

OPEN

Digital soil mapping including additional point sampling in Posses ecosystem services pilot watershed, southeastern Brazil

Bárbara Pereira Christofaro Silva , Marx Leandro Naves Silva, Fabio Arnaldo Pomar Avalos, Michele Duarte de Menezes & Nilton Curi

This study aimed to evaluate the performance of three spatial association models used in digital soil mapping and the effects of additional point sampling in a steep-slope watershed (1,200 ha). A soil survey was carried out and 74 soil profiles were analyzed. The tested models were: Multinomial logistic regression (MLR), C5 decision tree (C5-DT) and Random forest (RF). In order to reduce the effects of an imbalanced dataset on the accuracy of the tested models, additional sampling retrieved by photointerpretation was necessary. Accuracy assessment was based on aggregated data from a proportional 5-fold cross-validation procedure. Extrapolation assessment was based on the multivariate environmental similarity surface (MESS). The RF model including additional sampling (RF*) showed the best performance among the tested models (overall accuracy = 49%, kappa index = 0.33). The RF* allowed to link soil mapping units (SMU) and, in the case of less-common soil classes in the watershed, to set specific conditions of occurrence on the space of terrain-attributes. MESS analysis showed reliable outputs for 82.5% of the watershed. SMU distribution across the watershed was: Typic Rhodudult (56%), Typic Hapludult* (13%), Typic Dystrudept (10%), Typic Endoaquent + Fluventic Dystrudept (10%), Typic Hapludult (9.5%) and Rhodic Hapludox + Typic Hapludox (2%).

Knowledge of the geographic distribution of soils allows the assessment of environment-soil relationships at the landscape level¹. Such information is essential for agronomic assessment, soil and water management and land use planning^{1–3}. Digital soil mapping (DSM) techniques are supported by a well-known and widely accepted model in soil science: the factors of soil formation. Early efforts to bring a quantitative solution to this model could be traced back to the work of Jenny⁴. Currently, with the development of geographic information technology and data processing, a comprehensive digital framework could be applied for the production of digital soil maps^{1,5,6}.

The *scorpan* model, as formalized by McBratney *et al.*⁵, serves as a route for the production of digital soil maps. It establishes a quantitative model for spatial association between soil forming factors, evaluated as covariate maps, and the occurrence of soil classes or properties. However, since the spatial association complexity is highly variable among landscapes, there is still no consensus or a pre-established function or operational model (spatial inference models) that could relate soilscape features and the occurrence of soil classes. Hence, model selection still represents a challenge in DSM studies.

Spatial inference models could be divided into data-driven (pedometric approach) and knowledge-driven approaches⁷. The pedometric approach (automatic and quantitative, includes statistics, geostatistics, machine learning, and data mining) gives a predictive accuracy that is generally related to a dense sampling scheme⁸. In the knowledge-driven approach, the knowledge of pedologists is incorporated into spatial prediction⁹ and the soil maps are viewed as a representation of the pedologist's understanding of the soils¹⁰. Such tacit knowledge is mostly based on a paradigm of conceptual soil-landscape model, under the hypothesis that the location and distribution of soils in the landscape is predictable¹¹. Thus, the use of spatial inference models as fuzzy logics¹², or bayesian inference¹³, to name a few, take advantage of such concepts in predictive maps.

Departamento de Ciência do Solo, Universidade Federal de Lavras UFLA, Av. Doutor Sylvio Menicucci, 1001, Kennedy, Lavras, MG, Brazil. Correspondence and requests for materials should be addressed to B.P.C.S. (email: barbarapcsilva@gmail.com)

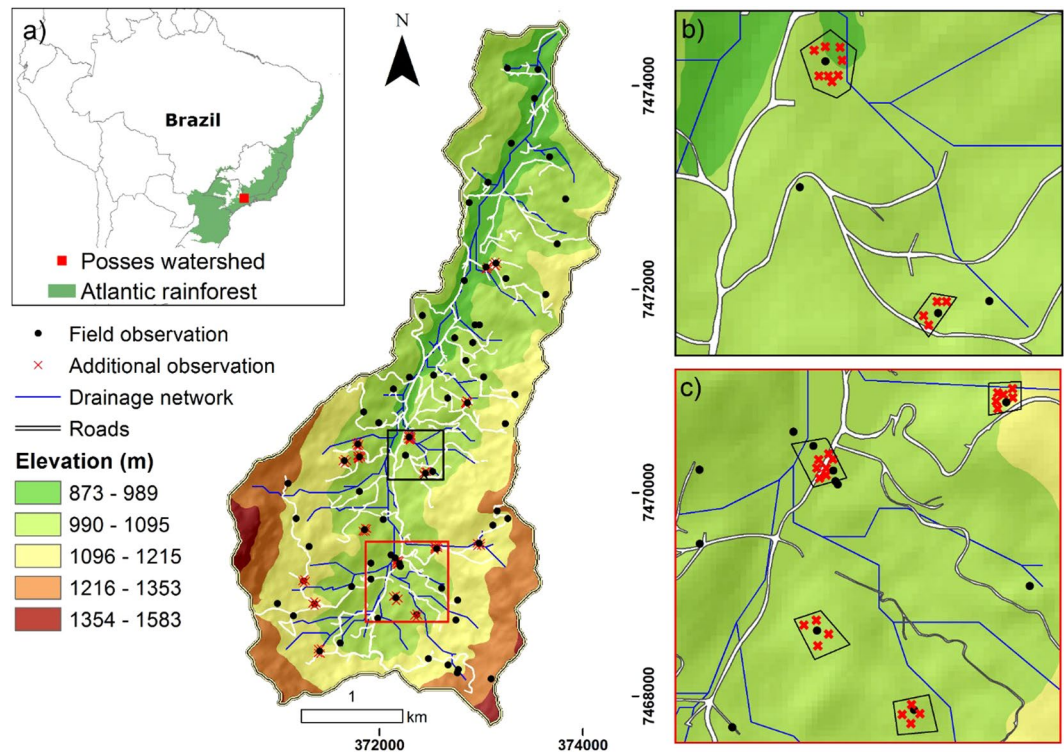


Figure 1. (a) Location of the study area and the soil survey sampling sites and their associated additional observations. (b,c) Detail of the field and additional observations clusters.

While current algorithms applied in DSM have shown acceptable performances^{14–16}, there is still a concern related to the occurrence of less-common soil classes across different landscapes. These soil classes can have significant effects on the process of model fitting, model selection, and output accuracies¹⁷. To overcome this issue, different strategies could be adopted, e.g. feature-space oversampling¹⁸ or geographic space oversampling¹⁹. Although these data-driven strategies have proved to be useful in specific circumstances¹⁷, some pedological aspects related to the occurrence of certain soil classes remain unsolved. Although pedometric and knowledge-driven approaches differ in philosophy and technical emphasis⁷, they are not mutually exclusive²⁰. In this sense, pedologists' knowledge about the study area (mental model) could also be incorporated into the digital mapping process, by identifying characteristic sites in aerial imagery associated to soil profiles previously surveyed²¹, which might be useful to account for specific conditions of less-common soil classes. In this way, a qualitative soil-landscape model would be translated into quantitative predictions supported by the spatial association between soil classes occurrence and environmental covariates²².

Different techniques of DSM have been widely assessed in studies in tropical soils^{23–27}. However, mountainous and complex relief areas still present challenges to any approach to soil mapping, due to the complex and scale-dependent interactions among soil forming factors and the cost-effort associated to survey inaccessible areas. Therefore, this study aimed to evaluate the performance of three algorithms used in digital soil mapping and the effects of additional sampling in the presence of less-common soil classes in a steep-slope watershed.

Material and Methods

Study area. This work was carried out at the PosSES watershed, southeastern Brazil (Fig. 1). This watershed covers an area of 1,200 ha, has an altitude range of 945 to 1,435 m, and a relief dominated by steep (>20%) to very steep slopes (>60%). According to Köppen's climate classification system, it corresponds to the Cfb class, i.e., mesothermic with no dry season and a warm summer²⁸. The average annual temperature is 18 °C, the hottest and coldest months have average temperatures of 25.6 °C and 13.1 °C, respectively, and the average annual precipitation is 1,447 mm. The predominant parental materials are alkaline granite and monzonite²⁹.

This water-production watershed is a pilot site for the “Conservador das Águas” project, which stands, since 2006, as a pioneering initiative to establish a payment system for ecosystem services in Brazil³⁰. Among the objectives of this project is the promotion of the sustainability of water resources for the Cantareira System, which is the major water supply system for the metropolitan region of São Paulo city.

Soil survey. The soil survey was carried out in accordance with Brazilian technical requirements of a semi-detailed soil survey³¹. For this study area, the requirements are: (a) mapping units composed by a single taxonomic unit; (b) a total of 16 samples/hectare, and (c) free-way sampling scheme, supported by the experience of the pedologists along with cartographic products (mainly digital terrain models). The later was used in order to capture the soil-landscape relationships and promote the sampling in different landforms (resulted from the

Soil mapping unit	Inclusion	FO	AO
Typic Dystrudept	Typic Humudept	6	19
Typic Endoaquent + Fluventic Dystrudept	—	4	14
Rhodic Hapludox + Typic Hapludox	—	3	20
Typic Hapludult (yellow)	—	10	16
Typic Rhodudult	Rhodic Kandiudult	36	0
Typic Hapludult* (red-yellow)	—	15	12

Table 1. Distribution of pedons per soil mapping unit in Posses watershed. FO = field observation; AO = additional observations.

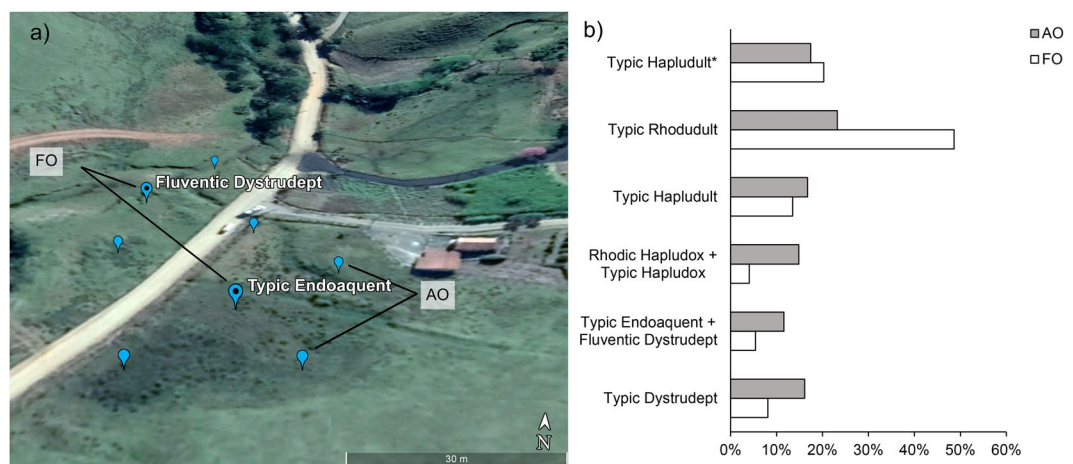


Figure 2. (a) Allocation of additional observations (AO) to field observations (FO) in Google Earth for the soil mapping unit (SMU) Typic Endoaquent + Fluventic Dystrudept. Source: Google Earth Pro 7.3, <https://www.google.com.br/earth/download/gep/agree.html>. (b) Original and final proportions of the training datasets by SMU.

interaction of soil forming factors). It has been considered as the traditional basis of variability capturing in soil surveys, historically based on dual well-accepted paradigms: soil forming factors and soil–landscape relationships (origin in the soil factor equation outlined by Dokuchaev³² and Hilgard³³).

Every point was visited by means of GPS/GNSS navigation and the coordinates of sampled sites were registered again with a nominal accuracy of ~10 m; such location accuracy allowed a spatial resolution mapping of 20 m³⁴. At each point soil profiles morphological description and collection of samples from horizons were carried out as support for soil taxonomic classification³⁵.

Soil mapping units. The soils surveyed in the study area were classified according to Soil Survey Staff³⁵ in: Typic Dystrudept, Typic Humudept, Typic Endoaquent, Fluventic Dystrudept, Rhodic Hapludox, Typic Hapludox, Typic Hapludult (yellow), Typic Rhodudult, Rhodic Kandiudult and Typic Hapludult* (red-yellow) (Table 1), which correspond to Haplic Cambisol, Humic Cambisol, Haplic Gleysol, Fluvic Cambisol, Red-Latosol, Red-Yellow Latosol, Yellow Argisol, Red Argisol, Red Nitosol, and Red-Yellow Argisol in the Brazilian Soil Classification System³⁶, respectively.

Additional observations. Although technical recommendations of soil survey were answered, for an proper algorithm performance, additional observations were created (AO), since the increasing of in-field sampling points is economical unfeasible. In addition, the proportion of sampling in less common landscapes (small proportions) is consequently smaller. In this sense, once the dominant soil mapping unit (SMU) was the Typic Rhodudult taxon (49% of the surveyed locations), this fact resulted in a highly imbalanced dataset (Table 1). To overcome this issue, AO were allocated for the less common soil mapping units by means of our ‘expert knowledge’ and photointerpretation from high-resolution imagery in Google Earth (Fig. 2). From the paradigm that support traditional soil surveys, according to Hudson³⁷ the understanding of soil–landscape paradigm leads to soil–landscape units concepts, which consists of natural terrains resulted from the interaction of soil forming factors. Generally, the more similar two landscape units are, the more similar their associated soil tend to be, and vice-versa. Such concepts guided the choose of AO points.

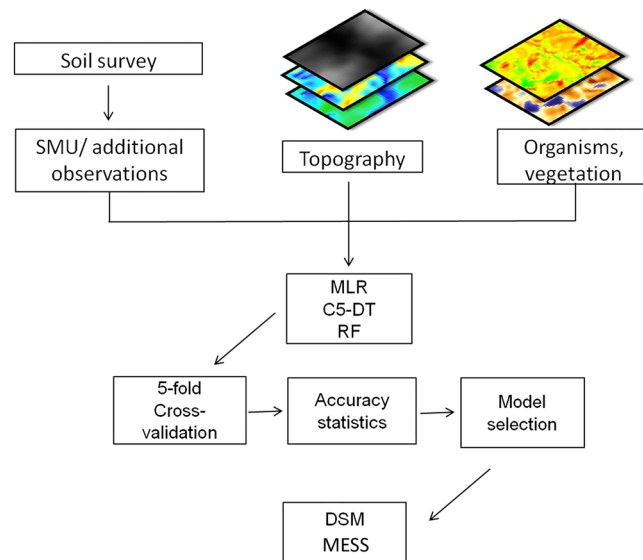


Figure 3. Soil map production flow chart. MLR: multinomial logistic regression, C5-DT: C5 decision tree, RF: random forest, DSM: digital soil map, MESS: multivariate environmental similarity surface.

The procedure of automatic allocation of AO to increase the sample size for model training has been used in some DSM applications with variable improvements in accuracy²⁷ since it depends on the degree of spatial autocorrelation among soil classes, which is specific to different landscapes. Consequently, an automatic approach should be applied prudently. Therefore, we adopted a ‘supervised’ approach based on the following criteria: AO points were prevented from occurring within 20 m of one another (mapping resolution) and within 20 m of previously surveyed pedon (field observation, FO) (Fig. 2a). The AO points were also constrained as to not exceed an imbalance greater than a proportion of 2 (dominant) to 1 (less common) (Fig. 2b). Photointerpretation is a conventional technique used in soil mapping that could be included in DSM operations to increase the representation of the feature-space of environmental covariates by attaching additional observations to the training data^{38,39}.

Soil covariates. The spatial prediction framework was based on the approach proposed by McBratney *et al.*⁵, which is an updated and spatially-explicit implementation of the early concepts of Dokuchaev’s soil forming factors and Jenny’s quantitative interpretation of them⁴. This technique assumes that a soil property or class is a function of a spatial representation of other soil property (s) or a soil-forming factor, namely: climate (c), organisms and vegetation (o), topography (r), parent material (p), time (a), and geographic location (n).

Climate (c) and parent material (p) proxies are not available at the scale of analysis of this study (10¹ km), and the coarse resolution of current information does not represent spatial variation that could be used for digital mapping. Therefore, only the available maps of relief (r) and organisms and vegetation (o) factors were assessed.

The relief factor (r) proxy was a high-resolution ALOS PALSAR digital elevation model (DEM)⁴⁰ accessed through www.asf.alaska.edu. The DEM’s spatial resolution was 12.5 m and was upscaled to the earlier mentioned spatial resolution mapping of 20 m by bilinear interpolation. Subsequently, it was hydrologically corrected, and a total of 28 terrain attributes (see Supplementary Table S1), that are commonly tested for digital soil mapping, were derived.

Sentinel 2 imagery data (top of the atmosphere reflectance), downloaded from scihub.copernicus.eu, were assessed as the organisms and vegetation factor (o). In addition to reflectance data, normalized difference vegetation index (NDVI)^{41,42} and normalized difference water index (NDWI)⁴³ were calculated and evaluated (Supplementary Table S1). Processing of o and r factors was carried out in SAGA GIS software⁴⁴ and the model analysis was run in R⁴⁵.

Model selection and soil map production. Aiming to fit the best spatial association model that relates soil-forming factor proxies and the occurrence of soil mapping units, three models were tested: Multinomial Logistic Regression (MLR), C5 Decision Tree (C5-DT), and Random Forest (RF).

The MLR analysis allows to model categorical responses by fitting linear combinations of a set of covariates (predictors) to the natural logarithm of the odds of every level of a response variable (logit transformation)^{46,47}. MLR has been widely applied and tested for classification and mapping of soil features^{15,48}. Therefore, the set of soil covariates were evaluated as predictors of the probability of occurrence of the soil mapping units. MLR analysis has been implemented following the best subset procedure that involved testing every combination of the soil covariates for the prediction of the soil mapping units. The best MLR model was selected on the basis of the Bayesian Information Criterion (BIC), being that the model with the lowest BIC was selected.

Classification and Regression Trees (CART) are non-parametric and rule-based models that have a tree-structure and are based on the mining of relationships between a target variable and the feature-space of a set of covariates⁴⁹. The C5-DT algorithm (an updated version of C4.5⁵⁰), in a similar way to other CART methods,

Subgroups	C5*	RF*	MLR*	C5 ^c	RF ^d	MLR ^c
	Producer's accuracy					
Typic Dystrudept	44	50	44	14	8	13
Typic Endoaquent + Fluventic Dystrudept	43	53	71	21	16	36
Rhodic Hapludox + Typic Hapludox	35	46	41	0	0	5
Typic Hapludult (yellow)	42	49	39	18	24	21
Typic Rhodudult	34	41	24	43	66	29
Typic Hapludult (red-yellow)	31	39	19	22	18	14
	User's accuracy					
Typic Dystrudept	34	45	43	20	0	17
Typic Endoaquent + Fluventic Dystrudept	37	37	57	10	0	24
Rhodic Hapludox + Typic Hapludox	32	42	33	0	0	5
Typic Hapludult (yellow)	39	49	30	18	22	12
Typic Rhodudult	42	52	32	43	50	39
Typic Hapludult (red-yellow)	31	37	22	22	23	13
OA% ^a	40	49	37	29	39	22
SE% ^b	1.2	1	1.7	0.8	0.7	0.62
Kappa	0.22	0.33	0.20	0.04	0.07	-0.02

Table 2. Accuracy statistics from repeated 5-fold cross-validation. ^aOverall accuracy. ^bStandard error. ^cC5 Decision Tree. ^dRandom Forest. ^eMultinomial logistic regression. *Model including additional point sampling.

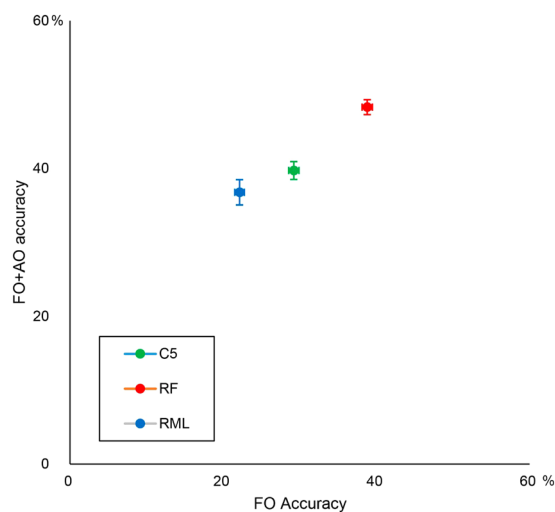


Figure 4. Mean overall accuracy of the tested models and the effects of additional observations (AO accuracy) in relation to field observations (FO accuracy). Bars are standard errors from 5-fold cross-validation after 100 repetitions. C5-DT: C5 decision-tree, RF: random forest, MLR: multinomial logistic regression.

uses a recursive process that relies on the partition of the feature-space and the separation of the observed classes until a stopping criterion is met (e.g. Gini index, Shannon entropy⁴⁹). The performance of C5-DT in soil mapping has been proven as efficient¹⁶, notably when non-linear and complex soilscape relationships are intended to be modeled. A C5-DT model was fitted with the default parameters in the C5.0 R's package⁵⁰.

The RF algorithm is also a CART method that records remarkable performances in DSM^{14,15}; it differentiates from other CART by being an ensemble and bootstrapping method^{51,52}. Since the most sensitive parameter in RF fitting is *mtry*, i.e., the number of candidate covariates tested on each split⁵², a test was run to find the *mtry* value that produced the lesser out-of-bag classification error. Subsequently, in order to facilitate interpretation and eliminate redundant covariates, a subset of the most important covariates, as calculated by the mean decrease in accuracy in the RF output^{51,52}, was selected; then, from this subset, high correlated covariates (Spearman rank correlation >0.85) were removed for subsequent analysis.

Model selection criteria and the effects of additional sampling were evaluated on accuracy statistics (global accuracy, kappa index and producer's and user's accuracies) aggregated by the mean values from a repeated 5-fold cross-validation procedure, which is recommended for imbalanced and relatively small samples⁵³; every fold in this operation precluded the inclusion of the cluster formed by the FO and its associated AO (Fig. 1b,c), being that accuracy statistics estimates are affected by spatial autocorrelation¹⁹. Subsequently, after 100 repetitions, the

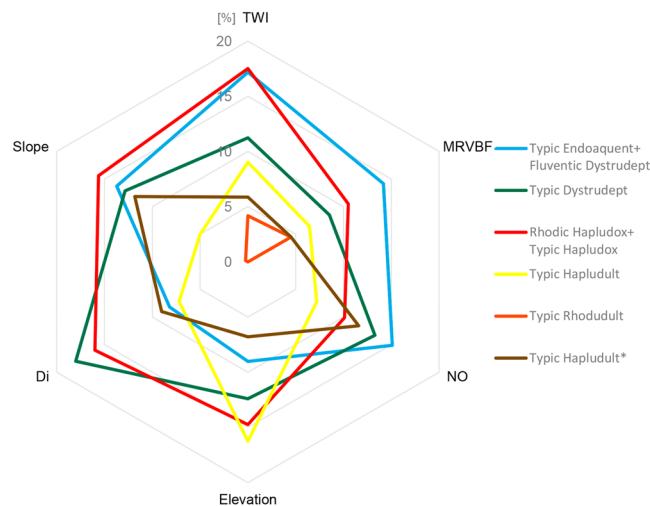


Figure 5. Importance of selected covariates by soil mapping unit. TWI: topographic wetness index; DI: direct insolation; NO: negative openness; MRVBF: multi-resolution index of valley bottom flatness.

model with best accuracy metrics was applied for the production of the digital soil mapping and further soilscape analysis. In order to test the performance of this approach, a comparison with a repeated 5-fold cross-validation procedure using only field observations was performed.

Multivariate environmental similarity surface (MESS). Map quality related to the level of representation of the feature-space by the point-samples (level of extrapolation) was evaluated by calculating the Multivariate Environmental Similarity Surface (MESS), according to Elith *et al.*⁵⁴ (Eq. 1).

$$P(V_i) = \begin{cases} \frac{(p_i - \min_i V_i)}{(\max_i V_i - \min_i V_i)} \times 100, & f_i = 0 \\ 2 \times f_i, & 0 < f_i \leq 50 \\ 2 \times (100 - f_i), & 50 \leq f_i < 100 \\ \frac{(\max_i V_i - p_i)}{(\max_i V_i - \min_i V_i)} \times 100, & f_i = 100 \end{cases} \quad (1)$$

where: $P(V_i)$ is the multivariate similarity (MES) of a point; \min_i = minimum value of the covariate V_i over the reference point set, \max_i = maximum value of variable V_i over the reference point set; p_i = value of the covariate V_i at point P ; f_i = percent of reference points whose value of covariate V_i is smaller than p_i . The MES of P is the minimum of its similarity with respect to each covariate.

MES surface is produced by applying Eq. 1 to every single pixel in the model output. This method quantifies the similarity between the training samples and the selected covariates; values less than zero indicate prediction locations, both in feature and geographic spaces, that have not been accounted for by the training samples. By doing this, it enables to represent spatially the level of extrapolation related to the sampling pattern of the study area. This approach has been used widely in species distribution modeling^{54,55} but its use is still scarce in DSM¹⁴. The evaluated feature-space was that made up by the selected covariates and the training samples of the best model.

The flow chart of the soil map production is shown in Fig. 3.

Results and Discussion

Accuracy assessment and model performance. The performance of the models to predict the soil mapping units registered in the Pesses watershed is summarized in Table 2. Additional point sampling improved the accuracy of all models (Fig. 4).

The models without additional observations were rarely able to predict less-common soils. This finding is consistent with data presented in other works (e.g. Barthold *et al.*⁵⁶; Brungard *et al.*¹⁷; Jafari *et al.*⁵⁷) and it is related to soil classes with very limited presence in the study area, in comparison with the dominating soil classes that are better represented in geographic and feature spaces.

Additional point sampling by photointerpretation enabled to capture specific conditions of occurrence of less-common soils and, in consequence, it was showed a substantial improvement in classification accuracy of the minority soil classes. This knowledge-driven technique could be regarded as an alternative to the data-driven approach and has already been integrated into predictive modeling with remarkable improvements in accuracy^{26,46,58,59}. Nevertheless, there is still a need for a measure of their quality, as it depends only on the expertise of the photo interpreter, which restricts its application.

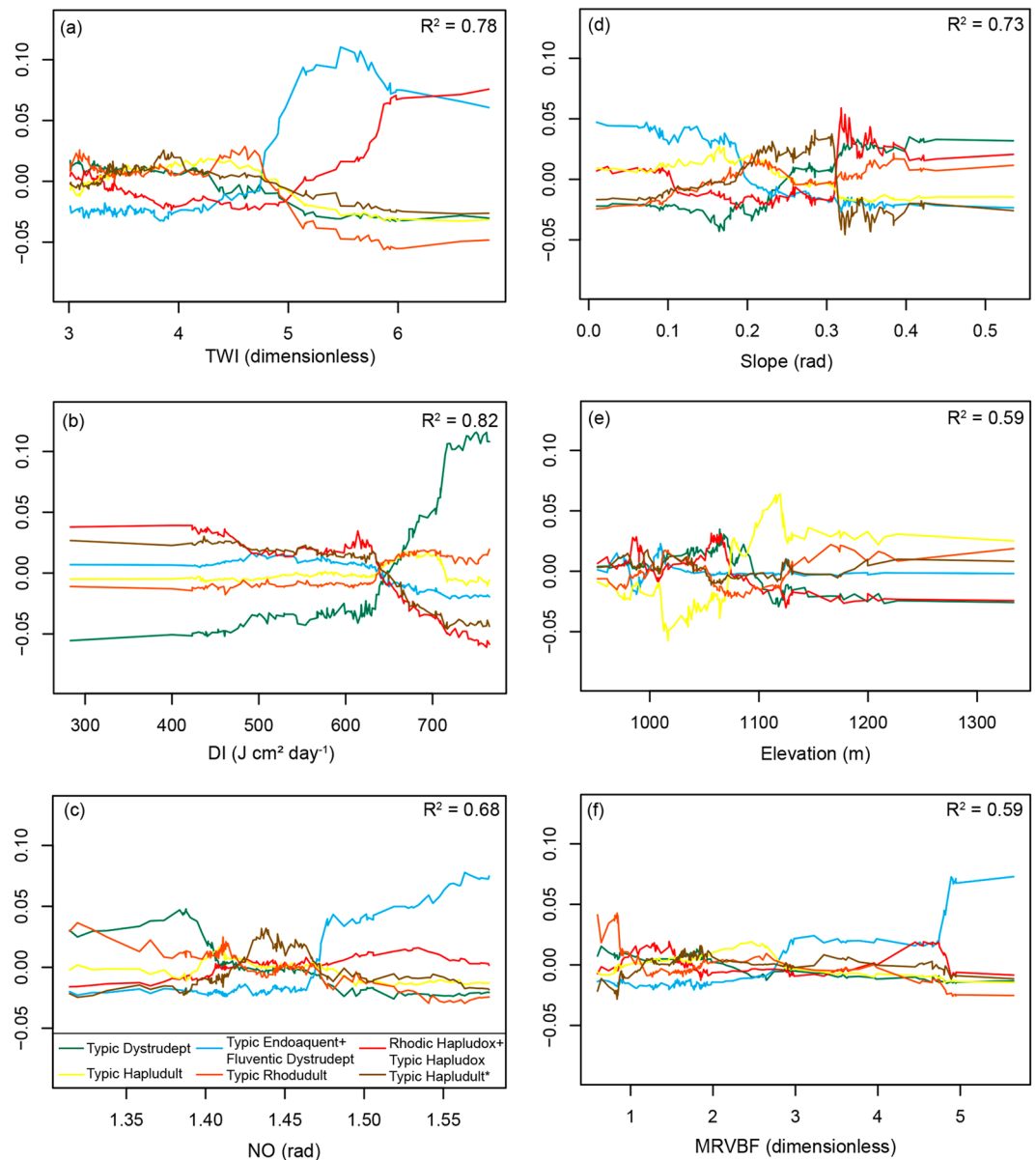


Figure 6. Occurrence signatures of soil mapping units projected on the feature-space of selected covariates. X-axis: covariate value, Y-axis: change of predicted class probability. R-squared values indicate goodness-of-fit between feature and prediction space projections (solid curves). TWI: topographic wetness index; DI: direct insolation; NO: negative openness; MRVBF: multiresolution index of valley bottom flatness.

Accuracy statistics from the repeated 5-fold cross-validation demonstrates that the RF* model consistently outperformed the other models (overall accuracy = 49%, kappa index = 0.33), while MLR had the lowest overall accuracy and kappa index (Table 2 and Fig. 4). For the purposes of digital soil mapping, it has been frequently observed that complex models, such as RF, are better classifiers than generalized linear models such as MLR^{14,17,60}. With regards to the accuracy rank in Table 2 (RF > C5-DT > MLR), it seems that the most sophisticated model fitted more complex relationships, as suggested by Heung *et al.*⁶⁰. Since MLR is a generalization of the logistic regression, it is highly dependent on the presumed sigmoidal model for class discrimination⁴⁶, which makes it the less flexible model. The C5-DT is a data-driven and non-parametric discrimination model, and, considering its single tree structure, it tends to show worse performances when compared to more sophisticated models. RF is also a data-driven and non-parametric model; however, its bootstrapping and bagging features are more efficient in revealing consistent and complex patterns⁶⁰, especially when dealing with complex spatial associations, as occurred in this watershed, helping to explain its performance as the most accurate model.

The spatial association model obtained from RF* had also the highest producer's and user's accuracies (Table 2). In order to support the interpretation of the RF* output, we applied the 'forest floor' method⁶¹, which enables to visualization of the link between feature and prediction spaces, which we refer to as occurrence signatures.

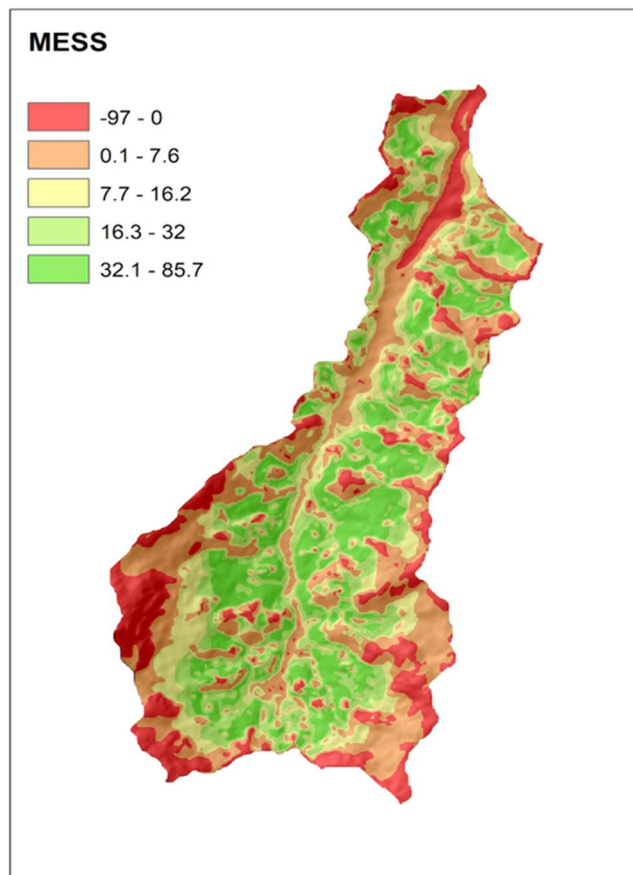


Figure 7. Multivariate environmental similarity surface (MESS) of the chosen model (random forest with additional observations), extrapolation level increase along with negative values.

Analysis of covariates. The most important and the less correlated covariates, as determined by the mean decrease in accuracy (MDA) in the RF* model (MDA < 15%, see Supplementary Fig. S3) and the correlation criterion ($\rho \leq 0.85$, see Supplementary Fig. S2), are presented in Fig. 5.

The topography proxy (r), as demonstrated by the mean decrease in accuracy (variable importance rank) in the RF* output, correlation analysis, and classification accuracies, had the best set of predictors, remarkably: SAGA topographic wetness index (TWI), slope, elevation, direct insolation (DI) and the multiresolution index of valley bottom flatness (MRVBF) (Fig. 5); while the vegetation and organisms (o) covariates showed very low importance (Supplementary Fig. S3) and were, consequently, removed from the analysis.

Regarding the global discrimination of the soil mapping units, the topographic wetness index (TWI) registered the highest importance score (Fig. 5). This terrain attribute is a substitute measure of water flux in the landscape, since it shows the tendency of a site to be water saturated and the possible drainage systems of a watershed⁶². TWI has often been reported as a potential predictor in digital soil mapping^{56,63–66}.

The TWI stands out as a specific predictor for the Typic Endoaquent + Fluventic Dystrudept mapping unit. Its influence on the discrimination of this soil mapping unit is verifiable for values above 4.5, where an increase of 10% in probability of occurrence is observed (Fig. 6a). The occurrence of Rhodic Hapludox + Typic Hapludox mapping unit is also influenced by TWI. These two soils classes could not be distinguished in the final map because they occur randomly distributed in the same landscape position. Silva *et al.*⁶⁶ also reported the influence of TWI on the distinction of Oxisols in Brazil. Higher TWI values (>5, i.e., poorly drained areas), supported also better discrimination among the Typic Rhodudult, Typic Endoaquent + Fluventic Dystrudept, and Rhodic Hapludox + Typic Hapludox mapping units (Fig. 6a).

The slope was the next important covariate for global classification. Flatter slopes are associated with a higher probability of occurrence of Typic Endoaquent + Fluventic Dystrudept mapping unit (Fig. 6d), which is expected to occur in flat slopes along with high values of TWI⁴⁸. These two soil classes could not be separated in the final map due to their intricate occurrence pattern in the watershed landscape. Slope values around 30 radians influenced the discrimination between Typic Dystrudept and Rhodic Hapludox + Typic Hapludox (Fig. 6d). The slope covariate also registered the highest importance regarding Typic Hapludult*, where slope values above 30 radians showed an inflection point in the occurrence signature of this soil mapping unit (Fig. 6d).

Direct insolation (DI), ranked as third in importance for global classification, represents a pixel-based calculation for solar radiation⁶⁷. This covariate has received little investigation in tropical conditions^{68,69}. The DI covariate was a remarkable predictor for Typic Dystrudept. Values above 600 J cm² day⁻¹ were associated with an

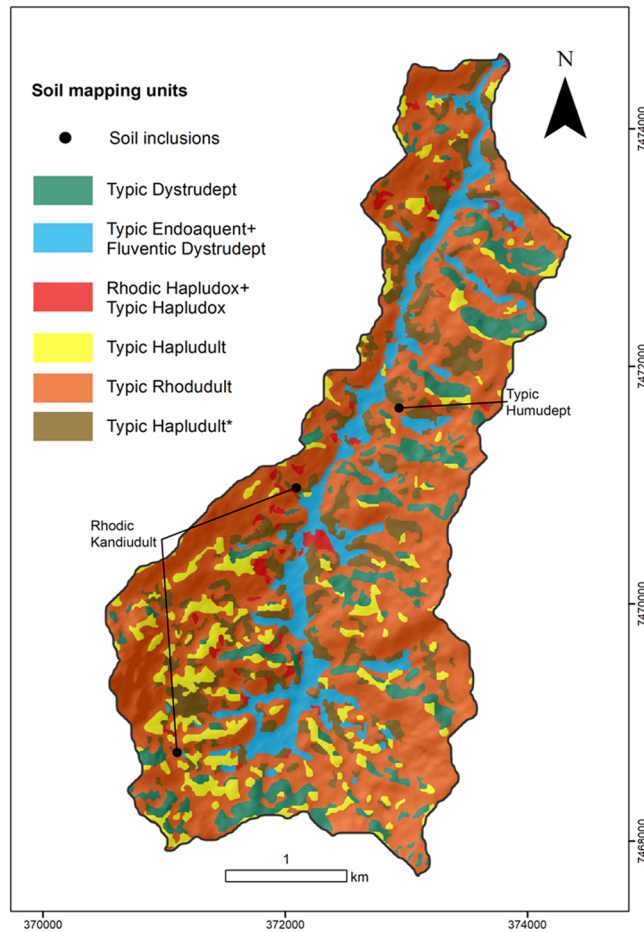


Figure 8. Soil mapping units distribution in the watershed by the RF* model.

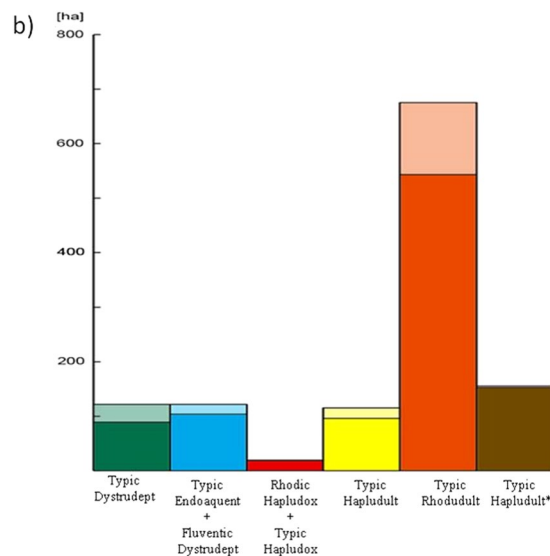


Figure 9. Extrapolation level of the digital soil map (lighter colors) by soil mapping unit.

increase in the probability of occurrence of this mapping unit, while lower values were associated to the occurrence of the Rhodic Hapludox + Typic Hapludox and Typic Hapludult* (Fig. 6b). The elevation covariate was relevant for Typic Hapludult distinction; Fig. 6c presents the occurrence signature of this soil mapping unit, where a distinctive pattern is observed between 1000 and 1200 m range.

While the RF* model satisfactorily discriminated the soil mapping units, a certain degree of mixture is still present in model outputs (Fig. 6). This could be an effect of the dominant soil and its wide range of occurrence conditions in the watershed. On the other hand, mapping units with a lesser degree of ‘impurity’ are those related to specific conditions across the landscape (Typic Dystrudept and Typic Endoaquent + Fluventic Dystrudept).

The level of detail in the topographic data analyzed, both spatial and informative, allowed the identification of specific relief-soil interactions. Topography, as a surficial hydrologic driver, could also reflect other soil-forming factors that have a larger scale domain, by modifying locally the effects of bioclimate and parent material⁶.

Multivariate environmental similarity surface (MESS). In Fig. 7 it is possible to see the spatial distribution of the level of representation of the model prediction by the training data. The MESS measures the similarity of any given point to a reference set of points, with respect to the chosen predictor variables in the RF* model. Negative values indicate the presence of at least one variable out of reach of the reference points in feature-space, while positive values indicate greater similarity to the set of reference points. The higher the score, the more common the point is and more reliable is the prediction⁵⁴. This feature enables to interpret the MESS as a quality measure of the sample configuration. The negative values are located in areas of lower accessibility (high altitudes and sharp slopes) and near to the watershed outlet (17.5% of the watershed area), whereas 82.5% of the area registered positive and, therefore, more reliable model outputs. It’s worth pointing out that the MESS method could be applied regardless of the chosen model, since all of three algorithms evaluated in this study offers the possibility for the selection of a subset of ‘most important covariates’, which could be of paramount importance to assess the map quality related to point-sampling, particularly when dealing with constrained sampling.

The map showing the spatial distribution of the different soil mapping units predicted in the watershed is presented in Fig. 8. The majority of the soils in the Poses watershed were predicted as Typic Rhodudult, which proportion is in agreement with the distribution derived from the field measurements and additional point sampling. Figure 9 highlights the extrapolation level of each soil mapping unit. Typic Dystrudept (120 ha) and Typic Rhodudult (672 ha) mapping units have 27% and 20% of their areas in less reliable regions, respectively; while the other soil mapping units, with mapping areas of 120 ha - Typic Endoaquent + Fluventic Dystrudept, 156 ha - Typic Hapludult*, 108 ha - Typic Hapludult, and 24 ha - Rhodic Hapludox + Typic Hapludox, do not show an expressive uncertain area (<15%).

Conclusions

A comparison among different models and the effects of additional point sampling for digital mapping of less-common soil classes, based on the spatial association with soilscape covariates, was performed. Tested models were Multinomial Logistic Regression, C5 Decision-Tree, and Random Forest. Additional point sampling retrieved by photointerpretation was necessary to improve the prediction performance of each model. The accuracy metrics were found to be higher for the Random Forest model with additional point sampling method, resulting in the best model. At the spatial resolution analyzed (20 m), the terrain-attribute covariates (relief data), related to surficial water distribution and direct insolation, were the most efficient to discriminate among soil mapping units and for producing a feasible digital soil map.

References

- Brevik, E. C. *et al.* Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma* **264**, 256–274 (2016).
- Brevik, E. C. & Hartemink, A. E. Early soil knowledge and the birth and development of soil science. *CATENA* **83**, 23–33 (2010).
- Keys, K., Sterling, S. M. & Guan, Y. Using historic soil survey data to map water erosion hazard for land-use planning in Nova Scotia. *Can. J. Soil Sci.* **95**, 299–304 (2015).
- Jenny, H. *Factors of Soil Formation: A System of Quantitative Pedology.* (McGraw-Hill, 1941).
- McBratney, A., Mendonça Santos, M. & Minasny, B. On digital soil mapping. *Geoderma* **117**, 3–52 (2003).
- Miller, B. A. & Schaetzl, R. J. History of soil geography in the context of scale. *Geoderma* **264**, 284–300 (2016).
- Shi, R., Long, R., Dekett, R. & Phillip, R. Integrating different types of knowledge for digital soil mapping. *Soil Sci. Soc. Am. J.* **73**, 1682–1692 (2009).
- Zhu, Q. & Lin, H. S. Comparing Ordinary Kriging and Regression Kriging for Soil Properties in Contrasting Landscapes. *Pedosphere* **20**, 594–606 (2010).
- Zhu, A. X. A similarity model for representing soil spatial information. *77*, 217–242 (1997).
- Bui, E. N. Soil survey as a knowledge system. *Geoderma* **120**, 17–26 (2004).
- Hudson, B. D. The Soil Survey as Paradigm-based Science. *Soil Sci. Soc. Am. J.* **56**, 836 (1992).
- Godinho Silva, S. H., Owens, P. R., Duarte de Menezes, M., Reis Santos, W. J. & Curi, N. A Technique for Low Cost Soil Mapping and Validation Using Expert Knowledge on a Watershed in Minas Gerais, Brazil. *Soil Sci. Soc. Am. J.* **78**, 1310 (2014).
- Cook, S. E., Corner, R. J., Grealish, G., Gessler, P. E. & Chartres, C. J. A Rule-based System to Map Soil Properties. *Soil Sci. Soc. Am. J.* **60**, 1893 (1996).
- Camera, C. *et al.* A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. *Geoderma* **285**, 35–49 (2017).
- Jeune, W. *et al.* Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil Classes in Western Haiti. *Rev. Bras. Ciência do Solo* **42**, 1–20 (2018).
- Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B. & Triantafyllis, J. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma* **253–254**, 67–77 (2015).
- Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A. & Edwards, T. C. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma* **239–240**, 68–83 (2015).
- Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
- Millard, K. & Richardson, M. On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sens.* **7**, 8489–8515 (2015).
- Grunwald, S. In *Environmental soil-landscape modeling: Geographic information technologies and pedometrics* (ed. Grunwald, S.) 3–36 (Taylor & Francis, 2006).
- Minasny, B. & McBratney, A. B. Digital soil mapping: A brief history and some lessons. *Geoderma* **264**, 301–311 (2016).

22. Menezes, M. D., de, Silva, S. H. G., Mello, C. R., de, Owens, P. R. & Curi, N. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. *Sci. Agric.* **71**, 316–323 (2014).
23. Meier, M., Souza, E., de, Francelino, M. R., Fernandes Filho, E. I. & Schaefer, C. E. G. R. Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area. *Rev. Bras. Ciência do Solo* **42**, 1–22 (2018).
24. Chagas, CdaS. *et al.* Data mining methods applied to map soil units on tropical hillslopes in Rio de Janeiro, Brazil. *Geoderma Reg.* **9**, 47–55 (2017).
25. Giasson, E., Clarke, R. T., Inda Junior, A. V., Merten, G. H. & Tornquist, C. G. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. *Sci. Agric.* **63**, 262–268 (2006).
26. Menezes, M. D. D., Silva, S. H. G., Owens, P. R. & Curi, N. Digital soil mapping approach based on fuzzy logic and field expert knowledge. *Ciência e Agrotecnologia* **37**, 287–298 (2013).
27. Pelegrino, M. H. P. *et al.* Mapping soils in two watersheds using legacy data and extrapolation for similar surrounding areas. *Ciência e Agrotecnologia* **40**, 534–546 (2016).
28. Alvares, C. A., Stape, J. L., Sentelhas, P. C., de Moraes Gonçalves, J. L. & Sparovek, G. Köppen's climate classification map for Brazil. *Meteorol. Zeitschrift* **22**, 711–728 (2013).
29. Regional, C.-S. geológico do B. Mapa geológico do estado de Minas Gerais. at, <http://rigeo.cprm.gov.br/jspui/handle/doc/5016> (2003)
30. Pereira, P. H., Cortez, B. A., Trindade, T. & Mazochi, M. N. Conservador das Águas, 5 anos (2010).
31. IBGE. *Manual Técnico de Pedologia*. (Coordenação de Recursos Naturais e Estudos Ambientais, 2015).
32. GLINKA, K. D. & Marbut, C. F. The great soil groups of the world and their development. *Nat. Publ. Gr.* **126** (1927).
33. Jenny, H. E. W. Hilgard and the Birth of Modern Soil Science. *Soil Sci.* **93**, 73 (1961).
34. Hengl, T. Finding the right pixel size. *Comput. Geosci.* **32**, 1283–1298 (2006).
35. Soil Survey Staff. *Keys to Soil Taxonomy*. (USDA-Natural Resources Conservation Service). at, https://www.nrcs.usda.gov/wps/PA_NRCSCConsumption/download?cid=stelprdb1252094&ext=pdf (2014).
36. dos Santos, H. G. *et al.* *Sistema Brasileiro de Classificação de Solos*. (Embrapa). at <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/181677/1/SiBCS-2018-ISBN-9788570358172.epub> (2018).
37. Hudson, H. D. The Soil Survey as Paradigm-based Science. *Soil Sci. Soc. Am. J.* **56**, 836–841 (1992).
38. Corcoran, J., Knight, J. & Gallant, A. Influence of Multi-Source and Multi-Temporal Remotely Sensed and Ancillary Data on the Accuracy of Random Forest Classification of Wetlands in Northern Minnesota. *Remote Sens.* **5**, 3212–3238 (2013).
39. Abbaszadeh Afshar, F., Ayoubi, S. & Jafari, A. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma* **315**, 36–48 (2018).
40. Tadono, T. *et al.* Generation of the 30 m-mesh global digital surface model by alos prism. *Isprs - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLI-B4**, 157–162 (2016).
41. Rouse, J. W. J., Haas, R. H., Schell, J. A. & Deering, D. W. Monitoring vegetation systems in the Great Plains with ERTS. In *3rd Earth Resource Technology Satellite (ERTS) Symposium* 48–62 (1974).
42. Tucker, C. J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **8**, 127–150 (1979).
43. Gao, B.-C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **58**, 257–266 (1996).
44. Conrad, O. *et al.* System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Development* 1991–2007, <https://doi.org/10.5194/gmd-8-1991-2015> (2015).
45. R-Core-Team. R: A language and environment for statistical computing. at, www.R-project.org (2017).
46. Kempen, B., Brus, D. J., Heuvelink, G. B. M. & Stoorvogel, J. J. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma* **151**, 311–326 (2009).
47. Lane, P. W. Generalized linear models in soil science. *Eur. J. Soil Sci.* **53**, 241–251 (2002).
48. Campling, P., Gobin, A. & Feyen, J. Logistic Modeling to Spatially Predict the Probability of Soil Drainage Classes. *Soil Sci. Soc. Am. J.* **66**, 1390 (2002).
49. Strobl, C., Malley, J. & Tutz, G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **14**, 323–348 (2009).
50. Quinlan, J. R. C4.5: *Programs for machine learning*. Morgan Kaufmann 5 (1993).
51. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
52. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
53. Kuhn, M. & Johnson, K. *Applied predictive modeling*. (Springer, 2013).
54. Elith, J., Kearney, M. & Phillips, S. The art of modelling range-shifting species. *Methods Ecol. Evol.* **1**, 330–342 (2010).
55. Zurell, D., Elith, J. & Schröder, B. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Divers. Distrib.* **18**, 628–634 (2012).
56. Barthold, F. K. *et al.* Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. *J. Arid Environ.* **88**, 194–205 (2013).
57. Jafari, A., Khademi, H., Finke, P. A., Van de Wauw, J. & Ayoubi, S. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. *Geoderma* **232–234**, 148–163 (2014).
58. Hengl, T. *et al.* Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutr. Cycl. Agroecosystems* **109**, 77–102 (2017).
59. Rossiter, D. G., Zeng, R. & Zhang, G.-L. Accounting for taxonomic distance in accuracy assessment of soil class predictions. *Geoderma* **292**, 118–127 (2017).
60. Heung, B. *et al.* An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* **265**, 62–77 (2016).
61. Welling, S. H., Refsgaard, H. H. F., Brockhoff, P. B. & Clemmensen, L. H. Forest Floor Visualizations of Random Forests. (2016).
62. Beven, K. & Kirkby, N. A physically based variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* **43**–69 (1979).
63. Bagheri Bodaghabadi, M. *et al.* Using Canonical Correspondence Analysis (CCA) to identify the most important DEM attributes for digital soil mapping applications. *CATENA* **86**, 66–74 (2011).
64. Machado, I. R. *et al.* Spatial Disaggregation of Multi-Component Soil Map Units Using Legacy Data and a Tree-Based Algorithm in Southern Brazil. *Rev. Bras. Ciência do Solo* **42**, 1–14 (2018).
65. Nauman, T. W. & Thompson, J. A. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma* **213**, 385–399 (2014).
66. Silva, S. *et al.* Proximal Sensing and Digital Terrain Models Applied to Digital Soil Mapping and Modeling of Brazilian Latosols (Oxisols). *Remote Sens.* **8**, 614 (2016).
67. Boehler, J. & Antonic, O. In *Geomorphometry - Concepts, Software, Applications* 195–226 (Elsevier, 2009).
68. Chagas, C., da, S., Fernandes Filho, E. I. & Bhering, S. B. Relação entre atributos do terreno, material de origem e solos em uma área no noroeste do estado do Rio de Janeiro. *Soc. Nat.* **25**, 147–162 (2013).
69. Ferreira, F. P., Azevedo, A. C., Kanieski, A. J., Girelli, D. & Pedrotti, J. Solar Exposure and Soil Properties in Santa Maria – RS. *Rev. Bras. Agrocência* **11**, 377–381 (2005).

Acknowledgements

This research was funded by Coordination of Superior Level Staff Improvement – CAPES – Finance Code 001, National Council of Technological and Scientific Development – CNPq (Process nº306511/2017-7 e 202938/2018-2) and Minas Gerais State Research Foundation – FAPEMIG (Process nºAPQ-00802-18 and CAG-APQ-01053-15).

Author Contributions

Marx Leandro Naves Silva managed the activities of the study. Soil survey and sample collection were carried out by Bárbara Pereira Christofaro Silva, Michele Duarte de Menezes and Nilton Curi. Data analyses were conducted by Fabio Pomar Avalos and Bárbara Pereira Christofaro Silva. All authors reviewed and contributed to the writing of the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-50376-w>.

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019