BioDinamica: a toolkit for analyses of biodiversity and biogeography on the Dinamica-EGO modelling platform

Ubirajara Oliveira[1], Britaldo Soares-Filho[1], Rômulo Fernandes Machado Leitão[1], Hermann Oliveira Rodrigues[1]

1- Centro de Sensoriamento Remoto, Instituto de Geociências, Universidade Federal de Minas Gerais – UFMG, Av. Antonio Carlos 6627, CEP 31270-901, Belo Horizonte, MG, Brasil.

Abstract

Biogeography and macroecology are at the heart of the debate on ecology and evolution. We have developed the BioDinamica package, a suite of user-friendly graphical programs for analysing spatial patterns of biogeography and macroecology. BioDinamica consists of a sub-library of Dinamica-EGO operators developed by integrating EGO native functions with R scripts. The BioDinamica operators can be assembled to create complex analytical and simulation models through the EGO graphical programming interface. In addition, we make available "Wizard" tutorials for end users. BioDinamica can be downloaded free of charge from the Dinamica EGO submodel store. The tools made available in BioDinamica not only facilitate complex biodiversity analyses, they also help develop state-of-the-art spatial models for biogeography and macroecology studies.

Keywords: Spatial patterns, GIS, modelling, species distribution models, beta-diversity, phylogenetic spatial analyses.

Biogeographical and macroecological studies have multiplied largely over the last decade (Ladle et al., 2015). Proportionally, novel methods of analyses have also been developed. Many of these methods focus on spatial pattern representation, such as areas of endemism, species richness and beta-diversity (Vilhena & Antonelli, 2015; Oliveira, Brescovit & Santos, 2015), while others aim to predict these patterns (Graham & Hijmans, 2006; Ferrier et al., 2007). Similarly, there has been an increasing number of studies including phylogenetic trees due to the growing availability of data (Hinchliff et al., 2015). Hence, biogeographical and macroecological methods that apply phylogenetic data to understand evolutionary geographical patterns have also multiplied, e.g. phylogenetic beta diversity, phylogenetic endemism and phylodiversity (Graham & Fine, 2008; Donnellan & Cook, 2009). In addition to biogeographical and macroecological analyses, all these novel methods are extremely important for conservation studies (Whittaker et al., 2005; Mcgoogan et al., 2007; Fenker et al., 2014).

Given the growing interest in spatial analyses in biogeographic and macroecology, we have developed a set of user-friendly tools embedded in the Dinamica-EGO software. We coupled a series of Dinamica EGO operators with R code to build more than 50 biogeographic and macroecological analytical functions (Table 1), all of which with a user-friendly graphical interface. These functions are stored in a sub library of Dinamica-EGO, named BioDinamica, thus allowing the user to build complex biodiversity models in a single integrated environment. In addition to the direct application of these tools to biogeography, biodiversity and macroecology (e.g. phylodiversity, species distribution models, phylogenetic endemism, areas of endemism, etc), some of the available functions, such as generalized linear models (GLM), geographically weighted regression (GWR), and raster PCA projection (principal components

[I1] Comentário: I would suggest including some examples of analyses included in the package.

[I2] Comentário: Present? Include?

[I3] Comentário: Not only because of the availability of data, but rather because the phylogenetic trees bring an explicit evolutionary component to biogeography analyses.

Excluído: Phylogenetic

Excluído: ,

[I4] Comentário: Before you introduce your own software, it would be enlightening to present a short review on already available packages and software. Surely most of the analysis included in BioDinamica can also be performed elsewhere. You should explain if these packages/software have any shortcomings, and how BioDinamica is a solution. Are they paid software? Are they all command-line-based? Are the needed functions dispersed in many different platforms and software?

[I5] Comentário: Please introduce Dinamica-EGO – what is it?

[I6] Comentário: Citation?

analysis) are also applicable to several other study fields. BioDinamica takes advantage of Dinamica EGO high performance computing, nonetheless, requiring computer resources as those available on common laptop computers, such as a minimum of 4GB of RAM and Windows or Linux operating system.

Overview

Functions provided include areas of endemism, species richness, phylodiversity, beta-diversity endemicity, species distribution models (SDMs), beta-diversity predictive models (GDM), interpolators, spatial analysis of ordination (PCA, PCR, NMDS), spatial statistical analysis (GLM, LM) and tools for conservation analysis, such as the Minimum Convex Hull (Table 1). All functions include R codes as well as specific R packages which are enveloped by the Dinamica EGO Operator called "Calculate R Expression". Although functions can be broken up for inspection, reuse, or further development, the users do not need to deal with the R code; instead they only need to configure or connect the parameters of these new hybrid operators by visually editing their inputs and outputs ports.

To facilitate the use of Biodinamica functions, we have standardized the operators' inputs (Figure 1). Thus, functions for analysis of spatial diversity patterns (species richness, beta-diversity, areas of endemism, etc.) have as input a table in csv format with points of occurrence of species in three columns: sp, x and y (species name, longitude and latitude in decimal degrees) and a mask of the study area in shapefile format (Figure 1). Analyses using phylogenetic data (phylodiversity, phylogenetic beta-diversity, phylogenetic-GDM, etc) include a phylogenetic tree in newick format, along with the inputs for analyses of diversity pattern (species points and mask, as mentioned above). Analyses that rely on predictor variables (such as GDM, SDMs, interpolation and prediction by GLM, LM, SAR) use as input raster files in the GeoTiff format. Spatial interpolation needs only a table in csv format with input variable values and respective geographic coordinates (columns: dependent, x and y). Predictor-based interpolations (GLM, LM, GWR, SAR) use as input a table in csv format including the values of dependent variable and their coordinates (dependent, x, y), together with the raster predictor variables. All analyses outputs textual logs including specific statistics (Figure 1). For analyses of spatial patterns, the functions output figures and graphs as well aimed to facilitate interpretation of results (Figure 1). To use BioDinamica, one only needs to install Dinamica-EGO (http://csr.ufmg.br/dinamica/) and the package BioDinamica. Complete documentation is available at (http://csr.ufmg.br/dinamica/dokuwiki/doku.php?id=biodinamica). The online supplementary material of BioDinamica comes with BioDinamica installation guide and a guide that provides a brief explanation of its functions.

Mapping spatial biodiversity patterns

BioDinamica includes several functions for spatial analyses of diversity patterns, such as beta-diversity, phylogenetic beta-diversity, endemicity, species richness, phylodiversity, and phylogenetic endemism. All these functions employ hexagonal tiles (equal area hexagons) as sample units, but also allow continuous interpolating of point data by using spatially explicit models. The interpolation models available in BioDinamica are the Spline method that derives a smooth prediction curve as a function of distance from observed points, nearest neighbour and the kriging, which applies a spatial interpolation according to a variogram distribution. Also available are analyses that predict spatial patterns (e.g. species richness, endemicity, phylodiversity) using predictor variables (e.g. climate variables) through generalized linear models (GLM), spatial autoregressive models (SAR), and universal kriging.

[I7] Comentário: A wide array of operating systems?

Excluído: tre

Excluído: (newick)

[I8] Comentário: Does it accept rasters in asc format?

[I9] Comentário: I was also asked to install an Enhancement Plugin. Some functions required additional submodels – could you please clarify this?

Analyses of beta-diversity and phylogenetic beta-diversity patterns allow beta-diversity partitioning into two components, turnover and nestedness. These components can be represented by using either hexagonal tiles or continuous interpolation in order to visualize the spatial variation of each component. To map beta-diversity patterns, we have implemented GDM (Ferrier et al., 2007). This model predicts the beta-diversity patterns by using environmental predictors. Our implementation of GDM also allows applying the beta-diversity model to scenario modelling (past, or future, for example). Some diversity variables are more affected by sampling density and bias, such as species richness (Oliveira et al., 2016). To cope with that, we have implemented a rarefaction technique (Oliveira et al. in press) that allows quantifying the relative richness between areas by standardizing sampling effort.

Excluído: to apply

Implementation of new methods

We also included novel analytical methods in BioDinamica. The Geographic Interpolation of Endemism (GIE) method identifies areas of endemism (AoE) (Oliveira, Brescovit & Santos, 2015). This method had not yet been fully implemented into a single integrated software environment. Hence, our GIE implementation needs not additional GIS software. The AoE outputs include raster maps for all scales and consensus, figures with AoE identification, tables describing how many and which species occur in each AoE, and a report with statistical information.

To identify spatial patterns of beta-diversity, we have implemented a new method named Species Composition Interpolation (SCI) (Oliveira, Vasconcelos & Santos, 2017). This method spatially interpolates beta-diversity patterns by using values of a NMDS of the beta-diversity index matrix. Our implementation generates a raster map for each axis of the specialized NMDS, and a multiband raster cube for visualization of the axes through a RGB composite. The model also tests the spatial autocorrelation of the values of NMDS, which is a premise for this analysis. As another option, the user can classify the resulting maps into discrete regions (biogeographic regions) through the k-means classification. We have also implemented a version of this analysis for phylogenetic beta-diversity (Oliveira et al. 2018 in press).

Evolutionary Spatial Patterns

BioDinamica provides a set of analytical tools for spatial mapping of evolutionary patterns. Using phylogenetic data, it is possible to create maps of phylogenetic beta-diversity (Graham & Fine, 2008), phylogenetic diversity, phylogenetic endemism (Rosauer et al., 2009), and to plot phylogenies on a map by using spatial interpolation. In addition, we have implemented Phylogenetic (GDM) (Rosauer et al., 2014). Finally, BioDinamica enables to perform scenario projections based on phylogenies analyses by using predictor interpolation (GLM, LM and SAR).

[I10] Comentário: ?

Species Distribution models

Today, species distribution models (SDM) are one of the most widely used biogeographic tools. We have implemented a set of SDM algorithms and tools in BioDinamica. In addition to SDMs themselves, we have created a set of tools for pre-processing and post-processing SDMs' inputs and outputs. For pre-processing, we have implemented two ways of creating pseudo-absences: the traditional one, which draws random points out of the presence samples of the species; and another based on sample evidence. The pseudo-absences based on sample evidence are obtained by sampling the pseudo-absences in the best sampled regions of the study group. In this way, the user provides sampling points for the study group and the

function generates a sampling effort map. From this map, the model draws samples (pseudoabsences) for areas where there is no record of the species and with more densely sampled places. In addition, various techniques for data validation and partitioning come with the SDMs package (table 1).

Interpolation

BioDinamica provides two forms of spatial interpolation: interpolation based on spatial data structure (spline, nearest neighbour and kriging); and statistical interpolation using predictive models (GLM, LM, SAR, GWR and universal kriging). Several biodiversity environmental data sets have an irregular spatial distribution and hence sampling gaps. For example, biodiversity variables, such as species richness, generally contain large sampling gaps. To cope with that, spatial interpolation is used to produce continuous surfaces of these phenomena. For example, by using the Spline and Kriging spatial interpolation tools, we can interpolate continuous variables based only on their spatial autocorrelation structure as a predictor. For more complex problems, and where there is information on possible predictors, we can interpolate the spatial distribution by using generalized linear models (GLM). In addition to these methods, we have implemented hybrid models that employ the spatial structure of the predictive variables (Spatial autoregressive model: SAR). The BioDinamica analyses of biodiversity patterns (species richness, phylodiversity, endemicity, beta-diversity and phylogenetic beta-diversity) interpolate results by using either spline, nearest neighbour or kriging as an option. In addition, these patterns can also be interpolated by predictive models (GLM, SAR and universal kriging).

[I11] Comentário: Most, if not all...

[I12] Comentário: This sentence is basically repeating the previous one. I suggest removing it.

Statistical and Ordination

Statistical analysis and ordering are central to biogeography and macroecology. To validate predictive models (such as SDM), we have built a binary map validation function. This function performs tests for accuracy, precision, sensitivity, specificity, Kappa and true skill statistics (TSS). For continuous value maps, we have included the area under the curve (AUC). For the analysis of spatial patterns, we have included the analysis of Moran I and the Spatial Variogram.

One common problem in spatial modelling (including SDMs) is the high correlation between variables. To analyse the correlation between raster maps, BioDinamica includes a map correlation test and the Clustering of Variables analysis (Chavent et al., 2012). Another strategy to avoid correlation between predictor variables is by means of principal component analysis (PCA). In BioDinamica, we have implemented a function that creates a raster cube of the axes of PCA. These raster maps can be used as predictors because while they still represent the original variables, there is no more correlation between them. In order to use PCA raster in models designed for scenario projection (such as climate change scenarios), we have included the PCA projection option. This option employs the PCA model generated with the current variables to produce a PCA raster under a different scenario from the one whereby the variables were generated. Another implemented ordering technique produces raster maps of axes that are free of correlation assigning different weights to the variables to maximize their predictive ability. These ordination methods are: principal component regression PCR; partial least squares regression PLSR; and canonical powered partial least squares CPPLS. In all of these techniques, the option of projection is available for scenario modelling. The raster maps generated from PCA, PCR and PLSR can be used as substitutes for the predictive variables in analyses in which the dependent variable has continuous values. The raster

[I13] Comentário: Yes; however, I think that as the first axes constrain a greater portion of the variance, not all axes should be used as predictors? Maybe you could rapidly comment on this.

generated from PCA and CPPLS can be used as predictor variables in analyses in which the dependent variable has discrete values. Furthermore, we have implemented spatialization by non-metric multidimensional scaling NMDS. This function can be used to spatialize genetic data (genetic, phylogenetic or phylogeographic distance matrix) or even morphometric data (by the morphological distance matrix).

Proof of concept

For exemplifying the potential of BioDinamica, we use the software to explore three patterns of bird diversity in the Amazon: beta-diversity, species richness, and endemism by using as input the distribution polygons of bird species from Birdlife International (http://www.birdlife.org). Oliveira, Vasconcelos & Santos (2017) have already explored biogeographic patterns of Amazonian birds using museum data. Here, we test the congruence of beta-diversity patterns as observed in Oliveira, Vasconcelos & Santos (2017) with those obtained from using occurrence polygons from the aforementioned dataset. We also investigate other bird geographical patterns (richness and index of endemism), which were not explored by Oliveira, Vasconcelos & Santos (2017). In addition, we investigate the use of predictive models based on environmental variables (GLM and GDM) to spatially predict these biodiversity patterns.

The polygons of distribution are converted into sample points through the function "Create samples" in BioDinamica. We employ 179,188 records of 446 species of birds endemic to the Amazon. To spatially interpolate the sample data, we apply "Species composition interpolation" (SCI), "Species richness interpolation" (SR) and "Endemism by weighing endemism" (WE). All methods consist of spatial interpolation techniques. In addition, we apply "Generalized Dissimilarity model" (GDM) for beta diversity and "Generalized linear model" (GLM) for predicting species richness and endemism (SRM and WEM, respectively). For GDM analysis, we use hexagons as sample units (1 degree side) and the geographic distance from sample units for estimating the effects of the environmental covariates. In GLM, we use the Gaussian distribution for model estimation. In this analysis, we employ all the 19 climatic variables from Wordclim (http://www.worldclim.org/) as environmental predictors. For that, we convert these variables (related to temperature and rainfall) into axes of a principal component analysis (PCA) to remove the correlations between them and to reduce the number of variables. This analysis is performed through the "Principal component" function.

[I14] Comentário: This part is unclear. You used polygons from Birdlife to randomly generate 179188 points using the function "Create samples"? If yes, why this particular number of records? How many records were created for each species, and was it weighted by the area of occurrence of each species?

Or are there 179188 records from museum samples from Oliveira et al. 2017?

[I15] Comentário: Does this mean that you have not used all of the 19 axes? If yes, how many axes did you use?

Beta-diversity results show spatial patterns very similar to those observed for Amazonian birds through collection data (Oliveira, Vasconcelos & Santos, 2017). Interpolation (SCI) and prediction using environmental variables (GDM) are quite similar as well (Figure 2). This is stressed by the high explanation of the model given the environmental variables (65% of explanation). This may indicate that the beta diversity geographic patterns as associated with the water basins of large rivers by Oliveira, Vasconcelos & Santos (2017) are, in fact, related to climatic conditions throughout the Amazon. Although all of these analyses are relatively complex, they are performed in a relatively short time. In a notebook with a 2.70GHz Core i7 - 7500U dual-core processor and 16GB of RAM, GDM runs in 3 minutes and 42 seconds and SCI in 27 minutes and 44 seconds.

Excluído:

[I16] Comentário: Could the rivers be responsible for the 35% of the turnover of beta-diversity unexplained by climate?

The analysis of species richness shows different results between interpolation and prediction by GLM (Figure 2). The interpolated results more closely resemble the raw data of Birdlife International. However, this type of analysis requires validation with independent data to determine which patterns best reflect reality. These analyses also run in a short time; the

Excluído: ,

interpolation of species richness runs in 7 minutes and 50 seconds and the prediction by GLM in 4 minutes and 23 seconds.

The patterns of endemism (WE index) are consistent with that observed by Oliveira, Vasconcelos & Santos (2017) for areas of endemism. The areas identified with the highest number of species with the most restricted distribution (Figure 2) are coincident with the smaller endemism areas identified by aforementioned authors who used a different set of data. The interpolation via WE runs in 15 minutes and 18 seconds and the prediction through GLM in 16 minutes.

[I17] Comentário: You should mention that the environmental prediction analysis indicated the AoE around the Guyana shield, which was recovered by the 2017 analysis, but was not indicated in the interpolation of polygon data (thus demonstrating the strength of using environmental data as predictors). (on the other hand, it also over-estimated endemism areas in southeastern Amazonia; you should mention this as well).

The short computer time demonstrates the efficiency of BioDinamica in processing large datasets. In addition, BioDinamica allows compressing the dimensionality of predictor variables through the "PCA function". Many other biogeographic patterns analyses are also doable using this same dataset and other BioDinamica functions, such as GIE, phylogenetic endemism, phylogenetic beta diversity, etc. This in turn demonstrates the software versatility in exploring geographical patterns of biological data.

Wizard: a tutorial interface

All the functions of BioDinamica are available as graphical operators of Dinamica-EGO. In addition, we provide model examples containing wizard tutorial. In this way, the user is guided through an illustrated tutorial that helps setting up and running the BioDinamica functions. Not only wizard tutorial illustrates applications, it also facilitates access to literature references (Figure 1). Lastly, BioDinamica installation comes with sample datasets for training. Also available is an online guidebook with a comprehensive tutorial on all BioDinamica functions (link: http://csr.ufmg.br/dinamica/dokuwiki/doku.php?id=biodinamica).

Final remarks

BioDinamica encompasses a wide variety of tools for spatial analyses of biodiversity, biogeography, macroecology and evolution. Developed using Dinamica-EGO freeware, BioDinamica delivers high performance on a user-friendly interface. In particular, the Dinamica-EGO platform allows the use of all functions into more complex models that includes loops, iterations and bifurcation pipelines. In this way, the BioDinamica functions become components of advanced models for conservation analyses and environmental simulations developed by using EGO graphical programming language.

Excluído:

[I18] Comentário: You could stress that the program is easy to use, but at the same time it is flexible enough so that more advanced users can modify models and scripts at their will.

Acknowledgments

We would like to thank Adalberto J. Santos, Ignacio Avila, Leonardo Sousa Carvalho, Marcelo Leandro Bueno, William Leles de Souza Costa. Ubirajara Oliveira and Britaldo Soares Filho received support from CNPQ. We also thank the support from FAPEMIG, Climate and Land Use Alliance, and the Alexander Humboldt Foundation.

References

Chavent M., Kuentz-Simonet V., Liquet B., Saracco J. 2012. ClustOfVar : An R Package for the Clustering of Variables. *Journal of Statistical Software* 50:1–16. DOI: 10.18637/jss.v050.i13.

Donnellan SC., Cook LG. 2009. Phylogenetic endemism : A new approach for identifying geographical concentrations of evolutionary history Phylogenetic endemism : a new approach for identifying geographical concentrations of evolutionary history. DOI: 10.1111/j.1365-294X.2009.04311.x.

Fenker J., Tedeschi LG., Pyron RA., Nogueira CDC. 2014. Phylogenetic diversity, habitat loss and conservation in South American pitvipers (Crotalinae: Bothrops and Bothrocophias). *Diversity and Distributions*:n/a-n/a. DOI: 10.1111/ddi.12217.

Ferrier S., Manion G., Elith J., Richardson K. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions* 13:252–264. DOI: 10.1111/j.1472-4642.2007.00341.x.

Graham CH., Fine PVA. 2008. Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecology Letters* 11:1265–1277. DOI: 10.1111/j.1461-0248.2008.01256.x.

Graham CH., Hijmans RJ. 2006. A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography* 15:578–587. DOI: 10.1111/j.1466-822x.2006.00257.x.

Hinchliff CE., Smith SA., Allman JF., Burleigh JG., Chaudhary R., Coghill LM., Crandall KA., Deng J., Drew BT., Gazis R., Gude K., Hibbett DS., Katz LA., Laughinghouse HD., McTavish EJ., Midford PE., Owen CL., Ree RH., Rees JA., Soltis DE., Williams T., Cranston KA. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* 112:12764–12769. DOI: 10.1073/pnas.1423041112.

Ladle RJ., Malhado ACM., Correia RA., Santos JG., Santos AMC. 2015. Research trends in biogeography. *Journal of Biogeography* 42:2270–2276. DOI: 10.1111/jbi.12602.

Mcgoogan K., Kivell T., Hutchison M., Young H., Blanchard S., Keeth M., Lehman SM. 2007. Phylogenetic diversity and the conservation biogeography of African primates. :1962–1974. DOI: 10.1111/j.1365-2699.2007.01759.x.

Oliveira U., Brescovit AD., Santos AJ. 2015. Delimiting Areas of Endemism through Kernel Interpolation. *PLOS ONE* 10:e0116673. DOI: 10.1371/journal.pone.0116673.

Oliveira U., Paglia AP., Brescovit AD., de Carvalho CJB., Silva DP., Rezende DT., Leite FSF., Batista JAN., Barbosa JPPP., Stehmann JR., Ascher JS., de Vasconcelos MF., De Marco P., Löwenberg-Neto P., Dias PG., Ferro VG., Santos AJ. 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions* 22:1232–1244. DOI: 10.1111/ddi.12489.

Oliveira U., Vasconcelos MF., Santos AJ. 2017. Biogeography of Amazon birds: rivers limit species composition, but not areas of endemism. *Scientific Reports* 7:2992. DOI: 10.1038/s41598-017-03098-w.

Rosauer DF., Ferrier S., Williams KJ., Manion G., Keogh JS., Laffan SW. 2014. Phylogenetic generalised dissimilarity modelling: a new approach to analysing and predicting spatial turnover in the phylogenetic composition of communities. *Ecography* 37:21–32. DOI: 10.1111/j.1600-0587.2013.00466.x.

Rosauer D., Laffan SW., Crisp MD., Donnellan SC., Cook LG. 2009. Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. *Molecular Ecology* 18:4061–4072. DOI: 10.1111/j.1365-294X.2009.04311.x.

Vilhena DA., Antonelli A. 2015. delimiting biogeographical regions. *Nature Communications* 6:1–9. DOI: 10.1038/ncomms7848.

Whittaker RJ., Araújo MB., Jepson P., Ladle RJ., Watson JEM., Willis KJ. 2005. Conservation biogeography: Assessment and prospect. *Diversity and Distributions* 11:3–23. DOI: 10.1111/j.1366-9516.2005.00143.x.

Figure 1: Graphical interface, inputs and outputs of BioDinamica operators.

Figure 2: Amazonian bird diversity patterns based on Birdlife International data analysed through BioDinamica functions. a: species composition interpolated by nearest neighbour, RGB represents the three axes of NMDS and b: predicted by GDM; c: species richness interpolated by nearest neighbour and d predicted by GLM; e: Weight endemism index corrected interpolated and f: predicted by GLM.

Table 1: Description of main functions, inputs and outputs of BioDinamica