Aplicações da Linguagem R em Análises de Vegetação

Écio Souza Diniz
Pedro Manuel Villa
(Organizadores)

Atena Editora
Ano 2020

Aplicações da Linguagem R em Análises de Vegetação

Écio Souza Diniz
Pedro Manuel Villa
(Organizadores)

Atena Editora
Ano 2020

Editora Chefe: Profª Drª Antonella Carvalho de Oliveira
Diagramação: Geraldo Alves
Edição de Arte: Lorena Prestes
Revisão: Os Autores

Dados Internacionais de Catalogação na Publicação (CIP)
(eDOC BRASIL, Belo Horizonte/MG)

A642 Aplicações da linguagem R em análises de vegetação [recurso eletrônico] / Organizadores Écio Souza Diniz, Pedro Manuel Villa. – Ponta Grossa, PR: Atena, 2020.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-86002-35-5
DOI 10.22533/at.ed.355200903

1. Desenvolvimento sustentável. 2. R (Linguagem de programação de computador). 3. Recursos vegetais – Manejo. I.Diniz, Écio Souza. II. Villa, Pedro Manuel.

CDD 333.7511

Elaborado por Maurício Amormino Júnior – CRB6/2422

Atena Editora
Ponta Grossa – Paraná - Brasil
www.atenaeditora.com.br
contato@atenaeditora.com.br

Atena Editora
Ano 2020

APRESENTAÇÃO

Os diferentes tipos de vegetação ao redor do globo, principalmente as florestas tropicais, se destacam por fornecer importantes bens e serviços ecossistêmicos para a humanidade como, por exemplo, regulação climática, provisão de alimentos e diversas fontes de energia. Contudo, as crescentes e rápidas mudanças no meio ambiente causadas por sua intensa exploração no século 21 têm promovido reduções drásticas de importantes vegetações distribuídas em distintos Biomas. O Brasil como um país de dimensão continental e rico em recursos vem atravessando profundas transformações em seus Biomas, o que é destacadamente devido aos usos intensos da terra sem técnicas adequadas de manejo para a sua exploração.

Diante desse panorama de significativas transformações do meio natural, se faz necessário e urgente o estudo de diferentes tipos de comunidades vegetais e ecossistêmicas para definir estratégias de manejo e conservação, assim como pesquisas que visem a otimização de produções agrícolas de forma sustentável. A união de compreensão ecológica precisa e adequadas técnicas de manejo permitem uma exploração sustentável a longo-prazo dos recursos vegetais, assegurando manutenção de diversidade e provisões para o futuro.

A execução de estudos robustos para alcançar essa interface entre conservação e exploração demanda o uso de eficientes ferramentas analíticas. Dentre essas ferramentas, as linguagens de programação têm se sido importantes aliadas para obtenções de predições e resultados estatísticos confiáveis e informativos. A linguagem contida no software R é a mais amplamente utilizada para processamento de dados e análises de vegetação. O R engloba diversos pacotes importantes para análises de dados de plantas em diversos contextos ecológicos e agrários. Com seus diversos pacotes, o R permite a busca mais apurada pela compreensão de padrões e processos ecológicos, avaliação de impactos antrópicos sobre vegetação, monitoramentos e previsões de condições do solo para plantios e predições de efeitos de mudanças climáticas em florestas. Essa gama de possibilidades analíticas amplifica o acerto em tomadas de decisão com relação ao uso dos nossos recursos naturais de forma geral.

Este livro tem como objetivo trazer uma compilação de algumas potencialidades do software R para análise de vegetação, contribuindo para o aumento da capacidade técnica de diversos profissionais das áreas de Ciências da Terra ou Naturais no uso dessa poderosa ferramenta analítica. Para tal, os capítulos aqui presentes discorrem de forma aplicada sob temas em contextos ecológicos e agrários. Todos os capítulos possuem links de compartilhamento livre de dados e scripts com códigos para execução das análises que eles abordam no R. Assim, desejamos que o conteúdo aqui presente auxilie você leitor (a) em sua tarefa analítica, amplificando a obtenção de resultados informativos e potenciais de aplicação prática.

Écio Souza Diniz

Pedro Manuel Villa

SUMÁRIO

CAPÍTULO 1

BIOVEG – A PROTOCOL TO LEARN AND TEACH STATISTICS IN R USING VEGETATION DATA

Data de aceite: 12/02/2020

Écio Souza Diniz
Laboratório de Geoprocessamento e Sensoriamento Remoto, Departamento de Engenharia Florestal, Universidade Federal de Viçosa, 36570-000, Viçosa, MG, Brazil.
eciodiniz@gmail.com

Jan Thiele
Thünen Institute for Biodiversity, Federal Research Institute for Rural Areas, Forestry and Fisheries, 38116, Braunchweig, Niedersachsen, Germany.
jan.thiele@thuenen.de

ABSTRACT: The wide range of vegetation types that exist throughout the world, and particularly tropical rainforests, are crucial for the provision of ecosystem goods and services. Efficient and robust analytical tools are required to extract valuable information from vegetation data, thereby enhancing our comprehension of the interaction of plants with the environment in which they occur. Such information is increasingly valuable in the context of rapid global change. The growing use of programming languages such as R, in computational statistics, has facilitated the development of ever more complex analyses, thus delivering significant advances in our understanding of vegetation dynamics. In this chapter, we present Bioveg, a freely available set of R scripts and datasets for the teaching and learning of statistical analysis using R. We use vegetation data to illustrate a range of statistical analyses, from commonly used basic statistics to many more advanced and complex procedures. The data we explore in this text was collected from temperate Vegetation in Germany and Brazilian Atlantic Forests. The datasets are freely available to download. We hope to contribute to the distribution of understanding of the proper use of statistical analyses in R, and their application for vegetation analysis.

KEYWORDS: R language, Statistics, Vegetation Data, Computational Programming

BIOVEG - UM PROTOCOLO PARA APRENDER E ENSINAR ESTATÍSTICAS EM R COM DADOS DE VEGETAÇÃO

RESUMO: A grande variedade de tipos de vegetação em todo o mundo, destacando as florestas tropicais, é crucial no fornecimento de bens e serviços ecossistêmicos. Devido à sua importância, os dados da vegetação exigem ferramentas analíticas eficientes e robustas para extrair informações valiosas, capazes de melhorar nossa compreensão da interação das plantas com o ambiente em que ocorrem, especialmente em tempos de

rápidas mudanças globais. O crescente uso de estatísticas computacionais por meio de linguagens de programação possibilitou alcançar níveis mais altos de precisão na aquisição de tais conhecimentos, o que destaca o uso do software R. Neste capítulo, apresentamos o Bioveg, um conjunto gratuito de scripts para R e conjuntos de dados para aprender e ensinar estatística em R usando dados de vegetação para exemplificar uma gama significativa de análises estatísticas distintas, das básicas às avançadas. Os conjuntos de dados fornecidos abrangem dados registrados pelos autores na Vegetação Temperada Alemã e Florestas Atlânticas Brasileiras e podem ser baixados gratuitamente. Esperamos contribuir para a difusão do conhecimento geral sobre estatística em R e sua aplicação na análise de vegetação.
PALAVRAS-CHAVE: Linguagem R, Estatística, Dados da Vegetação, Programação Computacional

1 I INTRODUCTION

Globally, all types of vegetation, and particularly tropical rainforests, are essential providers of ecosystem goods and services (CHAZDON, 2008; CHAZDON et al., 2009) such as carbon sequestration (PAN et al., 2011), protection of water supplies and flood control (ELLISON; FUTTER; BISHOP, 2012), soil maintenance (DITT et al., 2010) and many others. In times of rapid environmental change, understanding the interactions of plants with the ecosystem in which they occur is crucial. It is of particular importance to understand the likely consequences of environmental change on ecosystem dynamics, in order to generate plans to mitigate the negative impacts of environmental change on the provision of ecological goods and services. Modern computational tools are fundamental to delivering significant advances in this field.

The R environment is an integrated suite of software facilities for the manipulation of data using R programming language. Within the last two decades it has become one of the most commonly used programming languages to process data and execute a huge range of statistical analyses, for both plants and other organisms (CRAWLEY, 2012). There is a vast range of packages which have been developed for numerous statistical approaches, including many analytical specificities for multiple organisms, plant ecology and evolution (e.g. BORCARD; GILLET; LEGENDRE, 2011; CASSINI, 2013; GARAMSZEGI, 2014; SWENSON, 2014; PARADIS; CLAUDE; STRIMMER, 2004). The constant development of R packages and functions has facilitated important enhancements in our comprehension of plant communities (e.g. BARTON, 2016; BATES et al., 2015; FOX; WEISBERG, 2010; JIN; QIAN, 2019; KECK et al., 2016; KEMBEL et al., 2010; PARADIS; CLAUDE; STRIMMER, 2004; PINHEIRO et al., 2017; RITZ; STREIBIG, 2008).

Despite the significant benefits of using R and its associated packages for data processing and statistical analyses, there are regions (e.g. the tropics) where its use remains few explored. We present this text in order to contribute to the dissemination of knowledge of the use of R for vegetation data analyses. Bioveg is a freely available set of R scripts and data, with examples for a wide range of statistical analyses, covering basic statistical analyses and more advanced statistical techiniques. The Bioveg content comes with real data from Brazilian Atlantic Neotropical Forests and German Temperate Vegetation. This data forms the basis of the examples used to explain each type of analysis, and can be used to practice running analyses in R.

2 I DATA DETAILS

The data provided in the Bioveg dataset file are real data collected in the field and provided by the authors. The original data was randomized in order to avoid overlapping with published data or creating a conflict of interest. The datasets regarding Brazilian Atlantic Forest encompass three distinct forest types; semideciduos, Ombrophilous (rain) and gallery forest. Data collection took place between 2011 and 2018, using a standard sampling design of plots and subplots (10 per 10 m), at 3 sites in Southeast and Northeast Brazil.

The geographical coordinates of each study site were also altered to prevent overlap with published work. For each subplot within study sites, the diameter at breast height (DBH) and basal of each tree ≥ 10 and 15 cm were measured. Abundance and richness of each species within each subplot were also registered. Data collection was repeated after four and five years (depending on the study site location) to enable the annual rates of mortality, recruitment, gain and loss of basal area, for tree individuals within the subplots to be calculated. Climatic data (mean precipitation and temperature) were retrieved from WorldClim and extracted to each site using ArcMap from ArcGIS. Further environmental data regarding soil parameters (pH, P, K, Ca, Mg and Al), and species and vegetation types matrices are provided, to demonstrate how to perform multivariate analyses.

For the German Temperate Vegetation datasets, the field sampling took place in sites invaded by the species Giant Hogweed (*Heracleum mantegazzianum* Sommier & Levier), a monocarpic perennial herbaceous flowering plant in the family Apiaceae that is a real problem in Europe (BOOY et al., 2005) and North America (PYSEK et al., 2007) due to its aggressively invasive nature and the phytotoxic chemicals found in its sap. The data was recorded between 2002 and 2003 in South and West Germany. The sampling design was based on the Braun-Blanquet's Relevés method. The subplots were classified into broad habitat categories (Habitat type) based on physiognomy, type of disturbance and land use.

The habitat categories were agricultural grassland, which includes meadows and managed pastures; ruderal grassland, referring to meadows or pastures abandoned or irregularly maintained; tall herb vegetation, which includes several nitrophilous tall herb communities (of the Galio-Urticetea class, perennial or biannual vegetation types of mesic to wet habitats); woodland, for areas with more or less closed tree layer; and wasteland, which refers to areas of open vegetation that do not belong to either pasture or tall grass communities. Land use was classified in grassland, with regular cutting or grazing; maintenance, cut once or twice a year to prevent shrub invasion; 'none', no signs of cutting, grazing or other anthropic uses of the land. Disturbance types considered were deposition of organic materials ('deporg'), flooding by rivers ('flood'), 'none', disturbance of the topsoil ('soildis') and removal of trees or shrubs ('treeshrub').

For each study area, five soil samples were collected at a depth of 20 cm and homogenized and sieved, which was used to determine the content of major plant nutrients: total nitrogen (N-Perc), plant phosphorus availability in mg/100g (P_mg_100g) and potassium in mg/100g (K_mg_100g). Further, the altitudes of the study areas were obtained using GPS, and inclination and exposure were obtained with the use of inclinometer and compass. Additionally, height and percentage of Giant Hogweed cover were measured.

Overall, the content of Bioveg includes six dataset files: "Bioveg3.csv" containing most of the data on Brazilian Atlantic Forest, supported by the matrices "matrixspp.txt", "matrixtype.txt" and "matrixenv.txt" that relate to the tree community composition, type of vegetation and the soil parameters of these communities, respectively. Further, the file "Logit.csv" includes binary data (success or failure) describing tree recruitment. Finally, the file "Hogweed.csv" contains the complete German Temperate Vegetation dataset.

3 I ANALYSIS CONTENT

The examples in this text are all performed using the data provided in the Bioveg data file (6 datasets). The function and application of each of the analyses, and the particular relevance of these techniques in vegetation analysis, is explained. A complete list of the statistical analyses covered in Bioveg can be found in Table 1. Many of the examples can be executed using built-in functions in R, but some require prior installation of specific packages (Table 1).

ANALYSIS TYPE	STATISTICS APPLIED	R PACKAGES
Normality test	Shapiro Wilk test and residuals	Built-in Function
Data Transformation	Log, square, inverse and Arcsine square root	Built-in Function
One sample test	*Parametric*: T-test; *Nonparametric*: Wilcoxon Signed Rank Test	Built-in Function
Two sample tests	*Independent samples*: Nonparametric: Wilcoxon Rank-Sum Test (Mann-Whitney U-test); Parametric: T-test; *Dependent sample*: Wilcoxon Rank-Sum; Paired T-test	Built-in Function
Analyses of variance	*Parametric*: ANOVA One Way and Tukey post-hoc; *Nonparametric*: Kruskal-Wallis and Dunn's post-hoc test	Built-in Function and dunn.test
Independence test	Chi-squared	Built-in Function and MASS
Correlation analysis	Pearson, Spearman and Variance Inflation Factor (VIF)	Built-in Function and PerformanceAnalytics
Regression modeling	*Linear*: LM (Linear Model), GLM (Generalized Linear Models, GLMM (Generalized linear mixed models), Logistic regression, LME (Linear mixed models), GLS (Generalized Least squares) and post-hoc Tukey HSD and LmerTest tests. *Nonlinear*: Quadratic Model, Generalized Non-Linear Models (GNM), GAM (Generalized Additive Models) and GAMM (Generalized Additive Mixed Models). Note: There are examples treating spatial and temporal correlation.	Lme4, nlme, gnm, mgcv, multcomp, lmerTest, MASS, car, gmodels and ROCR
Model selection	Akaike information criterion (AIC) and AIC model average	MuMIn
Linearity diagnostic	Based on residuals of linear regression models	car
Spatial autocorrelation	Based on Moran's I statistics applied on linear regression models	lctools, ncf, nlme,
Supervised learning	Random Forest	randomForest
Multivariate analysis	NMDS, ANOSIM, CCA, PCA, PERMANOVA and INDVAL	BiodiversityR, vegan, indicspecies, permute and lattice

Table 1: Analyses, statistical methods and their related packages present in Bioveg. Built-in Function means package independent functions.

Regression analysis is used to examine relationships between variables in a wide range of scientific areas. In regression modelling we test how accurately the response variable is predicted by the explanatory variable. This type of statistical analysis is by far the most commonly used analysis in vegetation science and a large part of Bioveg is allocated to explain the various types of regression model (linear and non-linear) that can be used. Further detail and examples about some of these models can be found in Diniz and Thiele (2018). We focus on generalized models (fixed and mixed effects, GLM and GLMM) given the multitude of applications they offer and their ability to account for the variety of error distributions that are likely to be encountered in the dependent variable (e.g. Poisson, Binomial, Gaussian, Gamma).

It is vital when performing regression analysis that a number of criteria are met, violation of these criteria undermines the validity of the conclusions which are drawn from the model. Therefore, a number of important pre-tests must be conducted when constructing a model are explained in this text: a test for normality of the distribution; linearity diagnostics to verify whether there is a linear relationship between the predictor and dependent variables; correlation tests to check for collinearity among predictors and removal of variables that are found to highly correlate. We also describe how the model selection approach (Akaike Criterion) can be used to select the most useful and informative model from a global model which includes all possible predictor variables

We also provide examples of the use of smoothing functions for cases of generalized models which contain non-linear relationships (i.e. GAM, GAMM and GNM) between predictor and dependent variables. Lastly, we introduce the topic of machine learning by means of the decision tree approach using Random Forest. Random Forest is one of the most robust approaches of supervised machine learning used in predictive modelling; we therefore provide an example of this technique.

4 I DEMONSTRATIVE EXAMPLE

4.1 Predicting species richness with Random Forest

The algorithms of Random Forest constitute an ensemble learning method allowing for both classification and regression trees (BREIMAN, 2001). That is, these algorithms construct a multitude of decision trees during training and output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees (BREIMAN, 2001; HO, 1998). The decision tree is a type of modeling that operates with information gain on each node, classifying data points with greater information increment on each node (DIETTERICH, 2000). When the information gain ratings for all nodes are depleted, the model achieves its optimal performance result. Thus, the final decision is made, based on the majority, i.e. average, of the trees and is chosen by the random forest (DIETTERICH, 2000). In Random Forest, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error, since this is internally estimated using the out-of-bag (OBB) error for each data point recorded and averaged over the forest (JAMES et al., 2013). The averaging of the difference in out-of-bag error before and after the permutation over all trees provides an importance score for the predictors, i.e., which are more or which are less contributive to the overall model performance (HO, 1998). Moreover, the adjusted decision trees in Random Forest enhance its efficiency in dealing with overfitting in the training set data (BREIMAN, 2001).

In our example for Random Forest in the R script of Bioveg, we want to predict species richness from the Hogweed dataset on German Vegetation, utilizing all the other variables (16) as predictors. Thus, we used the function randomForest of the package randomForest (LIAW; WIENER, 2002) and carried out a regression in the training set (70% of the data). For further detail on this modeling technique, its steps and codes for reproduction of this example, please, refer to the Random Forest section in our freely available script. Figure 1 below shows in descending order the importance of each of the 16 predictors for the model performance, in predicting species richness.

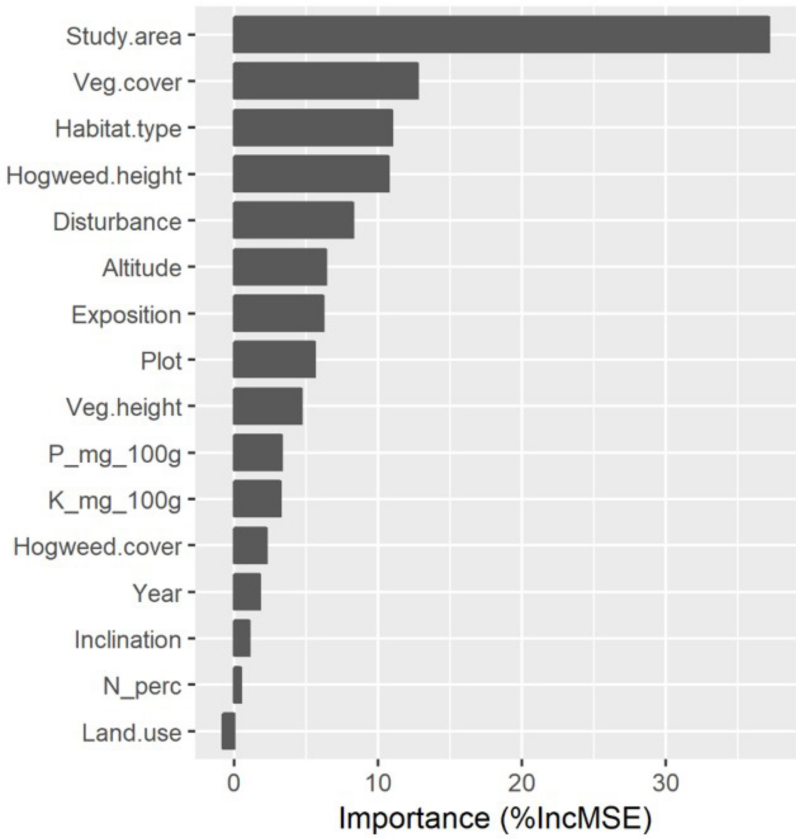

Figure 1: Variable Importance of predictors for species richness from a Random Forest regression model.

In accordance with the output of the Figure 1, study area, vegetation cover (Veg.

cover), vegetation type (Habitat.type) and hogweed cover are the most important variables for the model performance in predicting species richness, while from disturbance the importance starts to decline to all other predictors. The level of importance from potassium (K_mg_100g) to land use could be considered negligible for the model predictive power, i.e. withdrawing these variables from the model would have little effect on the performance of the model.

Bioveg details each analysis and tests it contains, providing a theoretical and practical overview of application for the user. The syntax an d code provided for each analysis may be adapted by the user to apply, test or exercise with their own data.

5 I ACCESSIBILITY TO BIOVEG

The whole content of Bioveg (datasets and script) can be freely accessed and download in the links:

Research Gate: http://doi.org/10.13140/RG.2.2.12488.26889

GitHub: https://github.com/eciodiniz/Bioveg

With the DOI (10.13140/RG.2.2.12488.26889) provided by license of Research Gate, Bioveg can be citable, please where possible support and give credit to this initiative by citing it in your studies.

6 I FINAL CONSIDERATIONS

Statistics and computer programming are central skills in the 21st century, and further progress in this field is vital if we are to address the increasingly complex problems that are arising in the modern world. Among these problems, environmental degradation and biodiversity loss demand urgent solutions. Developments in statistical analysis software are promising huge potential gains in the fundamental knowledge of the relationships and mechanisms maintaining ecosystem health. With greater knowledge we can begin to develop interventions to address these problems, but professionals need training in order to exploit the opportunities these new techniques offer. Our intention is that Bioveg will be used by professional and trainee researchers in the relevant fields (e.g. biologists, ecologists and foresters), to develop their use of R as a venue for data analysis and take advantage of the many benefits this platform can provide in the analysis of vegetation data.

REFERENCES

BARTON, K. **MuMIn: Multi-Model Inference.**, 2016. Disponível em: <https://cran.r-project.org/package=MuMIn>

BATES, D. et al. Fitting Linear Mixed-Effects Models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 51, 2015.

BOOY, O. et al. **The giant hogweed best practice manual. Guidelines for the management and control of an invasive alien weed in Europe**. Hoersholm: Forest and Landscape Denmark, 2005.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CHAZDON, R. L. Beyond deforestation: restoring forests and ecosystem services on degraded lands. **science**, v. 320, n. 5882, p. 1458–1460, 2008.

CHAZDON, R. L. et al. The potential for species conservation in tropical secondary forests. **Conserv Biology**, v. 23, n. 6, p. 1406–1417, 2009.

CRAWLEY, M. J. **The R book, 2nd Edition**. Chichester: Wiley, 2012.

DIETTERICH, T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. **Machine Learning**, v. 40, n. 2, p. 139–57, 2000.

DINIZ, E. S.; THIELE, J. **Modelagem de regressão no R: aplicações em análises de vegetação**. Saarbrücken: Novas Edições Acadêmicas, 2018.

DITT, E. H. et al. Forest Conversion and Provision of Ecosystem Services in the Brazilian Atlantic Forest. **Land Degradation & Development**, v. 21, n. 6, p. 591–603, 2010.

ELLISON, D.; FUTTER, M. N.; BISHOP, K. On the forest cover–water yield debate: from demand- to supply-side thinking. **Global Change Biology**, v. 18, p. 806–820, 2012.

FOX, J.; WEISBERG, S. **An R companion to applied regression**. Thousand Oaks: Sage, 2010.

HO, T. K. The Random Subspace Method for Constructing Decision Forests. **EEE Transactions on Pattern Analysis and Machine Intelligence**, v. 20, n. 8, p. 832–844, 1998.

JAMES, G. et al. **An Introduction to Statistical Learning**. New York: Springer, 2013.

JIN, Y.; QIAN, H. V.PhyloMaker: an R package that can generate very large phylogenies for vascular plants. **Ecography**, v. 42, n. 8, p. 1353–1359, 2019.

KECK, F. et al. phylosignal: an R package to measure, test, and explorethe phylogenetic signal. **Ecology and Evolution**, v. 6, n. 9, p. 2774–2780, 2016.

KEMBEL, S. W. et al. Picante: R tools for integrating phylogenies and ecology. **Bioinformatics**, v. 26, n. 11, p. 1463–1464, 2010.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002.

PAN, Y. et al. A large and persistent carbon sink in the world's forests. **Science**, v. 333, n. 6045, p. 988–993, 2011.

PARADIS, E.; CLAUDE, J.; STRIMMER, K. APE: analyses of phylogenetics and evolution in R language. **Bioinformatics**, v. 20, n. 2, p. 289–290, 2004.

PINHEIRO, J. et al. **nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-**

131, 2017. Disponível em: <https://cran.r-project.org/package=nlme>

PYSEK, P. et al. **Ecology and management of giant hogweed (Heracleum mantegazzianum)**. [s.l.] CABI, 2007.

RITZ, C.; STREIBIG, J. C. **Nonlinear Regression with R**. New York: Springer, 2008.

ÍNDICE REMISSIVO

A

B

C

D

E

F

G

I

K

M

N

P

R

S

Atena
Editora
2020