Reconstructing continental-scale variation in soil $\delta^{15}$N: a machine learning approach in South America

JOÃO PAULO SENA-SOUZA 1,2,†, BENJAMIN Z. HOULTON 3, LUÍZ ANTONIO MARTINELLI 4, and GABRIELA BIELEFELD NARDOTO 5

1Programa de Pós-Graduação em Ciências Ambientais, Universidade de Brasília (UnB), Campus de Planaltina, Planaltina, Distrito Federal 73345-010 Brazil
2Departamento de Geociências, Universidade Estadual de Montes Claros (Unimontes), Campus Professor Darcy Ribeiro, Montes Claros, Minas Gerais 39401-089 Brazil
3Department of Land, Air and Water Resources, University of California, Davis, California 95616 USA
4Departamento de Ecologia Isotópica, Centro de Energia Nuclear da Agricultura (CENA), Universidade de São Paulo (USP), Campus de Piracicaba, Piracicaba, São Paulo 13416-000 Brazil
5Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade de Brasília (UnB), Campus Universitário Darcy Ribeiro, Asa Norte, Brasília, Distrito Federal 70910-900 Brazil

Citation: Sena-Souza, J. P., B. Z. Houlton, L. A. Martinelli, and G. Bielefeld Nardoto. 2020. Reconstructing continental-scale variation in soil $\delta^{15}$N: a machine learning approach in South America. Ecosphere 11(8):e03223. 10.1002/ecs2.3223

Abstract. Soil nitrogen isotope composition ($\delta^{15}$N) is an essential tool for investigating ecosystem nitrogen balances, plant–microbe interactions, ecological niches, animal migration, food origins, and forensics. The advancement of these applications is limited by a lack of robust geospatial models that are capable of capturing variation in soil $\delta^{15}$N (i.e., isotopic landscapes or isoscapes). Due to the complexity of the nitrogen cycle and general scarcity of isotopic information, previous approaches have reconstructed regional to global soil $\delta^{15}$N patterns via highly uncertain linear regression models. Here, we develop a new machine learning approach to ascertain a finer-scale understanding of geographic differences in soil $\delta^{15}$N, using the South American continent as a test case. We use a robust training set spanning 278 geographic locations across the continent, spanning all major biomes. We tested three different machine learning methods: cubist, random forest (RF), and stochastic gradient boosting (GBM). 10-fold cross-validation revealed that the RF method outperformed both the cubist and GBM approaches. Variable importance analysis of the RF framework pointed to biome type as the most crucial auxiliary variable, followed by soil organic carbon content, in determining the model performance. We thereby created a biogeographic boundary map, which predicted an expected multiscale spatial pattern of soil $\delta^{15}$N with a high degree of confidence, performing substantially better than all previous approaches for the continent of South America. Therefore, the RF machine learning framework showed to be a great opportunity to explore a broad array of ecological, biogeochemical, and forensic issues through the lens of soil $\delta^{15}$N.

Key words: isoscape; machine learning; nitrogen cycle; random forest; soil nitrogen; stable nitrogen isotopes.

Received 6 May 2019; revised 9 December 2019; accepted 24 April 2020; final version received 17 June 2020.

Corresponding Editor: Lucas N. Joppa.

Copyright: © 2020 The Authors. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

† E-mail: jpsenasouza@gmail.com
INTRODUCTION

Coherent patterns of soil nitrogen (N) isotope composition (\(^{15}\)N/\(^{14}\)N) have been used to constrain an array of N cycling processes, including N fixation, mineralization, nitrification, denitrification, and leaching, leading to improved understanding of global biogeochemical fluxes and forecasts (e.g., Robinson 2001, Houlton and Bai 2009, Vitousek et al. 2013, Houlton et al. 2015, Zhu and Riley 2015, Denk et al. 2017). Moreover, \(^{15}\)N isotopic landscapes, or isoscapes, can provide critical spatial baseline information for examining animal migration (Hobson et al. 2012, Garcia-Perez and Hobson 2014), forensic drug trafficking, and tracing food origins (Chesson et al. 2018). However, uncertainty over spatial patterns of soil \(^{15}\)N is a significant obstacle to further advancing various N isotopic modeling approaches, with widespread implications for diverse fields of inquiry (Amundson et al. 2003, Bai and Houlton 2009, Bai et al. 2012).

The first spatially explicit soil \(^{15}\)N map was created by using simple regressions between \(^{15}\)N and climate, based on a select number of the data extrapolated to continents and the globe (Amundson et al. 2003). More recently, studies have included clay and organic carbon content in such global soil \(^{15}\)N analysis (Craine et al. 2015). While these past approaches have proven insightful, they are yet to provide a robust, spatial representation of N isotope composition across diverse ecosystems and landscapes, where soil \(^{15}\)N can vary from \(-4\%\) to \(20\%\) in temperate ecosystems to \(20\%\) in tropical ecosystems (Martinelli et al. 1999). Furthermore, previous global models tend to mask critical regional variations of soil \(^{15}\)N within biomes (Freitas et al. 2010). Within savanna ecosystems, for example, the spatial pattern of soil \(^{15}\)N at regional scale does not adhere to predictions of global models (Cheng et al. 2009, Freitas et al. 2010, Chen et al. 2018). Therefore, methods that integrate discrete and categorical variables can be essential tools for advancing our ability to reconstruct soil \(^{15}\)N variation, both within and among ecosystems.

Here, we explore the use of machine learning approaches in the assessment of spatial patterns of soil \(^{15}\)N in South America. Ensemble learning algorithms are a class of machine learning techniques that combine several models to build a robust final prediction. Unlike other approaches, ensemble learning algorithms have the distinct advantage of integrating digital mapping of soil attributes as both discrete and continuous variables (Hengl et al. 2017, Gomes et al. 2019). Among the most widely used ensemble learning algorithms are the cubist (Adhikari et al. 2014, Gray et al. 2015, Viscarra Rossel et al. 2016), stochastic gradient boosting (Yang et al. 2016), and random forest (Yang et al. 2016, Bataille et al. 2018, Hengl et al. 2018, Hounkpatin et al. 2018, Gomes et al. 2019) varieties. These methods are also emerging as a powerful alternative for mapping isotopic variations in the environment (Bataille et al. 2018). Despite the full range of applications in mapping soil attributes, machine learning approaches are yet to be systematically applied to the problem of soil \(^{15}\)N from regional to continental scales.

Our objective was to generate a useful soil \(^{15}\)N isoscape for South America by applying a variety of machine learning approaches. Although we test three machine learning techniques, we hypothesize that random forest algorithm will have the best performance in predicting soil \(^{15}\)N. This hypothesis is based on recent papers where random forest algorithm produced more accurate results in mapping soil attributes when compared to neural networks, linear models, and other ensemble learning algorithms (Bataille et al. 2018, Hengl et al. 2018, Gomes et al. 2019).

METHODS

Study area and soil \(^{15}\)N dataset

We used a total of 278 georeferenced locations to create the training dataset and validation of our machine learning model, of which 135 are from published soil \(^{15}\)N, and 143 data points are new data from fieldwork. All georeferenced locations are in unmanaged vegetation. We used the Web Plot Digitizer to extract values from plots when data were not available in tables (Rohatgi 2018). The data locations in our study are distributed across Brazil, Argentina, Chile, Peru, Venezuela, Ecuador, and Colombia. We converted the geographic coordinate system to decimal latitude and longitude WGS-84. We considered only soil sampled above 20 cm of
depth. For locations with multiple depth values between 0 and 20 cm, we standardized δ\(^{15}\)N by calculating weighted means of N percent. We did not select deeper soil because there is a strong correlation between surface and deep soil δ\(^{15}\)N (Amundson et al. 2003). The geographic distribution of data encompassed seven of the nine biomes in South America, according to the World Wild Foundation biomes boundaries (Olson et al. 2001; Fig. 1).

We aggregated geographical locations ≤1 km distance (~0.01 degree) and calculated mean soil δ\(^{15}\)N among aggregated points by using the aggregate function in R package sp, thus reducing over-representation of specific sites and any spatial biases.

Fig. 1. South America continent showing its biomes (see legend) and geographical locations of the compiled dataset (black circles).
Soil sampling and isotopic analysis

We collected some sets of samples for this study in fragments of the original vegetation. We sampled at 0–20 cm depth interval by using a Dutch soil sampler. Soil samples were air-dried and sieved using a 2-mm mesh to a level of air-dried fine earth (ADFE). A subsample of 30–35 mg of soil was placed into tin capsules for N and isotopic analysis using an elemental analyzer (Carlo Erba, model 1110, Milan, Italy). Combustion products were purified in a gas chromatography column and introduced directly into a mass spectrometer for isotopic analysis (ThermoQuest-Finnigan Delta Plus, Finnigan-MAT, California, USA). The natural abundance of \(^{15}\text{N}/^{14}\text{N}\) is expressed as deviations per mil (‰), an internationally recognized standard, through the equation \(\delta = (R_{\text{sample}}/R_{\text{standard}} - 1) \times 1000\), where \(R\) is the \(^{15}\text{N}/^{14}\text{N}\) ratio of the sample and standard. The standard for N is the atmospheric air with an N isotopic composition of 0‰. Internal, certified laboratory standards are routinely interspersed with target samples during analysis runs. Long-term standard deviations of internal standards (sugarcane leaves and tropical soil) used at the Ecology Isotope Laboratory (CENA/University of São Paulo) are 0.2‰ for \(\delta^{15}\text{N}\). Isotopic analyses were performed at the Isotope Ecology laboratory at the Center for Nuclear Energy in Agriculture (CENA), University of São Paulo, Piracicaba, Brazil.

Auxiliary variables

We used predictive variables from global geospatial datasets in our model (Table 1). Climatic variables are from the WorldClim database (Hijmans et al. 2005). We used the mean annual temperature (MAT) and mean annual precipitation (MAP), the aridity index (ai), and potential evapotranspiration (pet) in our model.

At the global scale, soil \(\delta^{15}\text{N}\) has been shown to correlate with both climatic variables and pedological variables, such as soil organic carbon content and clay content (Craine et al. 2015). For soil variables, we used the Soilgrid database for South America with a spatial resolution of 250 m (Hengl et al. 2017). In addition to soil organic carbon and clay content, we used bulk density, cation exchange capacity, pH (H$_2$O and KCl extractable), silt, and sand content in our analysis (Table 1). For elevation, we used a digital elevation model (DEM) with a 1 km spatial resolution from the Shuttle Radar Topography Mission (SRTM) (Jarvis et al. 2008).

We used the net primary production (NPP) from MODIS/Terra data (MOD17A3; Zhao et al. 2005). We specifically aggregated annual NPP grids over the period between 2000 and 2015 to estimate annual averages for NPP. All NPP grids are from the Numerical Terradynamic Simulation Group at the University of Montana. The biome and ecoregion maps used in our analysis were obtained from the World Wild Foundation (Olson et al. 2001). We transformed the biome and ecoregion original shapefile format to raster using ESRI ArcGIS 10.3 (ESRI 2016).

Modeling

We based the spatial analysis of soil \(\delta^{15}\text{N}\) on the caret package on R version 3.5.0 (Kuhn 2008). We adapted the script from Bataille et al. (2018) to our dataset (see Data S1). First, we projected all auxiliary data into the pattern projection WGS-84. Then, we resampled all auxiliary data at 1-km grid cells, by using the bilinear interpolation method for continuous variables and nearest-neighbor interpolation for categorical variables. After standardizing the spatial resolution of auxiliary data, we extracted the values for each geographical location to build the regression matrix, which we used to train the spatial prediction models in the caret package.

To select a group of representative auxiliary variables, we applied the recursive feature elimination (RFE) method, also present in the caret package (Kuhn 2008). RFE is an algorithm for retroactive selection of predictors. The RFE technique first builds a model with all predictors and calculates the importance of each one to the final model; then, it removes the least predictor and rebuilds the model (Kuhn and Johnson 2019). One of the advantages of this method is that RFE reduces the effect of the correlation on the importance measure (Gregorutti et al. 2016). In this study, we ran the RFE with thirteen subsets of variables: 4, 5, 6... 16. We selected the best subset of covariates based on the lower root mean squared error (RMSE) after a 10-fold cross-validation.

We stacked the chosen covariates to assess the accuracy and uncertainty of model outcomes. We compared three regression methods available in
the caret package (Kuhn 2008): cubist, random forest (RF), and stochastic gradient boosting (GBM; Fig. 2).

RF and GBM methods are similar to classification and regression trees (CART), which analyze variables that best appropriate values into two distinct groupings. The appropriation of values in distinct groups is repeated for subsets until a regression tree is constructed. The RF model is formed by a large number of random trees during the model-training process. All random trees are merged into a single prediction to reduce the noise and increase the accuracy of results. The random trees are averaged, resulting in low prediction variance. This approach is known as ensemble learning (bagging; Kuhn 2008). Model error is estimated by using the out-of-bag data (OOB). The main parameter of the RF method is the number of features tested in each node (m.try argument), which is used for model optimization (Liaw and Wiener 2002, Hengl et al. 2018). In this study, we used m.try = 4.

In addition to RF, GBM is also an ensemble learning approach. The main difference is that trees in the GBM method are built following a sequence as opposed to the random pattern approach of RF. After each regression tree is created, a new subsample is selected at random from the training dataset, which is used in the following interactions (Friedman 2002). The errors generated by the previous trees are corrected in the subsequent tree created in the GBM. Standard parameters used for tuning GBMs are the number of trees or the amount of boosting interactions (n.trees argument), the depth of trees (interaction.depth argument), and the learning rate (shrinkage argument; Greenwell et al. 2019). We applied the follow parameters: n.trees = 500; interaction.depth = 30; shrinkage = 0.1.

Cubist is a rule-based model where, for each node of the model tree, linear regression is generated, which is smoothed by the linear regression of the previous node. Initially, the model tree is defined by a set of rules, which in turn can be combined or eliminated for model simplification. The committee is the main tuning parameter of the cubist method. It works as a boosting scheme with the creation of interactive trees in sequence, in addition to establishing the number of trees (Kuhn and Quinlan 2018). Subsequent model trees are created by adjusting the error of the previous model trees. The final prediction is established by the outcome linear regression model, with influence from previous models along the same path through the model tree.

We predicted soil δ^{15}N by applying each of the three tested models using chosen covariates to select the most accurate one. We separated the

| Covariates (raster name) | Meaning | Resolution | Type | Reference |
|--------------------------|---------|------------|------|-----------|
| tmp                      | Mean Annual Temperature WorldClim (°C) | 1 km | C | Hijmans et al. (2005) |
| pre                      | Annual Precipitation WorldClim (mm) | 1 km | C | Hijmans et al. (2005) |
| ai                       | Aridity Index | 1 km | C | Zomer et al. (2008) |
| pet                      | Potential Evapotranspiration (annual mean) | 1 km | C | Zomer et al. (2008) |
| density                  | Bulk density (fine earth) in kg/m³ at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| cec                      | Cation exchange capacity of soil in cmol/kg at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| clay                     | Clay content (0–2 µm) mass fraction in % at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| orc                      | Soil organic carbon content (fine earth fraction) in g/kg at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| phH2O                    | Soil pH × 10 in H₂O at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| phKcl                    | Soil pH × 10 in KCl at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| silt                     | Silt content (2–50 µm) mass fraction in % at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| sand                     | Sand content (50–2000 µm) mass fraction in % at depth 0.05 m | 250 m | C | Hengl et al. (2017) |
| rpp                      | MODIS17 A3 Net Primary Production mean 2000–2015 (g C/m²·yr⁻¹) | 30-arc sec | C | Zhao et al. (2005) |
| alt                      | Elevation (m) | 1 km | C | Jarvis et al. (2008) |
| biome                    | Major biome boundaries for S. America | | D | Olson et al. (2001) |
| ecoreg                   | Ecoregions | | D | Olson et al. (2001) |

Note: C, continuous; D, discrete.
dataset randomly into a training subset (80% of the entire dataset) and a testing holdout subset for validation (20% of the total dataset; Fig. 2). We used 10-fold cross-validation with five repetitions with the training data to access the variability of the predictions. Hence, each model was repeated five times, using 90% of the training data, while the remaining 10% (OOB) was used for validation. That procedure gave the average value of root mean squared error (RMSE), mean absolute error (MAE), and R-square ($R^2$), as statistical metrics for model performance. To evaluate the final model’s performance, we applied the fitted model to the testing holdout data (20%) and used the statistical metrics RMSE, MAE, and $R^2$ to validate overall model efficacy.

We visualized the importance of each covariate to the final model by calculating the percentage increase in the mean standard error (%IncMSE) of each covariate. Besides, we assessed how the

---

**Fig. 2.** Workflow used for modeling soil $\delta^{15}$N in South America. The covariates are listed in Table 1.
essential covariates influenced the target variable by using partial dependence plots, where the y-axes indicate the mean marginal effect of each value of a given covariable to the final prediction (Friedman 2001). This approach has the advantage of graphically displaying the correlation between a given feature and the target variable.

**Spatial prediction of soil \( \delta^{15}N \) isoscape**

We applied the whole dataset (\( n = 278 \) geographical locations) on the best performing model to create the final soil \( \delta^{15}N \) isoscape map for South America. We performed the spatial predictions by using the prediction function in a raster package on R and applying the chosen model to the stacked covariate raster. There is no function to predict the spatial uncertainty from the random forest model. An alternative way is to make several model predictions to calculate the coefficient of variation (CV % = (standard deviation/mean) \( \times \) 100; Gomes et al. 2019). For this research, we applied 20 spatial predictions to RF models.

**RESULTS**

**Model performances and comparisons**

The highest performance on the RFE selection was the model with the sixteen covariates (Fig. 3). Cross-validation values of RMSE, \( R^2 \), and MAE from the RF, cubist, and GBM algorithms showed that RF was best able to capture geospatial variation in soil \( \delta^{15}N \) in South America (Fig. 4). Although the models were not statistical different (Kruskal-Wallis test, \( P \) value > 0.05), the average \( R^2 \) was slightly higher in RF (\( R^2 = 0.64 \)) vs. cubist (\( R^2 = 0.61 \)) and GBM (\( R^2 = 0.60 \); Fig. 4A) models. For RMSE, the average was lower in RF algorithm (1.85‰) than in cubist (1.92‰) and GBM (1.97‰; Fig. 4B). The same pattern was observed for MAE, where RF had the lowest average value (1.37‰), followed by cubist (1.40‰) and GBM (1.52‰; Fig. 4C).

Soil \( \delta^{15}N \) predicted by the RF algorithm explained 94% of the observations in the training data, with an RMSE of 0.63‰ (\( P \) value < 0.01; Fig. 5A). After 10-fold cross-validation, the predicted \( \delta^{15}N \) values explained 64% of the values observed in the training data, with an RMSE of 1.85‰ (\( P \) value < 0.01; Fig. 5B). However, the most robust performance evaluation approach involves a comparison between model-predicted \( \delta^{15}N \) and the holdout testing data that we did not use in the model (\( n = 50 \)). In this case, the predicted values explained 62% of the variation in the holdout testing data, with an RMSE of 1.72‰ (\( P \) value < 0.01; Fig. 5C).

**Importance of covariates**

Biome type was the most important predictor variable for geospatial soil \( \delta^{15}N \) in South America. Covariate analysis showed that biome contributed 17.5% to the decrease in mean accuracy (Fig. 6). Soil organic carbon was the second most important variable, with a 16.4% contribution, followed by climatic variables related to precipitation (pre, ai), bulk density, and NPP (Fig. 6).

Figure 7 shows the influence of the six most crucial covariates for predicting soil \( \delta^{15}N \). Each biome had a distinct signature range in soil \( \delta^{15}N \). Deserts and xeric shrublands showed the highest values of soil \( \delta^{15}N \) between the observed data (11.9 ± 3.6‰). Tropical moist forests had the second highest soil \( \delta^{15}N \) value (7.4 ± 2.0‰). Conversely, temperate biomes showed the lowest values of soil \( \delta^{15}N \): Temperate mixed forests showed the lowest mean value of soil \( \delta^{15}N \) between all represented biomes (1.1 ± 3.0‰), followed by temperate grasslands, savannas, and shrublands (3.4 ± 2.6‰). Tropical dry forests, tropical savannas, and shrublands, and flooded grasslands were intermediate, with soil \( \delta^{15}N \)

---

**Fig. 3. Models’ performance in the process of selecting covariates by using the Recursive Feature Elimination method. The black dot circled in dark gray indicates the subset with lowest RMSE after the cross-validation.**
ranging from 6.8 ± 1.9‰, 5.5 ± 1.6‰, and 5.0 ± 1.4‰, respectively (Fig. 7A).

Soil organic carbon content showed a negative correlation with soil δ¹⁵N. We found a sharp drop in soil δ¹⁵N between 0 and 25 g/kg of organic carbon. The negative trend was much less between 25 and 100 g/kg of organic carbon and again had a sharp drop between 100 and 150 g/kg of organic carbon (Fig. 7B). Low MAP was related to higher soil δ¹⁵N, following a
negative correlation trend up to 1100 mm of precipitation. Between 1100 and 1900 mm MAP, the predicted marginal values were much smaller than the values for other MAP ranges; and it was also possible to observe an increase in the partial value of predicted $\delta^{15}$N as precipitation increases, followed by a sharp rise at 2000 mm. After this value, the inverse correlation returned, with higher values of $\delta^{15}$N (Fig. 7C). The same was seen for the aridity index since it is a covariate correlated with MAP (Fig. 7D).

Bulk density exhibited a general pattern of positive correlation with $\delta^{15}$N, with a sharp increase in $\delta^{15}$N values between 900 and 1000 g/m$^3$. Partial dependence continued to rise as bulk density values range from 1000 to 1400 g/m$^3$, but less strongly (Fig. 7E). The influence of NPP on predicted soil $\delta^{15}$N was more complex, with a negative correlation between 200 and 600 gC.m$^{-2}$.yr$^{-1}$ of NPP, a positive correlation between 600 and 700 gC.m$^{-2}$.yr$^{-1}$, and a negative between 700 and 900 gC.m$^{-2}$.yr$^{-1}$. The correlation was again positive between 1000 and 1200 gC.m$^{-2}$.yr$^{-1}$ before stabilizing (Fig. 7F).

Uncertainty of the model
Regression residuals between observed and predicted soil $\delta^{15}$N did not exhibit spatial autocorrelation (Moran’s index = −0.02; $P$ value = 0.77). The extreme values of residuals, both positive and negative, were concentrated in the areas with maximum and minimum observed soil $\delta^{15}$N, specifically in the northeastern region of Brazil and in Patagonia in Argentina (Fig. 8A). The coefficient of variation map calculated from the predictions showed larger inconsistencies in the southern area of the continent (Fig. 8B). CV higher than 40% was observed for the mountainous area along the Andes. On the other hand, the lowest CV was concentrated in the northern region corresponding to tropical forests, in the semi-arid tropical region, and in the tropical savannas (Fig. 8B). The spatial pattern of the CV followed the spatial distribution of available data, with higher values in places with little or no sampling.

Prediction of soil $\delta^{15}$N for South America
The highest soil $\delta^{15}$N was observed for the northeastern region of Brazil, where modeled values of soil $\delta^{15}$N (0–20 cm) reached approximately 15‰ (Fig. 9). High soil $\delta^{15}$N values were also mapped in the northern region of Brazil (between 8‰ and 11‰, approximately), in the Amazon Forest region, and the southern areas of Bolivia and northwestern Argentina. The lowest values appeared in the extreme southwest and south of South America, with negative isotopic ratios reaching approximately −2‰. Negative or near zero soil $\delta^{15}$N also appeared in the northwest region of South America, following the mountainous Andes between Peru and Colombia. Intermediate values, between 3‰ and 7‰, were mainly in the central region of Brazil, where the tropical savannas occur, and to the south in a continuous range that follows from the central region of Paraguay to the south of Argentina, where the temperate savannas predominate (Fig. 9).

Discussion
Source of uncertainty and model performance
Our machine learning was capable of reconstructing patterns of soil $\delta^{15}$N for the South American continent, providing an improved...
prediction of observed soil $\delta^{15}$N variation (down to 20 cm) compared to previous regression-based approaches (Amundson et al. 2003). This result is consistent with the development of machine learning algorithms in digital soil mapping and biogeochemical predictions (Hengl et al. 2017, Bataille et al. 2018, Gomes et al. 2019). Among machine learning approaches, the RF model performed best, followed by the cubist and GBM algorithms. Upon 10-fold cross-validation, the RF model produced an $R^2 = 0.64$ vs. observed soil $\delta^{15}$N data. This significant correlation was

Fig. 7. Plots showing the influence of the six most essential variables to soil $\delta^{15}$N prediction using the random forest algorithm. (A) Plot showing observed mean ± SD values of observed soil $\delta^{15}$N by biome as categorical variable, reordered from the higher to the lower mean value: 1—Deserts and Xeric Shrublands; 2—Tropical and Subtropical Moist Broadleaf Forest; 3—Tropical and Subtropical Dry Broadleaf Forest; 4—Tropical and Subtropical Grasslands, Savannas and Shrublands; 5—Flooded Grasslands, Savannas and Shrublands; 6—Temperate Grasslands, Savannas and Shrublands; and 7—Temperate Broadleaf and Mixed Forest. (B, C, D, E, F) are partial dependence plots for some essential variables: (B) soil organic carbon; (C) mean annual precipitation; (D) aridity index; (E) bulk density; (F) net primary production. For detailed information about covariates, check Table 1. Marks on the x-axis indicate the data distribution.
roughly similar to our validation with the subset of data not used in the RF model (i.e., $R^2 = 0.62$ vs. the testing subset), indicating that the model was not significantly affected by overfitting.

RF yielded the most efficient algorithm in our study is consistent with previous RF applications to the mapping of different soil variables (Nawar and Mouazen 2017, Gomes et al. 2019). Similarly, in an attempt to map strontium isoscape to central Europe, Bataille et al. (2018), after comparing the efficiency of several methods besides the GBM and cubist algorithms, among them ordinary kriging, neural network, and the generalized linear model, found that RF had also the best performance.

The highest CV values concentrated in areas with limited data in our study followed the same pattern in the literature with the assumption that spatial gaps in the dataset are the primary sources of error in the modeling of any soil variable (Malone et al. 2018). In particular, areas between the central and southern regions of South America are the most underrepresented among soil $\delta^{15}$N observations. Filling these spatial gaps can further improve the accuracy of models using RF (Hengl et al. 2018).

**Soil $\delta^{15}$N spatial distribution in South America**

Our study identified biome type as the essential factor for understanding spatial patterns of soil $\delta^{15}$N at regional to continental scales. Areas of xeric shrublands in northeastern Brazil had the highest modeled soil $\delta^{15}$N, which is consistent with isotopic fractionation during ammonia volatilization, that leaves high residual soil $^{15}$N/$^{14}$N ratios in areas with water deficits and

---

**Fig. 8. Spatial patterns of model uncertainty: (A) residuals from observed vs predicted regression of the whole dataset; (B) coefficient of variation (CV).**
high pH soils (Högberg et al. 1997, Menezes et al. 2012). The same mechanisms are likely to explain the high soil δ¹⁵N (between 10‰ and 12‰) predicted for the Atacama Desert region (Ehleringer et al. 1992) and in the semi-arid areas of northwestern Argentina and Paraguay.

Conversely, the lowest predicted soil δ¹⁵N is apparent in temperate climates and along the mountainous Andes region in the northwest of South America. Temperate ecosystems have a slower rate of soil organic matter decomposition and N mineralization, which reduces isotopic...
fractionation in the surface soil, resulting in soil $\delta^{15}N$ convergence to litter inputs from plants (Martinelli et al. 1999). Further, N gaseous losses are relatively low in the mountainous Andes region compared to tropical forests, leading to less isotopic fractionation during gaseous N losses. Peri et al. (2012) found similar $\delta^{15}N$ of surface soil and leaves along a precipitation gradient in the Patagonia region of Argentina, indicating that there is small isotopic fractionation during organic matter decomposition in the surface soil.

In tropical rainforest areas, the RF algorithm was capable of capturing finer (regional) scale soil $\delta^{15}N$ variation than regression models based on climate alone. Such improved capacity was seen along a tropical gradient, mainly in the Amazon forest, where the predicted soil $\delta^{15}N$ decreased from east to west. This gradient of $\delta^{15}N$ has been widely observed in previous empirical studies (Nardoto et al. 2008, 2014). The systematic decline in soil $\delta^{15}N$ within the Amazon is also apparent in partial dependence plots with MAP, where a decrease in the modeled $\delta^{15}N$ occurs between MAP 2000 and 2500 mm. Due to the high soil N emissions in the eastern Amazonia (Davidson et al. 2004), isotopically fractionating denitrification losses are probably the main driver of the increase of soil $\delta^{15}N$ in this region (Houlton et al. 2006, Pérez et al. 2006).

The extreme positive $\delta^{15}N$ values in the semi-arid areas probably accentuated these correlations, where there is a low quantity of soil organic carbon, and negative values in temperate forests, with the highest amount of organic carbon. However, soil organic carbon was unable to explain the spatial distribution of the intermediate values of soil $\delta^{15}N$ in South America (from 3‰ to 7‰), which appear mainly in the tropical and temperate savanna areas of the model. Those three large patterns described above are partially (though not wholly) captured in the global trends found by Craine et al. (2015), in which latitudinal gradients of soil $\delta^{15}N$ were shown to be indirectly related to the MAP and MAT, which in turn influenced clay content and soil organic carbon. Craine et al. (2015) suggested that in the hot and/or dry regions, there is a higher amount of mineral organic matter, enriched in $^{15}N$, than less enriched non-mineral organic matter. It might explain the high influence of the organic carbon model on the prediction of soil $\delta^{15}N$, with the lowest values of organic carbon related to the highest predicted values of soil $\delta^{15}N$.

Savannas from South America have been substantially underrepresented in previous $\delta^{15}N$ soil models (Amundson et al. 2003, Craine et al. 2015). Our study adds an array of new data, thereby making it possible to examine correlations with soil $\delta^{15}N$ in this biome. The partial dependence plot shows that within the savanna’s precipitation range (between 1000 and 2000 mm), predicted soil $\delta^{15}N$ decreases considerably, followed by increasing soil $\delta^{15}N > 2000$ mm. Low predicted soil $\delta^{15}N$ in savannas compared to semi-arid and forest regions are consistent with observations in the central region of Brazil (Bustamante et al. 2004). Savannas of South America have low N availability compared to tropical rainforests, but with less severe water deficit than tropical semi-arid regions. Particularly, the savannas of Central Brazil present low decomposition rates due to high C:N ratios (~60:1), low net N mineralization, although with high NH$_4$ production in the soil (Nardoto and Bustamante 2003, Bustamante et al. 2006). This low N availability in the plant–soil system seemed to have been captured by low soil $\delta^{15}N$ that the model presented here.

Spatial variation of soil $\delta^{15}N$ is scale-dependent, given the difficulties encountered by global models in explaining important patterns at some regional scales (Cheng et al. 2009, Freitas et al. 2010, Chen et al. 2018). Even so, the Amundson et al. (2003) model has proven useful in several global scale studies of soil and plant $\delta^{15}N$. The difficulty in representing spatial regions comes from the fact that regional geographic or ecosystem limits have not been included in the analysis. Studies show that even in global biogeochemical models, the boundaries of biomes or climatic regions are important auxiliary variables (Terzer et al. 2013). When comparing our model with Amundson’s et al. (2003), the importance of regional boundaries for future modeling of $\delta^{15}N$ is evident (Fig. 10). Our model was able to delimit a large amplitude of soil $\delta^{15}N$ variation, showing spatial patterns of non-linear correlation with MAP (Fig. 10A). On the other hand, Amundson’s global model captured a smaller amplitude of variation of $\delta^{15}N$ compared to our model (Fig 10B).
Applicability of the machine learning soil $\delta^{15}N$ isoscape model

Spatial models of soil $\delta^{15}N$ have been identified as an essential benchmarking tool for improving N-carbon forecasts in global climate change (Houlton et al. 2015). Soil $\delta^{15}N$ isoscapes at large scales aggregate information on N balance and loss to the atmosphere from the soil–plant system (Brenner et al. 2001, Houlton and Bai 2009). Our model, therefore, offers a more robust constraint on models and for understanding local to global biogeochemical fluxes. Bai et al. (2012) demonstrated the strength of spatial models of soil $\delta^{15}N$, adapting the equation proposed by Amundson et al. (2003) to estimate the gaseous N losses of terrestrial ecosystems. The authors identified high NO emissions in Africa and high N$_2$ emissions in the southeastern United States, where there are few N emission measurements (Bai et al. 2012). The addition of the regional factor and the increase in the accuracy of the base models, as presented here, will increase the power of these predictions.

Soil $\delta^{15}N$ isoscapes also hold demonstrated utility for conservation studies, providing a pathway for tracking wild animal movement (Hobson et al. 2012, Garcia-Perez and Hobson 2014). These studies generally have a multi-isotopic approach, aggregating oxygen and hydrogen stable isotopes from water together with stable isotopes over the animal’s trophic niche, such as carbon and nitrogen. By relating these isoscapes to the animal tissue isotopic ratios, it is possible to create probability maps of origin (Garcia-Perez and Hobson 2014). The isoscape presented here has the potential to reduce the probability area of assignment studies since it accessed regional aspects of the variation of $\delta^{15}N$. This same principle applies to the use of the model in forensic investigations (Chesson et al. 2018), marking our approach of global cross-disciplinary importance.

ACKNOWLEDGMENTS

João Paulo Sena-Souza received a doctoral scholarship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil (CAPES), Finance Code 001. He also received a scholarship of the Programa de Doutorado Sanduíche no Exterior (PDSE)
from CAPES (process number 88881.189834/2018-01).
The fieldwork and laboratory analysis were financed by
the project number 23038.006832/2014-11- Edital
CAPES 25/2014 – Pró-Forenses.

LITERATURE CITED

Adhikari, K., A. E. Hartemink, B. Minasny, R. Bou
Kheir, M. B. Greve, and M. H. Greve. 2014. Digital
mapping of soil organic carbon contents and stocks
in Denmark. PLOS ONE 9:e105519.

Amundson, R., A. T. Austin, E. A. G. Schuur, K. Yoo,
V. Matzek, C. Kendall, A. Uebersax, D. Brenner,
and W. T. Baisden. 2003. Global patterns of the iso-
topic composition of soil and plant nitrogen. Glo-
bal Biogeochemical Cycles 17:1031.

Bai E., and B. Z. Houlton. 2009. Coupled isotopic and
process-based modeling of gaseous nitrogen losses
from tropical rain forests. Global Biogeochemical
Cycles 23:GB2011

Bai, E., B. Z. Houlton, and Y. P. Wang. 2012. Isotopic
identification of nitrogen hotspots across natural
terrestrial ecosystems terrestrial ecosystems. Bio-
geosciences 9:3287–3304.

Bataille, C. P., I. C. C. von Holstein, J. E. Laffoon, M.
Willmes, X. M. Liu, and G. R. Davies. 2018. A
bioavailable strontium isoscpe for Western Euro-
pe: a machine learning approach. PLOS ONE 13:
-e0197386.

Brenner, D. L., R. Amundson, W. T. Baisden, C. Ken-
dall, and J. Harden. 2001. Soil N and 15N variation
with time in a California annual grassland ecosys-
 tem. Geochimica Et Cosmochimica Acta
65:4171–4186.

Bustamante, M. M. C., L. A. Martinelli, D. A. Silva,
P. B. Camargo, C. A. Klink, T. F. Domingues, and R.
V. Santos. 2004. 14N Natural Abundance in Woody
Plants and Soils of Central Brazilian Savannas
(Cerrado). Ecological Applications 14:200–213.

Bustamante, M. M. C., E. Medina, G. P. Asner, G. B.
Nardoto, and D. C. Garcia-Montiel. 2006. Nitrogen
cycling in tropical and temperate savannas. Bio-
geosciences 79:209–237.

Chen, C., Y. Jia, Y. Chen, I. Mehmood, Y. Fang, and G.
Wang. 2018. Nitrogen isotopic composition of
plants and soil in an arid mountainous terrain:
south slope versus north slope. Biogeosciences
15:369–377.

Cheng, W., Q. Chen, Y. Xu, X. Han, and L. Li. 2009. Cl-
imate and ecosystem 15N natural abundance along
a transect of Inner Mongolian grasslands: contrast-
ing regional patterns and global patterns. Global
Biogeochemical Cycles 23:GB2005.

Chesson, L. A., et al. 2018. Applying the principles of
isotope analysis in plant and animal ecology to
forensic science in the Americas. Oecologia
187:1077–1094.

Craine, J. M., et al. 2015. Convergence of soil nitrogen
isotopes across global climate gradients. Scientific
Reports 5:8280.

Davidson, E. A., F. Y. Ishida, and D. C. Nepstad. 2004.
Effects of an experimental drought on soil emis-
sions of carbon dioxide methane, nitrous oxide,
and nitric oxide in a moist tropical forest. Global
Change Biology 10:718–730.

Denk, T. R. A., J. Mohn, C. Decock, D. Lewicka-Szczer-
biak, E. Harris, K. Butterbach-Bahl, R. Kiese, and B.
Wolf. 2017. The nitrogen cycle: a review of isotope
effects and isotope modeling approaches. Soil Biol-
ogy and Biochemistry 105:121–137.

Ehleringer, J. R., H. A. Mooney, P. W. Rundel, R. D.
Evans, B. Palma, and J. Delatorrell. 1992. Lack of
nitrogen cycling in the Atacama Desert. Nature
359:316–318.

ESRI. 2016. ArcGIS desktop, release 10.3.1. Environ-
mental Systems Research Institute, Redlands, Cali-
fornia, USA.

Freitas, A. D. S., E. V. de Sa Barretto Sampaio, R. S. C.
Menezes, and H. Tiessen. 2010. 15N natural abun-
dance of non-fixing woody species in the Brazilian
dry forest (caatinga). Isotopes in Environmental
and Health Studies 46:210–218.

Friedman, J. H. 2001. Greedy function approximation:
a gradient boosting machine. Annals of Statistics
29:1189–1232.

Friedman, J. H. 2002. Stochastic gradient boosting.
Computational Statistics & Data Analysis 38:
367–378.

Garcia-Perez, B., and K. A. Hobson. 2014. A multi-is-
toipe (δ2H, δ13C, δ15N) approach to establishing
migratory connectivity of Barn Swallow (Hirundo
rustica). Ecosphere 5:1–12.

Gomes, L. C., R. M. Faria, E. de Souza, G. V. Veloso,
C. E. G. R. Schaefer, and E. I. F. Filho. 2019. Modelling
and mapping soil organic carbon stocks in Brazil.
Geoderma 340:337–350.

Gray, J. M., T. F. A. Bishop, and X. Yang. 2015. Prag-
matic models for the prediction and digital mapping
of soil properties in eastern Australia. Soil
Research 53:24–42.

Greenwell, B., B. Boehmke, J. Cunningham, and G. B.
M. Developers. 2019. gbm: generalized Boosted
Regression Models. R Package Version 2.1.5.
https://CRAN.R-project.org/package=gbm

Gregorutti, B., B. Michel, and P. Saint-Pierre. 2016. Cor-
relation and variable importance in random for-
ests. Statistics and Computing 27:659–678.

Hengl, T., et al. 2017. SoilGrids250m: global gridded
soil information based on machine learning. PLOS
ONE 12:e0169748.
Hengl, T., M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. Peef 6:e5518.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965–1978.

Hobson, K. A., S. L. van Wilgenburg, L. I. Wassenaar, R. Powell, C. J. Still, and J. M. Craine. 2012. A multi-isotope ($\delta^{13}C$, $\delta^{15}N$, $\delta^{2}H$) feather isoscape to assign Afrotropical migrant birds to origins. Ecosphere 3:1–20.

Högberg, P. 1997. $^{15}N$ natural abundance in soil-plant systems. New Phytologist 137:179–203.

Houlton, B. Z., and E. Bai. 2009. Imprint of denitrifying bacteria on the global terrestrial biosphere. Proceedings of the National Academy of Sciences of the United States of America 106:21713–21716.

Houlton, B. Z., A. R. Marklein, and E. Bai. 2015. Representation of nitrogen in climate change forecasts. Nature Publishing Group 5:398–401.

Houlton, B. Z., D. M. Sigman, and L. O. Hedin. 2006. Isotopic evidence for large gaseous nitrogen losses from tropical rainforests. Proceedings of the National Academy of Sciences of the United States of America 103:8745–8750.

Hounkpatin, O. K. L., F. O. Hip, A. Y. Bossa, G. Welp, and W. Amelung. 2018. Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso). Catena 166:298–309.

Jarvis, A., H. I. Reuter, A. Nelson, and E. Guevara. 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m. http://srtm.csi.cgiar.org

Kuhn, M. 2008. Building predictive models in R using the caret Package. Journal of Statistical Software 28:1–26.

Kuhn, M., and K. Johnson. 2019. Feature engineering and selection: a practical approach for predictive models. First edition. Taylor & Francis Group, New York, New York, USA.

Kuhn, M., and R. Full. 2018. Cubist: rule- and instance-based regression modeling. R package version 0.2.2. https://CRAN.R-project.org/package=Cubist

Liaw, A., and M. Wiener. 2002. Classification and regression by random forest. R News 2:18–22.

Malone, B. P., N. P. Odgars, U. Stockmann, B. Minasny, and A. B. McBratney. 2018. Digital mapping of soil classes and continuous soil properties. Pages 373–413 in A. B. McBratney, B. Minasny, and U. Stockmann, editors. Pedometrics. Springer, Dordrecht, The Netherlands.

Martinelli, L. A., M. C. Piccolo, A. R. Townsend, P. M. Vitousek, E. Cuevas, W. McDowell, G. P. Robertson, O. C. Santos, and K. Treseder. 1999. Nitrogen stable isotopic composition of leaves and soil: tropical versus temperate forests. Biogeochemistry 46:45–65.

Menezzes, R. S. C., E. V. S. B. Sampaio, V. Giongo, and A. M. Pérez-Marin. 2012. Biogeochemical cycling in terrestrial ecosystems of the Caatinga Biome. Brazilian Journal of Biology 72:643–653.

Nardoto, G. B., et al. 2014. Basin-wide variations in Amazon forest nitrogen-cycling characteristics as inferred from plant and soil $^{15}N$: $^{14}N$ measurements. Plant Ecology & Diversity 7:173–187.

Nardoto, G. B., and M. M. D. C. Bustamante. 2003. Effects of fire on soil nitrogen dynamics and microbial biomass in savannas of Central Brazil. Pesquisa Agropecuaria Brasileira 38:955–962.

Nardoto, G. B., J. P. H. B. Ometto, J. R. Ehleringer, N. Higuchi, M. M. D. C. Bustamante, and L. A. Martine, 2008. Understanding the Influences of Spatial Patterns on N Availability Within the Brazilian Amazon Forest. Ecosystems 11:1234–1246.

Nawar, S., and A. M. Mouazen. 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR Spectroscopy measurements of soil total nitrogen and total carbon. Sensors 17:2428.

Olson, D. M., et al. 2001. Terrestrial ecoregions of the world: a new map of life on Earth. BioScience 51:933–938.

Pérez, T., D. Garcia-Montiel, S. Trumbore, S. Tyler, P. de Camargo, M. Moreira, M. Piccolo, and C. Cerri. 2006. Nitrous oxide nitrification and denitrification $^{15}N$ enrichment factors from Amazon Forest soils. Ecological Applications 16:2153–2167.

Peri, P. L., B. Ladd, D. A. Pepper, S. P. Bonser, S. W. Laffan, and W. Amelung. 2012. Carbon ($^{13}C$) and nitrogen ($^{15}N$) stable isotope composition in plant and soil in Southern Patagonia’s native forests. Global Change Biology 18:311–321.

Robinson, D. 2001. $\delta^{18}N$ as an integrator of the nitrogen. Trends in Ecology & Evolution 16:153–162.

Rohatgi, A. 2018. WebPlotDigitizer. Austin, Texas, USA. https://automeris.io/WebPlotDigitizer

Terzer, S., L. I. Wassenaar, L. J. Araguás-Araguás, and P. K. Aggarwal. 2013. Global isoscapes for $^{18}O$ and $^{2}H$ in precipitation: improved prediction using regionalized climatic regression models. Hydrology and Earth System Sciences 17:4713–4728.

Viscarra Rossel, R. A., D. J. Brus, C. Lobsey, Z. Shi, and G. McLachlan. 2016. Baseline estimates of soil organic carbon by proximal sensing: comparing design-based, model-assisted and model-based inference. Geoderma 265:152–163.

Vitousek, P. M., D. N. L. Menge, S. C. Reed, and C. C. Cleveland. 2013. Biological nitrogen fixation: rates,
patterns and ecological controls in terrestrial ecosystems. Philosophical Transactions of the Royal Society B 368:20130119.
Yang, R. M., G. L. Zhang, F. Liu, Y. Y. Lu, F. Yang, F. Yang, M. Yang, Y. G. Zhao, and D. C. Li. 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators 60:870–878.
Zhao, M., F. A. Heinsch, R. R. Nemani, and S. W. Running. 2005. Improvements of the MODIS terrestrial gross and net primary production global data set. Remote Sensing of Environment 95:164–176.
Zhu, Q., and W. J. Riley. 2015. Improved modelling of soil nitrogen losses. Nature Climate Change 5:705–706.
Zomer, R. J., A. Trabucco, D. A. Bossio, and L. V. Verchot. 2008. Climate change mitigation: a spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agriculture, Ecosystems and Environment 126:67–80.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.3223/full