Climate and soils at the Brazilian semiarid and the forest-Caatinga problem: new insights and implications for conservation

Guilherme de Castro Oliveira1,2, Márcio Rocha Francelino1, Daniel Meira Arruda1, Elpídio Inácio Fernandes-Filho1 and Carlos Ernesto Gonçalves Reynaud Schaefer1

1 Departamento de Solos, Universidade Federal de Viçosa, Viçosa, Brazil
2 Instituto de Geociências, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
3 Author to whom any correspondence should be addressed.

E-mail: guilhermecastrol86@gmail.com

Supplementary material for this article is available online

Abstract
This study aimed to test two hypotheses: (i) on the Brazilian semiarid territory, the climate has greater weight as a driver of vegetation than the soil and; (ii) the arboreal Caatinga is a vegetation whose environmental attributes are similar to the Dry Forest, in terms of soil and climate attributes. We analyzed attributes of the superficial horizon of 156 standardized profiles distributed throughout the Brazilian semiarid region. Bioclimatic variables were obtained from the WorldClim platform and extracted to profiles location. The main vegetation types in the region were considered: Caatinga, arboreal Caatinga, Dry Forest and Cerrado. Variable selection was performed with hierarchical correlation dendrogram and recursive feature elimination algorithm. Linear Discriminant Analysis and Random Forest (RF) algorithm were used for modeling the edaphic and climate niche and predict the vegetation with the selected variables. Climate and soil, individually, were able to separate the vegetation, but the climate was no better predictor than the soil. Therefore, we reject the first hypothesis. However, the better prediction was attained with the combined use of soil and climate attributes. The parsimonious RF model had good performance, with Kappa 0.61 ± 0.10 and 70.9% ± 7.7% accuracy. The combination of soil and climate predictors resulted in better separation of vegetation in the Brazilian semiarid region. Soil attributes are key variables in large-scale biogeographic modeling. The so-called arboreal Caatinga is distributed over a wide edaphic and climatic range, with strong similarity to the Dry Forest distribution, confirmed by the great overlap in the multivariate space, which confirms the second hypothesis. The results point towards an urgent review of the Atlantic Forest Law. The environments where the arboreal Caatinga and the Dry Forest occur are very similar, so that the former may represent a degraded phase of the Atlantic Forest, currently without the due legal protection.

1. Introduction

The geographical distribution of natural vegetation and biodiversity on Earth can be understood at different scales (Willis and Whittaker 2002). On a continental scale, climate explains the vegetation distribution and defines transitions between biomes (Woodward et al 2004, Arruda et al 2017, Langan et al 2017, Dionizio et al 2018, Casalini et al 2019). On a local scale, one can expect for a lower climate explanatory power and a greater importance of soil and topographic factors (Trejo and Dirzo 2002, Arruda et al 2015).

Soil attributes have long been recognized as an important factor in explaining plant distribution. However, the magnitude of its contribution to the improvement of vegetation distribution models compared to climate or topography, is little addressed (but see Santos et al 2012, Dubuis et al 2013, Neves et al 2015). With regional particularities, different soil variables are crucial for determining the plant community.
In addition to the deciduous forest physiognomies, the Caatinga domain also includes humid forests (semi-deciduous and ombrophilous), Cerrado (savannah) and the Caatingastrictosensu (or steppic-savannahsensu IBGE 2004, 2012a, 2012b). Despite peculiarities derived from small-scale analyses, on a large scale, the Caatinga domain and the Dry Forest are included along with the Neotropical Seasonally Dry Forests. (Prado 2000, Neves et al 2015, DRY-FLOR 2016, Pennington et al 2018). Although the official Brazilian classification does not adopt the terms ‘arboreal Caatinga’ and ‘Dry Forest’, the present study opted for this nomenclature for the best compatibility with the international literature.

In this context, the present study assessed environmental aspects of the main vegetation of the semi-arid domain and modeled their distributions based on environmental suitability. Considering the extensive area of the semi-arid region of Brazil (993,604 km²) and the possibility that the arboreal Caatinga is a physiognomic variation of the Dry Forest, we tested two hypotheses: (i) on a large scale, the climate has a greater contribution as a driver of plant formations than the soil and, (ii) the arboreal Caatinga is a vegetation whose environmental attributes are similar to the Dry Forest, in terms of soil and climate attributes.

2. Materials and methods

2.1. Data collection

Soil data were obtained from the selection of 156 standardized soil profiles scattered throughout the climatic domain of the semi-arid region (figure 1). These profiles were obtained from the Brazilian Soil Information System (EMBRAPA 2019). The standardization was made by EMBRAPA to make compatible the analyses made in different periods with different methods. We analyzed 18 topsoil attributes, namely: percentage of fine earth (TF, %), contents of clay, silt and sand (g kg⁻¹), pH in H₂O, K, Ca²⁺ + Mg²⁺, Na⁺, H⁺, Al³⁺, cation exchange capacity (µ, cmolc kg⁻¹), sum of bases (sb, %), base saturation (v, %), total organic carbon (co, g kg⁻¹), Al³⁺ saturation (m, %), Fe₂O₃ content and K (SiO₂/Al₂O₃) and Kr (SiO₂/Al₂O₃ + Fe₂O₃) ratios. The A horizon was considered as a whole regardless of its depth. Therefore no interpolation method was applied.

The criterion for selecting soil profiles from the large database was the clear identification of the local vegetation. Profiles whose associated vegetation was reportedly dubious in their description, or related to any transitional vegetation type, were not considered. Four types of vegetation were analyzed: (i) Dry Forest (deciduous forest with two to three strata and partly continuous to continuous canopy), (ii) arboreal Caatinga (hypoxerophilous low forest with two strata and discontinuous canopy), (iii) Caatinga (savannah-like hypoxerophilic physiognomy, with partially developed

(Ferreira-Júnior et al 2007, Neri et al 2012, Saporetti-Junior et al 2012, Dubuis et al 2013, Catorci et al 2014, Arruda et al 2015a, Nunes et al 2015, Silva et al 2015, Almeida et al 2018), such as sum of bases, acidity, organic matter, aluminum saturation, texture and others. The regional particularities of soil-vegetation relationship mean that their predictive potential is also restricted to each region, as opposed to climate attributes, that can serve to outline large-scale patterns (e.g. Holdridge 1967, Woodward et al 2004), as annual temperature range, annual precipitation, precipitation seasonality and others. Hence, understanding the regional pattern of the soil-vegetation relationship is fundamental to allow the adjustment of more complete vegetation prediction models. These models can provide a great contribution to understanding the geographical distribution of complex vegetation, where intricate interactions between soil, climate, topography and human action occur.

The Caatinga is a complex domain present in the Brazilian semi-arid climate (Velloso et al 2002, IBGE 2004, Santos et al 2011). This complexity is the result of its wide climatic, soil and topographic heterogeneity, as well as the cumulative impacts of human activity since pre-Columbian times (Andrade-Lima 1981, Velloso et al 2002). More precisely, the earliest human occupations in the northeastern region of Brazil date from the late Pleistocene, more than 20,000 y B.P. (Lahaye et al 2015, 2019). Among the various faces of Caatinga vegetation (see Andrade-Lima 1981), the herbaceous stratum is the one that concentrates the greatest diversity of species (Costa et al 2007, Linares-Palomino et al 2015). However, most of the studies on the environment-vegetation relationship focus in the tree stratum. Therefore, they provide great terminological confusion regarding the forest physiognomies of Caatinga. These are described as tall Caatinga forest (Andrade-Lima 1981), hypoxerophilic Caatinga (Brandão 1994, Arruda et al 2015b, Araújo Filho et al 2017), crystalline Caatinga (Queiroz 2006), arboreal Caatinga (Santos et al 2012, Moro et al 2016), deciduous stiff-leaved dwarf forest (Oliveira-Filho 2009) and forested steppic-savannah (IBGE 2004, 2012a, 2012b).

However, studies have shown that the arboreal Caatinga resembles floristically and environmentally the deciduous seasonal forest, or Dry Forest (Andrade-Lima 1981, Santos et al 2012, Arruda et al 2015b), the most hitherland and seasonal of the Atlantic Forest domain (Oliveira-Filho et al 2006). This similarity raises the question of whether arboreal Caatinga is a successional stage of Dry Forest (Arruda et al 2015b). If this hypothesis is confirmed, there are obvious implications for biodiversity protection. This issue is of paramount importance in the enforcement of the Atlantic Forest Law (Brasil 2006), which grants protection to this biome, but does not consider zones of ecological tension where the arboreal Caatinga and the Dry Forest are found, closely associated (IBGE 2012b).
non-grass herbaceous stratum), and (iv) Cerrado (savannic physiognomy with well-developed grassy herbaceous stratum). To avoid changes resulting from land use, we included only profiles described under natural vegetation.

From each soil profile coordinates, 19 bioclimatic attributes were extracted from WorldClim version 2 (Fick and Hijmans 2017): mean annual temperature, mean diurnal range, isothermality, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter and precipitation of coldest quarter. The data represent mean values recorded in the 1970–2000 period by meteorological stations and interpolated globally. The high density of stations present in the area of this study ensures good data reliability (see figures 2 and 4 in Fick and Hijmans 2017). The extraction of climate data was performed with ArcMap software, version 10.3 (ESRI 2016). The soil and climate data used are at different scales. While soil profiles were sampled in the field, climate data were interpolated to a 1 km resolution.

2.2. Data analysis

All analyses were processed using R version 3.5.1 (R Core Team 2016). Initially, a hierarchical correlation dendrogram was used to find groups of highly correlated variables (|r| > 0.95) with the varclus function from ‘Hmisc’ package (Harrell Jr 2019). For each of these clusters, the variable with the lowest correlation (i.e. less redundancy) with all variables present in the data set was selected.

The data were divided into three sets: (a) soil variables, (b) climate variables, and (c) soil and climate together. Next, the recursive feature elimination—RFE algorithm was used with the rfe function from ‘caret’ package (Kuhn 2008) to eliminate collinear variables and variables with little explanatory power (Genuer et al 2010, Ramasubramanian and Singh 2017). This procedure was performed individually for the three datasets. To create a parsimonious model, we selected only the best variables considering a 10% tolerance threshold. In short, tolerance expresses how much is lost in performance to build a parsimonious model, with a smaller number of variables.
First, we used the Linear Discriminant Analysis (LDA) to maximize the separation of vegetation in the multivariate space (Williams 1983). The discriminant function was also used to predict the vegetation classes. The LDA was performed for the three data sets. The results allowed comparing the contribution of the two groups of predictors individually (only soil and only climate) and jointly (soil and climate combined), and evaluating how the vegetation is separated. The second approach was accomplished with the Random Forests (RF) algorithm (Breiman 2001), which was also used to predict the vegetation classes with the combined soil and climate dataset.

The LDA was used to test the contribution of soil and climate predictors and better representation of data in multivariate space. The LDA seeks to determine the extent to which a set of independent variables can explain the groups (Borcard et al. 2018), maximizing the separability among pre-defined classes. Multivariate homogeneity of group dispersions (Anderson 2006) was verified with the betadisper function and tested with the permustest function, both from ‘vegan’ package. Variances within not homogeneous groups were standardized with the decostand function from ‘vegan’ package. The Wilk’s lambda test (Todorov and Filzmoser 2010) was applied to verify if the explanatory variables had different means. For this, we used the Wilks.test function from ‘rcov’ package. Finally, the LDA was performed with the ldev function from ‘MASS’ package. With the standardized data for each group of predictors, the discriminant functions for classifying the vegetation were computed.

LDA cross-validation was performed with Jackknife (leave-one-out) method and the proportion of correct classifications by class was compared for each set of predictors.

The selected soil and climate variables were also used to predict the vegetation with the RF classifier (Breiman 2001). The data were split into training (80%) and validation (20%) sets. We repeat the procedures of sample selection, model training, prediction and validation 100 times to get the model parameters from the confusion matrix. The model performance were evaluated by the mean Kappa—K, accuracy, sensitivity and specificity parameters. The K statistic is a measure of agreement between the predictions of a model and the observed value, in comparison to what one could expect mathematically by chance. Accuracy measures the overall success rate of the classifier. Sensitivity represents the percentage of correctly classified presences, while specificity indicates the percentage of correctly classified absences (Cutler et al. 2007). These parameters should be used together to better interpret the model performance (Ramasubramanian and Singh 2017).

3. Results
From the hierarchical correlogram analysis, six variables from the first 37 were eliminated, leaving 16 soil and 15 bioclimatic attributes (figure S1, appendix S1 in supporting information is available online at stacks.iop.org/ERL/14/104007/mmedia). Following the 10% tolerance criterion over RFE results, we selected six variables from climate predictors, nine from soil and eight from the joint dataset (figure S2). The selected variables are presented in table 1.

### 3.1. Linear discriminant analysis
The test for homogeneity of variances within the groups showed that the groups of predictors were not homogeneous. Subsequently, the data were standardized, allowing homogeneity to be confirmed by the permutational ANOVA. The Wilk’s lambda test showed that the explanatory variables have different means among the vegetations for the three data sets. Thus, all assumptions for the LDA were met (Williams 1983, Borcard et al. 2018).

For the three sets of predictors, the first two functions of LDA have contributed to explaining more than 90% of the variance between the classes. The LDA coefficients are presented in tables S1–S3. As for the climatic niche, Cerrado, Caatinga and Dry Forest were organized into well-defined groups (figure 2(a)). Caatinga and Cerrado are distinguished by annual precipitation and by precipitation in the wettest month (LDA axis 1). The Dry Forest is distinguished from these two mainly by the seasonality of precipitation, annual precipitation and temperature seasonality (LDA axis 2). Climatically, the arboreal Caatinga is
very similar to Dry Forest, being closer to it than to any other vegetation.

Concerning soil attributes (figure 2(b)), the first axis of LDA separated the vegetations in terms of saturation of exchangeable bases and aluminum saturation (table S2). Thus, Caatinga and Dry Forest are positioned in the negative dimension of axis 1 (eutrophic and low Al$^{3+}$ saturation) and the Cerrado to the positive dimension (dystrophic and high Al$^{3+}$ saturation). The second LDA axis separates vegetation mainly in terms of organic carbon content and soil pH. Thus, there is an increase of soil pH and a reduction of organic carbon content in the Dry Forest–Caatinga direction. Arboreal Caatinga holds an intermediate position in this edaphic gradient between the two vegetation. This subtle gradient is also controlled by soil texture, more clayey in the Dry Forest and more sandy in the Caatinga. However, the sand content had a relatively small contribution to the model (see table S2), as it does not serve to differentiate the arboreal Caatinga from other vegetation.

The results of LDA cross-validation (table 2) show that classification success with climate predictors was slightly greater than with soil predictors (55% and 53%, respectively), although the values are very close. However, the best result was achieved with the combination of the two predictors, showing a synergistic interaction between them. While the soil was better predictor for Cerrado and Dry Forest, the climate was better for Caatinga and arboreal Caatinga. Thus, the different groups of predictors are complementary (Arruda et al. 2017), allowing a better separation of vegetation without increasing the number of variables.

Combining soil and climate predictors increased the distance between the centroids of the vegetation classes while maintaining dispersion within the groups (figure 2(c)). It means that the combination of these predictors forms better-defined samples, allowing better separation of classes. With this deeper view of the ecological niche, the similarity between arboreal Caatinga and Dry Forest is evident.

### 3.2. Random forests

Once the best group of predictors was determined (figure 3), they were used for classification with the RF algorithm. The parsimonious RF model for vegetation class prediction with combined soil and climate predictors obtained an average $K = 0.61 \pm 0.10$. The overall accuracy was $70.9\% \pm 7.7\%$ (table S4 and figure S3). The arboreal Caatinga had the lowest balanced accuracy per class and lower sensitivity among the classes, revealing the model’s confusion between this class and the Dry Forest (table 3). The low sensitivity means that the model was not able to place the arboreal Caatinga correctly since there are no environmental attributes that are characteristic of this vegetation (figure 3).

For Caatinga, Cerrado, and Dry Forest classes, the model obtained an accuracy of 87.6%, 87.5% and 81.4%, respectively (table S4). As for specificity, the parsimonious model had high value (>0.88) for all classes. The RF model corroborates the results of the LDA, confirming the similarity between arboreal Caatinga and Dry Forest. However, the RF classification was more successful, achieving more precise separation of the other classes than LDA.

### 4. Discussion

The Brazilian semiarid is the most diverse and complex geographic space in the country (IBGE 2004, Santos et al. 2011, Schaefer 2013). Considering the spatial scale covering roughly 15° of latitude, the large climate contribution cannot be underestimated to
explain the vegetation classes. However, the strong distinction between soils present in the semiarid nucleus and those of the peripheral zones are conditions that favor the use of soil attributes as a factor for enhancing the vegetation differentiation. Thus, although the soil has a slightly lower contribution, we consider that the difference presented (2%) is not ecologically relevant. With that, we reject the first hypothesis and assume that the use of the two groups of predictors best represents the environmental variability among classes of vegetation.

Soil attributes increased the multidimensional distance between groups while maintaining dispersion within them. Thus, allowed a better separation of vegetation classes (Figure 2 and Table 2). Therefore, the importance of soil is not limited to the landscape scale.
unlike what Willis and Whittaker (2002) suggest, but is also an environmental factor that should not be neglected at regional scales (Arruda et al 2017, Langan et al 2017, Dionizio et al 2018, Casalini et al 2019). In the context of this study, $\text{Al}^{3+}$ saturation, base saturation and $K$ were the most important soil attributes. Apparently, the high $\text{Al}^{3+}$ saturation associated with low soil fertility of Cerrado represents a barrier for Caatinga and Dry Forest, since these vegetation were not present on soils with this characteristic (figure 3). Potassium content is related to a more efficient nutrient cycling in Dry Forest (Jaramillo and Sanford 1995). In the Caatinga, it is probably due to the high natural fertility of less weathered soils, where the presence of rich primary minerals and 2:1 clay minerals provides the replacement of this nutrient (Barré et al 2008).

The Cerrado stands out from other vegetation classes for its affinity to soils with high aluminum saturation and higher annual rainfall, associated with a moderate seasonality of precipitation (figure 3), i.e. better-distributed rainfall. The Caatinga is matched by chemically rich soils with low aluminum saturation and warm semiarid climate. However, total precipitation does not differentiate it from other vegetation. The high precipitation seasonality, the lower precipitation in the driest period and the high temperatures in the wettest quarter were the best climatic predictors for this vegetation. Although the Dry Forest also presents affinity with eutrophic soils, the climatic aspects distinguish it from the Caatinga.

The similarity between the samples of arboreal Caatinga and Dry Forest, shown by both LDA and RF, proves the second hypothesis. Thus, it is likely that the difference between these vegetations is mainly structural (Arruda et al 2015b). The environmental similarity between the two vegetation can be explained in at least two ways. One possibility is that the arboreal Caatinga is a variation of the Dry Forest that has suffered chronic anthropogenic disturbance and is now in a stage of stagnant regeneration (Arruda et al 2015b), being incorrectly associated with the Caatinga biome only by its misleading structure. This does not mean that Dry Forest has not suffered such pressure. Nevertheless, minor environmental differences (figure 3), such as slightly higher soil fertility, clay content and humidity in the Dry Forest give greater resilience to this vegetation, allowing its regeneration more quickly than in the arboreal Caatinga. The anthropic disturbance has more impact as the aridity increases (Rito et al 2017).

Another possibility is that the arboreal Caatinga is reminiscent of seasonal forests that covered a larger area in the Brazilian semiarid region during the Pleistocene, and wetter phases in the Holocene (Behling et al 2000, Werneck et al 2011, Arruda et al 2018, Bouimetarhan et al 2018, Costa et al 2018, Medeiros et al 2018). Nowadays, they are the result of the combination of vestigial Dry Forest with species of the Caatinga, which have advanced with the dryness increase in the region (Silva and Souza 2018), which justifies its greater distribution in the border region of the Caatinga biome, in contact with Dry Forests.

The two possibilities above mentioned are not mutually exclusive. It is plausible that colonization by xerophytic shrub species has been favored by human activity (Sagar et al 2003). Pereira et al (2005) found that arboreal individuals in the highest height class ($>3$ m) were only present in the least disturbed environment. In the environments that suffered the greatest disturbance, they observed a reduction in vegetation size, lower floristic diversity and higher density of disturbed area indicator species. Grazing on natural vegetation, a common activity in the Caatinga, also reduces the plant diversity. Areas subjected to more intense grazing become less diverse and more homogeneous among themselves (Schulz et al 2019).

Regarding the floristic pattern of the tree component, different studies have shown that arboreal Caatinga is very similar to Dry Forest (Santos et al 2012, Moro et al 2016). Among nine biogeographic sub-regions of the Caatinga biome identified by Silva and Souza (2018), those located to the south, where they border the Dry Forest, were grouped into a single cluster. The authors recognize that the expansion and contraction of the surrounding Atlantic forest may influence the floristic of this group. Moro et al (2016) also observed that the arboreal Caatinga of northern Minas Gerais and southern Bahia are distinct from the other Caatinga vegetation, suggesting that this group should be appropriately treated as a peripheral subgroup within the crystalline Caatinga.

In addition to the historical and natural processes, human interventions should also be taken into consideration as a driver for differentiation of communities and vegetation physiognomies. Ever since the pre-colonial period, human dispersed plant species of interest in the tropics, enhancing the dominance of some species (Ter Steege et al 2013). In the same period, farmers using slash-and-burn methods for clearing vegetation in seasonal climates began landscape conversion. Thus, the use and alteration of the landscape by man, mainly after 1850, led to an intense fragmentation of dry forests in Brazil (Costa 2005, Espírito-Santo et al 2009), strongly affecting the species turnover between the remaining paths of different landscapes. As a result, a mosaic of vegetation of different successional stages remains in the landscape. Thus, we assume that soil attributes can be good proxies for estimating the original cover of these anthropized environments since the edaphic variability associated with different stages of the same vegetation class should not be sufficiently broad.

In this context, in the Middle São Francisco region (north of Minas Gerais and west of Bahia), arboreal Caatinga and Dry Forest also have great similarities.
with subtle difference only in soil texture (Arruda et al 2015b). The former is associated with soils slightly more sandy than the latter. This suggests that the difference between arboreal Caatinga and Dry Forest is locally controlled by a pedoclimatic gradient, as shown by Santos et al (2012). However, the same authors also showed that this gradient does not differentiate these vegetation in terms of floristic.

For the set of environmental predictors used in this study, the model also shows that the arboreal Caatinga is more similar to Dry Forest than to the Caatinga stricto sensu. Hence, there is strong evidence that the arboreal Caatinga is not a climax formation of the Caatinga domain, but an indistinguishable continuum of the Atlantic Dry Forest, which points to an urgent review of the protection offered by the Atlantic Forest Law (IBGE 2012b). This discussion needs to be fostered with more data for a conclusion and, in this regard, the work on ecological niche modeling, floristic and phytosociology can make a great contribution. We highlight the importance of works that fill the sampling gaps present for the Caatinga in southern Bahia and northern Minas Gerais, as emphasized by Moro et al (2016).

Studies aiming to model the niche of vegetation in this region must consider that errors in the vegetation classification may generate noise in the models. Thus, disfiguring the multidimensional space that defines these niches and leading to inaccurate results. We built a parsimonious model with only eight variables (three soil and five climate variables), from which it was possible to predict the vegetation with 71% average accuracy using the RF algorithm. The methodological framework applied in this study allowed the creation of a comprehensible model to explain the occurrence of the main vegetation found in the Brazilian semiarid region.

5. Conclusion

The hypothesis that the climate is the main driver to the vegetation class pattern found in the Brazilian semiarid was rejected. Although the contribution of the soil was slightly smaller than the climate, the combination of both provided better class separation, showing that ecological niche modeling, even at a large scale, should not neglect the interaction between the two groups of predictors.

The striking environmental similarity of arboreal Caatinga to Dry Forest allows us to assume that this may not be a distinct vegetation unit from deciduous forest formations. This point towards an urgent review of the Atlantic Forest Law, which currently does not protect the arboreal Caatinga, and faces increasing pressures from deforestation. The hypothesis that this vegetation represents the effect of long-term human activity should be addressed in future work.

Acknowledgments

We acknowledge the financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)-Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) agencies during the development of this Project. We also thank the anonymous reviewers for their important contributions to the first version of the manuscript.

Data availability statement

The data that support the findings of this study are openly available at https://doi.org/10.25412/iop.9447518.v1 and https://doi.org/10.25412/iop.9447515.v1. The compiled data used in this work are available in the supplementary material.

ORCID iDs

Guilherme de Castro Oliveira @ https://orcid.org/0000-0002-9926-0587
Márcio Rocha Francelino @ https://orcid.org/0000-0001-8837-1372
Daniel Meira Arruda @ https://orcid.org/0000-0002-3541-4786
Elpídio Inácio Fernandes-Filho @ https://orcid.org/0000-0003-2440-8329
Carlos Ernesto Gonçalves Reynaud Schaefer @ https://orcid.org/0000-0001-7060-1598

References

Almeida H A, Bahia T, de O, Gélvez-Zúñiga I and Fernandes G W 2018 Together yet separate: variation in soil chemistry determines differences in the arboreal-shrub structure of two contiguous rupesrestrial environments Acta Bot. Brasileira 32 578–87
Anderson M J 2006 Distance-based tests for homogeneity of multivariate dispersions Biometrics 62 245–53
Andrade-Lima D 1981 The caatingas dominium Rev. Bras. Botânica 4 149–63
Araújo Filho J C, Ribeiro M R, Burgos N and Marques F 2017 Solos da Caatinga Pedologia—Solos Dos Biomas Brasileiros ed N Curi et al (Viçosa: Sociedade Brasileira de Ciência do Solo) pp 227–60
Arruda D M, Fernandes-Filho E I, Solar R R C and Schaefer C E G R 2017 Combining climatic and soil properties better predicts covers of Brazilian biomes Sci. Nat. 104
Arruda D M, Schaefer C E G R, Corrêa G R, Rodrigues P M S, Duque Brasil R, Ferreira-JR W G and Oliveira-Filho A T 2013a Landforms and soil attributes determine the vegetation structure in the Brazilian semiarid Folia Geobot. 50 175–84
Arruda D M, Schaefer C E G R, Fonseca R S, Solar R R C and Fernandes-Filho E I 2018 Vegetation cover of Brazil in the last 21 ka: new insights into the Amazonian refugia and Pleistocene arc hypotheses Glob. Ecol. Biogeogr. 27 47–56
Arruda D M, Schaefer C E G R and Moraes M I. B 2013b Relações entre atributos do solo e vegetações da região ecotonal do Médio Rio São Francisco, Brasil Rev. Bras. Cienc. do Solo 39 1524–32
Barré P, Velde B, Fontaine C, Catel N and Abbadie L 2008 Which 2:1 clay minerals are involved in the soil potassium reservoir? Insights from potassium addition or removal experiments on...
three temperate grassland soil clay assemblages Geoderma 146 216–23
Behling H, Arz H W, Patzold J and Wefer G 2000 Late Quaternary
tropical forest vegetation in the Amazon Basin, Brazil, as inferred from marine core GeoB 3104–1 Quat. Sci. Rev. 19 981–94
Borcard D, Gillet F and Legendre P 2018 Numerical Ecology with R
(Cham: Springer)
Bouimetarhan I, Missi C M, González-Arango C, Dupont L, Vioigt I, Prange M and Zonnenveld K 2018 Intermittent development of forest corridors in northeastern Brazil during the last deglaciation: climatic and ecologic evidence Quat. Sci. Rev. 192 86–96
Brandão M 1994 Área Mineira do Polígono das Secas - cobertura
vegetal Inf. Agropecuário 17 5–9
Brasil 2006 Lei N.º 11.428, de 22 de Dezembro de 2006 Diário Of. da
União 246 1
Breiman L 2001 Random forests Mach. Learn. 45 5–32
Casalini A I, Bouza P J and Bisigato A J 2019 Geomorphology, soil
and vegetation patterns in an arid ecotone Cactea 174 353–61
Catarci A, Tardella F M, Velasquez J L, Cesaretti S, Malatesta L and
Noronha S E de and Sampaio C A 2009 Sustainability of
a Brazilian Atlantic tropical semideciduous forest For. Ecol. Manage. 258 237–47
Cutler D R, Edwards T C, Beard K H, Cutler A, Hess K T, Gibson J and Lwaller J 2007 Random forests for classification in ecology Ecology 88 2783–92
Dionizio E A, Costa M H, Castanho A D A, Pires G F, Marimon B S, Marimon-Junior B H, Lenza E, Pimenta F M, Yang X and Jain A K 2018 Influence of climate variability, fire and phosphorus limitation on vegetation structure and dynamics of the Amazon-Cerrado border Biogeosciences 15 919–36
DRYFLOR 2016 Plant diversity patterns in neotropical dry forests and their conservation implications Science 353 1383–7
Dubuis A, Giovanetti S, Pellisser L, Potier J, Vittor P and Gusmao A 2013 Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables J. Veg. Sci. 24 593–606
EMBRAPA 2019 Sistema de Informação de Solos Brasileiros (https://bsdos.cnptia.embrapa.br/consultas_publica.html)
Espírito-Santo M M, Sevilha A C, Anaya F C, Barbosa R, Espírito-Santo M M, Sevilha A C, Anaya F C, Barbosa R, Ferreira-Júnior W G, Silva A F, Schaefer C E G R, Meira Neto J A A, Costa R C, Araújo F S de and Lima-Verde L W 2007 Flora and life-Climatol. 8
Environ. Res. Lett. 14 (2019) 104007
Fick S and Hijmans R J 2017 WorldClim 2: new 1 km spatial
resolution climate surfaces for global land areas Int. J. Climatol. 37 4302–15
Genuer R, Poggi J M and Tuleau-Malot C 2010 Variable selection using random forests Pattern Recognit. Lett. 31 2225–36
Harrell F E Jr 2019 CRAN—Package Hmisc Hmisc Harrell Misc (http://cran.r-project.org/package=Hmisc)
Holdridge L R 1967 Life Zone Ecology (Costa Rica: San Jose)
IBGE 2004 Mapa de Vegetação do Brasil Inst. Bras. Geogr. e Estatística (http://geoftp.ibge.gov.br/informacoes_ambientais/vegetacao/ mapas/brazil/vegetacao.pdf)
IBGE 2012a Manual Técnico da Vegetação Brasileira (Rio de Janeiro: IBGE) (https://biblioteca.ibge.gov.br/visualizacao/livros/ liv63011.pdf)
IBGE 2012b Mapa da Área de Aplicação da Lei n° 11.428 de 2006 See http://www2.mma.gov.br/images/arquivos/biosfera/mata_atlantica/mapa_mata_atlantica_lei_11428_2006_e_decreto6660_2008.pdf)
Jaramillo V and Sanford R L J 1995 Nutrient cycling in tropical deciduous forests Seasonally Dry Tropical Forests ed S H Bullock et al (Cambridge: Cambridge University Press) pp 346–61
Kuhn M 2008 Building predictive models in R using the caret package J. Stat. Softw. 28 1–26
Lahaye C et al 2015 New insights into a late-Quaternary human occupation in America: the Vale da Pedra Furada complete chronological study Quat. Geochronol. 30 445–51
Lahaye C et al 2019 Another site, same old song: the pleistocene-holocene archaeological sequence of Toca da Janela da Barra do Antonião-North, Piauí, Brazil Quat. Geochronol. 49 223–39
Langan L, Higgins S I and Scheiter S 2017 Climate-biomes, pedo-
biomes or pyro-biomes: which world view explains the tropical forest–savanna boundary in South America? J. Biogeogr. 44 2319–30
Linares-Palomino R, Aybar D and Morales Ruiz E J 2015 Floristics of neotropical seasonal dry forests: a perspective from complete vascular plant inventories Phytocoenologia 45 251–67
Medeiros V B, Oliveira P E, Santos R A, Barreto A M F, Oliveira M A T and Pinaya J I 2018 New holocene pollen records from the Brazilian Caatinga An. Acad. Bras. Cienc. 90 2011–23
Moro M F, Nic Lughadha E, de Araújo F S and Martins F R 2016 A phytogeographical metaanalysis of the semiarid caatinga domain in Brazil Bot. Rev. 82 91–148
Neri A V, Schaefer C E G R, Silva A F, Souza A L, Ferreira-Junior W G and Meira-Neto J A A 2012 The influence of soils on the floristic composition and community structure of an area of Brazilian Cerrado vegetation Edinburgh J. Bot. 69 1–27
Neves D M, Dexter K G, Pennington R T, Bueno M L and Oliveira Filho A T 2015 Environmental and historical controls of floristic composition across the South American Dry Diagonal J. Biogeogr. 42 1566–76
Nunes J A, Schaefer C E G R, Ferreira Júnior W G, Neri A V, Correa G R and Enright N J 2015 Soil-vegetation relationships on a banded ironstone ‘island’, Carajás Plateau, Brazilian Eastern Amazonia An. Acad. Bras. Cienc. 87 2097–110
Oliveira-Filho A T 2009 Classificação das fitofisionomias da América do Sul cisanterior tropical e subtropical: proposta de um novo sistema—prático e flexivel - ou uma injeção a mais de caos? Rodriguesia 60 237–58
Oliveira-Filho A T, Jarenkow J A and Rodal M J N 2006 Floristic relationships of seasonally dry forests of eastern South America based on tree species distribution patterns Neotropical savannas and dry forests: Plant diversity, biogeography and conservation ed R T Pennington et al (Boca Raton: Taylor and Francis) pp 159–92
Pennington R T, Lehmann C E R and Rowland L M 2018 Tropical savannas and dry forests Carr. Biol. 28 541–5
Pereira I M, Andrade L A, de, Costa J R M and Dias J M 2005 Regeneração natural em um remanescente de Caatinga sob diferentes níveis de perturbação, no agreste paraibano Acta Bot. Brasilia 15 413–26
Prado D E 2000 Seasonally Dry Forests of tropical South America: from forgotten ecosystems to a new phytogeographic unit Edinburgh J. Bot. 57 437–61
Queiroz L F 2006 The Brazilian caatinga: phytogeographical patterns inferred from distribution data of the Leguminosae Neotropical Savannas and Seasonally Dry Forests: Plant
Diversity, Biogeography, and Conservation ed
R T Pennington and J A Ratter (Boca Raton, FL: CRC Press) p 38
Ramasubramanian K and Singh A 2017 Machine Learning Using R—
A Comprehensive Guide to Machine Learning (Berkley: Apress)
(https://doi.org/10.1007/978-1-4842-2334-5)
R Core Team 2016 R: A language and environment for statistical
computing R Found. Stat. Comput. (http://r-project.org/)
Rito K F, Arroyo-Rodríguez V, Queiroz R T, Leal I R and Tabarelli M
2017 Precipitation mediates the effect of human disturbance
on the Brazilian Caatinga vegetation J. Ecol. 105 828–38
Sagar R, Raghubanshi A S and Singh J S 2003 Tree species
composition, dispersion and diversity along a disturbance
gradient in a dry tropical forest region of India For. Ecol.
Manage. 186 61–71
Santos H G, Carvalho Jr W, Dart R O, Aglio M L D, Sousa J S,
Pares J G, Fontana A, Martina A L S and Oliveira A F 2011 O
novo mapa de solos do Brasil: legenda atualizada (Rio de
Janeiro: Embrapa Solos) (https://ainfo.cnptia.embrapa.br/
digital/bitstream/item/123772/1/DOC-130-O-novo-
mapa-de-solos-do-Brasil.pdf)
Santos R M, Oliveira-Filho A T, Eisenlohr P V, Queiroz L P,
Cardoso D B O S and Rodal M J N 2012 Identity and
relationships of the Arboreal Caatinga among other floristic
units of seasonally dry tropical forests (SDTFs) of north-
eastern and Central Brazil Ecol. Evol. 2 409–28
Saporetti-Junior A W, Schaefer C E G R, Souza A L, Soares M P,
Araújo D S D and Metra-Neto I A A 2012 Influence of soil
physical properties on plants of the mussununga ecosystem,
Brazil Folia Geobot. 47 29–39
Schaefer C E G R 2013 Bases físicas da paisagem brasileira: estrutura
geológica, relevo e solos Tópicos em Ciência do Solo ed
A Araújo and B Alves (Viçosa: Sociedade Brasileira de Ciência
do Solo) pp 1–69
Schulz K, Guschal M, Kowarik I, Almeida-Cortez J S,
Sampaio E V S B and Cierjacks A 2019 Grazing reduces plant
species diversity of Caatinga dry forests in northeastern Brazil
Appl. Veg. Sci. 22 348–59
Silva A C and Souza A F 2018 Aridity drives plant biogeographical
sub-regions in the Caatinga, the largest tropical dry forest and
woodland block in South America PLoS One 13 e0196130
Silva R B M, Francelino M R, Moura PA, Moura T A,
Pereira M G and Oliveira C P de 2015 Relação solo/vegetação
em ambiente de cerrado sobre influência do grupo urucuia
Cienc. Florest. 25 363–73
Ter Steege H et al 2013 Hyperdominance in the Amazonian tree
flora Science 342 1243992
Todorov V and Filzmoser P 2010 Robust statistic for the one-way
MANOVA Comput. Stat. Data Anal. 54 37–48
Trejo I and Dirzo R 2002 Floristic diversity of Mexican seasonally
dry tropical forests Biodivers. Conserv. 11 2063–84
Velloso A L et al 2002 Ecorregiões Propostas Para o Bioma Caatinga ed
A L Velloso et al (Recife: The Nature Conservancy do Brasil;
Associação Plantas do Nordeste)
Werneck F P, Costa G C, Colli G R, Prado D E and Sites J W 2011
Revisiting the historical distribution of Seasonally Dry Tropical
Forests: new insights based on palaeodistribution modelling
and palynological evidence Glob. Ecol. Biogeogr. 20 272–88
Williams B K 1983 Some observations on the use of discriminant
analysis in ecology Ecology 64 1283–91
Willis K J and Whittaker R J 2002 Species diversity–scale matters
Science 295 1245–8
Woodward F I, Lomas M R and Kelly C K 2004 Global climate and the
distribution of plant biomes Phil. Trans. R. Soc. B 359 1465–76