Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 1229–1234 doi: 10.1111/2041-210X.12401

APPLICATION

letsR: a new R package for data handling and analysis in macroecology

Bruno Vilela[1,2]* and Fabricio Villalobos[1]

[1]*Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás, CxP. 131, 74001-970 Goiânia, Goiás, Brazil; and* [2]*Departamento de Ciencias de la Vida, Universidad de Alcalá, 28805 Alcalá de Henares, Madrid, Spain*

Summary

1. The current availability of large ecological data sets and the computational capacity to handle them have fostered the testing and development of theory at broad spatial and temporal scales. Macroecology has particularly benefited from this era of big data, but tools are still required to help transforming this data into information and knowledge.

2. Here, we present '*letsR*', a package for the R statistical computing environment, designed to handle and analyse macroecological data such as species' geographic distributions (polygons in shapefile format and point occurrences) and environmental variables (in raster format). The package also includes functions to obtain data on species' habitat use, description year and current as well as temporal trends in conservation status as provided by the IUCN RedList online data base.

3. '*letsR*' main functionalities are based on the presence–absence matrices that can be created with the package's functions and from which other functions can be applied to generate, for example species richness rasters, geographic mid-points of species and species- and site-based attributes.

4. We exemplify the package's functionality by describing and evaluating the geographic pattern of species' description year in tailless amphibians. All data preparation and most analyses were made using the '*letsR*' functions. Our example illustrates the package's capability for conducting macroecological analyses under a single computer platform, potentially helping researchers to save time and effort in this endeavour.

Key-words: biodiversity gradients, geographic distribution, presence–absence matrix, spatial analysis, species richness

Introduction

The rise of 'big data' is leading to transformations in the way research is conducted in different knowledge areas and ecology is not an exception (Hampton *et al.* 2013). Such amount and variety of data and the computational capacity to process it have turned ecology into a data-intensive science (Michener & Jones 2012), enabling theory testing and development at broader spatial and temporal scales and with more resolution than any time before. For instance, the macroecological research programme with its focus on the statistical regularities emerging from studying ecological systems at large scales (Marquet 2009) has particularly benefited from this big data revolution. Indeed, macroecology has steadily grown and become central in ecological research (Beck *et al.* 2012) contributing answers to unsolved questions in ecology and evolution, some of which were stated centuries ago (Hawkins 2001; Gaston, Chown & Evans 2008).

This current era of big data in ecology also brings big challenges. These challenges are mainly related to the manipulation of large data sets necessary to transform such data into information and knowledge (Schadt *et al.* 2010; Michener & Jones 2012). Converting large data sets into information and knowledge demands different steps including data gathering, organization, preparation, analysis and presentation (modified from Liew 2007). Moreover, without proper tools, conducting these steps may be exceedingly time-consuming and easily subject to human errors. Thus, current efforts are being conducted to develop tools that can help in this endeavour (Rangel, Diniz-Filho & Bini 2010).

In macroecology, typical data involve species' geographic distributions in polygon format or occurrence records (e.g. spatial data made available by IUCN, BirdLife or GBIF), species traits measurements (e.g. Jones *et al.* 2009), phylogenetic hypotheses on species' evolutionary relationships (e.g. Piel *et al.* 2000) and spatial environmental layers (e.g. Hijmans *et al.* 2005). Macroecological analyses require these data to be integrated and manageable. A practical way to integrate basic macroecological information is the presence–absence matrix (PAM) (Arita *et al.* 2008; Gotelli *et al.* 2009) that summarizes species' geographic distributions and diversity, the two fundamental units of biogeography (Arita *et al.* 2008). Conventionally, in a PAM, rows

*Correspondence author. E-mail: brunovilelasilva@hotmail.com.

represent species, columns represent sites, and elements record the occurrence of a given species in a given site, either as binary (i.e. presence: 1, or absence: 0) or quantitative data (i.e. abundance or trait values) (Bell 2003; Arita *et al*. 2008). Such PAMs can be generated from ecological surveys of localities or from overlaying a grid of cells onto the area of study, with the latter method being standard in macroecological analyses (Gotelli *et al*. 2009). Moreover, further data can be integrated into the PAM as additional rows representing site descriptors such as location (e.g. geographic coordinates), environmental conditions or other descriptors. In this way, a PAM allows species- and site-based analyses either summarizing data solely by species or sites (R-mode and Q-mode analyses; Simberloff & Connor 1979) or simultaneously considering information from both species and sites (Rq-mode and Qr-mode; Arita *et al*. 2008).

Here, we present a new R package, '*letsR*', for obtaining, handling and analysing data for macroecological research. The main functions of this package allow the user to generate a PAM from species' geographic distributions (polygon and point occurrence data) and merge it with species' traits and spatial environmental layers. Other package's functions provide tools to summarize and visualize the information contained in a PAM including transformation and direct analyses such as the estimation of distance matrices or spatial autocorrelograms. In addition, the package contains functions to acquire species data from the IUCN Red List online data base such as description year (the year in which a species was described by the species authority), current conservation status and its temporal trend, and habitat use, among others. We exemplified the package's use, functionality and capacity to convert large amounts of data into information and knowledge by analysing the geographic pattern of description year of tailless amphibians.

letsR package

The '*letsR*' package is written in the R language and its first version was released on CRAN (The Comprehensive R Archive Network) in May 2014 and can be used under R version 2.1 or higher (R Core Team 2015). We are constantly updating the package and encourage interested users to look for the latest package version on GitHub (https://github.com/macroecology/letsR). The functions available in '*letsR*' depend on other packages: '*raster*' (Hijmans 2015), '*XML*' (Lang 2013), '*geosphere*' (Hijmans 2014), '*maptools*' (Bivand & Lewin-Koh 2015), '*maps*' (Becker *et al*. 2014), '*sp*' (Bivand *et al*. 2008), '*fields*' (Nychka, Furrer & Sain 2014) and '*rgdal*' (Bivand, Keitt & Rowlingson 2014). Every function of '*letsR*' starts with the prefix '*lets*.' to avoid conflict with other R functions (see Table 1 for the main functions and their descriptions). The package was named after the Theoretical Ecology and Synthesis Lab (LETS; Portuguese acronym) of the Federal University of Goiás for their contribution to the advancement of macroecology in Brazil and world-wide.

Table 1. List of the '*letsR*' package main functions and their description

Type	Function	Description
Presence–absence functions	*lets.presab*	Creates a presence–absence matrix of species' geographic ranges within a grid
	lets.presab.birds	Creates a presence–absence matrix of species' geographic ranges within a grid for the BirdLife spatial data
	lets.presab.points	Creates a presence–absence matrix based on species' point occurrences
Spatial functions	*lets.addpoly*	Adds polygon coverage to a *PresenceAbsence* object
	lets.addvar	Adds variables (in raster format) to a *PresenceAbsence* object
	lets.classvar	Calculates the frequency distribution of a variable within a species' range
	lets.correl	Computes correlogram based on Moran's *I* index
	lets.distmat	Computes a geographic distance matrix
	lets.field	Creates species' values based on the species co-occurrence within focal ranges
	lets.gridirizer	Fits a *PresenceAbsence* object into a grid in shapefile format
	lets.maplizer	Creates a matrix summarizing species' attributes within cells of a *PresenceAbsence* object
	lets.midpoint	Computes species' geographic range mid-points
	lets.overlap	Computes pairwise species' geographic overlap
	lets.pamcrop	Crops a *PresenceAbsence* object based on a input shapefile
	lets.rangesize	Computes species' geographic range size
	lets.shFilter	Filters species' shapefiles based on its presence, origin and season
	lets.summarizer	Summarizes variable(s) values within species' ranges based on a presence–absence matrix
Data download	*lets.iucn*	Downloads species' information from the IUCN RedList online database
	lets.iucn.ha	Downloads species' habitat information from the IUCN Red List online database
	lets.iucn.his	Downloads species' temporal trend in conservation status from the IUCN RedList online database

PRESENCEABSENCE CLASS

The package basic functions work mainly with a new S3 object class called '*PresenceAbsence*'. The class '*PresenceAbsence*' is generated by the three '*letsR*' functions that create a PAM based on a user-defined grid cell system (see presence–absence function in Table 1). The new object class is a list consisting of three objects: (I) a sites by species matrix indicating the

presence (1) or absence (0) of a given species in a given site (note that this PAM arrangement is the transpose version – with sites in the rows and species in the columns – of the conventional one described above) with the first two columns containing the coordinates corresponding to the cells' centroids; (II) a raster containing species richness values per cell; and (III) a vector with species' names contained in the matrix. Any of these three internal objects of the '*PresenceAbsence*' class can be obtained in the standard way for objects of '*list*' class (i.e. using list-subsetting operators: '*[[*'or '*$*').

The class '*PresenceAbsence*' is fundamental for other '*letsR*' functions, mainly because it contains information beyond the PAM itself such as the user-defined grid cell system, including its resolution, projection, datum and extent as described by the extreme coordinates, as attributes of the raster object (the second object mentioned above). These attributes are indispensable for other analysis and they cannot be stored in a simple PAM. The '*PresenceAbsence*' class also allows using generic functions such as '*plot*', '*summary*' and '*print*', which facilitate its description and visualization. Nevertheless, the presence–absence functions (Table 1) also include an argument allowing users to choose between getting the PAM as an object of class '*matrix*' or '*PresenceAbsence*'. Finally, some '*letsR*' functions allow the user to input customary R objects (e.g. '*vector*', '*matrix*', '*data.frame*').

MEMORY AND TIME CONSUMPTION

Generating and manipulating a PAM can consume a large amount of random-access memory (RAM), mainly when handling data of large spatial extents and high resolution. The specific amount of RAM needed will depend on the grid extent, cell resolution and the number of species in the analysis. For example, to create a PAM for all anuran species of the world (5597) with a 1° grid cell resolution (see the example analysis below), R itself can consume up to 4·7GB of RAM (note that this number can grow exponentially with the increase in cell resolution). In this example, the whole process took around 1 h 30 min on a laptop computer with an Intel Core i7-4500U @ 1·80 GHz processor and running under Windows 8·1 of 64 bits. To help the user keep track of the analysis relative running time, namely how many runs are left to finish the analysis, some functions include the '*count*' argument to open a separate window containing the run countdown.

Example: spatial pattern of description year in tailless amphibians (Amphibia: Anura)

To illustrate the utilities of '*letsR*', we applied some of its functions to examine the geographic pattern of description year of anuran species (Fig. 1; the detailed R code to recreate the example is available in .Rmd and .pdf formats as supporting information). More specifically, we aimed to (i) generate a global map of the description year for the species of the Anura order; (ii) evaluate whether geographically closer species show similar description year; (iii) assess the effect of the species' geographic range size and the maximum value of human footprint (i.e. population density, land transformation, human access and power infrastructure) within their ranges on their description years. Instead of attempting an exhaustive exploration on this subject, our main goal was to exemplify the package's functionalities applied to a simple macroecological question using a spatially explicit approach that had only been conducted for limited areas (Diniz-Filho *et al*. 2005).

We obtained data on species' geographic distributions, in the shapefile format, by manually downloading them from the IUCN online data base (IUCN 2012; http://www.iucnredlist.org/technical-documents/spatial-data; Fig. 1, step 1) and transformed it into a presence–absence matrix (PAM) by using a global grid of 1° resolution using the function '*lets.presab*'. It should be noted that the grid itself is created by the function with its properties (e.g. resolution and extent) being chosen by the user. Applying the function '*plot*' on the resulting '*PresenceAbsence*' object allows the visualization of the species richness pattern (Fig. 1, step 2), whereas the function '*summary*' compiles the key information stored in the object and prints it in the console:

Class: PresenceAbsence

_ _
Number of species: 5597
Number of cells: 13 594
Cells with presence: 13 594
Cells without presence: 0
Species without presence: 0
Species with the largest range: Bufo bufo

_ _
Grid parameters
Resolution: 1, 1 (x, y)
Extention: −180, 180, −90, 90 (xmin, xmax, ymin, ymax)
Coord. Ref.: +proj=longlat +datum=WGS84

We gathered the description year for the 5597 species in our data set from the IUCN online data base (IUCN 2012) using the function '*lets.iucn*' (the '*PresenceAbsence*' object can be used directly as input). We obtained the mean description year of species within each cell by applying the function '*lets.maplizer*' (although we used the mean, other summary statistics can be applied; e.g. median, standard deviation), which returns either a vector with the cells' coordinates and the summarized attribute or this vector and a raster containing this information. Then, we used this raster to map the geographic pattern of description year of tailless amphibians (Fig. 1, step 3). Our results indicated that Europe concentrates the oldest described species, whereas tropical regions (mainly the west region of Australia, sub-Saharan Africa and the northern Andes) contain more recently described anuran species. This pattern can be expected considering that Europe was the cradle of most early naturalists and taxonomists that pioneered the application of Linnaeus' binomial nomenclature to species classification.

Exploring the geographic pattern of description year from a species-oriented perspective also provides information on the factors affecting the trend species discovery. Therefore, we obtained the species' distributional mid-point (i.e. the geographic range centroid) using the function '*lets.midpoint*' and

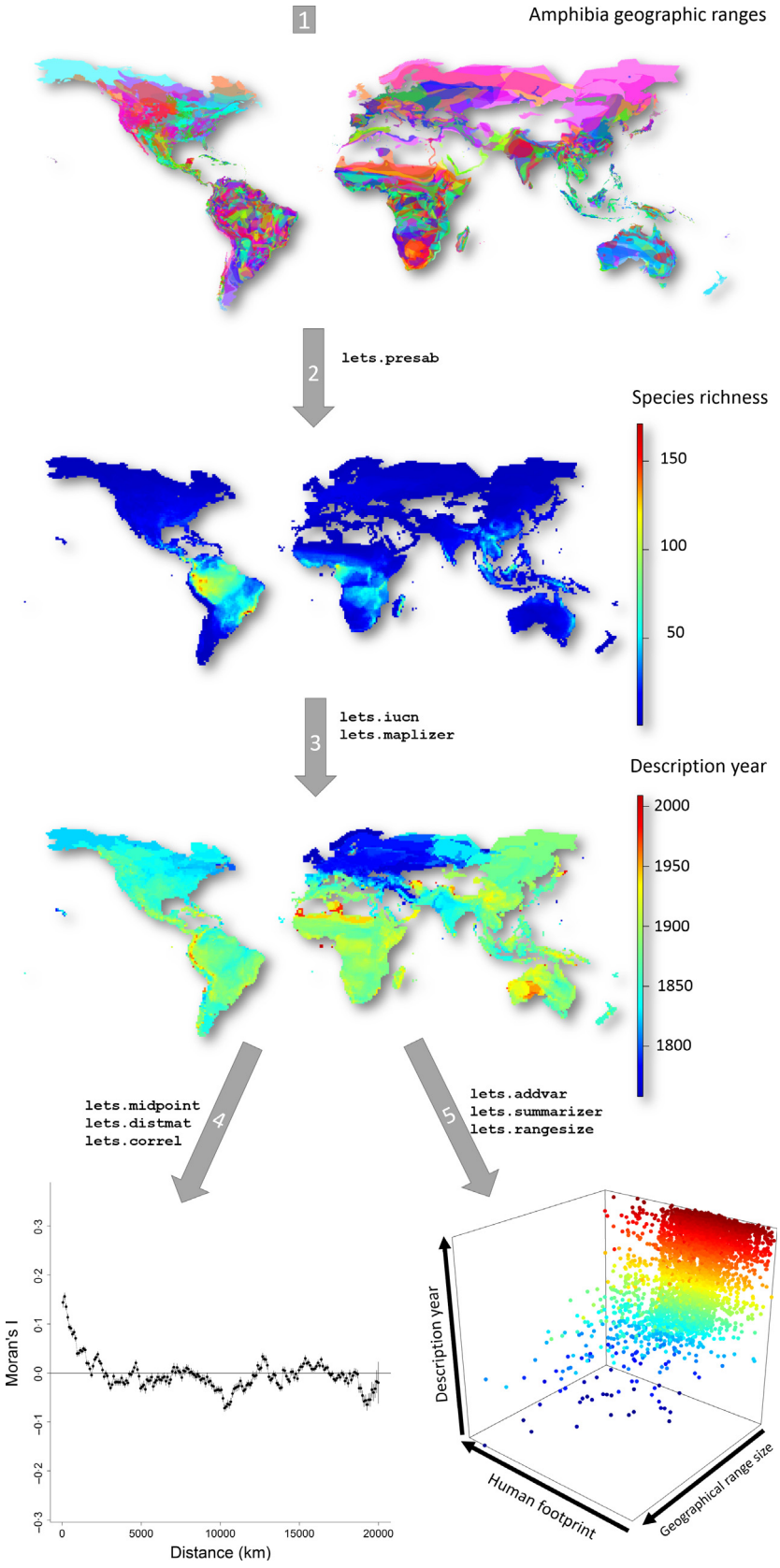

Fig. 1. Graphical sequence of the analysis to describe the geographic pattern of description year in Anura (Amphibia) using the '*letsR*' package. (1) Obtain the geographic ranges of tailless amphibians. (2) Transform this data into a presence–absence matrix (PAM) of 1° of cell resolution for the whole world. (3) Get the description year for each species from the IUCN online data base and summarize it by cells. (4) Calculate the species geographic midpoints and compute the pairwise distance between them, and use this information to generate a Moran's *I* correlogram for 200 equidistant classes. (5) Add the human footprint variable to the PAM and calculate the maximum value within each species' range, compute the geographic range size for each species and evaluate the effect of both variables (human footprint and range size) on description year variation among species. Functions used at each step of the process are listed at the side of each arrow representing these steps.

generated a geographic distance matrix between species using the function '*lets.distmat*'. This distance matrix and the description year of each species were used to calculate a Moran's *I* correlogram using the function '*lets.correl*' for 200 equidistant classes – which can be equiprobable, depending on the user's preference (Fig. 1, step 4). The results of this

species-oriented analysis support the idea that geographically closer species were described in proximate years. Moreover, species that are separated by a distance of 2000 km (or higher) had independent description years.

We also tested whether the species' description year pattern was determined by range size or human footprint level within species' ranges. To do so, we calculated the species range size using the function '*lets.rangesize*', which calculates the number of cells in which each species occurs (this function can also generate other range size metrics such as the polygon area or the summed area of cells). Then, we obtained a global layer of human footprint at a resolution of 30 arc-second cell size from the NASA Socioeconomic Data and Applications Center (WCS & CIESIN 2005). To assign a value of this variable to each species, we first upscaled the variable to a 1° cell size by averaging the values within each cell and then added it to the PAM as an additional column, using the function '*lets.addvar*'. After, we extracted the maximum value of human footprint within each species' range using the function '*lets.summarizer*' that transfers the spatial information at the cell level to the species level. Finally, we did a multiple regression with both range size and maximum human footprint within ranges as explanatory variables and species' description year as the response variable. A Monte Carlo simulation with 999 repetitions was used to test the model's significance. Results of this analysis indicated that both aspects of species' ranges influence the variation of description year among anuran species, jointly explaining nearly 30% of the variance ($p < 0{\cdot}001$).

Conclusion

Here, we have presented the '*letsR*' package and illustrated its functionality for conducting macroecological analyses under a single computer platform. '*letsR'* allows handling large data quantities commonly used in macroecology and applying different tools to process and analyse such data. For instance, to our knowledge, this is the first effort showing a global map of species' description year for an entire order of vertebrates. In addition, this also highlights the package's potential to help researchers in testing macroecological theories. '*letsR*' is constantly being improved to attend new demands as the field of macroecology continuous to develop.

Acknowledgements

We thank A.S. Melo, L. Sgarbi, S. Varela, J.A.F. Diniz-Filho, L. Jardim, F. V. Faleiro, M.S. Lima-Ribeiro, L.C. Terribile, M.A. Rodríguez, R. Dobrovolski and S. Gouveia for useful suggestions on the package's code and theoretical background. We also thank T. Lucas and S. Chamberlain for detailed suggestions that greatly improved our package's code and manuscript clarity. We are grateful to all the people of LETS (Laboratório de Ecologia Teórica e Síntese – UFG) for testing earlier versions of the package. B.V. was supported by a CAPES grant for doctoral studies and F.V. by a CNPq 'Science without borders' fellowship.

Data accessibility

Species distribution data are available in polygon format on the IUCN Red List of Threatened Species data base (http://goo.gl/UOhD7Q). The global layer of human footprint at a resolution of 30 arc-second cell size is available from the NASA Socioeconomic Data and Applications Center (http://www.ciesin.columbia.edu/repository/wildareas/data/hfp_global_geo_grid.zip). The '*letsR*' package is available on CRAN (http://cran.r-project.org/web/packages/letsR/index.html) and on GitHub (https://github.com/macroecology/letsR).

The R scripts to recreate the example analysis, including the function to gather the description year data (available from the IUCN Red List of Threatened Species, http://www.iucnredlist.org/), are uploaded as supporting information.

References

Arita, H.T., Christen, J.A., Rodríguez, P. & Soberón, J. (2008) Species diversity and distribution in presence-absence matrices: mathematical relationships and biological implications. *American Naturalist*, **172**, 519–532.

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S.A., Gruber, B. *et al*. (2012) What's on the horizon for macroecology? *Ecography*, **35**, 673–683.

Becker, R.A., Wilks, A.R., Brownrigg, R. & Minka, T.P. (2014) maps: Draw Geographical Maps. R package version 2.3-9. http://CRAN.R-project.org/package = maps.

Bell, G. (2003) The interpretation of biological surveys. *Proceedings of the Royal Society of London B: Biological Sciences*, **270**, 2531–2542.

Bivand, R., Keitt, T. & Rowlingson, B. (2014) rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.9-1. http://CRAN.R-project.org/package = rgdal.

Bivand, R. & Lewin-Koh, N. (2015) maptools: Tools for Reading and Handling Spatial Objects. R package version 0.8-34. http://CRAN.R-project.org/package = maptools.

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V. & Pebesma, E.J. (2008) *Applied Spatial Data Analysis With R*. Springer, New York, NY, USA.

Diniz-Filho, J.A.F., Bastos, R.P., Rangel, T.F., Bini, L.M., Carvalho, P. & Silva, R.J. (2005) Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. *Global Ecology and Biogeography*, **14**, 469–477.

Gaston, K.J., Chown, S.L. & Evans, K.L. (2008) Ecogeographical rules: elements of a synthesis. *Journal of Biogeography*, **35**, 483–500.

Gotelli, N.J., Anderson, M.J., Arita, H.T., Chao, A., Colwell, R.K., Connolly, S.R. *et al*. (2009) Patterns and causes of species richness: a general simulation model for macroecology. *Ecology Letters*, **12**, 873–886.

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S. & Porter, J.H. (2013) Big data and the future of ecology. *Frontiers in Ecology and the Environment*, **11**, 156–162.

Hawkins, B.A. (2001) Ecology's oldest pattern? *Trends in Ecology & Evolution*, **16**, 470.

Hijmans, R. (2015) raster: Geographic data analysis and modeling. R package version 2.3-24. http://CRAN.R-project.org/package = raster.

Hijmans, R.J. (2014) geosphere: Spherical Trigonometry. R package version 1.3-11. http://CRAN.R-project.org/package = geosphere.

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

IUCN (2012) *IUCN red List of Threatened Species. Version 2012.2*. International Union for the Conservation of Nature Gland, Switzerland.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L. *et al*. (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology*, **90**, 2648.

Lang, D.T. (2013) XML: Tools for parsing and generating XML within R and S-Plus. R package version 3.98-1.1. http://CRAN.R-project.org/package = XML.

Liew, A. (2007) Understanding data, information, knowledge and their inter-relationships. *Journal of Knowledge Management Practice*, **8**, 1–16.

Marquet, P.A. (2009) Macroecological perspectives on communities and ecosystems. *The Princeton guide to ecology*, 386.

Michener, W.K. & Jones, M.B. (2012) Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, **27**, 85–93.

Nychka, D., Furrer, R. & Sain, S. (2014) fields: Tools for spatial data. R package version 7.1. http://CRAN.R-project.org/package = fields.

Piel, W.H., Donoghue, M., Sanderson, M. & Netherlands, L. (2000) TreeBASE: a database of phylogenetic information. *Proceedings of the 2nd International Workshop of Species 2000*.

Rangel, T.F., Diniz-Filho, J.A.F. & Bini, L.M. (2010) SAM: a comprehensive application for spatial analysis in macroecology. *Ecography*, **33**, 46–50.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. & Nolan, G.P. (2010) Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, **11**, 647–657.

Simberloff, D. & Connor, E. (1979) Q-mode and R-mode analyses of biogeographic distributions: null hypotheses based on random colonization. *Contemporary Quantitative Ecology and Related Ecometrics*, **12**, 123–138.

WCS & CIESIN (2005) *Last of the Wild Project, Version 2 (LWP-2): Global Human Footprint Dataset (IGHP)*. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY. See http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-footprint-ighp (accessed 01 September 2014).

Received 16 April 2015; accepted 22 April 2015
Handling Editor: Timothée Poisot

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Data S1. R code description for analyzing the spatial pattern of description year in tailless amphibians.