmap-GSP. Currently, SOC information across Latin America is derived from global models such as the SoilGrids system, or the Harmonized World Soil Database (Hengl et al., 2017; Köchy et al., 2015), which lack quantification of uncertainty and large areas are parameterized with limited country-specific information. This challenge is not unique for Latin America as many regions around the world (e.g., Africa, Siberia) have limited SOC information to parameterize models to predict SOC. To inform future SOC mapping efforts, this study addresses two specific questions: a) Which environmental variables (derived from publicly available information) have the highest correlations with country-specific SOC information?; and b) Which is the best method (i.e., predictive algorithm) to represent SOC across Latin America and within each country? The ultimate aim of this study is to contribute with the discussion about the importance of integrating country-specific information for representing and predicting soil-related variables (e..g., SOC) to improve regional-to-global predictions.

2 Methods

2.1 SOC observations

Soil organic carbon information was extracted from the WoSIS soil profile database . This dataset includes local-to-national soil profile collections with a sampling strategy generally based on morphological soil attributes (Batjes et al., 2017). The goal of the GSOCmap-GSP is to produce global information for the first 30 cm; thus, we generated synthetic horizons for this depth using a mass preserving spline approach (Bishop et al., 1999). We applied a pedotransfer function if the bulk density (BLD) information was missing (Yigini et al., 2017), and assumed a value of 0% of coarse fragments when information on coarse fragments (CRFVOL) was missing. The organic carbon stock for 0 to 30 cm was estimated using Global Soil Information Facilities R, GSIF following a standardized SOC calculation method (D.W. and L.E., 1982):

$$SOC_{stock} = \frac{ORC}{1000} \times \frac{H}{100} \times BLD \times \frac{(100 - CRFVOL)}{100} \quad (2)$$

where ORC is SOC density ($\mathrm{g \cdot kg^{-1}}$) and H is soil depth ($30\ \mathrm{cm}$). Finally, each country-specific dataset was transformed to its natural logarithm to reduce the right-skewed distribution of SOC values and because exploratory analysis showed that this

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

transformation can improve the prediction capacity of further modeling methods. To analyze the statistical distribution of SOC values, a probability distribution function was plotted and a Shapiro-Wilk test of normality was conducted on each dataset.

2.2 Soils prediction factors

We used environmental information from WorldGrids (worldgrids.com), which is an initiative of ISRIC-World Soil Information. We downloaded and masked 118 environmental layers (i.e., prediction factors) for each country to quantitatively represent the soil forming environment. The prediction factors were harmonized into a 1x1km global grid by the WorldGrids project from three main information sources: remote sensing, climate surfaces, and digital terrain analysis (http://worldgrids.org/doku.php/wiki:layers). Additional terrain parameters (e.g., terrain slope, aspect, catchment area, channel network base level, terrain curvature, topographic wetness index, length-slope factor) from elevation data were calculated in SAGA GIS for each country following the standard implementation for basic terrain parameters (Conrad et al., 2015). We re-sampled the prediction factors into a 5x5km pixel size grid to reduce the computational demand required to make predictions and facilitate the reproducibility of this DSM framework without the need of High Performance Computing.

2.3 Prediction of SOC and model evaluation

First, the relationship between SOC and prediction factors was explored using simple correlation analysis. Second, the 10 prediction factors with highest correlations with SOC data were selected for each country and used for further analyses. Third, we implemented Regression-Kriging (based on a multiple linear regression model (RK) and partial least squares regression (PLS)), and three machine learning models: support vector machines (SVM), random forests (RF), and kernel weighted nearest neighbors (KK) to generate SOC maps for each country. A brief explanation for each modeling approach is provided in Appendix A1.

We also analyzed the influence of the maximum allowed prediction limits for each prediction algorithm. The units of the SOC estimates are $\mathrm{kg} \cdot \mathrm{m}^{-2}$. The sensitivity of the total SOC stock related to the model prediction limit was tested by changing the maximum prediction limit from 2.7 $\mathrm{kg} \cdot \mathrm{m}^{-1}$ (1 in a log scale) to 2980.95 $\mathrm{kg} \cdot \mathrm{m}^{-2}$ (8 in a log scale).

To generate a combined SOC map, we used a weighted average of the country-specific predictions. The weights of this average were defined by the relationship between the errors (measured as the RMSE) and the correlation (EC_r). We propose this EC_r as an approach to better understand the agreement between the correlation (calculated by the means of cross validation) and the RMSE (derived from the unbiased residuals of cross validation). Before calculating the RMSE/correlation ratio, the RMSE and the correlation between observed and predicted were standardized (by its maximum and minimum values) to a range between 0 and 1 using:

$$\mathrm{RMSE}_{std} = \frac{\mathrm{RMSE}_i - min(\mathrm{RMSE})}{range(\mathrm{RMSE})} \tag{3}$$

$$\mathrm{corr}_{std} = \frac{\mathrm{corr}_i - min(\mathrm{corr})}{range(\mathrm{corr})} \tag{4}$$

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

$$EC_r = \frac{\text{RMSE}_{std}}{\text{corr}_{std}} \quad (5)$$

Where EC_r is the proposed ratio between errors and correlation between observed and predicted (derived by cross-validation); RMSE_i is the observed RMSE for the ith model; $min(\text{RMSE})$ is the minimum observed value of RMSE, and $range(\text{RMSE})$ is the difference between the maximum and minimum observed values of RMSE; corr_i is the observed correlation for the ith model; $min(\text{corr})$ is the minimum observed value of correlation, and $range(\text{corr})$ is the difference between the maximum and minimum observed values of correlation

If the value of the EC_r was close to 0, then there is a stronger agreement between high RMSE and low correlation, or low RMSE and high correlation. If this value deviated from 0 (up to 1 or more), then the RMSE would tend to be high while the correlation was also high, suggesting that the method represents the variability of SOC but with high bias. Finally, the uncertainty (represented by the variance of the different prediction approaches) was divided by the mean and multiplied by 100 to provide an interpretable standardized visualization of uncertainty (i.e., in percent). Country-level SOC stocks are reported as the sum of all 5x5km pixels of all SOC predicted values (i.e., weighted average of SOC) within each country. All analyzes were performed using the R software. (R Core Team, 2017).

3 Results

3.1 Descriptive statistics

SOC across the different countries showed a wide diversity of data-scenarios (Table 1). Costa Rica (with a mean of 11.05 $\text{g} \cdot \text{kg}^{-1}$), Chile (with a mean of 9.88 $\text{g} \cdot \text{kg}^{-1}$) and Colombia (with a mean of 8.15 $\text{g} \cdot \text{kg}^{-1}$) are the countries with the highest SOC values. Brazil (n=5616) and Mexico (n=4321) were the countries with highest data availability. In contrast, Honduras (n=11), Guatemala (n=20) and Belize (n=21) were the countries with less density of of SOC estimated values (Table 1). With the original (untransformed) dataset, the only countries that showed a normal distribution after the Shapiro- Wilk test of normality with an alpha of 0.05 were Belize, Guatemala, Honduras and Suriname.

3.2 Correlation of SOC and its predictors

Best correlated predictors were not the same across countries. We found higher correlations with the original data sets transformed to its natural logarithm, as data had a right-skewed distribution and did not follow a normal distribution (i.e., log-normal). Highest correlations of available SOC data and its environmental predictors were associated with temperature-related-variables across Honduras, Costa Rica, Peru, Chile, Guatemala and Suriname (the r^2 varied from from 0.35 to 0.58). However, there were a low number of available SOC observations across these countries in the WoSIS system (between 11 to 34). Similarly, across countries with high data availability (e.g., Mexico and Brazil) the strongest correlations between SOC and prediction factors were associated with temperature-related variables (Table 2). In all cases, the relationship between SOC

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

Table 1. Descriptive statistics of SOC estimates $kg \cdot m^2$ and total land area for each analyzed country. N is the number of observations. We provide quantiles, median, mean and the standard deviation of SOC data. The columns p and plog represent the probability values derived from the Shapiro-Wilk test of normality before (p) and after (plog) the log transformation of SOC values. When p is larger than plog, the log transformation of the data did not increased the probability of normality in the dataset. ARG=Argentina, BLZ=Belize, BOL=Bolivia, BRA=Brazil, CHL=Chile, COL=Colombia, CRI=Costa Rica, CUB=Cuba, ECU=Ecuador, ESP=Espana, GTM=Guatemala, HND=Honduras, JAM=Jamaica, MEX=México, NIC=Nicaragua, PAN=Panama, PER=Peru, SUR=Suriname, SLV=El Salvador, URY=Uruguay, VEN=Venezuela.

Country	n	Land Area (km2)	Min	1st Q	Med	Mean	3rd Q	Max	SDev	p / plog
ARG	231	2736690	0.34	1.88	3.21	5.65	5.96	86.85	9.33	<0.001 / 0.03
BLZ	21	22970	1.84	4.49	6.72	7.71	9.99	19.48	4.32	0.08 / 0.99
BOL	76	1083301	0.64	1.83	2.56	2.64	3.20	7.65	1.21	<0.001 / 0.08
BRA	5616	8358140	0.07	1.99	2.67	3.23	3.34	573.76	9.18	<0.001 / <0.001
CHL	44	743812	0.43	3.58	5.19	9.88	16.52	31.87	8.86	<0.001 / 0.01
COL	166	1038700	0.66	3.44	5.78	8.15	9.95	52.62	7.35	<0.001 / 0.96
CRI	43	51060	2.27	4.07	7.23	11.05	10.85	82.57	14.90	<0.001 / 0.001
CUB	48	109820	0.36	2.85	3.61	4.32	5.73	10.98	2.23	0.004 / <0.001
ECU	77	276841	0.99	2.37	3.65	5.15	4.36	24.36	5.15	<0.001 / <0.001
GTM	20	107159	2.60	5.66	8.48	7.73	9.75	12.41	3.11	0.14 / 0.007
HND	11	111890	2.69	5.25	6.48	6.71	8.32	12.38	2.78	0.72 / 0.39
JAM	76	10831	1.29	3.01	3.99	4.35	4.83	12.90	1.99	<0.001 / 0.72
MEX	4321	1943945	0.00	1.73	2.49	2.56	3.25	35.55	1.49	<0.001 / <0.001
NIC	26	119990	2.93	3.94	7.31	7.50	9.04	15.91	3.78	0.05/0.09
PAN	25	74177	3.39	4.90	7.53	7.59	9.13	19.89	3.76	0.003 / 0.49
PER	145	1279996	0.19	1.89	2.93	2.92	3.55	8.35	1.42	0.005 / <0.001
SUR	27	156000	1.38	2.60	3.35	3.37	4.07	6.01	1.20	0.69 / 0.51
URY	130	175015	0.82	2.70	3.38	4.34	3.90	46.54	4.67	<0.001 / <0.001
VEN	164	882050	0.31	2.58	4.14	5.92	6.57	44.35	6.37	<0.001 / 0.11

and temperature-related variables was negative. In contrast, SOC had a positive relationship with elevation-derived terrain parameters (r^2 varied from 0.43 to 0.59) such as terrain curvature, potential incoming solar radiation, and slope of terrain.

Lower correlations of SOC data with prediction factors were found across Brazil, Bolivia, Uruguay, Cuba, Panama, Venezuela and Argentina (e.g.r^2 <0.2). The correlation analysis was useful to formulate a working hypothesis about the major drivers of the spatial variability of SOC across countries based on our DSM conceptual framework (e.g. $SOC_{ARG} = f$ [px4wcl3a + px3wcl3a + evmmod3a + l07igb3a + px2wcl3a + ...]). For example, the best correlated predictors with SOC for Argentina were precipitation-related variables (px4wcl3a, px3wcl3a, px2wcl3a), remote sensing based vegetation indexes (evmmod3a), and a probability-based shrubland map (l07igb3a) (Table 2) (see sources of this maps in http://worldgrids.org/doku.php/wiki:layers).

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

Table 2. Best correlated predictors and its frequency across the analyzed data country-scenarios, given available data in the WOSIS system. See the predictor codes in http://worldgrids.org/doku.php/wiki:layers. ARG=Argentina, BLZ=Belize, BOL=Bolivia, BRA=Brazil, CHL=Chile, COL=Colombia, CRI=Costa Rica, CUB=Cuba, DOM=Dominican Republic, ECU=Ecuador, ESP=Espana, GTM=Guatemala, HND=Honduras, JAM=Jamaica, MEX=México, NIC=Nicaragua, PAN=Panama, PER=Peru, SUR=Suriname, SLV=El Salvador, URY=Uruguay, VEN=Venezuela

Var	factor	subfactor	freq	Country
gachws3a	Soil	Soil type	2	CUB, SUR
garhws3a	Soil	Soil type	2	PER, URY
ghshws3a	Soil	Soil type	2	BLZ, URY
gphhws3a	Soil	Soil type	2	CUB, JAM
gplhws3a	Soil	Soil type	2	BLZ, BOL
gvrhws3a	Soil	Soil type	2	JAM, URY
tdmmod3a	Climate	Temperature	11	ARG, BOL, BRA, CHL, COL, CRI, CUB, ECU, MEX, PER, VEN
tx1mod3a	Climate	Temperature	10	ARG, BOL, BRA, COL, CUB, ECU, JAM, NIC, PER, URY
tx4mod3a	Climate	Temperature	10	BRA, CHL, CRI, CUB, ECU, GTM, JAM, MEX, PER, VEN
tx5mod3a	Climate	Temperature	9	BOL, BRA, CHL, CUB, ECU, JAM, MEX, PER, VEN
tx6mod3a	Climate	Temperature	9	ARG, BOL, BRA, CHL, COL, CRI, ECU, MEX, VEN
tnhmod3a	Climate	Temperature	8	BLZ, COL, CRI, GTM, HND, JAM, PAN, VEN
tnmmod3a	Climate	Temperature	8	BLZ, COL, CRI, GTM, HND, PAN, URY, VEN
tx3mod3a	Climate	Temperature	7	BRA, CHL, CUB, ECU, PAN, PER, VEN
tdhmod3a	Climate	Temperature	6	ARG, CUB, ECU, JAM, MEX, URY
tdlmod3a	Climate	Temperature	6	BRA, CHL, COL, ECU, GTM, JAM
tnsmod3a	Climate	Temperature	5	ARG, MEX, NIC, PAN, SUR
tx2mod3a	Climate	Temperature	4	ARG, ECU, PER, URY
tdsmod3a	Climate	Temperature	3	MEX, PAN, SUR
tnlmod3a	Climate	Temperature	3	BLZ, COL, GTM
px2wcl3a	Climate	Precipitation	2	BOL, PAN
px3wcl3a	Climate	Precipitation	2	CHL, MEX
px4wcl3a	Climate	Precipitation	2	BRA, CHL
etmnts3a	Climate	ET	2	ARG, MEX
evmmod3a	Organism	Vegetation	5	ARG, ECU, HND, MEX, VEN
l07igb3a	Organism	Vegetation	2	ARG, CHL
DEMSRE3a	Topography		5	COL, CRI, GTM, HND, SUR
twisre3a	Topography		5	BRA, JAM, NIC, PAN, SUR
ChannNetworkBLevel	Topography		4	COL, HND, PAN, SUR
l3pobi3b	Topography		4	COL, CRI, PAN, VEN
inssre3a	Topography		3	BLZ, HND, SUR
opisre3a	Topography		3	CRI, NIC, SUR
SLPSRT3a	Topography		3	CRI, NIC, SUR
AnalyticalHillshading	Topography		2	BLZ, CUB
Aspect	Topography		2	BLZ, BOL
CovergenceIndex	Topography		2	BOL, HND
inmsre3a	Topography		2	CRI, GTM
ValleyDepth	Topography		2	BLZ, JAM
geaisg3a	Age		3	CHL, NIC, SUR

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

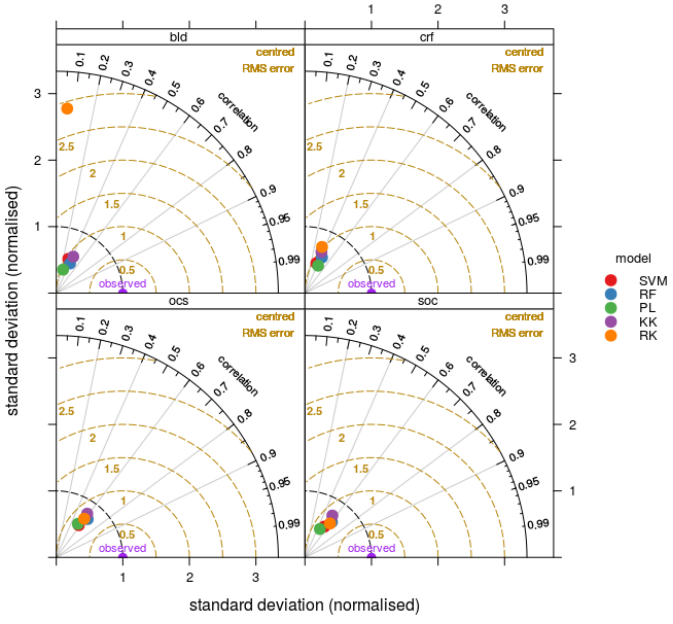

Figure 1. Taylor diagram showing the performance of the 5 models evaluated for SOC (the SOC stock, OCS) and its related soil attributes (OCR, BLD and CRFVOL) using all available data across Latin America.

3.3 SOC related properties

Correlations between ORCDR and prediction factors were higher with maximum and mean night-time temperature, where Costa Rica and Chile had the highest correlations (r^2 varied from 0.61 to 0.71). The best correlated variables with BLD were terrain parameters: relative slope position, vertical distance to channel network, flow accumulation areas, and potential incoming solar radiation. These correlations were stronger across Guatemala, Belize and Panama (r^2 varied from 0.52 to 0.67). We found that terrain slope and the standard deviation of temperature were the variables with highest correlations with CRFVOL; where Nicaragua, Honduras and Argentina had the highest correlations (r^2 varied from 0.40 to 0.55). We did not found a dominant algorithm to predict SOC related properties. Slightly higher correlations between observed and predicted values were achieved with RF, but in most cases different methods showed similar prediction capacity. The highest prediction error was found with RK for CRFVOL, but for all other output variables all prediction algorithms had a similar range of errors (Fig. 1). The PLS and SVM had the lowest variance for prediction of each one of the four soil properties. The correlations for predicting the combined SOC related properties (ORCDR, CRFVOL and BLD) for each prediction algorithm where: RK(0.82 to 0.87), RF(0.75 to 0.86), SVM (0.57 to 0.84), PL (0.68 to 0.83) and KK (0.41 to 0.80). Lower data availability and sparse distribution had a stronger effect for SVM and RK algorithms resulting in lower data-model agreement.

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

3.4 Country-specific SOC predictions

We did not find a dominant algorithm to predict SOC in a country-specific basis (Fig. 2). Overall, machine learning prediction algorithms generated similar results. Higher agreement of machine learning prediction algorithms was found in small countries where environmental conditions and land cover/use characteristics tend to be more homogeneous (e.g. Jamaica, Suriname). RK showed higher discrepancies in countries where data distribution was sparse (e.g., Suriname, Chile, Guatemala), but was effective across countries with higher and/or well distributed data availability (e.g., Mexico, Brazil). Machine learning SOC predictions were conservative compared with RK (RK generated the higher density of extreme and unreliable SOC values). PL had comparable results with machine learning algorithms (i.e., KK, SVM, RF). Higher correlation between observed and predicted data was found for Costa Rica (0.76; n=21) using SVM while the lowest error was found Suriname (0.36; n =37) using PL. In contrast, algorithms had lower prediction capacity for countries with large areas (e.g., Brazil, Mexico) despite the large data availability.

The correlation between the r^2 and rmse for RF, PL, KK and RK was positive (0.18, 0.35, 0.32 0.1; respectively). In contrast, this correlation was stronger for SVM (but negative; -0.65) where increasing the explained variance resulted in a lower error. These results suggest a low level of agreement between these two information criteria (r^2 and rmse) commonly used on DSM to assess performance of prediction algorithms.

Agreement between the rmse and r^2 was found only in 12 of the 19 countries, resulting in country-specific "recommended" prediction algorithms. Here we list the prediction algorithms that generated the best correlation and the best rmse for each country: ARG (RK, RK), BLZ (RF, RK), BOL (SVM, KK), BRA (RF, RF), CHL (PL, PL), COL (RF, RF), CRI (SVM, SVM), CUB (PL, PL), ECU (RK, RK), GTM (KK, RF), HND (SVM, KK), JAM (RF, RF), MEX (RK , RK), NIC (RF, RF), PAN (PL, KK), PER (KK, KK), SUR (SVM, PL), URY (RF, RK) and VEN (RK, RK) (see country codes in Table 1).

High discrepancy was found across the SOC predictions because algorithms use available data in different ways (Fig. 3B). The higher EC_r was found with PL (0.96) followed by RF (0.54) and KK (0.43), informing that these predictive algorithms do not minimize prediction bias while increasing the explained variance. SVM (with 0.008) and RK (with 0.003) had the lowest EC_r (inset histogram in Fig. 4), informing that they maximize the explained variance while minimizing prediction bias.

3.5 Estimated SOC stocks and uncertainties

We found a strong linear relationship (r^2 0.84) between SOC stocks and the area of each country (Fig. 4 A). The relationship between SOC predicted values by unit area and SOC prediction variance was negative (Fig. 4 B). Higher uncertainty and a relatively low density of SOC per unit area was found across Mexico, Bolivia, Brazil Honduras, Peru Suriname and Cuba. The standardized uncertainty of the total stocks reached values over 300% for countries such as Mexico and Bolivia (Fig. 4 B). In contrast, countries with higher SOC per unit area and a relatively low prediction variances were Panama, Guatemala, Costa Rica, Nicaragua and Belize.

Overall, we found a median prediction variance of 53% across countries in Latin America. Areas with high uncertainty were across northern Mexico, Central America, limits between Colombia and Brazil, and limits between Chile and Argentina (Fig

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

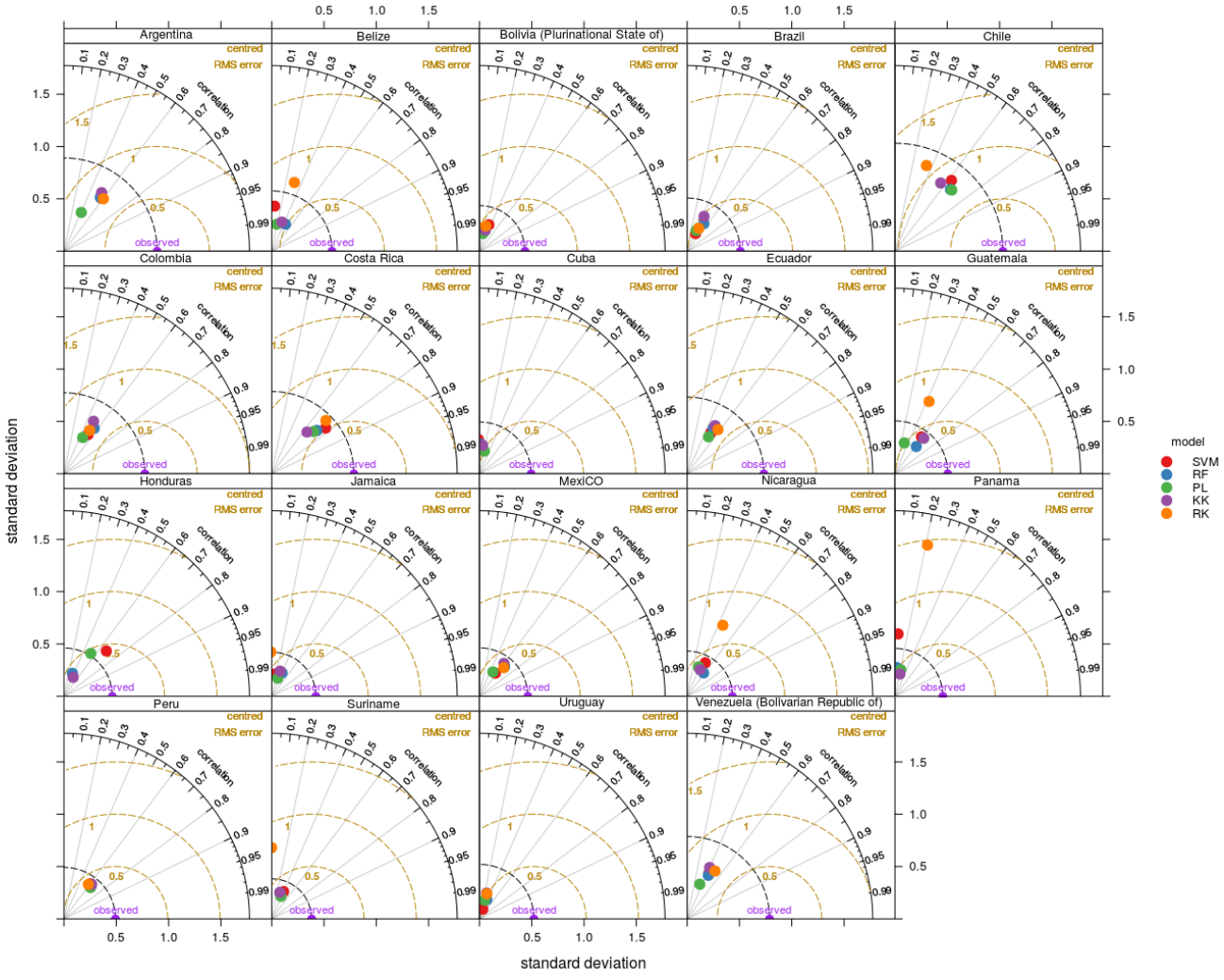

Figure 2. Taylor diagram showing the performance of the 5 models evaluated for country-specific SOC estimates across Latin America.

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

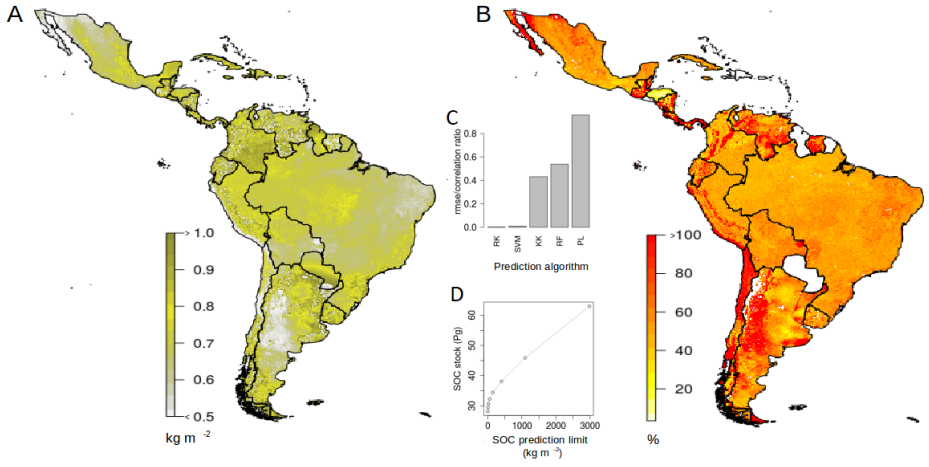

Figure 3. Mosaic of country specific SOC maps ($\mathrm{kg} \cdot \mathrm{m}^2$) and SOC prediction variance. In A we show a weighted average of predictions in which the weights were the EC_r. In the map we shrink the range of values between 0 and 1 to better illustrate the gradients of spatial variability of SOC. Note that no-data countries were filled using simple Geostatistics. The map in B shows the standardized uncertainty, which was generated by dividing the SOC variance by the mean. Red color areas suggest represent areas were the discrepancy of the models reached up to 100% or more. The inset histogram shows the median EC_r for each method. PL was the method with higher discrepancy between explained variance and bias. The inset scatter plot shows the relationship of SOC stock (in Pg) and the maximum limit of SOC prediction values, showing the sensitivity of the total estimated stock to the limit of maximum limit of predictions (1 to 8 in a log scale).

3A, B). Across countries, SOC stocks varied from 28.14 ±14.92 Pg (considering a maximum prediction limit of 2.71 $\mathrm{kg} \cdot \mathrm{m}^2$) to 62.99 ± 33.39 Pg (considering a maximum prediction limit of 2980 $\mathrm{kg} \cdot \mathrm{m}^2$; inset scatterplot in Fig. 4).

4 Discussion

We developed a reproducible DSM framework to characterize the spatial variability of SOC across Latin America. Our results suggest that several predictive algorithms can be used to better understand modeling bias, which can be associated with a) the property of interest (i.e., SOC), b) the environmental complexity and area/country of interest, and c) the characteristics of available data (e.g., spatial distribution and representativeness) to meet model-specific assumptions.

Our results incorporate a multi-model perspective for quantifying/evaluating the spatial variability of SOC. This effort is expected to increase the capacity of Latin American institutions to provide accurate baseline estimates of SOC with a country-specific perspective following recommendations of GSOCmap-GSP. Ultimately, these efforts will enhance the development of new guidelines for measuring, mapping, reporting, verification and monitoring SOC stocks at national level (Vargas et al., 2013). Accurate country-specific DSM frameworks for SOC are required to facilitate interoperability and inform environmental policy across developing countries (Vargas et al., 2017). Our results highlight that attention is needed to better understand the influence of model prediction limits (e.g., the full conditional distribution) for the predicted SOC stocks. Setting a unreliable

(excessive or low) prediction limit can have important effects (under or overestimating) on the overall estimated stocks (Fig. 3). Therefore, we argue that data science systems for DSM carbon assessments should be fundamentally based on SOC expert knowledge and informed by expert-based soil mapping systems.

Across Latin America we did not find a common predictive algorithm for SOC. These results suggest that country-specific environmental predictors and available data influence the applicability of different approaches. This assessment is needed to address the requirements from the GSOCmap-GSP with the official mandate to generate and update country-specific soil information by the means of DSM. Thus, we argue that the DSM form of each country should assess and incorporate country-specific available data and environmental predictors to select the best prediction algorithm. The FAO SOC mapping cookbook explores possibilities to derive country-specific SOC maps from a variety of prediction algorithms (Yigini et al., 2017), and multiple resources have described the state of the art of modeling methods focused on DSM of soil carbon (Minasny et al., 2013; Malone et al., 2017) including geostatistics (Hengl, 2009). Thus, data characteristics (e.g., spatial structure, representativeness) are specifically important for developing a DSM framework as legacy soil profile collections, generated with long-term soil inventory purposes, will determine data availability and spatial distribution within a country.

This country-specific approach to map regional SOC results in artifacts across geo-political borders. Therefore, data sharing, model validation and calibration experiments across borders (i.e., between countries) are required to better capture the spatial variability of SOC. The use of a natural-defined prediction domain (e.g., ecoregional or physiographic map) could reduce the border effects. However, we understand that geo-political limits are required for public policy decisions around country-specific needs. We highlight that there is a lack of publicly available country-specific data that ultimately influence the assessment of country-specific prediction algorithms (Fig. 3 A). The selection of a proper prediction algorithm for sparse scenarios of available data is then required to achieve the highest possible accuracy of country-specific SOC estimates. Our results highlight important uncertainty levels (>100%) across large areas of Latin America (Fig. 3B). The data contained in WoSIS has a low density distribution given the large area and environmental complexity of several analyzed countries. Thus, larger uncertainty dominates countries with larger carbon pools probably because available data does not capture the large spatial heterogeneity of SOC stocks. We highlight that the WoSIS dataset is a unique and invaluable effort that has proven to generate global SOC predictions (Hengl et al., 2017; Sanderman et al., 2017), but there is a global need to increase information and networking capabilities for SOC (Harden et al., 2017).

This study generated predictions of SOC across Latin America, but also provided information about the main relationships driving the spatial distribution of SOC. Machine learning (i.e., data driven) models have proven to be more efficient to model non-linear relationships of SOC (Hengl et al., 2015), but our results suggest that linear-based models (e.g., RK) could outperform machine learning methods under well distributed and representative SOC data scenarios. Similar results were found across productive landscapes of Brazil (Bonfatti et al., 2016). We argue that our capacity to meet modeling assumptions will determine the most suitable prediction algorithm. Machine learning models are usually conceived as black boxes and the influence of non-informative SOC prediction factors on machine learning-based SOC models has not been evaluated in detail. Therefore, we propose that the use of simple linear methods (i.e., correlation of available data and its predictors) can be a useful and parsimonious first step to inform data driven approaches and enhance the interpretability of machine learning models to

predict SOC. Furthermore, our data suggests that country-specific predictor factors are needed to better parameterize models but also could be useful for country-specific model interpretation. These results have important implications because it has been proposed that an extensive set of prediction factors are required to capture the large variance of the global SOC pool (Hengl et al., 2017). Thus, we propose that a limited but informative country-specific prediction factors should be explored to describe the local biophysical characteristics controlling SOC variability.

5 Conclusions

We provide a multi-model comparison approach to map SOC stocks across Latin America and found that there is not a dominant best prediction algorithm given available data. The relatively performance of the different methods vary from one place to another as well as the relatively correlation of SOC with the prediction factors given available data. We tested hypothesis driven approaches (e.g. linear Geo-statistics) and data driven algorithms (e.g. machine learning) which are used, respectively, to generate interpretable and predictable models of soil variability. We argue that models should not be conceived as competitors, because they have different assumptions (about the data itself, or about the empirical relationship between the response variable and its predictors). Therefore, different models will capture different portions of soil variability. There are no silver bullets on digital soil mapping across the 19 analyzed countries given available data in the WOSIS system. We highlight important levels of uncertainty SOC stocks associated with the maximum allowed prediction limit. Public data may not be representative across large areas and we call for the countries to strength digital soil mapping capacity building initiatives, reproducible research and data sharing. The use country-specific information and the use of different modeling approaches will enhance regional soil carbon mapping efforts, so we can easily identify where and the reasons why different modeling approaches generate different results.

This scientific paper shows that the initiative to build digital soil mapping capacities in Latinamerica offers very positive results. Each country has its particularities, the best methods or algorithms are not the same for everyone. It encourages us to keep increasing our capacities and collaboration at a regional level.

Code availability. The code used for this work will be available under the AGPL 3.0 license at https://github.com/DSM-LAC/NoSilverBulletsForDSM (Guevara et al., 2018)

Data availability. The soil dataset used is this paper is kindly provided by ISRIC. It can be downloaded from WOSIS http://www.isric.org/explore/wosis and correponds to the July 2016 snapshot (Batjes et al., 2017)

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

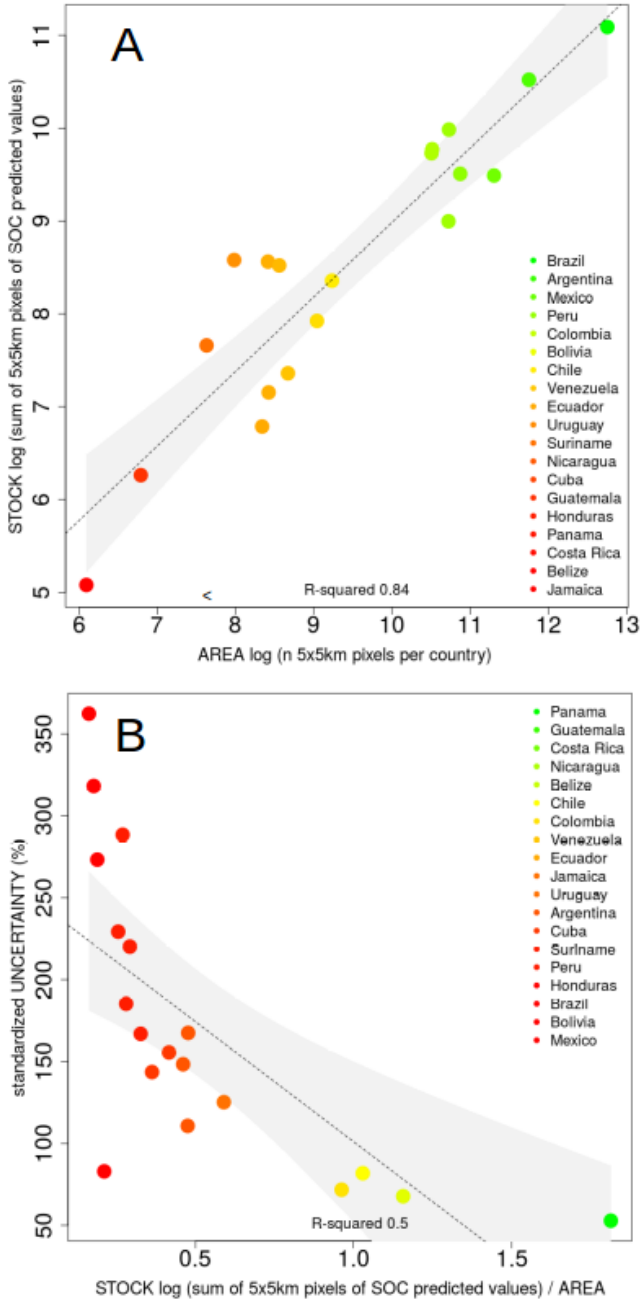

Figure 4. SOC stocks and uncertainties. In A we show the strong linear relationship between SOC stocks and the area of each country. In B we show the negative relationship between the estimated prediction variance (uncertainty component) and the SOC stocks per unit of area (density of SOC stock). Note how larger uncertainty and lower SOC stocks are associated with larger countries such as Mexico, Brasil, Bolivia or Peru.

SOIL Discuss., https://doi.org/10.5194/soil-2017-40
Manuscript under review for journal SOIL
Discussion started: 25 January 2018

Appendix A: Brief description of implemented methods

RK is a hybrid model with both, a deterministic and a stochastic component (Hengl et al., 2004). The regression part took form of a step-wise (back and forward) multiple linear regression to avoid statistical redundancy among the best prediction factors. The residual kriging was ordinary. The variogram parameters supporting the spatial interpolation were automatically fitted using the framework proposed by Hiemstra et al. (2008). RK was applied only to countries with 10 or more available observations.

PLS is a common method to deal with the presence of highly correlated predictors. The PLS algorithm integrates the compression and regression steps and it selects successive orthogonal factors that maximize the covariance between predictor and response variables (Wold, 1983; Viscarra Rossel et al., 2014). Most of its development and application is in the fields of chemometrics, but is used in several research areas to effectively solve regression and classification problems.

SVM apply a simple linear method to the data but in a high-dimensional feature space non-linearly related to the input space (Karatzoglou et al., 2006). It creates a hyperplane through n-dimensional spectral-space. Then, SVM separates numerical data based on a kernel function and parameters (e.g. gamma and cost) that maximize the margin from the closest point to the hyperplane that divides data with the largest possible margin, being the support vectors the points which fall within (Heumann, 2011). Then, linear models are fitted to the support vectors.

RF is an ensemble of regression trees based on bagging (Breiman, 1996). This machine learning algorithm uses a different combination of prediction factors to train multiple regression trees. Each tree is generated using a different subsets of available data (Breiman, 2001). The number of prediction factors to use on each tree is known as the mtry parameter. The final prediction is the weighted average of all individual trees.

KK is a pattern recognition technique which is based on the distances to training examples in the feature space (Silverman and Jones, 1989). The observations within the learning set, which are particularly close to the new observation (y, x), should get a higher weight in the decision than such neighbors that are far away from (y, x) (Hechenbichler and Schliep, 2004). The parameter k determines the number of neighbors from which information will be considered for prediction and a kernel function (eg. triangular, Gaussian among others) converts distances into weights which will be used for regression problems.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the Global Soil Partnership, the Central America, Caribbean and Mexico Soil Partnership and the South America Soil Partnership in collaboration with the Department of Plant and Soil Sciences at the University of Delaware. MG acknowledges support from a Conacyt fellowship. GFO is supported by the Argentinian government through the project INTA PNSUELO1134032. RV acknowledges support from NASA (80NSSC18K0173) and USDA (2014-67003-22070).

References

Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B., and Greve, M. H.: Digital mapping of soil organic carbon contents and stocks in Denmark, PLoS ONE, 9, https://doi.org/10.1371/journal.pone.0105519, 2014.

Batjes, N. H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., and de Jesus, J.: WoSIS: providing standardised soil profile data\hack{\newline} for the world, Earth System Science Data, 9, 1–14, https://doi.org/10.5194/essd-9-1-2017, https://www.earth-syst-sci-data.net/9/1/2017/, 2017.

Bishop, T., McBratney, A., and Laslett, G.: Modelling soil attribute depth functions with equal-area quadratic smoothing splines, Geoderma, 91, 27 – 45, https://doi.org/http://dx.doi.org/10.1016/S0016-7061(99)00003-8, http://www.sciencedirect.com/science/article/pii/S0016706199000038, 1999.

Bonfatti, B. R., Hartemink, A. E., and Giasson, E.: Comparing Soil C Stocks from Soil Profile Data Using Four Different Methods, in: Progress in Soil Science, pp. 315–329, Springer International Publishing, https://doi.org/10.1007/978-3-319-28295-4_20, https://doi.org/10.1007/978-3-319-28295-4_20, 2016.

Breiman, L.: Bagging Predictors, Machine Learning, 24, 123–140, https://doi.org/10.1023/A:1018054314350, https://doi.org/10.1023/A:1018054314350, 1996.

Breiman, L.: Random forests, Machine learning, pp. 5–32, https://doi.org/10.1023/A:1010933404324, http://link.springer.com/article/10.1023/A:1010933404324, 2001.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J.: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geoscientific Model Development, 8, 1991–2007, https://doi.org/10.5194/gmd-8-1991-2015, 2015.

Delgado-Baquerizo, M., Eldridge, D. J., Maestre, F. T., Karunaratne, S. B., Trivedi, P., Reich, P. B., and Singh, B. K.: Climate legacies drive global soil carbon stocks in terrestrial ecosystems, Science Advances, 3, e1602 008, https://doi.org/10.1126/sciadv.1602008, https://doi.org/10.1126/sciadv.1602008, 2017.

D.W., N. and L.E., S.: Total Carbon, Organic Carbon, and Organic Matter., 1982.

FAO: Fifth Meeting of the Global Soil Partnership Plenary Assembly, http://www.fao.org/3/a-bs973e.pdf, 2017.

Florinsky, I. V.: The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication), Eurasian Soil Science, 45, 445–451, https://doi.org/10.1134/S1064229312040047, http://dx.doi.org/10.1134/S1064229312040047, 2012.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H.: Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis, Geoderma, 146, 102–113, https://doi.org/10.1016/j.geoderma.2008.05.008, 2008.

Guevara, M., Olmedo, G. F., and Vargas, R.: DSM-LAC/NoSilverBulletsForDSM: No Silver Bullets - raw code, https://doi.org/10.5281/zenodo.1144079, https://doi.org/10.5281/zenodo.1144079, 2018.

Harden, J. W., Hugelius, G., Ahlström, A., Blankinship, J. C., Bond-Lamberty, B., Lawrence, C. R., Loisel, J., Malhotra, A., Jackson, R. B., Ogle, S., Phillips, C., Ryals, R., Todd-Brown, K., Vargas, R., Vergara, S. E., Cotrufo, M. F., Keiluweit, M., Heckman, K. A., Crow, S. E., Silver, W. L., DeLonge, M., and Nave, L. E.: Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter, Global Change Biology, pp. n/a–n/a, https://doi.org/10.1111/gcb.13896, http://dx.doi.org/10.1111/gcb.13896, 2017.

Hashimoto, S., Nanko, K., Ťupek, B., and Lehtonen, A.: Data-mining analysis of the global distribution of soil carbon in observational databases and Earth system models, Geoscientific Model Development, 10, 1321–1337, https://doi.org/10.5194/gmd-10-1321-2017, https://www.geosci-model-dev.net/10/1321/2017/, 2017.

Hechenbichler, K. and Schliep, K.: Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, Molecular Ecology, 399, 17, 2004.

Hengl, T.: A Practical Guide to Geostatistical Mapping, vol. 13, https://doi.org/10.1016/0277-9390(86)90082-8, http://book.spatial-analyst.net/system/files/cover{_}geostat{_}2009.pdf, 2009.

Hengl, T., Heuvelink, G. B., and Stein, A.: A generic framework for spatial prediction of soil variables based on regression-kriging, Geoderma, 120, 75–93, https://doi.org/10.1016/j.geoderma.2003.08.018, http://linkinghub.elsevier.com/retrieve/pii/S0016706103002787, 2004.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., De Jesus, J. M., Tamene, L., and Tondoh, J. E.: Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions, PLoS ONE, 10, 1–26, https://doi.org/10.1371/journal.pone.0125814, 2015.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, vol. 12, https://doi.org/10.1371/journal.pone.0169748, http://dx.plos.org/10.1371/journal.pone.0169748, 2017.

Heumann, B. W.: An object-based classification of mangroves using a hybrid decision tree-support vector machine approach, Remote Sensing, 3, 2440–2460, https://doi.org/10.3390/rs3112440, 2011.

Hiemstra, P., Pebesma, E., Twenh"ofel, C., and Heuvelink, G.: Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network, Computers and Geosciences, dOI: http://dx.doi.org/10.1016/j.cageo.2008.10.011, 2008.

Jackson, R. B., Lajtha, K., Crow, S. E., Hugelius, G., Kramer, M. G., and Piñeiro, G.: The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls, Annual Review of Ecology, Evolution, and Systematics, 48, 419–445, https://doi.org/10.1146/annurev-ecolsys-112414-054234, https://doi.org/10.1146/annurev-ecolsys-112414-054234, 2017.

Karatzoglou, A., Meyer, D., and Hornik, K.: Support Vector Algorithm in R, Journal of Statistical Software, 15, 1–28, 2006.

Köchy, M., Hiederer, R., and Freibauer, A.: Global distribution of soil organic carbon â€“ Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world, SOIL, 1, 351–365, https://doi.org/10.5194/soil-1-351-2015, https://www.soil-journal.net/1/351/2015/, 2015.

Kumar, S., Lal, R., and Liu, D.: A geographically weighted regression kriging approach for mapping soil organic carbon stock, Geoderma, 189-190, 627–634, https://doi.org/10.1016/j.geoderma.2012.05.022, http://dx.doi.org/10.1016/j.geoderma.2012.05.022, 2012.

Ließ, M., Schmidt, J., and Glaser, B.: Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches, PLoS ONE, 11, 1–22, https://doi.org/10.1371/journal.pone.0153673, 2016.

Malone, B. P., Minasny, B., and McBratney, A. B.: Using R for Digital Soil Mapping, Springer International Publishing, https://doi.org/10.1007/978-3-319-44327-0, https://doi.org/10.1007/978-3-319-44327-0, 2017.

Marchetti, A., Piccini, C., Francaviglia, R., and Mabit, L.: Spatial Distribution of Soil Organic Matter Using Geostatistics: A Key Indicator to Assess Soil Degradation Status in Central Italy, Pedosphere, 22, 230–242, https://doi.org/10.1016/S1002-0160(12)60010-1, http://dx.doi.org/10.1016/S1002-0160(12)60010-1, 2012.

Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., and Arrouays, D.: Spatial distribution of soil organic carbon stocks in France, Biogeosciences, 8, 1053–1065, https://doi.org/10.5194/bg-8-1053-2011, 2011.

McBratney, A., Santos, M. M., and Minasny, B.: On digital soil mapping, Geoderma, 117, 3 – 52, https://doi.org/http://dx.doi.org/10.1016/S0016-7061(03)00223-4, http://www.sciencedirect.com/science/article/pii/S0016706103002234, 2003.

Minasny, B., McBratney, A. B., Malone, B. P., and Wheeler, I.: Digital Mapping of Soil Carbon, in: Advances in Agronomy, pp. 1–47, Elsevier, https://doi.org/10.1016/b978-0-12-405942-9.00001-3, https://doi.org/10.1016/b978-0-12-405942-9.00001-3, 2013.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., and Van Meirvenne, M.: Predicting Soil Organic Carbon Stock Using Profile Depth Distribution Functions and Ordinary Kriging, Soil Science Society of America Journal, 73, 614, https://doi.org/10.2136/sssaj2007.0410, https://www.soils.org/publications/sssaj/abstracts/73/2/614, 2009.

Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., and Mukhopadhyay, A.: Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data, Egyptian Journal of Remote Sensing and Space Science, 20, 61–70, https://doi.org/10.1016/j.ejrs.2016.06.004, http://dx.doi.org/10.1016/j.ejrs.2016.06.004, 2017.

Nussbaum, M., Papritz, A., Baltensweiler, A., and Walthert, L.: Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging, Geoscientific Model Development, 7, 1197–1210, https://doi.org/10.5194/gmd-7-1197-2014, 2014.

Peng, G., Bing, W., Guangpo, G., and Guangcan, Z.: Spatial distribution of soil organic carbon and total nitrogen based on GIS and geostatistics in a small watershed in a hilly area of northern China, PLoS ONE, 8, 1–9, https://doi.org/10.1371/journal.pone.0083592, 2013.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2017.

Rossel, R. V. and Behrens, T.: Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma, 158, 46–54, https://doi.org/10.1016/j.geoderma.2009.12.025, https://doi.org/10.1016/j.geoderma.2009.12.025, 2010.

Sanderman, J., Hengl, T., and Fiske, G. J.: Soil carbon debt of 12, 000 years of human land use, Proceedings of the National Academy of Sciences, 114, 9575–9580, https://doi.org/10.1073/pnas.1706103114, https://doi.org/10.1073/pnas.1706103114, 2017.

Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., and Dai, Y.: Mapping the global depth to bedrock for land surface modeling, Journal of Advances in Modeling Earth Systems, 9, 65–88, https://doi.org/10.1002/2016MS000686, 2017.

Silverman, B. W. and Jones, M. C.: E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951), International Statistical Review / Revue Internationale de Statistique, 57, 233–238, http://www.jstor.org/stable/1403796, 1989.

Sreenivas, K., Dadhwal, V. K., Kumar, S., Harsha, G. S., Mitran, T., Sujatha, G., Suresh, G. J. R., Fyzee, M. A., and Ravisankar, T.: Digital mapping of soil organic and inorganic carbon status in India, Geoderma, 269, 160–173, https://doi.org/10.1016/j.geoderma.2016.02.002, http://dx.doi.org/10.1016/j.geoderma.2016.02.002, 2016.

Vargas, R., Paz, F., and de Jong, B.: Quantification of forest degradation and belowground carbon dynamics: ongoing challenges for monitoring, reporting and verification activities for REDD+, Carbon Management, 4, 579–582, https://doi.org/10.4155/cmt.13.63, https://doi.org/10.4155/cmt.13.63, 2013.

Vargas, R., Alcaraz-Segura, D., Birdsey, R., Brunsell, N. A., Cruz-Gaistardo, C. O., de Jong, B., Etchevers, J., Guevara, M., Hayes, D. J., Johnson, K., Loescher, H. W., Paz, F., Ryu, Y., Sanchez-Mejia, Z., and Toledo-Gutierrez, K. P.: Enhancing interoperability to facilitate implementation of REDD+: case study of Mexico, Carbon Management, 8, 57–65, https://doi.org/10.1080/17583004.2017.1285177, https://doi.org/10.1080/17583004.2017.1285177, 2017.

Viscarra Rossel, R. A., Webster, R., Bui, E. N., and Baldock, J. A.: Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change, Global Change Biology, 20, 2953–2970, https://doi.org/10.1111/gcb.12569, 2014.

Wold, H.: Systems Analysis by Partial Least Squares, Iiasa collaborative paper, IIASA, Laxenburg, Austria, http://pure.iiasa.ac.at/2336/, 1983.

Yang, R.-M., Zhang, G.-L., Yang, F., Zhi, J.-J., Yang, F., Liu, F., Zhao, Y.-G., and Li, D.-C.: Precise estimation of soil organic carbon stocks in the northeast Tibetan Plateau, Scientific Reports, 6, 21 842, https://doi.org/10.1038/srep21842, http://www.nature.com/articles/srep21842, 2016.

Yigini, Y. and Panagos, P.: Assessment of soil organic carbon stocks under future climate and land cover changes in Europe, Science of the Total Environment, 557-558, 838–850, https://doi.org/10.1016/j.scitotenv.2016.03.085, http://dx.doi.org/10.1016/j.scitotenv.2016.03.085, 2016.

Yigini, Y., Olmedo, G. F., Viatkin, K., Baritz, R., and Vargas, R. R., eds.: Soil Organic Carbon Mapping Cookbook, FAO, 2nd edn., http://www.fao.org/3/a-bs901e.pdf, 2017.