Received: 14 February 2018 | Accepted: 1 March 2019

DOI: 10.1111/2041-210X.13178

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Making the most of scarce data: Mapping soil gradients in data-poor areas using species occurrence records

Gabriela Zuquim[1] | Juliana Stropp[2] | Gabriel M. Moulatlet[1] | Jasper Van doninck[1,3] | Carlos A. Quesada[4] | Fernando O. G. Figueiredo[5] | Flávia R. C. Costa[5] | Kalle Ruokolainen[1,3] | Hanna Tuomisto[1]

[1]Department of Biology, University of Turku, Turku, Finland

[2]Instituto de Ciências Biológicas e da Saúde, Universidade Federal de Alagoas, Maceió, Brazil

[3]Department of Geography and Geology, University of Turku, Turku, Finland

[4]Coordenação de Dinâmica Ambiental, Instituto Nacional de Pesquisas da Amazônia, Manaus, Brazil

[5]Coordenação de Biodiversidade, Instituto Nacional de Pesquisas da Amazônia, Manaus, Brazil

Correspondence
Gabriela Zuquim
Email: gabriela.zuquim@utu.fi

Present address
Gabriel M. Moulatlet, Universidad Regional Amazónica IKIAM, Parroquia Muyuna, Tena, Napo, Ecuador.

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico; Ministério da Ciência e Tecnologia; Academy of Finland, Grant/Award Number: 273737; CNPq, Grant/Award Number: 152816/2016-0; University of Turku Graduate School

Handling Editor: Nigel Yoccoz

Abstract

1. Maps of environmental characteristics are needed to improve our understanding of species distributions and ecosystem dynamics. Despite the growing demand for digital environmental maps, scarcity of environmental field samples to be used as input data often constrains the accuracy of such maps, especially for soils.
2. We developed and tested a method that combines information on species–environment associations and the spatial distribution of indicator species (as retrieved from repositories such as GBIF) to improve mapping accuracy of environmental variables.
3. Our approach includes: (a) Compile field data on the environmental variable of interest (direct environmental data) and documented occurrences of the species to be used as indicators; (b) define species optima for the environmental variable; (c) use georeferenced records of the indicator species to calculate species-based environmental values (indirect environmental data); (d) generate maps using direct and indirect environmental data as input data for interpolation; (e) validate the maps. We applied the method to map the concentration of exchangeable base cations in Amazonian soils using fern and lycophyte species as indicators.
4. Including soil values that had been indirectly estimated using indicator species represented a 12-fold increase in the number of input data points used for mapping. At the same time, map accuracy improved considerably: the correlation between mapped soil cation concentration estimates and field-measured values from an independent validation dataset increased from $r = 0.48$ to $r = 0.71$.
5. Knowledge on species–environment relationships can be useful for modelling ecologically relevant environmental variables in areas where species occurrence data are more readily available than direct environmental measurements. The method works even with haphazard species occurrence points obtained from public repositories such as GBIF and can be applied to other environmental variables and other indicator groups, provided that the environmental variable of interest is relevant as a determinant of species occurrences in the indicator group. The Amazonian soil cation concentration maps produced (available at https://doi.

pangaea.de/10.1594/PANGAEA.879542) can be used as digital layers in species distribution and habitat modelling, and to guide conservation actions in Amazonia.

KEYWORDS
Amazonia, exchangeable base cations, ferns, GBIF, Kriging, plant occurrences, species distribution models, tropical forests

1 | INTRODUCTION

Environmental maps are useful for a wide range of studies, from landscape evolution and dynamics to species distribution modelling, with implications for conservation planning (Guisan & Zimmermann, 2000; Rylands, 1990). The reliability of the results of such studies depends to a large degree on the quality of available environmental maps (Araújo & Guisan, 2006; Carneiro, Lima, Machado, & Magnusson, 2016; Dawson, Jackson, House, Prentice, & Mace, 2011). For example, soils play a central role in shaping plant communities and constraining species distributions across scales (Phillips et al., 2003; Baldeck et al., 2012, Tuomisto et al. 1995, 2016), so thematically and spatially accurate soil maps are needed for many purposes. Even though considerable effort has been invested in producing digital soil maps that cover the entire world (Dijkshoorn, Huting, & Tempel, 2005; Hengl et al., 2017; Nachtergaele, van Velthuizen, Verelst, & Wiberg, 2012), in data-poor areas these maps still suffer from serious inaccuracies (Moulatlet et al., 2017).

One reason for map inaccuracy is that values in data-poor areas are estimated using spatial interpolation over large unsampled areas. Moreover, global digital soil maps are based on global models, although data availability varies considerably among regions (Hengl et al., 2014). Because predictive models cannot be optimized for all regions at the same time (Grunwald, Thompson, & Boettinger, 2011), producing maps specifically for a smaller area of interest and finding novel ways to increase the density of input data points for that area can contribute greatly to improving modelling accuracy (Grunwald et al., 2011).

If direct measurements of an environmental variable are scarce, surrogates can be used to alleviate data paucity. In Europe, the use of indicator plant species to infer site conditions dates back to the 1920s (Cajander, 1926; Ellenberg et al., 1992), and the general idea has been applied to distinguishing habitats or forest types also in Amazonia (Salovaara, Cárdenas, & Tuomisto, 2004; Tuomisto et al., 2003; Tuomisto & Ruokolainen, 1994). Palaeoecologists use indicator species (such as diatoms observed in sediment samples) to reconstruct past environmental conditions that cannot be directly measured (such as past pH in lakes) through calibration by weighted averaging (WA; Birks, 2003; ter Braak & Juggins, 1993). WA is based on the idea that the optimum value along a gradient for a species can be estimated as the abundance-weighted average of the environmental variable values at the sites where the species occurs, which also corresponds to the peak of a species-abundance curve along that environmental gradient. Thus, the presence of a species in a certain site implies that the environmental variable value at that site is close to the species-specific optimum.

Following these principles, knowledge of soil associations of plant species can be used to infer soil conditions: once the optimum of a given species for the soil variable of interest is known, it can be used as an estimate of the soil variable at the sites where the species occurs. The motivation for using indicator species stems from the fact that chemical analyses of soil samples are costly and, therefore, direct measurements of soil data are available from relatively few sites only. In contrast, taxonomical work is based on collecting plant specimens in as many sites as possible, and botanists tend to maximize the number of species collected (ter Steege, Haripersaud, Banki, & Schieving, 2011). The volume of georeferenced plant occurrence records available digitally through herbaria or other data repositories has increased dramatically, and so has our knowledge about species distributions (Lavoie, 2013). Consequently, localities with plant species records are usually much more numerous than localities with soil samples in any geographical area. The indicator species approach makes it possible to derive soil information for localities that have been sampled for plants but not soils, which substantially increases the number of points available as input data in soil mapping. Information on species optima along soil gradients is more difficult to obtain, but here we present such data for understorey ferns and lycophytes. Data for other plant groups, such as trees, may eventually become available through some of the standardized inventory efforts both in Amazonia (e.g. the Amazon Forest Inventory Network—RAINFOR, the Amazon Tree Diversity Network—ATDN, the Brazilian Program in Biodiversity—PPBio, and Forest Global Earth Observatory—ForestGEO) and elsewhere (Jürgens et al., 2012).

The use of understorey plants as indicators of soil properties has already been formally evaluated in Amazonia (Suominen, Ruokolainen, Tuomisto, Llerena, & Higgins, 2013; Zuquim et al., 2014) and even applied at the landscape scale (Sirén, Tuomisto, & Navarrete, 2013). Here, we describe how field data containing both species occurrence and environmental data can be combined with georeferenced species occurrence records downloaded from the Internet to map an environmental variable of interest over an extensive, data-poor area. As an example, we map the concentration of exchangeable base cations in Amazonian soils using ferns and lycophytes as indicator species. Soil properties are of interest because they are major factors determining ecosystem services, forest structure, carbon stocks and species distributions (Figueiredo et al., 2018; Quesada et al., 2010; Schaefer et al., 2008). Nevertheless, existing

soil maps covering Amazonia contain many problems, from spatial inaccuracies to failure to include variables that are relevant to many ecological questions (Moulatlet et al., 2017; Quesada et al., 2011). Our method can be applied to any environmental variables for which species affinities are quantifiable and sufficiently strong. Details such as spatial resolution and extent of the mapping, interpolation method and possible corrections for spatial bias can be adjusted to match specific interests.

2 | MATERIALS AND METHODS

2.1 | General framework

2.1.1 | Step 1: Compile available data on the environmental variable of interest and the potential indicator group

The first step is to compile as much data as possible on the environmental variable to be mapped and on the potential set of species to be used as indicators. Appropriate data may include quantitative plot data as well as data from other sources, such as public data repositories. Plot data often contain both environmental and species information for the same locations, and are, therefore, essential to determine species–environment relationships in step 2. Data obtained from online data portals typically provide information on either environmental variables (e.g. Global Soil Information Facility—GSIF) or species occurrence (e.g. Global Biodiversity Information Facility—GBIF, Specieslink, or Botanical Information Ecology Network—BIEN), but not both. These data can be used in step 3, as they typically represent both a wider geographical coverage and many more sites than the plot data do. Since data from data portals usually contain many identification and georeferencing errors (Maldonado et al., 2015), data cleaning is important to ensure adequate data quality.

2.1.2 | Step 2: Determine environmental optima for all species and verify their utility as indicators

The second step is to calculate the species-specific optima for the environmental variable of interest. The optimum value is obtained using data from locations for which both species and environmental data are available (plot data). Once the optima have been calculated, they can be used to estimate the environmental variable values at new sites with known species composition. Various transfer functions can be used for this purpose; here we use the Weighted Averaging calibration method (WA; ter Braak & van Dam, 1989). Transfer functions are widely used by palaeoecologists to infer past environmental conditions from fossil and extant species records, but here we apply them to predict current conditions. Care should be taken to ensure that a sufficient number of observations of the indicator species are used to calculate the species-specific optima (Zuquim et al., 2014). This will minimize the overall impact of confounding factors, such as other environmental variables or local species interactions, on the species-specific optimum values. In WA, species occurrence optima along the environmental gradient are calculated as the average (for presence–absence data) or weighted average (for abundance data, using species abundances as weights) of the environmental variable values in those plots where the species occurred (eq. 4 in ter Braak & van Dam, 1989). Tolerance is calculated as the root-mean squared error (RMSE) between the species optimum and the observed environmental variable values corresponding to each species observation (eq. 7 in ter Braak & van Dam, 1989). Due to the repeated taking of means, WA suffers from the tendency of the predicted values to be biased towards the overall mean value of the modelled variable (i.e. small values get overestimated and large values underestimated). To prevent this, several deshrinking methods were developed to restore the original variable range (ter Braak & Juggins, 1993).

Comparing the predicted environmental variable values with observed values (e.g. using leave-one-out cross-validation) gives a measure of the utility of the chosen indicator species group for the variable of interest. It is possible that prediction accuracy is adversely affected by generalist species. Making a second set of predictions such that species are downweighted in proportion to their tolerance can be used to assess the magnitude of this effect. If tolerance-weighted predictions are clearly more accurate than unweighted predictions, the species with broad tolerances (= the generalists) can be excluded from the final analyses.

2.1.3 | Step 3: Obtain species-derived estimates and combine with direct environmental measurements

In the third step, every geo-referenced occurrence location of a species (such as a record found in a public portal) is assigned the WA-estimated species environmental optimum value. Then the geo-referenced locations are rasterized to the desired grid cell size, and the average of the optima within a given grid cell is used as the estimate of the environmental variable value for that grid cell. Averaging provides more accurate estimates of the environmental variable than the individual species optima would. It also reduces spatial bias in sampling, as each grid cell is assigned exactly one environmental variable value no matter how many plant collections were available for it. The species-derived environmental data points thus obtained are then combined with the direct environmental measurements into a single dataset to be used in step 4.

2.1.4 | Step 4: Generate maps by interpolating between data points

The fourth step is to submit all the measured (direct) and species-derived (indirect) environmental data points as input data to a procedure that interpolates the values and produces an environmental map covering the whole area of interest. Various interpolation methods are available for this purpose; here we focus on Kriging but also provide results obtained with inverse distance weighting (IDW) for comparison.

2.1.5 | Step 5: Validate the maps

Finally, the obtained maps can be validated. Their accuracy can be tested by using an external validation dataset, or by splitting the existing data into a training set and a test set for cross-validation (Chatfield, 1995).

2.2 | Applied case: soil cation concentration map for Amazonia based on occurrences of ferns and lycophytes

We applied the approach outlined above to create maps of exchangeable base cations soil concentration in Amazonia (Ca + Mg + K measured in cmol(+)/kg; henceforth, soil cation concentration), using ferns and lycophytes as the indicator species group. We used ferns and lycophytes because earlier ecological studies provide a solid basis for the calculation of their soil cation concentration optima (Sirén et al., 2013; Tuomisto, Ruokolainen, & Yli-Halla, 2003; Tuomisto, Zuquim, & Cárdenas, 2014; Tuomisto et al., 2016; Zuquim et al., 2014).

2.2.1 | Step 1: Compile available data on soil cation concentration and occurrences of ferns and lycophytes in Amazonia

Points with both species and soil data (Plot data)

We compiled data from 1,353 quantitative fern and lycophyte inventory plots across Amazonia that also provided locally measured soil cation concentration (Figure 1a). Of these plots, 371 are part of the Brazilian Biodiversity Research Program database (PPBio) and 982 are part of the University of Turku Amazon Research Team database (UTU). The PPBio plots were 2 m wide and 250 m long and placed along terrain isoclines, following the guidelines of RAPELD (Portuguese acronym for Rapid Assessment—Long-term Ecological Research; Magnusson et al., 2005). The UTU plots were 150 m × 5 m in size and followed a predefined compass bearing. In every plot, ferns and lycophytes were inventoried and soils were collected and analysed as detailed in (Moulatlet et al., 2017).

Points with species data only (Herbarium data)

We searched for occurrence records within Amazonia (as delimited by Eva & Huber, 2005) for those fern and lycophyte species that were present in the plot data. Species occurrence records were downloaded from the Global Biodiversity Information Facility (GBIF; gbif.org) and SpeciesLink (http://www.splink.org.br) in November 2016 and both datasets were combined. Duplicate records of the same species with the same latitude and longitude were excluded. We also excluded species records with any of the following issues: (a) coordinates (in decimal degrees) were given with a precision of less than three decimal places; (b) coordinates landed in a different country than that indicated in the 'country' field of the specimen metadata; (c) coordinates coincided with the centre of a city or major village (places classified as 'administrative level 2' or 'populated places' in the GEONAMES database); (d) the record came from the UTU or PPBio plots and had already been used in the species optimum calculations. These four steps of data filtering removed 2,667 species occurrence records out of 33,604 and left 30,937 for the analyses (Figure 1b).

Points with soil data only (Public repositories)

Values of soil cation concentration were retrieved from the Harmonized World Soil Database v1.2 (HWSD) (Nachtergaele et al., 2012) and a Brazilian national database (BND) (Cooper, Mendes, Silva, & Sparovek, 2005). We used the data from those 347 soil samples that had geographic coordinates within Amazonia and had been collected at a maximum depth of no more than 30 cm. In addition, we used the soil cation concentration information from around 2,300 soil samples from the UTU and PPBio databases. In the PPBio plots, six surface soil samples (the top 5 cm of the mineral soil) were taken at every 50 m and bulked to obtain a single composite sample. The samples were analysed in the Soil Thematic Laboratory of Brazilian National Institute for Amazonian Research (LTSP-INPA) using the Mehlich I protocol (KCl 1 Normality method; Donagena, Campos, Calderano, Teixera, & Viana, 2011). Each UTU plot was represented by one composite surface soil sample (top 5 cm of the mineral soil) that consisted of five subsamples collected within an area of about 5 m × 5 m. These samples were analysed at MTT Agrifood Research (Jokioinen, Finland) using extraction in 1 M ammonium acetate (van Reeuwijk, 1993). Details of the soil sampling can be found in (Moulatlet et al., 2017).

2.2.2 | Step 2: Determine environmental optima for all fern and lycophyte species and verify their utility as indicators

We used the Weighted Averaging calibration method with monotonic curvilinear deshrinking to calculate optima and tolerances along the log-transformed (base 10) soil cation concentration gradient for each of the 282 fern and lycophyte species observed in the plots. The nonlinear deshrinking procedure applied is similar to the linear deshrinking described in eq. 5 of ter Braak and van Dam (1989). The curvilinear approach avoids nonlinear distortions such as edge effects (ter Braak & Juggins, 1993) and was developed by fitting a smooth monotonic function to the inverse deshrinking as implemented in the R package 'RIOJA' (Juggins, 2017).

To confirm that the species were good predictors, we estimated soil cation concentrations for the PPBio and UTU plots and compared the estimates with the observed values using leave-one-out cross-validation. RMSE and the coefficient of determination (R^2) between the predicted and the laboratory-analysed soil cation concentrations were used to quantify prediction accuracy. To test whether species with a wide tolerance decreased prediction accuracy, we repeated the WA calculations using the inverse of species tolerance as an additional weight. All calculations were carried out separately for presence–absence and abundance data.

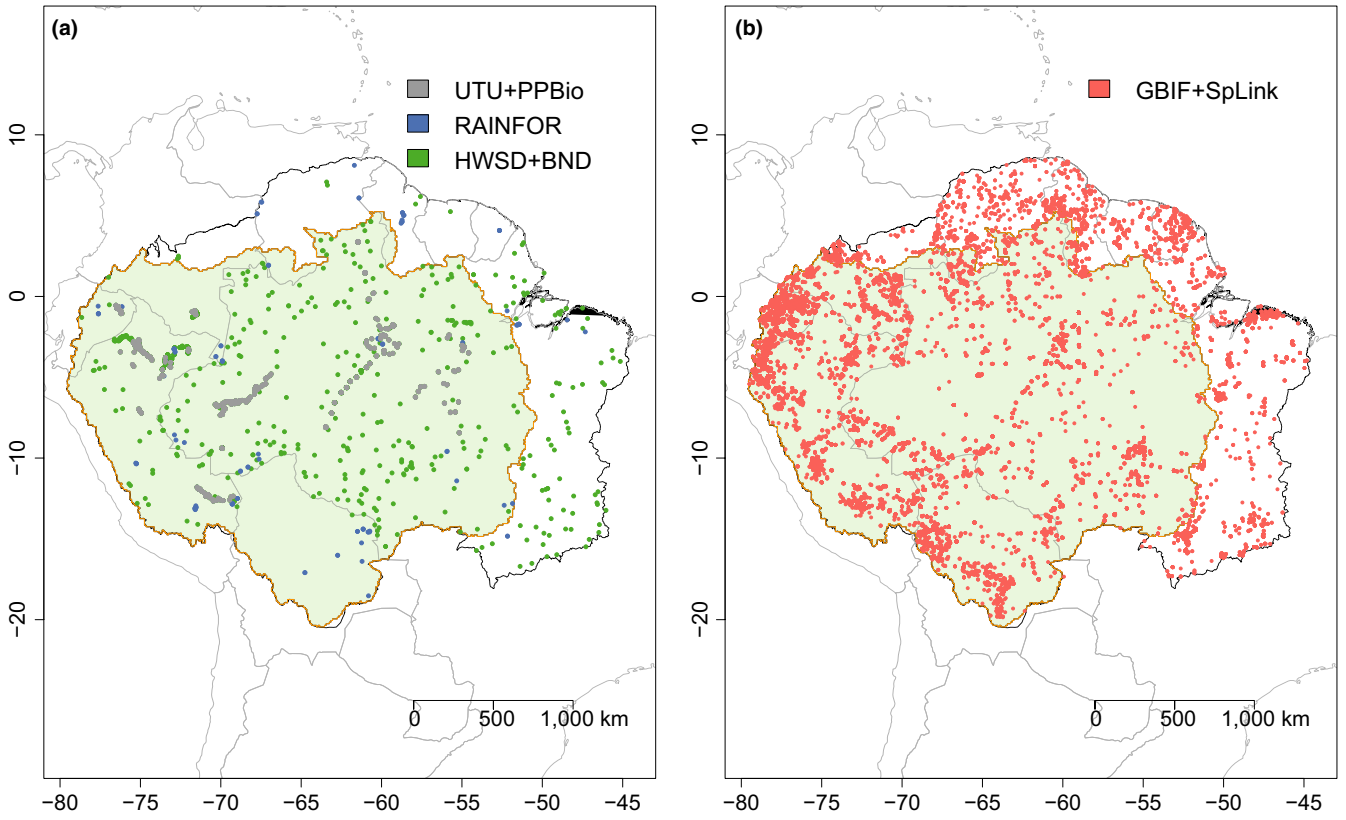

FIGURE 1 Distribution of data points used to produce and test estimates of exchangeable base cation concentration in surface soil across Amazonia. (a) Soil sample points from our own database (UTU+PPBio) and from external databases used as input data (HWSD+BND - Harmonized World Soil Database +Brazilian National Database) and as validation data (RAINFOR - Amazon Forest Inventory Network). (b) Fern and lycophyte species occurrence records retrieved from the Global Biodiversity Information Facility and SpeciesLink data portals (GBIF+SpLink). Analyses were done over all Amazonia as defined by Eva & Huber (2005; black line, white polygon) and over a subset limited by the wetlands map of Hess et al. (2015; orange line, pale green polygon)

Some closely related species are difficult to reliably distinguish in the field, and such complexes were lumped before calculating species optima. For simplicity, we make no distinction between these complexes and what are thought to be real species in the text. Calculations of species optima were done using the R package 'RIOJA' (Juggins, 2017). The analysis code is available at https://github.com/gabizuquim/paper_fern-soil_map/blob/master/spp_optima.R.

2.2.3 | Step 3: Obtain species-derived soil cation concentration estimates and combine with direct soil measurements

Each geo-referenced location of a species occurrence obtained from GBIF and SpeciesLink was assigned the optimum soil value of the corresponding species. We then aggregated the points to grid cells of 1 arcmin (~2 km at the equator) and used the average value per cell as a plant-derived (indirect) soil cation concentration value. Similarly, we averaged for each grid cell the directly measured soil data obtained from the PPBio and UTU plots and the HWSD and BND databases. A third soil dataset was obtained by combining the indirect and direct soil values; an average value was used if both indirect and direct soil data were available for the same grid cell. Each of the three soil datasets (indirect data only, direct data only, both combined) was used separately as input data in step 4.

2.2.4 | Step 4: Generate maps by interpolating between data points

A raster map of estimated soil cation concentration values covering all Amazonia was obtained by interpolation at the spatial resolution of 6 arcmin (~11 km at the equator). Even though there are general spatial trends in soil cation concentration across Amazonia (Quesada et al., 2010) that divides into major geochemical regions, abrupt changes in soil characteristics have also been documented (e.g. Higgins et al., 2011; Tuomisto et al., 2016). This means the stationarity assumption of simple Kriging does not hold. Therefore, we used ordinary Kriging, where parameters of the semi-variogram (model type, nugget, partial sill and range) were estimated by visual inspection and then improved by an automatic fitted variogram function. An associated layer indicating the Kriging standard deviation was also generated to illustrate the uncertainty associated with the interpolated soil cation concentration values. In order to evaluate to what

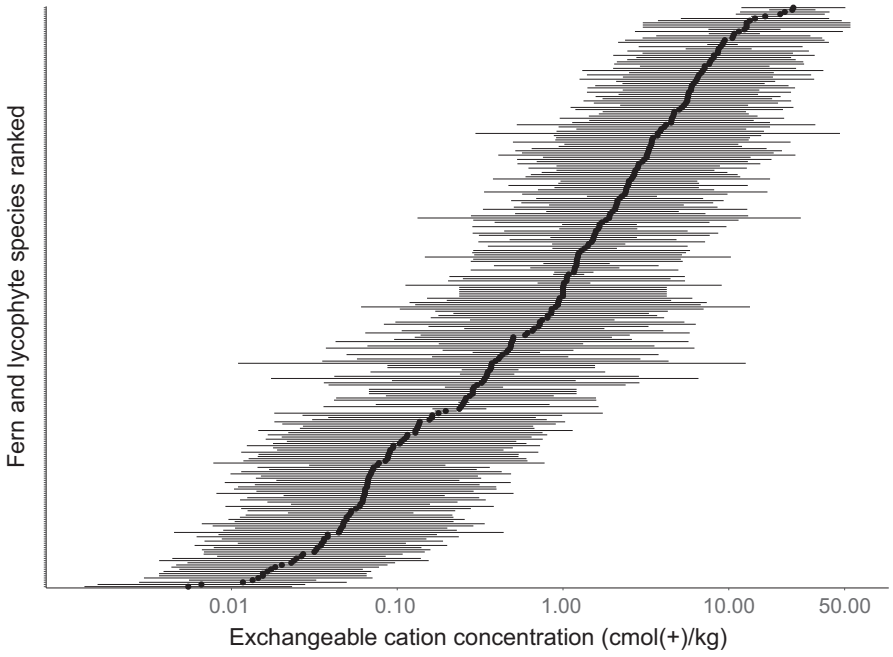

FIGURE 2 Fern and lycophyte species optima (black dots) and tolerances (grey horizontal bars; based on root mean squared error) for soil cation concentration (exchangeable Ca+Mg+K) as calculated using data from 1,353 inventory plots in lowland Amazonia. The species are ranked by their cation optimum.

degree the results depend on the selected interpolation method, we performed the spatial interpolations also using IDW with a weighting power of 2.

Both spatial interpolation methods are implemented in the R package (R Core Team, 2017) 'GSTAT' (Pebesma & Graeler, 2017). Automated fitting was carried out using the *fit.variogram* function. The codes for averaging the fern and lycophyte species optima values and Kriging are available at https://github.com/gabizuquim/paper_fern-soil_map/blob/master/script_krig_share.R.

To visualize the effect of input data on the final maps, we repeated the interpolations using different subsets of the available point data. These were as follows: (a) direct soil measurements only (1,033 grid cells with soil samples obtained from HWSD, BND, UTU and PPBio datasets); (b) plant-derived (indirect) soil estimates only (6,041 grid cells with species occurrence data obtained from GBIF and SpeciesLink); (c) direct and plant-derived soil data together (6,945 grid cells); and (d) direct and plant-derived soil data together, but analysis limited to the area covered by a recent wetlands map (Hess et al., 2015). The data points in subset 4 were first classified into *terra-firme* (non-inundated uplands) versus wetlands, and the interpolation was done separately for each landscape type. The wetlands map covers about 87% of Amazonia as defined by (Eva & Huber, 2005) (Figure 1).

2.2.5 | Step 5: Validate the maps

To evaluate the accuracy of the maps, we obtained external validation data on soil cation concentration taken at a maximum depth of 30 cm from 194 soil samples of the Amazon Forest Inventory Network (RAINFOR; http://www.rainfor.org) (Figure 1a). Laboratory methods applied by RAINFOR are described by Quesada et al. (2010).

We used the RAINFOR soil sample coordinates to match measured exchangeable cation concentrations with the estimated values from each of our maps and assessed the accuracy of the estimates by calculating the Pearson correlation between the estimated and measured values. Even though the validation dataset is spatially clustered, we expect independently collected validation data to be less prone to bias than cross-validation approaches would be, since the estimates used in validation are model-free (Brus, Kempen, & Heuvelink, 2011). Cross-validation also often over-estimates mapping accuracy (Chatfield, 1995).

For comparison, we also obtained the map of soil cation exchange capacity (CEC) values at maximum depth of 30 cm made available by the SoilGrids project (Hengl et al., 2017); www.soilgrids.org, accessed in December 2016) and calculated the correlation with CEC values from SoilGrids and the exchangeable cation concentrations from the RAINFOR soil data. CEC is a soil variable that is commonly used in ecological studies (Figueiredo et al., 2018; Levis et al., 2017; McMichael et al., 2014).

Next, we evaluated how much the pixel-by-pixel patterns in the maps depend on the input data and interpolation technique. To do this, we correlated all the maps generated by Kriging with each other and with the CEC map, as well as with the maps based on the same input data as the Kriged map but applying IDW for interpolation.

Finally, we took the map that performed best (= with higher correlation between mapped and observed values in the validation dataset) and assessed how its errors and uncertainties were related to the heterogeneity and distribution of input points. We extracted the Kriging standard deviation and the mapped cation concentration values for each point in the validation dataset and quantified the error of the map as the absolute difference between the mapped and measured soil values. We obtained the number of input points and the standard deviation of the soil values within buffers of 6 arcmin

(~11 km radius) around the location of the validation samples. We further accessed uncertainty in structurally random by obtaining the number of sampling points and standard deviations in one point or buffer randomly chosen inside each cell of a 20 × 20 arcmin grid to Amazonia. This assures that the uncertainty assessment has covered the whole Amazonia and was not biased by the locations of the validation sampling points.

3 | RESULTS

In the 1,353 plots across Amazonia, we recorded 63,104 individuals of ferns and lycophytes belonging to 268 species. Species that were confused in the field were lumped before the analyses, so these were based on 245 taxa (species, species complexes or genera; all referred to as species for simplicity). Field data covered a long gradient of exchangeable base cation concentration (median = 0.37, mean = 4.11, range = 0.03–54.67 cmol(+)/kg), and species optima were well spread along this gradient (Figure 2). The cation concentration optima based on species presence–absence and abundance data were highly correlated (r = 0.98; note that all correlations and other results reported here were obtained using the log-transformed values). The accuracy of soil cation concentration predictions based on species composition was also high (R^2 = 0.74–0.85) regardless of the input data type (presence–absence or abundance data) used in the estimations (Table 1). Moreover, it made little difference whether species tolerances were used as inverse weights in the WA calculations or not. This indicates that species with broad tolerance did not substantially decrease the accuracy of the estimated soil values. Since prediction accuracy was high overall, and taking species tolerances into account had little effect, we retained all species that had been observed in the plots in the indicator species pool to be used when predicting soil cation concentration for the final mapping exercise.

There were big differences among the maps depending on which input dataset was used. Soil cation concentration values extracted from the maps that were based on the combined direct and plant-derived soil data (soil data subset 3 as defined in Step 3 above) and interpolated using Kriging had the highest correlations with the measured values in the validation dataset (r = 0.71). The map based on species-derived soil data only (subset 2) performed slightly worse (r = 0.68 and 0.61 for Kriging and IDW, respectively) and the map based on direct soil data only (subset 1) performed clearly worse (r = 0.48 and 0.52 for Kriging and IDW, respectively). The lowest correlations with exchangeable soil cation concentration were obtained for the CEC values of the SoilGrids map (r = 0.30).

In all interpolated maps (Figure 3a,c,e,g), most of the cation-rich soils were predicted to be in western and southern Amazonia and most of the cation-poor soils in central Amazonia. However, large areas especially in eastern Amazonia had unstable model outputs that predicted either cation-rich or cation-poor soils, depending on the input data. Soil cation concentration maps using direct and plant-derived soil data together were more strongly correlated with the maps using only plant-derived estimates (r = 0.90–0.92) than with the maps using only direct soil measurements (r = 0.70; Figure 4). This is no surprise since more than 90% of the total input data points were obtained from plant-derived estimates based on species optima. Values of the CEC map were only weakly correlated with the values from the other maps (r = 0.10–0.28) (Figure 4). Kriging and IDW maps (Figure S1) that used the same input data were highly correlated (Figure 4) but the mapped values using Kriging tended to have a higher correlation with the measured soil cation concentrations of the validation dataset (compare r values from Figure 3 and Figure S1).

The maps of Kriging standard deviation (square root of Kriging variance; Figure 3b,d,f,h) obviously reflect the unequal density of sampling. In some areas, neighbouring soil data points were more than 200 km apart, such as in the large red patch in eastern Amazonia (Figure 3f). The uncertainty is generally the highest in the map based on direct soil measurements only, which has the fewest input data points (Figure 3b). The maps based on both direct and indirect soil data points were similar whether the wetlands mask was used or not, but the Kriging standard deviations were higher in the map with the wetlands mask. This is probably related to the loss of densely sampled areas in northern and eastern Amazonia and inaccurate georeferencing in GBIF points, leading to species occurrences being incorrectly assigned to wetlands versus *terra-firme*.

The difference between measured and mapped soil values was not linearly related to either density or heterogeneity of the soil samples within a local buffer (Figure 5a,b). This indicates that the number and distribution of sampling points may have weak or no effect on the overall accuracy of the estimated values. Kriging standard deviation significantly decreased with increasing density of sampling points regardless of whether it was assessed in the locations of the validation dataset (Figure 5c) or in random locations (Figure 5e). We found a negative relationship between the mapped Kriging standard deviation and the standard deviation estimated for all sample values within each buffer (Figure 5d,f).

TABLE 1 Accuracy of predictions of soil cation concentration using fern and lycophyte species as indicators. Accuracies are measured by root mean squared error (RMSE) and coefficient of determination (R^2) of the linear regressions between predicted and observed values. The combination of high R^2 and low RMSE indicates a good regression fit. The transfer method was weighted averaging (WA) with monotonic deshrinking. The inverse of species tolerance was used as weight where indicated. Values are based on leave-one-out cross-validation

Data	Tolerance weighing	RMSE	R^2
Abundance	No	0.536	0.742
	Yes	0.521	0.757
Presence–Absence	No	0.529	0.749
	Yes	0.492	0.782

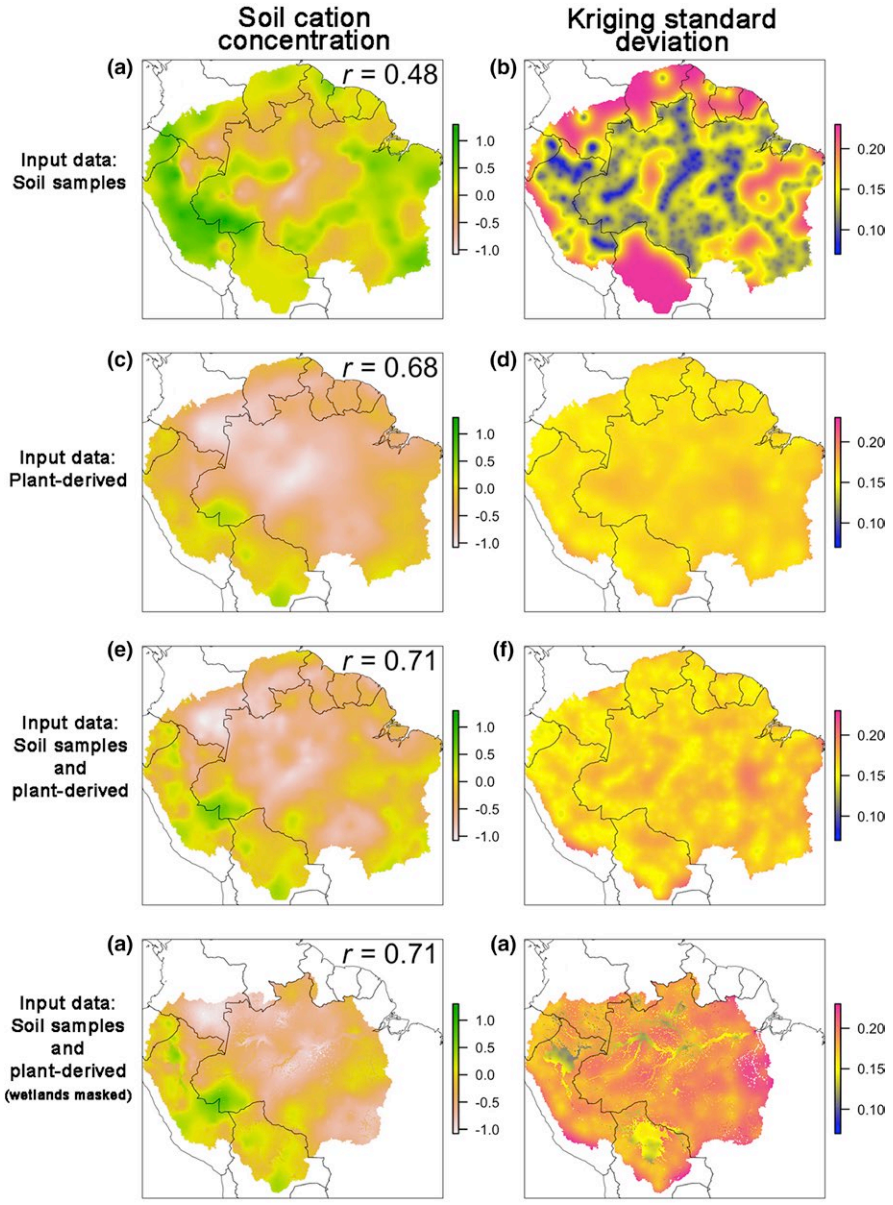

FIGURE 3 Maps of soil cation concentration (left column) and Kriging standard deviation (right column) as modelled using different sets of input data (rows). Pearson's correlation (*r*) between soil cation concentrations as read off the map and as measured in soil samples of the Amazon Forest Inventory Network (RAINFOR) is shown in the upper right corner. Input data used in the maps are: (a, b) direct soil data only (measured from soil samples); (c, d) indirect soil data only (plant-derived estimates); (e, f) direct and plant-derived soil data together; (g, h) direct and plant-derived soil data together with Kriging run separately for wetlands and *terra-firme* areas (map extent limited to the area covered by the available wetlands map). The scale bars show 10-based logarithms of base cation concentration (Ca + Mg + K) as expressed in cmol(+)/kg. High-resolution versions of the maps obtained by Kriging are available in Supplementary Material S1. Raster versions of maps e, f, g, and h are available in Pangaea (see Data Accessibility section)

4 | DISCUSSION

4.1 | Using indicator species to mitigate environmental data paucity

Using soil cation concentrations as estimated from the occurrences of indicator plant species gave us a nearly 12-fold increase in the volume of input data for modeling, compared to using soil data available from direct measurements only. This explains why the maps that included plant-based soil data performed considerably better than maps that were based on direct soil measurements only. An increase in the volume of input data not only tends to improve modelling accuracy as such (Grunwald et al., 2011) but it also reduces mapping sensitivity to different interpolation methods, which makes map accuracy less dependent on methodological choices (Chaplot et al., 2006; Hijmans, Cameron, Parra, Jones, & Jarvis, 2005).

Species-derived environmental values can be less accurate than directly measured values, so there may be a trade-off between data density and data accuracy (Chaplot et al., 2006). This is because species occurrences are affected by other factors than the variable of interest, such as biological interactions, different aspects of climate and various soil properties. The effects of such confounding factors can bias the species optima for the environmental variable of interest. Thus, although validation results suggest that the maps presented here mostly reflect the modelled environmental gradient, it is important to be aware that residual effects of other factors may also influence the mapped patterns.

If the modelled variable corresponds to the strongest environmental gradient that the species respond to, residual effects are likely to be small, but if the interest is in modelling a less important environmental variable, confounding factors are likely to be a real

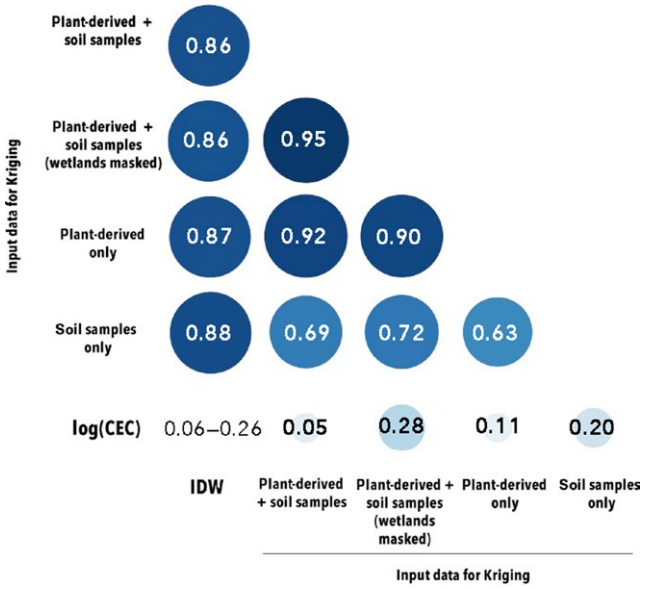

FIGURE 4 Pixel-by-pixel Pearson's correlations (*r*) between soil cation concentration maps obtained using four different input datasets and two different interpolation methods as well as the cation exchange capacity (CEC) map obtained from SoilGrids. Both cation concentrations and CEC were log-10-transformed before analysis

problem. In addition, species tolerances will be very broad along an environmental gradient that is irrelevant to them, which further reduces species predictive power. In our example, the increase in the number of data points amply compensated for any decrease in point accuracy there might have been due to the use of surrogate data. This is consistent with earlier studies having found that soil base cation concentration is a very strong determinant of fern and lycophyte species occurrence patterns in Amazonia (Tuomisto et al., 2016; Tuomisto, Ruokolainen, et al., 2003; Zuquim et al., 2014).

Many data-poor regions are similar to Amazonia in that directly measured soil data are much sparser than species occurrence records. Moreover, the latter are increasingly accessible thanks to biodiversity data portals such as GBIF and SpeciesLink, which facilitates the use of indicator species for modeling ecologically relevant environmental variables. We used ferns and lycophytes as indicators, because earlier studies have both shown them to be useful for this purpose and provided the field data needed to calculate species optima (Ruokolainen, Tuomisto, Macía, Higgins, & Yli-Halla, 2007; Tuomisto et al., 2003; Tuomisto, Ruokolainen, et al., 2003; Tuomisto et al., 2014, 2016; Zuquim et al., 2014). The method we propose here for modelling an environmental variable is not restricted to ferns and lycophytes, however; any biological group that has a strong affinity with the variable of interest and enough data for model calibration could be used. For soil variables, other plant groups that have indicator potential in the tropics include the Melastomataceae, palms and Zingiberales (Cámara-Leret, Tuomisto, Ruokolainen, Balslev, & Munch Kristiansen, 2017; Suominen et al., 2013; Tuomisto et al., 2016).

Our approach for modelling environmental variables using indicator species has the potential to inform both fundamental and applied research on ecology and biogeography in Amazonia and other poorly sampled areas. The currently available digital soil maps have problems of low spatial accuracy and lack of ecologically relevant variables (Moulatlet et al., 2017). This may have caused researchers to underestimate the role of soils in shaping tropical forests, as several studies have found only a weak relationship between map-derived soil information and the structure, composition and resilience of tropical forests (Albuquerque & Beier, 2015; Kissling et al., 2012; Levis et al., 2017; McPherson, 2014; Poorter et al., 2015; Thomas, Alcázar Caicedo, Loo, & Kindt, 2014). In stark contrast, such studies that have sampled soils in the field have found soil variables, including soil cation concentration, very important (Cámara-Leret et al., 2017; Higgins et al., 2015; Pansonato, Costa, de Castilho, Carvalho, & Zuquim, 2013; Phillips et al., 2003; Suominen et al., 2013; Tuomisto, Ruokolainen, et al., 2003; Tuomisto et al., 2014; Zuquim et al., 2014).

Although soil base cations are important plant nutrients and therefore directly linked to plant physiology and growth, broad-scale maps of their concentration in the soil are currently lacking for many areas, such as Amazonia. Our maps provide estimates of this information and can be incorporated in habitat and species distribution models across Amazonia. When using such plant-derived environmental data layers, it is important to be aware of which data they were based on in order to avoid circularity. For example, species distribution models (SDM) should never use as input occurrence data the same plant occurrence records that were already used to generate the soil map. Therefore, if our soil cation concentration map is used for SDMs of ferns or lycophytes, the species occurrence data that is currently included in GBIF or SpeciesLink should not be used as input data. However, SDMs concerning any other plant or animal group do not have this limitation.

4.2 | Transfer functions and indicator species optima

One can test if a given organism group is informative regarding an environmental variable of interest by applying a transfer function to first calculate species optima and then use these optima to reconstruct the variable at sites for which direct measurement data are available (Birks, 2003). Our tests using weighted averaging confirmed that ferns and lycophytes provide good predictors of soil cation concentration in Amazonia. We also found that, although species with broad tolerances along an environmental gradient are less informative than species with narrow tolerances, there was no need to exclude the generalist species as they did not noticeably increase the error in the environmental variable estimates. It is also noteworthy that we obtained similar results using presence–absence and abundance data, which is in agreement with an earlier study using a smaller dataset (Zuquim et al., 2014). This is good news both for the calculation of species optima and for applying them for modelling environmental conditions, because several existing datasets do not contain abundance information. Furthermore, presence–absence data are easier and faster to collect than abundance data.

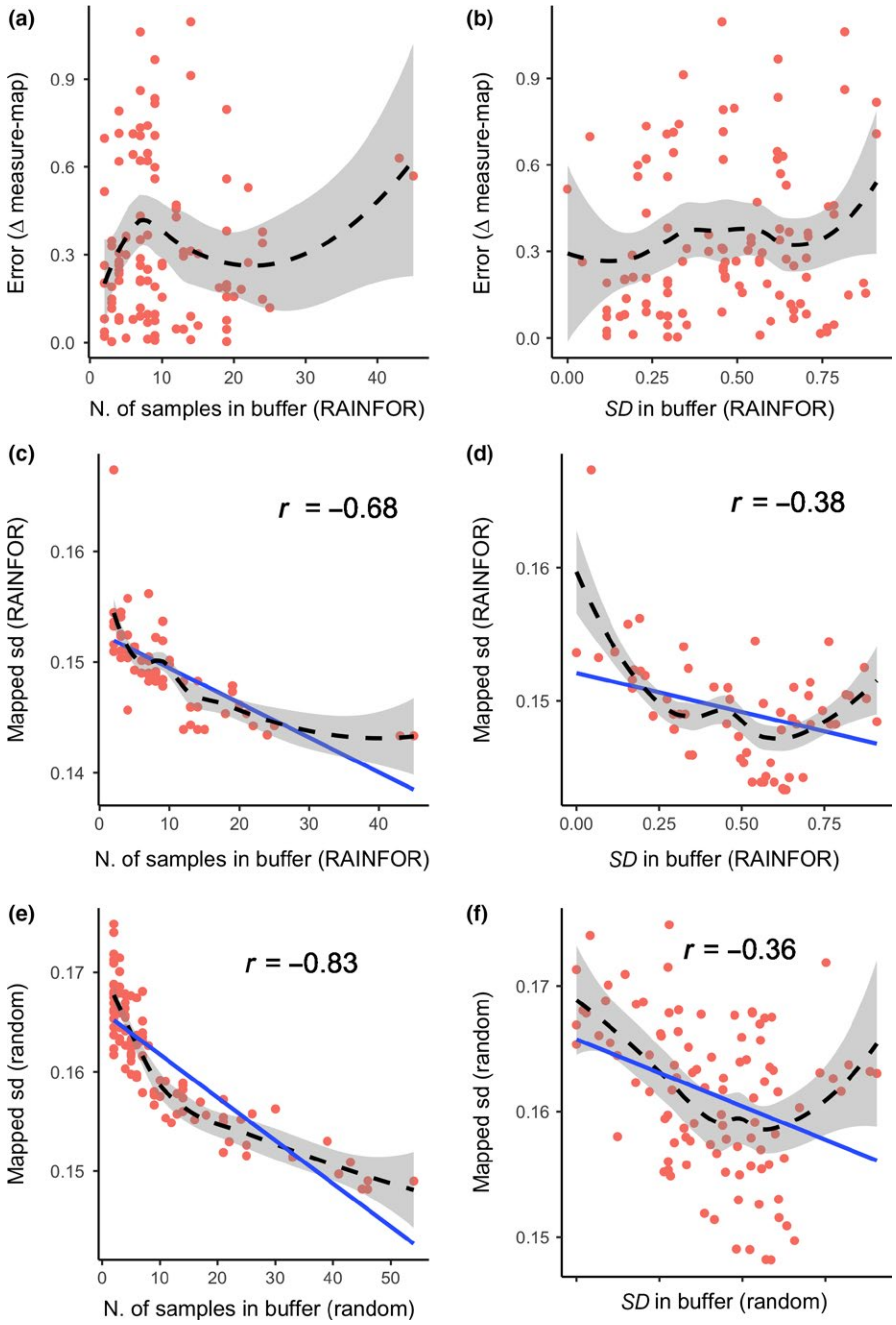

FIGURE 5 Relationships between error, uncertainty and sampling density in the soil cation concentration map shown in Figure 3e. Sampling density was defined as the number of data points within a buffer of 6 arcmin radius (~11 km) around each validation data point *(N. of samples in buffer [RAINFOR])* or around points randomly chosen in each of a 20 × 20 arcmin cells of an Amazonian grid *(N. of samples in buffer [random])*. Error was defined as the absolute difference between values observed in the RAINFOR validation dataset and mapped values *(Δ measure-map)*. Two kinds of indices were used to quantify uncertainty. The first was based on the standard deviation of input values inside each buffer *(sd in buffer [RAINFOR]; SD in buffer [random])* and the second on the mapped value of Kriging standard deviation for the centre points of the buffers *(Mapped sd [RAINFOR]; Mapped sd [random])*. Significant Pearson's correlation values (*r*) are shown in the upper right corner of each panel. When significant, linear regressions are shown with blue solid lines. Black dashed lines are loess smooth curves with two degrees of freedom and confidence intervals shown in gray

The accuracy of the calculated species optima is dependent on how representative the training data is of the environmental gradient of interest. If only a part of the gradient is sampled, or sampling density varies along the gradient, the optima may become biased. Our sampling covers a long edaphic gradient, but whether it is representative of all Amazonia is difficult to assess. Nevertheless, we believe that the optima are robust, because earlier studies have found soil associations of fern and lycophyte species to be consistent among non-overlapping study areas in different parts of Amazonia (Salovaara et al., 2004; Tuomisto & Poulsen, 1996; Tuomisto et al., 2002; Zuquim et al., 2014). Furthermore, there seems to be much redundancy in floristic communities, such that several species with similar optima coexist in any one locality. Consequently, relatively good predictions can be obtained even with superficial sampling that has many false absences, such as the entirely opportunistic species occurrence points that can be obtained from herbarium data through online portals such as GBIF and SpeciesLink.

When weighted averaging is used to calculate plant-derived environmental values, the repeated calculations of averages necessarily bias optimum values towards the overall mean of the variable, especially for data points at the extremes of the gradient. The resulting underestimation of the length of the environmental gradient can be mitigated with a deshrinking step (see Step 2), which restores the original gradient length and gives more reliable estimates (ter Braak & van Dam, 1989; Juggins, 2017).

4.3 | Interpolation and validation

In our example case, we used popular and simple interpolation methods that are widely applied in digital soil mapping (Cook, Jarvis, & Gonzalez, 2008; Grunwald, 2009). Because Kriging implements distance-based averaging, it gives relatively smooth and gradual changes in the modelled values, whereas IDW tends to respond more strongly to the values found at individual data points. With our data, the maps of soil cation concentration generated by Kriging were more accurate than those generated by IDW.

Digital mapping techniques are evolving rapidly, and more sophisticated and computation-intensive interpolation methods are becoming increasingly available (Hengl et al., 2017). These methods may use covariates and machine learning to refine the accuracy and resolution of maps. Modern machine-learning techniques have been applied to fit predictive models of edaphic conditions and create maps with global accuracy often close to a correlation of 60% (Hengl et al., 2017). However, high global accuracy does not mean that a map is uniformly accurate. On the contrary, map accuracy can vary drastically among areas, being lowest where data density is low (e.g. Amazonia, Africa).

Validation is an important step to evaluate the quality of maps. Surprisingly, it has been estimated that more than one-third of the published soil maps have not been validated at all (Grunwald, 2009). Validation can be done using an independent test dataset, or by cross-validation (Chatfield, 1995), even though the latter might overestimate the accuracy of the maps (Brus et al., 2011). We used an independent validation dataset with broad coverage. However, sampling in Amazonia is strongly biased towards accessible areas (Nelson, Ferreira, da Silva, & Kawasaki, 1990; Schulman, Toivonen, & Ruokolainen, 2007), so both the input dataset used for mapping and the validation dataset are spatially biased and large areas have very low data density. Yet, this bias seems to have a minor impact on our validation since the results obtained from random and sparsely-distributed points were similar.

Despite the advance in new modeling techniques, significant improvement in map accuracy for data-poor areas can only be expected if the quality of modeling input data increases. This can be achieved by including covariates in the models (Hengl et al., 2017) with the aid of remote sensed data (Van Doninck & Tuomisto, 2018) and/or increasing the number of input points. Our method tackles the latter by taking advantage of already existing information available in natural history museums, which makes it possible to obtain species-derived environmental variable estimates for sites lacking direct environmental measurements.

5 | CONCLUSIONS

Scarcity of input data is often a major constraint to the quality of thematic maps (Hengl et al., 2014; Lagacherie, 2008; Minasny, McBratney, & Lark, 2008). Here, we developed a method to alleviate data paucity and improve digital environmental mapping of data-poor areas. Our results demonstrate that georeferenced biological data can be used to interpolate environmental values over large, otherwise unsampled areas. We tested the method by using fern and lycophyte occurrences to map soil cation concentration in Amazonia, but the method is flexible and can be applied to other environmental variables and other organism groups. The prerequisite for success is that the variable to be mapped is strongly related to the occurrences of the selected indicator species group. This can be tested with transfer function analysis using field data that provide both species occurrence information and measurements of the environmental variable of interest at the same sites. Species with broad tolerances can be removed if their inclusion would reduce accuracy of predictions. Once the set of indicator species has been chosen, their optima along the gradient can be calculated and assigned to geographical locations where the species have been documented to occur. For this purpose, data from public repositories such as GBIF, SpeciesLink, VertNet, etc. can be used after appropriate cleaning. The values of the species-derived environmental variable can then be combined with actually measured values and used as input data points in interpolation to create a raster surface for the area of interest. Finally, the output digital maps can be used as environmental layers in studies such as species distribution modelling, habitat modelling, and development of habitat suitability scenarios under climate change. The results of such studies can contribute both to advancing future research and to conservation planning in poorly sampled areas.

ACKNOWLEDGEMENTS

The UTU and PPBio datasets have been accumulated over the years with funding from various sources, including the European Union, Academy of Finland, Brazilian National Council for Scientific and Technological Development (CNPq), Brazilian Ministry of Science and Technology (MCT) and Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM). We thank Mark Higgins for part of the UTU soil data, Jefferson Prado and Alan R. Smith for help in plant species identification, Pablo Pérez Chaves for comments and abstract in Spanish and numerous field assistants and colleagues for participation in fieldwork. ICMBio, CNPq and INRENA provided permits for fieldwork. We also thank those involved in making biological and environmental data accessible. G.Z. and JVd have been funded by Academy of Finland (grant 273737 to H.T.), J.S. by CNPq (#152816/2016-0) and GMM by the University of Turku Graduate School.

AUTHORS' CONTRIBUTIONS

G.Z., H.T. and J.S. conceived the ideas. C.A.Q., F.R.C.C., F.O.G.F., G.M.M., G.Z., H.T. and K.R. designed the field methods and collected the data; G.Z. developed the method and analysed the data with contributions from H.T., J.S., J.V.d. and G.M.M.; G.Z. led the writing with contribution from all authors.

DATA ACCESSIBILITY

Floristic and environmental data collected in the PPBio plots are available at https://search.dataone.org/#data/query=PPBio. The R codes used to calculate species optimum values and to perform spatial interpolation and the data used for these analyses are available at https://github.com/gabizuquim/paper_fern-soil_map (https://doi.org/10.5281/zenodo.2585607) The output maps generated using all the data and corresponding variance maps for whole Amazonia are available in raster format at https://doi.org/doi.pangaea.de/10.1594/pangaea.879542 (maps presented in Figure 3e–h).

ORCID

Gabriela Zuquim https://orcid.org/0000-0003-0932-2308
Juliana Stropp https://orcid.org/0000-0002-2831-4066
Jasper Van doninck https://orcid.org/0000-0003-2177-7882
Flávia R. C. Costa https://orcid.org/0000-0002-9600-4625
Hanna Tuomisto https://orcid.org/0000-0003-1640-490X

REFERENCES

Albuquerque, F., & Beier, P. (2015). Using abiotic variables to predict importance of sites for species representation. *Conservation Biology*, *29*(5), 1390–1400. https://doi.org/10.1111/cobi.12520

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Baldeck, C. A., Harms, K. E., Yavitt, J. B., John, R., Turner, B. L., Valencia, R., … Dalling, J. W. (2012). Soil resources and topography shape local tree community structure in tropical forests. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1753), 1–7. https://doi.org/10.1098/rspb.2012.2532

Birks, H. J. B. (2003). Quantitative palaeoenvironmental reconstructions from Holocene biological data. In A. Mackay, R. W. Battarbee, H. J. B. Birks, & F. Oldfield (Eds.), *Global change in the holocene* (pp. 107–123). London: Arnold.

Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, *62*(3), 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

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. *Journal of Ecology*, *105*(2), 367–381. https://doi.org/10.1111/1365-2745.12708

Carneiro, L. R. de A., Lima, A. P., Machado, R. B., & Magnusson, W. E. (2016). Limitations to the use of species-distribution models for environmental-impact assessments in the Amazon. *PLoS ONE*, *11*(1), e0146543. https://doi.org/10.1371/journal.pone.0146543

Chaplot, V., Darboux, F., Bourennane, H., Leguédois, S., Silvera, N., & Phachomphon, K. (2006). Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density. *Geomorphology*, *77*(1–2), 126–141. https://doi.org/10.1016/j.geomorph.2005.12.010

Chatfield, C. (1995). *Problem solving: A statician's guide* (2nd ed.). London: Chapman and Hall.

Cook, S. E., Jarvis, A., & Gonzalez, J. P. (2008). A new global demand for digital soil information. In A. E. Hartemink, A. McBratney & M. L. Mendonça-Santos (Eds.), *Digital soil mapping with limited data* (pp. 31–41). Dordrecht, the Netherlands: Springer.

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 Science Society of America Journal*, *69*(3), 649. https://doi.org/10.2136/sssaj2004.0140

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, *332*(6025), 53–58. https://doi.org/10.1126/science.1200303

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.

Donagena, G., Campos, D. V. B., Calderano, S. B., Teixeira, W. G., & Viana, J. H. M. (2011). *Manual de Métodos de Análise de Solo*, 2nd ed. Rio de Janeiro, Brazil: Embrapa Solos.

Ellenberg, H., Weber, H. E., Düll, R., Wirth, V., Werner, W., & Paulissen, D. (1992). Zeigerwerte von Pflanzen in Mitteleuropa. *Scripta Geobotanica*, *18*, 1–248.

Eva, H. D., & Huber, O. (2005). A proposal for defining the geographical boundaries of Amazonia. Synthesis of the results from an expert consultation workshop organized by the European Commission in collaboration with the Amazon Cooperation Treaty Organization—JRC Ispra, 7–8 June 2005. Office for Official Publications of the European Communities, Luxemburg.

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. *Journal of Biogeography*, *45*(1), 190–200. https://doi.org/10.1111/jbi.13104

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

Grunwald, S., Thompson, J. A., & Boettinger, J. L. (2011). Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. *Soil Science Society of America Journal*, *75*(4), 1201. https://doi.org/10.2136/sssaj2011.0025

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., … Kempen, B. (2017). SoilGrids250 m: Global gridded soil information based on machine learning. *PLoS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., … Gonzalez, M. R. (2014). SoilGrids1 km—Global soil information based on automated mapping. *PLoS ONE*, *9*(8), e105992. https://doi.org/10.1371/journal.pone.0105992

Hess, L. L., Melack, J. M., Affonso, A. G., Barbosa, C., Gastil-Buhl, M., & Novo, E. M. L. M. (2015). Wetlands of the lowland Amazon basin: Extent, vegetative cover, and dual-season inundated area as mapped with JERS-1 synthetic aperture radar. *Wetlands*, *35*(4), 745–756. https://doi.org/10.1007/s13157-015-0666-y

Higgins, M. A., Asner, G. P., Anderson, C. B., Martin, R. E., Knapp, D. E., Tupayachi, R., … Alonso, A. (2015). Regional-scale drivers of forest structure and function in Northwestern Amazonia. *PLoS ONE*, *10*(3), e0119887. https://doi.org/10.1371/journal.pone.0119887

Higgins, M. A., Ruokolainen, K., Tuomisto, H., Llerena, N., Cardenas, G., Phillips, O. L., … Räsänen, M. (2011). Geological control of floristic composition in Amazonian forests. *Journal of Biogeography*, *38*(11), 2136–2149. https://doi.org/10.1111/j.1365-2699.2011.02585.x

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Juggins, S. (2017). RIOJA: Analysis of quaternary science data, R package version (0.9-15). Retrieved from (http://cran.r-project.org/package=rioja).

Jürgens, N., Schmiedel, U., Haarmeyer, D. H., Dengler, J., Finckh, M., Goetze, D., … Zizka, G. (2012). The BIOTA Biodiversity Observatories in Africa—a standardized framework for large-scale environmental monitoring. *Environmental Monitoring and Assessment*, *184*(2), 655–678. https://doi.org/10.1007/s10661-011-1993-y

Kissling, W. D., Baker, W. J., Balslev, H., Barfod, A. S., Borchsenius, F., Dransfield, J., … Svenning, J.-C. (2012). Quaternary and pre-Quaternary historical legacies in the global distribution of a major tropical plant lineage. *Global Ecology and Biogeography*, *21*(9), 909–921. https://doi.org/10.1111/j.1466-8238.2011.00728.x

Lagacherie, P. (2008). Digital soil mapping: A state of the art. In A. E. Hartemink, A. McBratney & M. L. Mendonça-Santos (Eds.), *Digital soil mapping with limited data* (pp. 3–14). Dordrecht, the Netherlands: Springer.

Lavoie, C. (2013). Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics*, *15*(1), 68–76. https://doi.org/10.1016/j.ppees.2012.10.002

Levis, C., Costa, F. R. C., Bongers, F., Peña-Claros, M., Clement, C. R., Junqueira, A. B., … ter Steege, H. (2017). Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. *Science*, *355*(6328), 925–931. https://doi.org/10.1126/science.aal0157

Magnusson, W. E., Lima, A. P., Luizão, R., Luizão, F., Costa, F. R. C., de Castilho, C. V., & Kinupp, V. F. (2005). RAPELD: A modification of the Gentry method for biodiversity surveys in long-term ecological research sites. *Biota Neotropica*, *5*, 19–24. https://doi.org/10.1590/S1676-06032005000300002

Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., … Antonelli, A. (2015). Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? *Global Ecology and Biogeography*, *24*(8), 973–984. https://doi.org/10.1111/geb.12326

McMichael, C. H., Palace, M. W., Bush, M. B., Braswell, B., Hagen, S., Neves, E. G., … Czarnecki, C. (2014). Predicting pre-Columbian anthropogenic soils in Amazonia. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1777), 20132475–20132475. https://doi.org/10.1098/rspb.2013.2475

McPherson, T. Y. (2014). Landscape scale species distribution modeling across the Guiana Shield to inform conservation decision making in Guyana. *Biodiversity and Conservation*, *23*(8), 1931–1948. https://doi.org/10.1007/s10531-014-0696-4

Minasny, B., McBratney, A. B., & Lark, R. M. (2008). Digital soil mapping technologies for countries with sparse data infrastructures. In A. E. Hartemink, A. McBratney & M. L. Mendonça-Santos (Eds.), *Digital soil mapping with limited data* (pp. 15–30). Dordrecht, the Netherlands: Springer.

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. *Ecology and Evolution*, *7*(20), 8463–8477. https://doi.org/10.1002/ece3.3242

Nachtergaele, F., van Velthuizen, H., Verelst, L., & Wiberg, D. (2012). Harmonized World Soil Database version 1.2. (http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/ (Accessed 15 Setember 2016)

Nelson, B. W., Ferreira, C. A. C., da Silva, M. F., & Kawasaki, M. L. (1990). Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature*, *345*(6277), 714–716. https://doi.org/10.1038/345714a0

Pansonato, M. P., Costa, F. R. C., de Castilho, C. V., Carvalho, F. A., & Zuquim, G. (2013). Spatial scale or amplitude of predictors as determinants of the relative importance of environmental factors to plant community structure. *Biotropica*, *45*(3), 299–307. https://doi.org/10.1111/btp.12008

Pebesma, E. J., & Graeler, B. (2017). *Gstat: Spatial and spatio-temporal geostatistical modelling, prediction and simulation V.1.1-5*. https://cran.r-project.org/web/packages/gstat/index.html.

Phillips, O. L., Vargas, P. N., Monteagudo, A. L., Cruz, A. P., Zans, M.-E. C., Sánchez, W. G., … Rose, S. (2003). Habitat association among Amazonian tree species: A landscape-scale approach. *Journal of Ecology*, *91*(5), 757–775. https://doi.org/10.1046/j.1365-2745.2003.00815.x

Poorter, L., van der Sande, M. T., Thompson, J., Arets, E. J. M. M., Alarcón, A., Álvarez-Sánchez, J., … Peña-Claros, M. (2015). Diversity enhances carbon storage in tropical forests. *Global Ecology and Biogeography*, *24*(11), 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*(6), 1415–1440. https://doi.org/10.5194/bg-8-1415-2011

Quesada, C. A., Lloyd, J., Schwarz, M., Patiño, S., Baker, T. R., Czimczik, C., … Paiva, R. (2010). Variations in chemical and physical properties of Amazon forest soils in relation to their genesis. *Biogeosciences*, *7*(5), 1515–1541. https://doi.org/10.5194/bg-7-1515-2010

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Ruokolainen, K., Tuomisto, H., Macía, M. J., Higgins, M. A., & Yli-Halla, M. (2007). Are floristic and edaphic patterns in Amazonian rain forests congruent for trees, pteridophytes and Melastomataceae? *Journal of Tropical Ecology*, *23*(01), 13–25. https://doi.org/10.1017/S0266467406003889

Rylands, A. B. (1990). Priority areas for conservation in the Amazon. *Trends in Ecology & Evolution*, *5*(8), 240–241. https://doi.org/10.1016/0169-5347(90)90062-I

Salovaara, K. J., Cárdenas, G. G., & Tuomisto, H. (2004). Forest classification in an Amazonian rainforest landscape using pteridophytes as indicator species. *Ecography*, *27*(6), 689–700. https://doi.org/10.1111/j.0906-7590.2004.03958.x

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. *Environmental Monitoring and Assessment*, *140*(1–3), 279–289. https://doi.org/10.1007/s10661-007-9866-0

Schulman, L., Toivonen, T., & Ruokolainen, K. (2007). Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation: Amazonian collecting and range estimation. *Journal of Biogeography*, *34*(8), 1388–1399. https://doi.org/10.1111/j.1365-2699.2007.01716.x

Sirén, A., Tuomisto, H., & Navarrete, H. (2013). Mapping environmental variation in lowland Amazonian rainforests using remote sensing and floristic data. *International Journal of Remote Sensing*, *34*(5), 1561–1575. https://doi.org/10.1080/01431161.2012.723148

ter Steege, H., Haripersaud, P. P., Banki, O. S., & Schieving, F. (2011). A model of botanical collectors' behavior in the field: Never the same species twice. *American Journal of Botany*, *98*(1), 31–37. https://doi.org/10.3732/ajb.1000215

Suominen, L., Ruokolainen, K., Tuomisto, H., Llerena, N., & Higgins, M. A. (2013). Predicting soil properties from floristic composition in western Amazonian rain forests: Performance of *k* -nearest neighbour estimation and weighted averaging calibration. *Journal of Applied Ecology*, *50*(6), 1441–1449. https://doi.org/10.1111/1365-2664.12131

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*(1), 485–502. https://doi.org/10.1007/BF00028046

ter Braak, C. F., & van Dam, H. (1989). Inferring pH from diatoms: A comparison of old and new calibration methods. *Hydrobiologia*, *178*(3), 209–223. https://doi.org/10.1007/BF00006028

Thomas, E., Alcázar Caicedo, C., Loo, J., & Kindt, R. (2014). The distribution of the Brazil nut (*Bertholletia excelsa*) through time: From range contraction in glacial refugia, over human-mediated expansion, to anthropogenic climate change. *Bol Do Mus Para Emílio Goeldi Ciências Nat*, *9*, 267–291.

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. *Journal of Biogeography*, *43*(12), 2400–2411. https://doi.org/10.1111/jbi.12864

Tuomisto, H., & Poulsen, A. D. (1996). Influence of edaphic specialization on pteridophyte distribution in neotropical rain forests. *Journal of Biogeography*, *23*(3), 283–293. https://doi.org/10.1046/j.1365-2699.1996.00044.x

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

Tuomisto, H., & Ruokolainen, K. (1994). Distribution of Pteridophyta and Melastomataceae along an edaphic gradient in an Amazonian rain forest. *Journal of Vegetation Science*, *5*(1), 25–34. https://doi.org/10.2307/3235634

Tuomisto, H., Ruokolainen, K., Kalliola, R., Linna, A., Danjoy, W., & Rodriguez, Z. (1995). Dissecting amazonian biodiversity. *Science*, *269*(5220), 63–66. https://doi.org/10.1126/science.269.5220.63

Tuomisto, H., Ruokolainen, K., Poulsen, A. D., Moran, R. C., Quintana, C., Cañas, G., & Celi, J. (2002). Distribution and Diversity of Pteridophytes and Melastomataceae along Edaphic Gradients in Yasuní National Park, Ecuadorian Amazonia. *Biotropica*, *34*(4), 516–533.

Tuomisto, H., Ruokolainen, K., & Yli-Halla, M. (2003). Dispersal, environment, and floristic variation of western Amazonian forests. *Science*, *299*(5604), 241–244. https://doi.org/10.1126/science.1078037

Tuomisto, H., Zuquim, G., & Cárdenas, G. (2014). Species richness and diversity along edaphic and climatic gradients in Amazonia. *Ecography*, *37*(11), 1034–1046. https://doi.org/10.1111/ecog.00770

Van doninck, J., & Tuomisto, H. (2018). A Landsat composite covering all Amazonia for applications in ecology and conservation. *Remote Sensing in Ecology and Conservation*, *4*(3), 197–210. https://doi.org/10.1002/rse2.77

van, Reeuwijk, L. P. (1993). *Procedures for soil analysis*. 4th ed. International Soil Reference and Information Centre Technical Paper 9. Wageningen, The Netherlands: International Soil Reference and Information Centre.

Zuquim, G., Tuomisto, H., Jones, M. M., Prado, J., Figueiredo, F. O. G., Moulatlet, G. M., … Emilio, T. (2014). Predicting environmental gradients with fern species composition in Brazilian Amazonia. *Journal of Vegetation Science*, *25*(5), 1195–1207. https://doi.org/10.1111/jvs.12174

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zuquim G, Stropp J, Moulatlet GM, et al. Making the most of scarce data: Mapping soil gradients in data-poor areas using species occurrence records. *Methods Ecol Evol*. 2019;00:1–14. https://doi.org/10.1111/2041-210X.13178