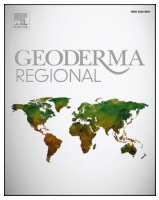
Geoderma Regional 33 (2023) e00645

Contents lists available at ScienceDirect

Geoderma Regional

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



Introducing a map of soil base cation concentration, an ecologically relevant GIS-layer for Amazonian forests

G. Zuquim [a,b,*], J. Van doninck [b,c], P.P. Chaves [b], C.A. Quesada [d], K. Ruokolainen [b], H. Tuomisto [b]

[a] Section for Ecoinformatics and Biodiversity, Department of Biology, Aarhus University, Denmark
[b] Department of Biology, University of Turku, Finland
[c] Department of Integrative Biology, Michigan State University, USA
[d] Coordination of Environmental Dynamics, National Institute for Amazonian Research (INPA), Manaus, Brazil

ARTICLE INFO

Keywords:
Nutrient concentration
Edaphic conditions
Species distribution models
Random forest
Soilgrids
Indicator species
Soil mapping
Histosols
Acrisols
Ferrasols
Arenosols
Machine learning
Digital map
Tropical forests
Multiple soil classes

ABSTRACT

Soil maps are crucial for habitat and species distribution modeling under present and future conditions, thereby providing relevant background information for conservation planning. In Amazonia, soil conditions are highly heterogeneous, which has important implications for the distribution and dynamics of the area's exceptional biodiversity. Unfortunately, available soil maps for this region suffer from inaccuracies and lack of ecologically relevant variables. Here, we develop a map of the sum of exchangeable base cation concentration (SB) in the surface soil by applying machine learning to a comprehensive set of over 10,000 field data points of SB values directly measured from soil samples or inferred using indicator plant species occurrences. As predictors, we used rasters of soil type probabilities, elevation, biomass and reflectance values from Landsat satellite images. Random Forest (RF) models were trained and tested using two different cross-validation strategies. We also assessed in which areas the map was more reliable using the area of applicability approach and compared the results with two other soil layers. The best predictors of SB variation were Landsat bands 7, 4 and 3, elevation, and probability of Histosols. The regional patterns observed across Amazonia were consistent with current geological understanding; lower SB values tended to occur in central Amazonian soils and higher values in western Amazonian soils, with considerable variation within each region. The model was found applicable over most of the Amazonian biome, especially in non-inundated (*terra-firme*) forest, but not over coastal areas, floodplains of major rivers and wetlands, which were poorly represented in the training data. Our new SB map over performed previous SB map and represent an accurate and ecologically meaningful variable. It is available as a digital GIS layer and can be used in habitat mapping and in modeling the current or future distributions of biological communities and species. This will advance general understanding of Amazonian biogeography and help in conservation planning.

1. Introduction

Soils are central for ecosystem functioning, as they define the growing conditions for organisms and regulate the productivity of habitats. In Amazonian forests, edaphic conditions shape species composition of plant and animal communities (Figueiredo et al., 2018; Dambros et al., 2020; Tuomisto et al., 2003a, 2003b, Tuomisto et al., 2016), as well as forest structure and dynamics (Dalagnol et al., 2021; Heinrich et al., 2021), which are related to forest productivity and carrying capacity. Therefore, maps of relevant soil properties could greatly facilitate understanding of the rainforest ecosystem and the spatial distributions of species in it, providing relevant background information for conservation planning. However, scant availability of information on the spatial heterogeneity of soil conditions remains a major challenge in Amazonia, which is both vast and data-poor.

Considerable effort has been invested in producing digital soil maps with global coverage (Dijkshoorn et al., 2005; Hengl et al., 2014, 2021), but over Amazonia they have problems with both accuracy and thematic relevance (Moulatlet et al., 2017). Out of the available soil variables, the most popular ones in ecological studies have been Cation Exchange Capacity (CEC) and soil types (e.g. Levis et al., 2017; Poorter et al., 2015). However, CEC only measures the capacity of soils to bind cations, and gives no indication on how much of that capacity is used up by the potentially toxic aluminum and how much by actual nutrient cations

* Corresponding author at: Department of Biology, University of Turku, Finland
E-mail address: gabriela.zuquim@utu.fi (G. Zuquim).

https://doi.org/10.1016/j.geodrs.2023.e00645
Received 29 August 2022; Received in revised form 26 April 2023; Accepted 3 May 2023
Available online 4 May 2023
2352-0094/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

(Quesada et al., 2011). Soil types, in turn, are defined on the basis of other characteristics, and can be very heterogeneous in their nutrient concentrations (Moulatlet et al., 2017). Moreover, soil properties can change considerably over short distances (Luizão et al., 2004; Quesada et al., 2011; Tuomisto et al., 2003b, 2016) but field data points on soils are sparse (Hengl et al., 2014; Zuquim et al., 2019). This is a major concern, because interpolation methods rely either on spatially dense sampling (e.g. ordinary kriging) or on representative sampling across a set of informative predictors (e.g. regression and machine learning approaches).

To alleviate paucity in field data, surrogates can be used. An established procedure is to use species composition of biotic communities to deduce values of environmental variables for which direct measurements are not available (Birks et al., 2010). Known as the indicator species approach, this has been applied for decades especially in vegetation science (Cajander, 1926; Ellenberg et al., 1992) and paleoecology (ter Braak and Juggins, 1993). In Amazonian forests, ferns have been found to be a good indicator group for the sum of base (SB) cation concentration in the soil (Cárdenas Ramírez et al., 2021; Tuomisto et al., 2003a; Sirén et al., 2013; Suominen et al., 2013). In fact, Zuquim et al. (2014) reported an R^2 of 0.75 in predictions of the concentration of SB of a focal place using fern community composition. Based on these findings, a coarse-resolution map (~11 km pixel size) of SB was produced by combining values measured from soil samples with values estimated using indicator species (Zuquim et al., 2019). Including the indicator species data led to a 12-fold increase in the number of input data points and greatly improved the resulting maps (Zuquim et al., 2019). However, the sampling was still not dense enough for ordinary kriging to adequately capture the high patchiness in soil properties that is typical for Amazonia (Quesada et al., 2011).

Our aim here is to use the best available field data and relevant predictor variables to map the concentration of soil base cations across Amazonia. The applied framework and the resulting SB map are relevant for a wide range of purposes from species distribution modeling and biogeographical hypothesis testing to conservation planning and assessing distribution of species in the present and in the future (Zuquim et al., 2020).

2. Material and methods

2.1. Study area

The study area covers lowland (< 600 m) Amazonian forests, broadly defined as the tropical rainforests of the Amazon and Orinoco basins and adjacent forested ecoregions. We excluded deforested areas as well as savannas and other non-forest ground cover types.

2.2. Sum of Bases (SB) data

Here we define Sum of Bases (SB) as the sum of the concentrations of calcium (Ca^{++}), magnesium (Mg^{++}) and potassium (K^{+}), expressed in cmol(+)/kg, in the surface soil (samples taken no deeper than 30 cm below the surface of mineral soil). We did not include sodium (Na^{+}), because its concentrations were below the detection limit in some of the laboratories used by the data providers. Na is a base cation but not an essential plant nutrient, so its omission is rather inconsequential for the applicability of the map.

We obtained geolocated SB values from two kinds of field data. Firstly, direct measurements from soil samples were available from several sources. These were the Harmonized World Soil Database v1.2 (HWSD) (Nachtergaele et al., 2012), the Brazilian national database (Cooper et al., 2005), and four databases produced by different research groups: the Amazon Research Team of University of Turku (www.utu.fi/amazon), the Brazilian Biodiversity Research Program (www.ppbio.org), the Aarhus University (Balslev et al., 2019) and the Amazon Forest Inventory Network (www.rainfor.org). These datasets contained a total of 6120 SB values.

Secondly, we downloaded a dataset (https://doi.org/10.5281/zenodo.2585607) containing fern-derived georeferenced SB values obtained by associating fern species occurrences from GBIF (www.gbif.org) with the SB optimum values of the species (Zuquim et al., 2019). The estimated SB values were log(10)-transformed and the coordinates of the data points were converted to decimal degrees and rounded to four decimals. The SB values of data points with identical coordinates were then averaged to obtain a single fern-derived SB estimate for each locality. Averaging reduced the effect of the fact that collecting intensity varies drastically: some localities can have dozens of fern observations while others have just one. Averaging reduced the number of fern-derived input data points from 14,313 to 4287.

2.3. Predictor data layers

As predictors, we used a set of 37 Amazon-wide digital data layers. Surface reflectance values of five bands in the visible and infrared wavelengths were obtained from a basin-wide Landsat TM/ETM+ composite (Van doninck and Tuomisto, 2018; Van doninck and Tuomisto, 2019). The composite was produced using a large number of multitemporal observations per pixel (Van Doninck and Tuomisto, 2017a and applying a correction for reflectance anisotropy that was calibrated for tropical forests (Van doninck and Tuomisto, 2017b). Therefore, it has a high radiometric consistency when compared to other products remote sensing over Amazonia (Van doninck and Tuomisto, 2018). This is essential, because atmospheric noise and biased scattering of light can cause larger pixel value differences within one Landsat scene than is observed between edaphically different forest types in Amazonia (Higgins et al., 2015; Toivonen et al., 2006; Muro et al., 2016). The composite contains Landsat bands 2, 3, 4, 5 and 7 at one arcsecond resolution (approximately 30 m near the equator). To extract values for modeling, we centered a square window of 15 × 15 pixels (approximately 450 m by 450 m) on the coordinates of each input data point and took the median value separately for each band after having excluded non-forest pixels.

For the other 32 predictor variables, we extracted the value corresponding to the pixel containing the input data point. The probabilities of thirty WRB soil types (World Reference Base for Soil Resources) were extracted from SoilGrids at 250-m resolution (https://files.isric.org/soilgrids/latest/data/wrb/; accessed on 24 Nov 2021). A soil type is a soil taxonomic unit with unique pedogenesis and other characteristics, and although it is not directly indicative of SB, it can be a relevant predictor in the models. Since soil properties often covary with elevation, we extracted elevation above sea level from the digital elevation model at 30-m resolution derived from Shuttle Radar Thematic Mission (SRTM; Rabus et al., 2003). Finally, since biomass and edaphic characteristics in Amazonia are related (Laurance et al., 1999), we also used Aboveground Biomass data at 300-m resolution (Spawn et al., 2020) - downloaded from https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1763 in January 2023).

2.4. Analyses

Data analyses were run with a total of 10,407 input data points, of which 59% obtained the response variable from direct laboratory analyses of soil samples and 41% from averaged optima of indicator species. All analyses used log(10)-transformed SB values as the response variable.

We used Random Forest models (RF; Breiman, 2001) to predict SB values across Amazonia. We first built an RF model with 500 trees using all 37 predictors. For the evaluation of model performance, we used 10-fold cross-validation (CV), which consists of iteratively dividing the dataset into 10 random folds and using nine folds for modeling and one fold to test the model. Given that the estimation of SB using ferns as indicators may be less accurate than measuring SB directly from soil

samples, we used two different approaches to selecting the testing set for model evaluation (Fig. 1). In the first approach, the input data points were randomly divided into a training set and a testing set independently of the origin of the SB data (hereafter, referred to as the All data CV). In the second approach, division to a training set and a testing set was otherwise similar, but input points whose SB values originated from fern data were excluded from the testing set and only input points with directly measured SB values were used (stratified CV). Models were evaluated on the basis of R^2, Root Mean Squared Error (RMSE) and concordance correlation coefficient (ccc), each averaged over the ten runs using different folds.

The importance of each variable was assessed by randomly permuting its values and comparing model performance when using the permuted vs. original variable values. The bigger the difference, the more important the variable. This was done separately using the All data CV approach and Stratified CV approach of model testing. Finally, we re-run RF using only the variables with importance values ≥25 to optimize computational time and avoid low variance in the final model evaluations.

To produce a soil SB map, we projected the final RF model over all Amazonia. Given the estimated georeferencing accuracy of the field data, the high local variability of Landsat reflectance values among neighboring pixels and the large area covered, this was done using the same pixel size as the Landsat reflectance data extraction (approximately 450 m). The Landsat composite was converted from 30 m to 450 m this resolution through a median value aggregation. The soil type probability layers were reprojected from 250 m to 450 m resolution using bilinear interpolation. To assess spatial variation in model reliability, we estimated the Dissimilarity Index (DI) for each pixel. This is the distance between the focal pixel and the most similar training data

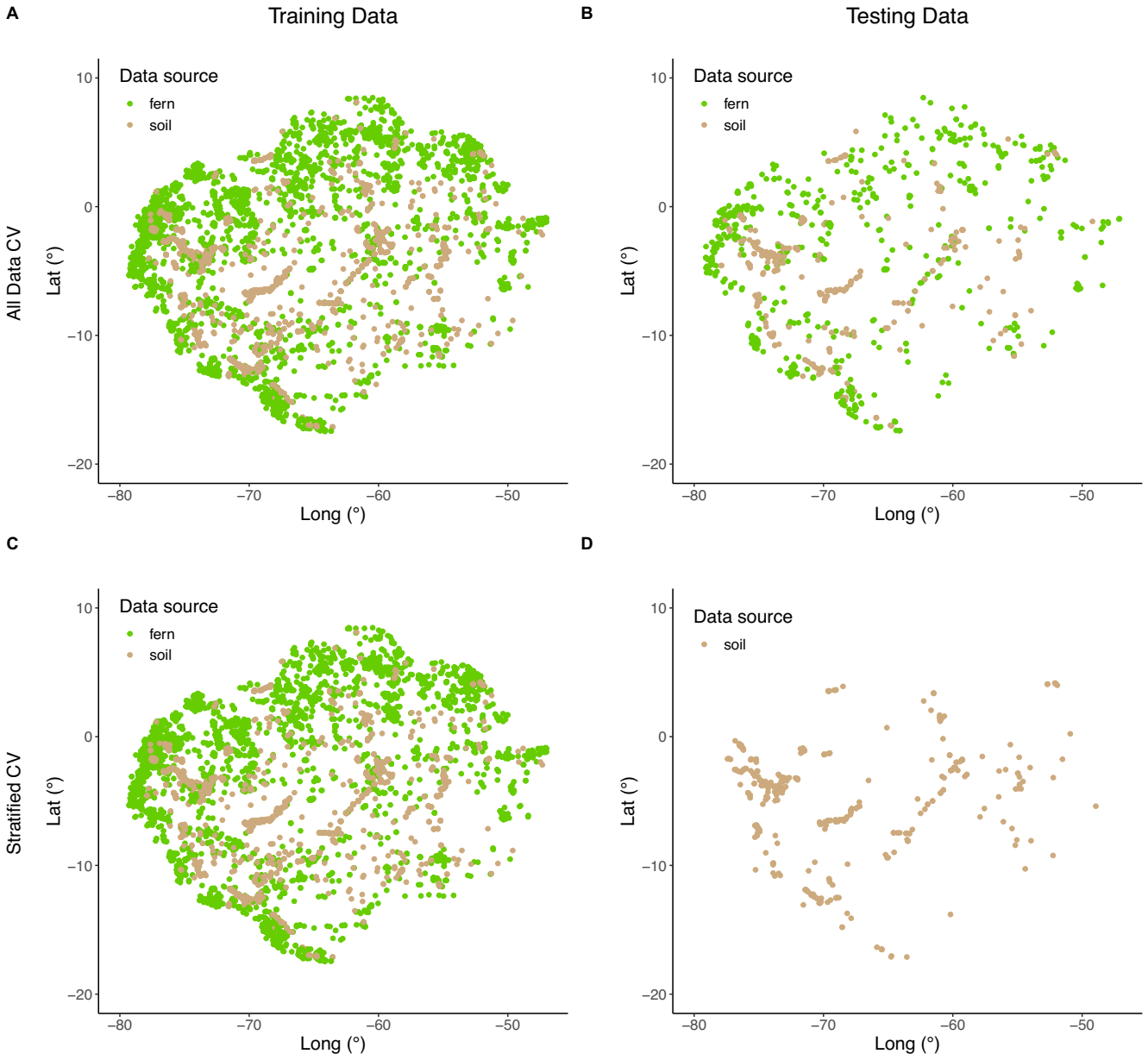


Fig. 1. Distributions of input data points for the training and testing of soil base cation concentration (SB) models in one of the ten random folds from each of two cross-validation approaches. In the first approach, the training data contains 9 folds (90%) of the input points (A) and the testing set the remaining fold (B). In the second approach, the training also contains 9 folds of the input points (C) and the testing fold only contains SB values input points obtained from direct soil measurements (D). Brown dots correspond to input data points with direct soil measurements and green dots to input data points whose SB values were estimated using SB optima of fern species.

pixel (in the standardized environmental space defined by the predictor variables of the final model) divided by the average distance between all training data pixels (Meyer and Pebesma, 2021). The DI was then used to derive pixel-wise binary values of the area of applicability (AOA) for the SB map. We applied three thresholds to transform the DI map to an AOA map. The first threshold was according to the default defined in the AOA framework, following Meyer and Pebesma (2021). Then we used two more restrictive thresholds (= lower DI) to visualize how it affects the final AOA map.

Finally, we compared the new SB map with two already available soil property maps. The first of these was the Soilgrids Cation Exchange Capacity (CEC) map (Hengl et al., 2017). Although CEC and SB do not represent the same soil chemical properties, both have been interpreted in the ecological literature, as indicators of soil nutrient availability and therefore, these two proxies should be comparable. We also investigated the relation of our map with an earlier SB map that was based on similar SB data but using ordinary kriging as the interpolation method (Zuquim et al., 2019). Both of these comparisons were done by randomly selecting 10,000 points across Amazonia and extracting the values corresponding to them from the respective soil maps. The values were then compared both over all Amazonia and separately over six Amazonian subregions. SB estimates from the old and new SB maps were also compared with the SB values from the input points with actual soil measurements by calculating Person's correlation coefficient (r) and concordance correlation coefficient (ccc).

All analyses were done in the R environment (R Core Team, 2022). The main packages used were: "raster" (Hijmans, 2021) for spatial data manipulation, "caret" (Kuhn, 2021) to carry out Random Forest models and variable importance calculations, "CAST" (Meyer, 2021) for estimations of DI and AOA, and "yardstick" for ccc (Kuhn et al., 2022) calculations.

3. Results

3.1. Sum of Bases map and best predictors

The spatial predictions of the Random Forest models using the most important variables revealed clear regional patterns in SB across

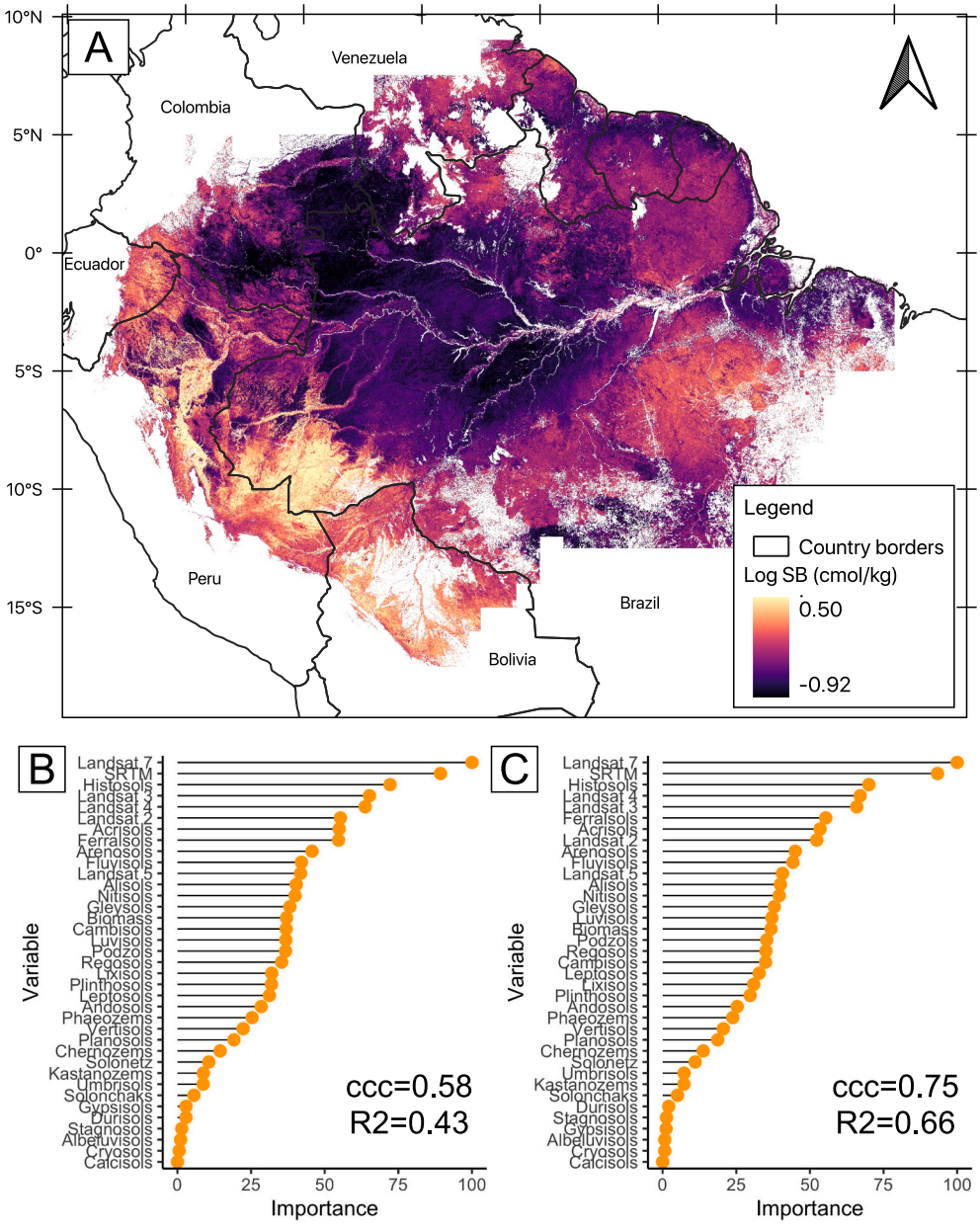


Fig. 2. (A) Map of predicted sum of exchangeable base cations concentration (SB) in forest soils across Amazonia. Modeling used the random forest methodology and included the 19 variables with Importance value >25 in (C). White areas in the map within Amazonia correspond to non-forested terrain and were masked out before analysis. (B) Variable importance when the cross-validation testing fold included all input data points. (C) Variable importance when the testing fold in cross-validation included only SB values obtained from direct soil measurements (see Fig. 1). The concordance correspondence coefficient (ccc) and R^2 are shown in (B) and (C).

Amazonia (Fig. 2A). In general, lower values tend to be found in central Amazonian soils and higher values in western Amazonian soils. The upper Rio Negro area and the interfluve between Purus and Madeira are dominated by very low values of SB. The top 5 most important variables in the models were three Landsat bands (3, 4 and 7), elevation, and probability of Histosols (Fig. 2B, C).

3.2. Model fitting and area of applicability

Evaluation of model performance varied considerably depending on which approach to testing data selection was adopted. When the testing sets included both direct soil measurements and fern-derived estimates, the performance of the full model with 37 predictors ($R^2 = 0.43$, RMSE = 0.58, ccc = 0.58) was clearly lower than when only direct soil measurements were used ($R^2 = 0.66$, RMSE = 0. 47, ccc = 0.75). The best RF model was obtained by allowing 19 predictors in each tree split, even thought this reduction in the number of predictors in the models made almost no difference to the performance metrics (differences are in the order of third decimals).

Spatial patterns in the Dissimilarity index (DI) suggested that the predicted SB values are least reliable in the extensive swamp forests of the Pastaza-Marañón basin in northern Peru, in the seasonally inundated forests along the lower Amazon River, and in several places along the Atlantic coast (Fig. 3A). These areas contain most of the pixels for which the AOA analysis deemed the model not to be applicable when a DI threshold of 0.46 is used (Fig. 3B). As the DI threshold was reduced, the model applicability was reduced from 96% of the pixels at DI < 0.46 to 89% at DI < 0.35 and 78% at DI < 0.30. (Fig. 3B, C, D).

3.3. Regional variation and comparison with other soil maps

The visual impressions of regional variation in the SB map (Fig. 1A) were confirmed by randomly selecting 10,000 points across Amazonia. The highest average soil SB was found in South-Western Amazonia and the lowest in Northern, Central and Central-Western Amazonia (Fig. 4). However, there was a wide range of variation within each region, especially in North-Western Amazonia.

Comparisons among maps based on 10,000 random points revealed that the SB values in the new map were moderately correlated with the SB values in the older map obtained by kriging ($r = 0.67$, ccc = 0.70; Fig. 4C) but were not correlated with the CEC values in the SoilGrids map ($r = 0.034$; Fig. 4D). When calculated separately within each

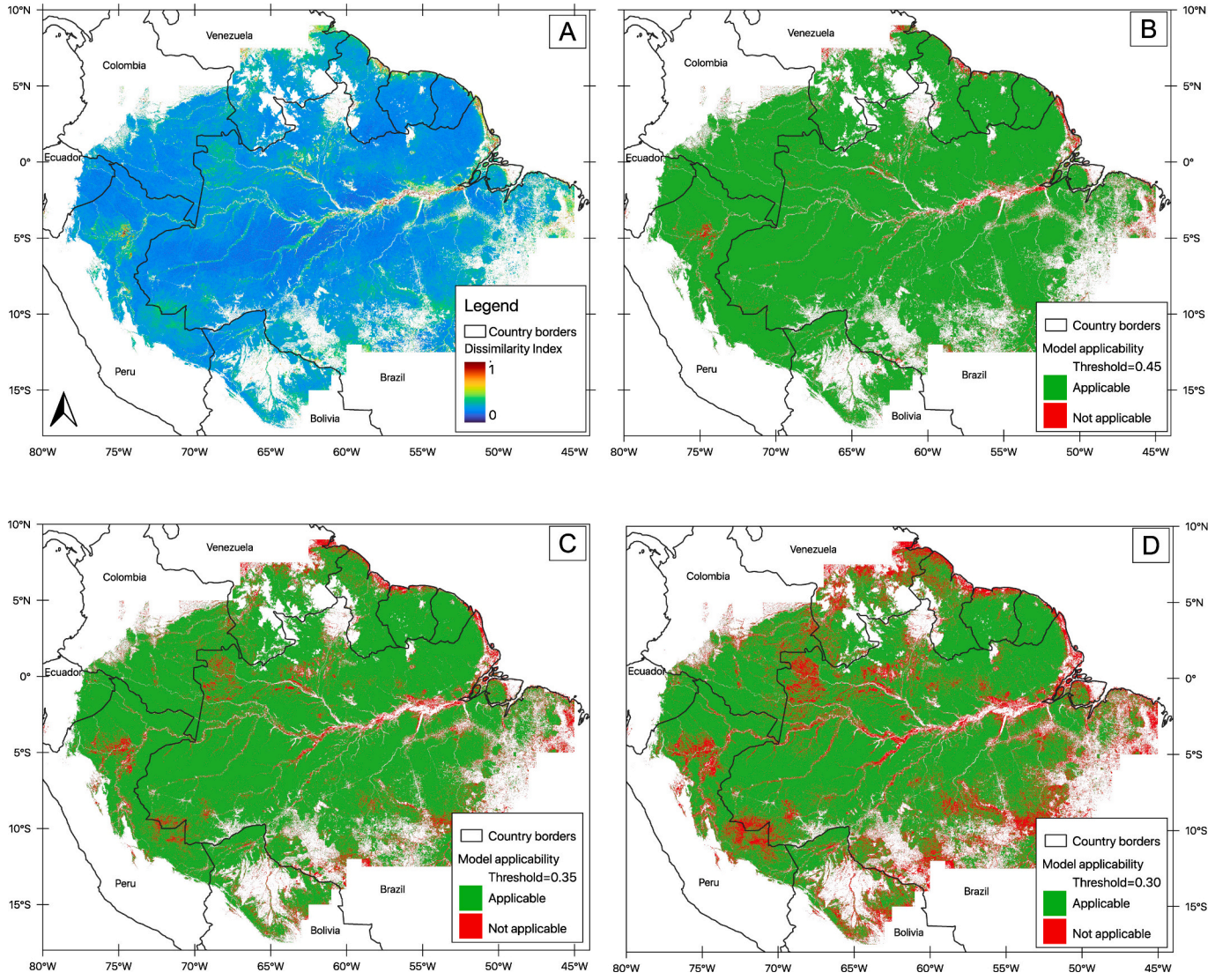


Fig. 3. (A) Map of the Dissimilarity index (DI), i.e. the degree of difference between each pixel and the most similar pixel with soil Sum of Bases (SB) data. Higher dissimilarities imply less reliable model predictions. (B—D) Area of applicability of the model when different thresholds of DI are used: (B) AOA default threshold = 0.46, (C) threshold = 0.35 and (D) threshold = 0.30. White areas within Amazonia correspond to non-forested terrain and were masked out before analysis.

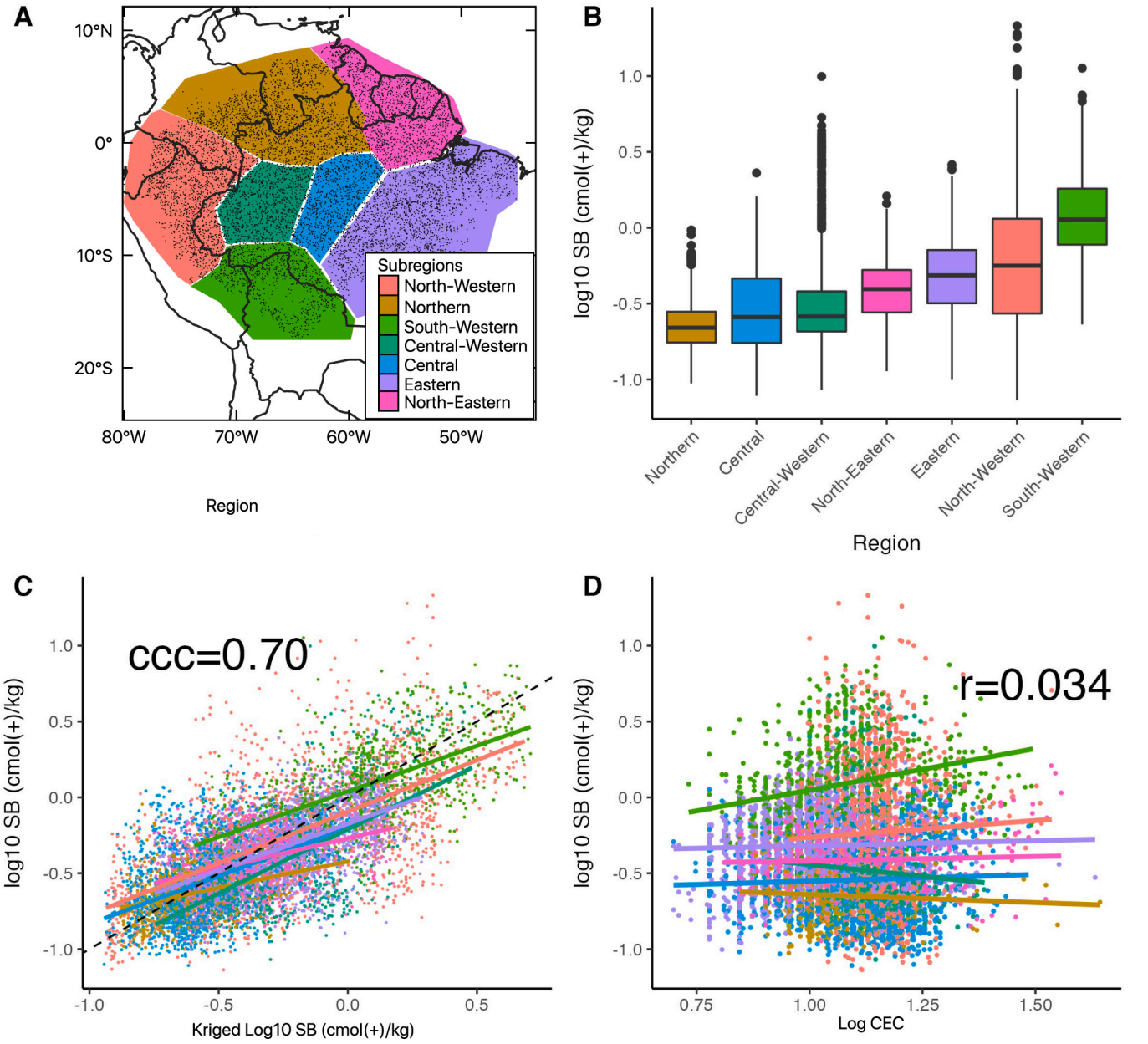


Fig. 4. (A) Distribution of 10,000 randomly selected points used to test how well the soil-related variables in different maps correlate with each other across Amazonia and within seven subregions. (B) Predicted soil base cation concentration (SB) within subregions. (C) Comparison between predicted SB values and SB values estimated by the map produced by kriging in Zuquim et al. (2019). The 1:1 relationship is shown with a dashed black line and the relationships within regions by colored lines. (D) Comparison between predicted SB values and CEC values from SoilGrids. The concordance (ccc) and Pearson's correlation coefficients (r) were calculated for the Amazon-wide comparison.

region, correlations between values in the two SB maps were highest in South-Western and North-Western Amazonia ($r = 0.70$ and $r = 0.60$, respectively), intermediate in Central-Western, Eastern, Northern and North-Eastern Amazonia ($r = 0.54$, r = 0.54, $r = 0.50$ and $r = 0.44$, respectively) and lowest in Central Amazonia ($r = 0.25$). Correlations between SB values in the new map and SoilGrids CEC values were low in all regions (the highest correlations were $r = 0.32$, $r = 0.23$ and $r = 0.20$ in South-Western, Northern and North-Eastern Amazonia, respectively).

The SB values obtained from field data had clearly higher concordance with estimates from the new SB map (ccc = 0.70) than with estimates from the SB map based on kriging (ccc = 0.50) (Fig. 5). The same was true within each Amazonian region separately (Table 1).

The differences between the current SB map and the older SB map based on kriging (Zuquim et al., 2019) showed some noteworthy spatial patterns (Fig. S1). The largest differences in estimated SB values between the two maps were found in South-Western Brazil and Northern Peru. In general, the differences between the two maps tended to be larger in Western than in Eastern Amazonia.

4. Discussion

4.1. Spatial heterogeneity of soils in Amazonia and SB map applications to biogeography and conservation

In Amazonia, soils are heterogeneous (Richter and Babbar, 1991) and shape the distribution and dynamics of the exceptional rainforest biodiversity (Cámara-Leret et al., 2017; Figueiredo et al., 2018; Schaefer et al., 2008; Tuomisto et al., 2003a, 2003b). One of the first attempts to

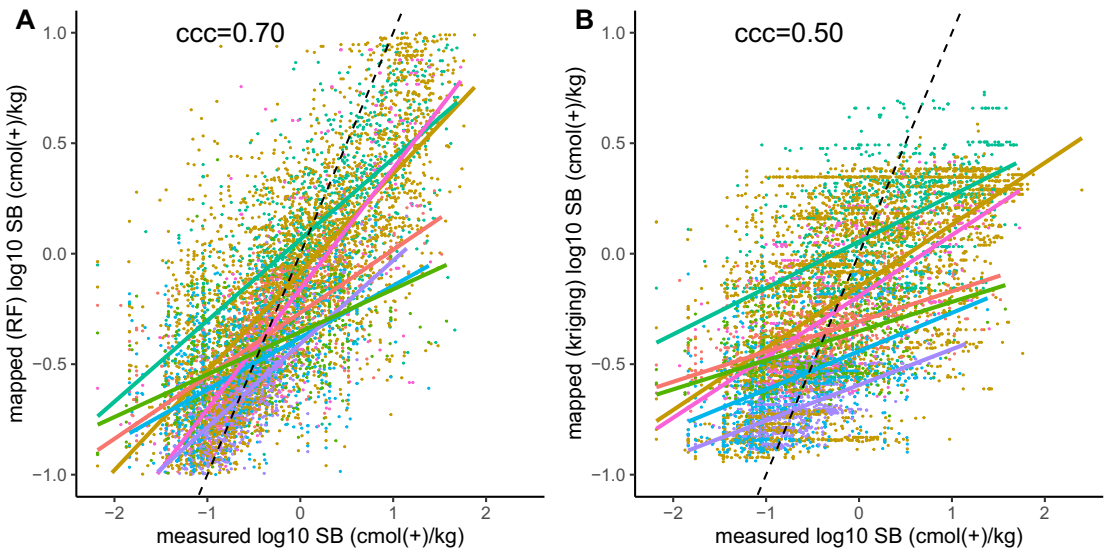


Fig. 5. Comparison between mapped and measured SB values for maps based on (A) the Random Forest model and (B) ordinary kriging (Zuquim et al., 2019). The 1:1 relationship is shown with a dashed black line and the relationships within regions with colored lines (see legend in Fig. 4A for the regional colors). The concordance correlation coefficient (ccc) was calculated for the Amazon-wide comparison.

Table 1
Pearson's (r) and concordance (ccc) correlation coefficients calculated between the soil measurements and mapped estimates in SB maps based on Random Forest models ("RF map") and ordinary kriging ("Kriged map") for whole Amazonia and within seven Amazonian subregions (defined as in Fig. 4A).

	RF map		Kriged map	
Region	r	ccc	r	ccc
North-Western	0,79	0,72	0,59	0,45
Northern	0,57	0,42	0,47	0,30
South-Western	0,71	0,59	0,52	0,36
Central-Western	0,80	0,75	0,65	0,47
Central	0,73	0,61	0,56	0,30
Eastern	0,69	0,51	0,47	0,25
North-Eastern	0,54	0,34	0,44	0,24
AMAZONIA	0,77	0,70	0,63	0,50

document the spatial patterns in Amazonian geochemistry was done by Fittkau et al. (1975), who distinguished western Amazonia with young nutrient-rich soils of Andean origin, central Amazonia with nutrient-poor highly weathered soils, and peripheral northern and southern regions with intermediate soils derived from cratonic rocks. On our map, the variation in SB roughly corresponds to this view. However, it is clear that within each geochemical region, there is wide local variation in SB. This is in line with earlier studies documenting a mosaic of different soils that have been formed by a variety of geological processes, including repeated depositional and erosional rearrangements by river dynamics, sedimentation in (semi)marine conditions, and in situ weathering (Rossetti et al., 2005, Hoorn et al., 2010, Quesada et al., 2011, Higgins et al., 2011). The high heterogeneity of SB, which stands out in our map, is a consequence of this complex geological evolution of Amazonian landscapes.

As previous studies have shown, SB is strongly linked to plant species composition (e.g. Dambros et al., 2020; Tuomisto et al., 2016; Cámara-Leret et al., 2017; Figueiredo et al., 2018; Schaefer et al., 2008; Tuomisto et al., 2003a, 2003b). Therefore, our map of SB can be used to predict species distribution patterns. However, it is important to bear in mind that the map is especially reliable in lowland terra-firme forests. Caution should be taken when applying this map to other physiognomies. When combined with other environmental layers, (e.g. climate), SB information is relevant to map habitats. Habitat maps are the basis for conservation planning as it indicates units harboring unique species assemblages. An SB map is useful to identify habitat types that are lacking protection in the existing network of conservation units (Fearnside and Ferraz, 1995), to identify conservation gaps, and to prioritize and plan new conservation areas. This is the framework of gap analysis and systematic conservation planning (Margules and Pressey, 2000; Scott et al., 1993), which have been extensively applied and require information on environmental conditions, landscape and vegetation characteristics.

When testing biogeographical hypothesis, habitat maps are also relevant. For example, turnover in species composition across a river might reflect the effect of the river as a dispersal barrier, as predicted by the riverine barrier hypothesis (Cracraft, 1985), but it might also reflect environmental determinism, in the case of changes in habitat types coincides with the different sides of the river (Colwell, 2000; Tuomisto and Ruokolainen, 1997). To disentangle the biogeographical effect of rivers as dispersal barriers vs. habitat distribution, reliable maps of environmental variability are needed (see de Maximiano et al., 2020).

Soils are also relevant when investigating the effects of climate change on species distributions and extinction probabilities. Future scenarios are usually based on climate-only models (Bellard et al., 2012; Velazco et al., 2017), but these may overestimate the amount of suitable area for a given species. Areas with suitable climate may be unsuitable due to edaphic reasons, and unfavorable soil characteristics may prevent species from tracking climatically suitable areas they need for survival (Figueiredo et al., 2018; Zuquim et al., 2020). Therefore, incorporating soil properties in species distribution models is crucial both for understanding the current species distribution patterns and to make realistic projections about their future under global warming.

4.2. *Comparison with other soil property maps*

Our SB map performed clearly better than the earlier available map (Zuquim et al., 2019) both at the Amazonian extent and within Amazonian subregions. The predicted values were generally closer to the actually measured values in the map based on the Random Forest approach than in the one using ordinary kriging. This reflects the better fine-tuning capabilities of models that include spatially explicit covariables, such as remote sensing data which have been shown to successfully predict edaphic patterns in Amazonia (Higgins et al., 2011; Van doninck and Tuomisto, 2018). Therefore, our model allowed abrupt spatial changes in SB values to be detected when surface reflectance

changed, even in the absence of field data.

The lack of relationship between CEC and our SB map was expected. CEC quantifies the capacity of the soil to bind any cations, whether they are relevant as plant nutrients or not. Therefore, CEC is a useful surrogate for soil nutrients only in conditions where the soils are not very acid and aluminum concentrations are low e.g. in most of North America, Southern Europe and continental Asia (www.soilgrids.org). These conditions do not hold over most of Amazonia, where soils are weathered, acidic and >90% of CEC can be occupied by aluminum (Quesada et al., 2011). In our dataset, the correlation between the CEC and SB values were weak, except in in South-Western Amazonia probably because of the proximity to Andes which implies that soils in this area tend to be derived from relatively young Andean sediments with higher pH values.

4.3. Developments, limitations and avenues for improvement

Soil mapping is a challenge worldwide. Three important avenues towards better soil maps are 1) improving training data; 2) improving the modeling approach and; 3) merging global and local models/maps (Hengl et al., 2017). We here addressed these points by 1) using the indicator species approach, which increased the number of training points; 2) adopting machine learning methods and taking advantage of remote sensing products; and 3) incorporating globally modeled soil type probabilities as predictors.

Our SB map was assessed to be widely applicable over lowland forests in Amazonia, but it has its limitations. Some habitats were poorly represented in the training data, especially those strongly determined by hydrological conditions, such as seasonally inundated forests, swamp forests, and coastal forests. Areas to which the model was not applicable using a more conservative threshold were mostly in these specific habitats and included the coastal areas, the floodplains of major rivers, the large wetland areas in northern Peru and northern Brazil, but also some parts of the bamboo-dominated forests in southwestern Amazonia and white-sand forests in the border between Brazil and Colombia. This is reflected in high DIs over these habitats, which suggests that the modeled SB values may have considerable error. Floodplain forests are particularly challenging to model because they are characterized by frequent disturbances, which causes strong temporal heterogeneity that is difficult to represent in stationary maps. The use of our SB map is best if restricted to lowland *terra firme* (non-inundated) forests, in which most of the training data was collected.

Nearly half of the final input data was derived from indicator species occurrences rather than from direct measurements of SB. While such surrogate data may be less accurate, it significantly increases data quantity and gives a better spatial coverage, both of which contribute to better models (Zuquim et al., 2019). In principle, these errors could be quantified and taken into account. Takoutsing et al., 2022 suggests to account for measument errors by assigning weights to training points inversely proportional to the estimated error. However, there are some caveats in such an attempt. For example, it assumes a known error variance (van der Westhuizen et al., 2022), which is not realistic in many cases. Moreover, some errors can be difficult to trace, for example errors associated with georeferencing (Moulatlet et al., 2017). Accounting for measurement errors was found to have little to no effect in model performances in two different African countries (Takoutsing and Heuvelink, 2022; van der Westhuizen et al., 2022), however, more studies are needed to verify the transferability of such results. In our case, quantifying the error was not possible. We believe that the most urgent action towards better maps in Amazonia is to improve the number and spatial coverage of direct soil measurements.

Another field for further development is the improvement of remote sensing products, both in terms of spectral resolution and the development of harmonized imageries that corrects for bidirectional reflectance effects and atmospheric interference. These are major sources of noise in spectral variation and can cause differences that are larger than the actual spectral differences between forest types (Toivonen et al., 2006). This effect is stronger in tropical forests and a major source of artifacts in many remote-sensing products over Amazonia (Muro et al., 2016). Such problems can be corrected but this is technically challenging and time-consuming, e.g., the Landsat TM/ETM+ imagery used here is based on pre-processing over 16,000 relatively cloud-free acquisitions (Van doninck and Tuomisto, 2018). Addressing these remote sensing images technical artifacts can expand the availability of imageries to more recent and spectrally refined remote sensors. Finally, even though RF has been considered the best single machine learning technique in soil mapping (Hengl et al., 2021), combining different modeling algorithms to generate ensemble predictions can also improve model performance (Hengl et al., 2021).

It is important to note that our modeling excluded non-forested areas. Deforestation has profound effects on both soil characteristics and Earth reflectance. In Amazonia, deforestation it is happening at a fast pace (INPE, 2022). The Landsat TM/ETM+ composite that we used was based on imagery from the years 2000–2009, and non-forested areas were excluded on the basis of their characteristic reflectance as of that time (Van doninck and Tuomisto, 2018). All soil data used were obtained in forest conditions. For areas that have been deforested since, the map should be read as modeling SB as it was under natural forest cover. Moreover, the map is provided at 450-m resolution, which means that it is not appropriate for studies that require high spatial or temporal resolution. Even though the map itself is static, its creation process can be repeated to obtain an updated version when more soil data or new predictors become available.

5. Conclusion

We generated a map of nutrient concentration in the surface soil (expressed as the sum of exchangeable base cations; SB) to provide information on SB variation across Amazonia. It provides an ecologically meaningful digital soil layer that represents a relevant background information to understand spatial variation in current and future species compositions, habitats and other properties of forests, especially Amazonian lowland *terra firme* forests. The models were considered to perform well, given the sparseness of training data. The Area of Applicability of our SB map covers 96% of Amazonia, however, some habitats (especially seasonally inundated and waterlogged ones) were underrepresent in the training dataset, and the corresponding SB values should be considered cautiously. Nevertheless, the SB map represents a major improvement in relation to previously available maps, as it incorporates more input data and uses more sophisticated modeling methods and meaningful covariates. The resulting map is available as a digital GIS layer and it can be applied in studies on Amazonian biogeography and in conservation planning.

Authors statement

GZ: Conceptualization, methodology, software, Data curation, formal analysis, Investigation, Writing - Original Draft, Visualization.

JVd: Methodology, software, Writing - review & editing.

PPC: Methodology, software, Writing - review & editing.

CAQ: Data curation, Writing - review & editing.

KR: Data curation, Writing - review & editing.

HT: Conceptualization, methodology, software, Data curation, Writing - editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Soil data: Raw data was not generated for this study. The map products (Sum of Bases map in digital GIS-layer format, file name: soil_map450res_topvariables.tif) and error assessment (AOA maps using different thresholds (0.46, file name: AOA_t=045.tif; O.35, file name: AOA_t=035.tif and 0.30, file name: AOA_t=030.tif in digital GIS-layer format) are deposited at fairdata.fi (https://doi.org/10.23729/61a2f9fa-8347-423d-911a-b0e606a8eb79)

Acknowledgments

We thank Mirkka Jones and Tom Hengl for fruitful discussions, all open source software and data providers, and the many field assistants and funding sources that allowed data accumulation and availability over the years. The analyses were carried out using the supercomputers of the CSC – IT Center for Science (Finland). Funding for this work from the following sources is gratefully acknowledged: CLAMBIO consortium funded through the BiodivERsA 2019-2020 Joint COFUND Call on "Biodiversity and Climate Change" (Academy of Finland grant #344733 to HT), Danish Council for Independent Research - Natural Sciences (grant #9040-00136B to Henrik Balslev), and Academy of Finland (grant #273737 to HT). The authors declare to have no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geodrs.2023.e00645.

References

Balslev, H., Kristiansen, S.M., Muscarella, R., 2019. Palm community transects and soil properties in western Amazonia. Ecology 100, e02841. https://doi.org/10.1002/ecy.2841.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., Courchamp, F., 2012. Impacts of climate change on the future of biodiversity: biodiversity and climate change. Ecol. Lett. 15, 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x.

Birks, H.J.B., Heiri, O., Seppä, H., Bjune, A.E., 2010. Strengths and weaknesses of quantitative climate reconstructions based on late-Quaternary biological proxies. Open Ecol. J. 3, 68–110.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cajander, A.K., 1926. The theory of forest types. Acta Forestalia Fennica 29, 1–108.

Cámara-Leret, R., Tuomisto, H., Ruokolainen, K., Balslev, H., Munch Kristiansen, S., 2017. Modelling responses of western Amazonian palms to soil nutrients. J. Ecol. 105, 367–381. https://doi.org/10.1111/1365-2745.12708.

Cárdenas Ramírez, G.G., Jones, M.M., Heymann, E.W., Tuomisto, H., 2021. Characterizing primate home-ranges in Amazonia: using ferns and lycophytes as indicators of site quality. Biotropica 53, 930–940. https://doi.org/10.1111/btp.12935.

Colwell, R.K., 2000. A barrier runs through it … or maybe just a river. Proc. Natl. Acad. Sci. 97, 13470–13472. https://doi.org/10.1073/pnas.250497697.

Cooper, M., Mendes, L.M.S., Silva, W.L.C., Sparovek, G., 2005. A national soil profile database for Brazil available to international scientists. Soil Sci. Soc. Am. J. 69, 649. https://doi.org/10.2136/sssaj2004.0140.

Cracraft, J., 1985. Historical biogeography and patterns of differentiation within the South American avifauna: areas of endemism. Ornithol. Monogr. 36, 49–84. https://doi.org/10.2307/40168278.

Dalagnol, R., Wagner, F.H., Galvão, L.S., Streher, A.S., Phillips, O.L., Gloor, E., Pugh, T.A.M., Ometto, J.P.H.B., Aragão, L.E.O.C., 2021. Large-scale variations in the dynamics of Amazon forest canopy gaps from airborne lidar data and opportunities for tree mortality estimates. Sci. Rep. 11, 1388. https://doi.org/10.1038/s41598-020-80809-w.

Dambros, C., Zuquim, G., Moulatlet, G.M., Costa, F.R.C., Tuomisto, H., Ribas, C.C., Azevedo, R., Baccaro, F., Bobrowiec, P.E.D., Dias, M.S., Emilio, T., Espirito-Santo, H.M.V., Figueiredo, F.O.G., Franklin, E., Freitas, C., Graça, M.B., d'Horta, F., Leitão, R.P., Maximiano, M., Mendonça, F.P., Menger, J., Morais, J.W., de Souza, A.H.N., Souza, J.L.P., da Tavares, C.V., do Vale, J.D., Venticinque, E.M., Zuanon, J., Magnusson, W.E., 2020. The role of environmental filtering, geographic distance and dispersal barriers in shaping the turnover of plant and animal species in Amazonia. Biodivers. Conserv. 29, 3609–3634. https://doi.org/10.1007/s10531-020-02040-3.

de Maximiano, M.F.A., d'Horta, F.M., Tuomisto, H., Zuquim, G., Doninck, J.V., Ribas, C.C., 2020. The relative role of rivers, environmental heterogeneity and species traits in driving compositional changes in southeastern Amazonian bird assemblages. Biotropica 52, 946–962. https://doi.org/10.1111/btp.12793.

Dijkshoorn, K., Huting, J., Tempel, P., 2005. Update of the 1: 5 Million Soil and Terrain Database for Latin America and the Caribbean (SOTERLAC; version 2.0). Report 2005/01. ISRIC - World Soil Information, Wageningen.

Ellenberg, H., Weber, H.E., Düll, R., Wirth, V., Werner, W., Paulissen, D., 1992. Zeigerwerte von Pflanzen in Mitteleuropa. In: Scripta Geobotanica, , 2nd ed.18, pp. 1–248.

Fearnside, P.M., Ferraz, J., 1995. A conservation gap analysis of Brazil's Amazonian vegetation. Conserv. Biol. 9, 1134–1147. https://doi.org/10.1046/j.1523-1739.1995.9051127.x-i1.

Figueiredo, F.O.G., Zuquim, G., Tuomisto, H., Moulatlet, G.M., Balslev, H., Costa, F.R.C., 2018. Beyond climate control on species range: the importance of soil data to predict distribution of Amazonian plant species. J. Biogeogr. 45, 190–200. https://doi.org/10.1111/jbi.13104.

Fittkau, E.J., Junk, J.W., Klinge, H., Sioli, H., 1975. Substrate and vegetation in the Amazon region. In: Dierschke, H., Tüxen, R. (Eds.), Vegetation Und Substrat. J. Cramer, Vaduz (Liechtenstein), pp. 73–90.

Heinrich, V.H.A., Dalagnol, R., Cassol, H.L.G., Rosan, T.M., de Almeida, C.T., Silva Junior, C.H.L., Campanharo, W.A., House, J.I., Sitch, S., Hales, T.C., Adami, M., Anderson, L.O., Aragão, L.E.O.C., 2021. Large carbon sink potential of secondary forests in the Brazilian Amazon to mitigate climate change. Nat. Commun. 12, 1785. https://doi.org/10.1038/s41467-021-22050-1.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — global soil information based on automated mapping. PLoS One 9, e105992. https://doi.org/10.1371/journal.pone.0105992.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., 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., Kempen, B., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12, e0169748. https://doi.org/10.1371/journal.pone.0169748.

Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M., McGrath, S.P., Acquah, G.E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F.B.T., Yemefack, M., Wendt, J., MacMillan, R.A., Wheeler, I., Crouch, J., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci. Rep. 11, 6130. https://doi.org/10.1038/s41598-021-85639-y.

Higgins, M.A., Ruokolainen, K., Tuomisto, H., Llerena, N., Cardenas, G., Phillips, O.L., Vásquez, R., Räsänen, M., 2011. Geological control of floristic composition in Amazonian forests. J. Biogeogr. 38, 2136–2149. https://doi.org/10.1111/j.1365-2699.2011.02585.x.

Higgins, M.A., Asner, G.P., Anderson, C.B., Martin, R.E., Knapp, D.E., Tupayachi, R., Perez, E., Elespuru, N., Alonso, A., 2015. Regional-scale drivers of forest structure and function in northwestern Amazonia. PLoS One 10, e0119887. https://doi.org/10.1371/journal.pone.0119887.

Hijmans, R.J., 2021. raster: Geographic Data Analysis and Modeling. R Package Version 3.4–10. https://CRAN.R-project.org/package=raster.

Hoorn, C., Wesselingh, F.P., ter Steege, H., Bermudez, M.A., Mora, A., Sevink, J., Sanmartín, I., Sanchez-Meseguer, A., Anderson, C.L., Figueiredo, J.P., Jaramillo, C., Riff, D., Negri, F.R., Hooghiemstra, H., Lundberg, J., Stadler, T., Särkinen, T., Antonelli, A., 2010. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. Science 330, 927–931. https://doi.org/10.1126/science.1194585.

INPE, 2022. Deforestation Monitoring of the Brazilian Amazon Rainforest and Cerrado Biome by Satellite.

Kuhn, M., 2021. caret: Classification and Regression Training. R Package Version 6.0–88. https://CRAN.R-project.org/package=caret.

Kuhn, M., Vaughan, D., Hvitfeldt, E., 2022. yardstick: Tidy Characterizations of Model Performance. R Package Version 1.1.0.

Laurance, W.F., Fearnside, P.M., Laurance, S.G., Delamonica, P., Lovejoy, T.E., Rankin-de Merona, J.M., Chambers, J.Q., Gascon, C., 1999. Relationship between soils and Amazon forest biomass: a landscape-scale study. For. Ecol. Manag. 118, 127–138. https://doi.org/10.1016/S0378-1127(98)00494-0.

Levis, C., Costa, F.R.C., Bongers, F., Peña-Claros, M., Clement, C.R., Junqueira, A.B., Neves, E.G., Tamanaha, E.K., Figueiredo, F.O.G., Salomão, R.P., Castilho, C.V., Magnusson, W.E., Phillips, O.L., Guevara, J.E., Sabatier, D., Molino, J.-F., López, D.C., Mendoza, A.M., Pitman, N.C.A., Duque, A., Vargas, P.N., Zartman, C.E., Vasquez, R., Andrade, A., Camargo, J.L., Feldpausch, T.R., Laurance, S.G.W., Laurance, W.F., Killeen, T.J., Nascimento, H.E.M., Montero, J.C., Mostacedo, B., Amaral, I.L., Vieira, I.C.G., Brienen, R., Castellanos, H., Terborgh, J., de Carim, M.J.V., da Guimarães, J.R.S., de Coelho, L.S., de Matos, F.D.A., Wittmann, F., Mogollón, H.F., Damasco, G., Dávila, N., García-Villacorta, R., Coronado, E.N.H., Emilio, T., de Filho, D.A.L., Schietti, J., Souza, P., Targhetta, N., Comiskey, J.A., Marimon, B.S., Marimon, B.-H., Neill, D., Alonso, A., Arroyo, L., Carvalho, F.A., de Souza, F.C., Dallmeier, F., Pansonato, M.P., Duivenvoorden, J.F., Fine, P.V.A., Stevenson, P.R., Araujo-Murakami, A., Baraloto, C., do Amaral, D.D., Engel, J., Henkel, T.W., Maas, P., Petronelli, P., Revilla, J.D.C., Stropp, J., Daly, D., Gribel, R., Paredes, M.R., Silveira, M., Thomas-Caesar, R., Baker, T.R., da Silva, N.F., Ferreira, L.V., Peres, C.A., Silman, M.R., Cerón, C., Valverde, F.C., Fiore, A.D., Jimenez, E.M., Mora, M.C.P., Toledo, M., Barbosa, E.M., de Bonates, L.C.M., Arboleda, N.C., de Farias, E.S., Fuentes, A., Guillaumet, J.-L., Jørgensen, P.M., Malhi, Y., de Miranda, I.P.A., Phillips, J.F., Prieto, A., Rudas, A., Ruschel, A.R., Silva, N., von Hildebrand, P., Vos, V.A., Zent, E.L., Zent, S., Cintra, B.B.L., Nascimento, M.T., Oliveira, A.A., Ramirez-Angulo, H., Ramos, J.F., Rivas, G., Schöngart, J., Sierra, R., Tirado, M., van der Heijden, G., Torre, E.V., Wang, O.,

Young, K.R., Baider, C., Cano, A., Farfan-Rios, W., Ferreira, C., Hoffman, B., Mendoza, C., Mesones, I., Torres-Lezama, A., Medina, M.N.U., van Andel, T.R., Villarroel, D., Zagt, R., Alexiades, M.N., Balslev, H., Garcia-Cabrera, K., Gonzales, T., Hernandez, L., Huamantupa-Chuquimaco, I., Manzatto, A.G., Milliken, W., Cuenca, W.P., Pansini, S., Pauletto, D., Arevalo, F.R., Reis, N.F.C., Sampaio, A.F., Giraldo, L.E.U., Sandoval, E.H.V., Gamarra, L.V., Vela, C.I.A., ter Steege, H., 2017. Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. Science 355, 925–931. https://doi.org/10.1126/science.aal0157.

Luizão, R.C.C., Luizão, F.J., Paiva, R.Q., Monteiro, T.F., Sousa, L.S., Kruijt, B., 2004. Variation of carbon and nitrogen cycling processes along a topographic gradient in a central Amazonian forest. Glob. Chang. Biol. 10, 592–600. https://doi.org/10.1111/j.1529-8817.2003.00757.x.

Margules, C.R., Pressey, R.L., 2000. Systematic conservation planning. Nature 405, 243–253. https://doi.org/10.1038/35012251.

Meyer, H., 2021. CAST: "caret" Applications for Spatial-Temporal Models. R Package Version 0.5.1. https://CRAN.R-project.org/package=CAST.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650.

Moulatlet, G.M., Zuquim, G., Figueiredo, F.O.G., Lehtonen, S., Emilio, T., Ruokolainen, K., Tuomisto, H., 2017. Using digital soil maps to infer edaphic affinities of plant species in Amazonia: problems and prospects. Ecol. Evol. 7, 8463–8477. https://doi.org/10.1002/ece3.3242.

Muro, J., Van Doninck, J., Tuomisto, H., Higgins, M.A., Moulatlet, G.M., Ruokolainen, K., 2016. Floristic composition and across-track reflectance gradient in Landsat images over Amazonian forests. ISPRS J. Photogramm. Remote Sens. 119, 361–372. https://doi.org/10.1016/j.isprsjprs.2016.06.016.

Nachtergaele, F., van Velthuizen, H., Verelst, L., Wiberg, D., 2012. Harmonized World Soil Database Version 1.2.

Poorter, L., van der Sande, M.T., Thompson, J., Arets, E.J.M.M., Alarcón, A., Álvarez-Sánchez, J., Ascarrunz, N., Balvanera, P., Barajas-Guzmán, G., Boit, A., Bongers, F., Carvalho, F.A., Casanoves, F., Cornejo-Tenorio, G., Costa, F.R.C., de Castilho, C.V., Duivenvoorden, J.F., Dutrieux, L.P., Enquist, B.J., Fernández-Méndez, F., Finegan, B., Gormley, L.H.L., Healey, J.R., Hoosbeek, M.R., Ibarra-Manríquez, G., Junqueira, A.B., Levis, C., Licona, J.C., Lisboa, L.S., Magnusson, W.E., Martínez-Ramos, M., Martínez-Yrizar, A., Martorano, L.G., Maskell, L.C., Mazzei, L., Meave, J. A., Mora, F., Muñoz, R., Nytch, C., Pansonato, M.P., Parr, T.W., Paz, H., Pérez-García, E.A., Rentería, L.Y., Rodríguez-Velazquez, J., Rozendaal, D.M.A., Ruschel, A. R., Sakschewski, B., Salgado-Negret, B., Schietti, J., Simões, M., Sinclair, F.L., Souza, P.F., Souza, F.C., Stropp, J., ter Steege, H., Swenson, N.G., Thonicke, K., Toledo, M., Uriarte, M., van der Hout, P., Walker, P., Zamora, N., Peña-Claros, M., 2015. Diversity enhances carbon storage in tropical forests. Glob. Ecol. Biogeogr. 24, 1314–1328. https://doi.org/10.1111/geb.12364.

Quesada, C.A., Lloyd, J., Anderson, L.O., Fyllas, N.M., Schwarz, M., Czimczik, C.I., 2011. Soils of Amazonia with particular reference to the RAINFOR sites. Biogeosciences 8, 1415–1440. https://doi.org/10.5194/bg-8-1415-2011.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria [WWW Document]. URL. https://www.r-project.org/ (accessed 2.28.19).

Rabus, B., Eineder, M., Roth, A., Bamler, R., 2003. The shuttle radar topography mission-a new class of digital elevation models acquired by spaceborne radar. ISPRS J. Photogramm. Remote Sens. 57, 241–262.

Richter, D.D., Babbar, L.I., 1991. Soil diversity in the tropics. In: Begon, M., Fitter, A.H., Macfadyen, A. (Eds.), Advances in Ecological Research. Academic Press, pp. 315–389. https://doi.org/10.1016/S0065-2504(08)60100-2.

Rossetti, D.F., Mann de Toledo, P., Góes, A., 2005. New geological framework for Western Amazonia (Brazil) and implications for biogeography and evolution. Quat. Res. 63 (1), 78–89. https://doi.org/10.1016/j.yqres.2004.10.001.

Schaefer, C.E.G.R., do Amaral, E.F., de Mendonça, B.A.F., Oliveira, H., Lani, J.L., Costa, L.M., Fernandes Filho, E.I., 2008. Soil and vegetation carbon stocks in Brazilian Western Amazonia: relationships and ecological implications for natural landscapes. Environ. Monit. Assess. 140, 279–289. https://doi.org/10.1007/s10661-007-9866-0.

Scott, J.M., Davis, F., Csuti, G., Noss, R., Butterfield, B., Groves, C., Anderson, H., Caicco, S., D'Erchia, F., Edwards, T., 1993. Gap analysis: a geographical approach to protection of biological diversity. Wildl. Monogr. 123, 1–41.

Sirén, A., Tuomisto, H., Navarrete, H., 2013. Mapping environmental variation in lowland Amazonian rainforests using remote sensing and floristic data. Int. J. Remote Sens. 34, 1561–1575. https://doi.org/10.1080/01431161.2012.723148.

Spawn, S.A., Sullivan, C.C., Lark, T.J., Gibbs, H.K., 2020. Harmonized global maps of above and belowground biomass carbon density in the year 2010. Sci. Data 7, 112. https://doi.org/10.1038/s41597-020-0444-4.

Suominen, L., Ruokolainen, K., Tuomisto, H., Llerena, N., Higgins, Mark A., 2013. Predicting soil properties from floristic composition in western Amazonian rain forests: performance of *k* -nearest neighbour estimation and weighted averaging calibration. J. Appl. Ecol. 50, 1441–1449. https://doi.org/10.1111/1365-2664.12131.

Takoutsing, B., Heuvelink, G.B.M., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma 428, 116192. https://doi.org/10.1016/j.geoderma.2022.116192.

Takoutsing, B., Heuvelink, G.B.M., Stoorvogel, J.J., Shepherd, K.D., Aynekulu, E., 2022. Accounting for analytical and proximal soil sensing errors in digital soil mapping. Eur. J. Soil Sci. 73 https://doi.org/10.1111/ejss.13226.

ter Braak, C.J., Juggins, S., 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia 269, 485–502.

Toivonen, T., Kalliola, R., Ruokolainen, K., Naseem Malik, R., 2006. Across-path DN gradient in Landsat TM imagery of Amazonian forests: a challenge for image interpretation and mosaicking. Remote Sens. Environ. 100, 550–562. https://doi.org/10.1016/j.rse.2005.11.006.

Tuomisto, H., Ruokolainen, K., 1997. The role of ecological knowledge in explaining biogeography and biodiversity in Amazonia. Biodivers. Conserv. 6, 347–357. https://doi.org/10.1023/A:1018308623229.

Tuomisto, H., Poulsen, A.D., Ruokolainen, K., Moran, R.C., Quintana, C., Celi, J., Cañas, G., 2003a. Linking floristic patterns with soil heterogeneity and satellite imagery in Ecuadorian Amazonia. Ecol. Appl. 13, 352–371. https://doi.org/10.1890/1051-0761(2003)013[0352:LFPWSH]2.0.CO;2.

Tuomisto, H., Ruokolainen, K., Yli-Halla, M., 2003b. Dispersal, environment, and floristic variation of Western Amazonian forests. Science 299, 241–244. https://doi.org/10.1126/science.1078037.

Tuomisto, H., Moulatlet, G.M., Balslev, H., Emilio, T., Figueiredo, F.O.G., Pedersen, D., Ruokolainen, K., 2016. A compositional turnover zone of biogeographical magnitude within lowland Amazonia. J. Biogeogr. 43, 2400–2411. https://doi.org/10.1111/jbi.12864.

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P., Poggio, L., 2022. Measurement error-filtered machine learning in digital soil mapping. Spat. Stat. 47, 100572 https://doi.org/10.1016/j.spasta.2021.100572.

Van doninck, J., Tuomisto, H., 2017a. Influence of compositing criterion and data availability on pixel-based Landsat TM/ETM+ image compositing over Amazonian forests. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 857–867. https://doi.org/10.1109/JSTARS.2016.2619695.

Van doninck, J., Tuomisto, H., 2017b. Evaluation of directional normalization methods for Landsat TM/ETM+ over primary Amazonian lowland forests. Int. J. Appl. Earth Obs. Geoinf. 58, 249–263. https://doi.org/10.1016/j.jag.2017.01.017.

Van doninck, J., Tuomisto, H., 2018. A Landsat composite covering all Amazonia for applications in ecology and conservation. Remote Sens. Ecol. Conserv. 4, 197–210. https://doi.org/10.1002/rse2.77.

Van doninck, J., Tuomisto, H., 2019. Amazonian Landsat TM/ETM+ Composite July–September 2000–2009 (Version 1). http://urn.fi/urn:nbn:fi:att:71ba2590-7112-4669-a4b3-a427c85c7a86.

Velazco, S.J.E., Galvão, F., Villalobos, F., De Marco Júnior, P., 2017. Using worldwide edaphic data to model plant species niches: an assessment at a continental extent. PLoS One 12, e0186025. https://doi.org/10.1371/journal.pone.0186025.

Zuquim, G., Tuomisto, H., Jones, M.M., Prado, J., Figueiredo, F.O.G., Moulatlet, G.M., Costa, F.R.C., Quesada, C.A., Emilio, T., 2014. Predicting environmental gradients with fern species composition in Brazilian Amazonia. J. Veg. Sci. 25, 1195–1207. https://doi.org/10.1111/jvs.12174.

Zuquim, G., Stropp, J., Moulatlet, G.M., Van Doninck, J., Quesada, C.A., Figueiredo, F.O. G., Costa, F.R.C., Ruokolainen, K., Tuomisto, H., 2019. Making the most of scarce data: mapping soil gradients in data-poor areas using species occurrence records. Methods Ecol. Evol. 10, 788–801. https://doi.org/10.1111/2041-210X.13178.

Zuquim, G., Costa, F.R.C., Tuomisto, H., Moulatlet, G.M., Figueiredo, F.O.G., 2020. The importance of soils in predicting the future of plant habitat suitability in a tropical forest. Plant Soil 450, 151–170. https://doi.org/10.1007/s11104-018-03915-9.