speciesgeocodeR: An R package for linking species occurrences, user-defined regions and phylogenetic trees for biogeography, ecology and evolution

Alexander Zizka[1]* & Alexandre Antonelli[1,2]

[1] *University of Gothenburg, Department of Biological and Environmental Sciences, Carl Skottsbergs gata 22B, P.O. Box 461, SE 405 30, Göteborg, Sweden*

[2] *Gothenburg Botanical Garden, Department of Biological and Environmental Sciences, Carl Skottsbergs gata 22A, SE 41319, Göteborg, Sweden.*

*corresponding author: alexander.zizka@bioenv.gu.se

Abstract

1. Large-scale species occurrence data from geo-referenced observations and collected specimens are crucial for analyses in ecology, evolution and biogeography. Despite the rapidly growing availability of such data, their use in evolutionary analyses is often hampered by tedious manual classification of point occurrences into operational areas, leading to a lack of reproducibility and concerns regarding data quality.

2. Here we present speciesgeocodeR, a user-friendly R-package for data cleaning, data exploration and data visualization of species point occurrences using discrete operational areas, and linking them to analyses invoking phylogenetic trees.

3. The three core functions of the package are 1) automated and reproducible data cleaning, 2) rapid and reproducible classification of point occurrences into discrete operational areas in an adequate format for subsequent biogeographic analyses, and 3) a comprehensive summary and visualization of species distributions to explore large datasets and ensure data quality. In addition, speciesgeocodeR facilitates the access and analysis of publicly available species occurrence data, widely used operational areas and elevation ranges. Other functionalities include the implementation of minimum occurrence thresholds and the visualization of coexistence patterns and range sizes. SpeciesgeocodeR accompanies a richly illustrated and easy-to-follow tutorial and help functions.

Keywords

Geographic information system (GIS), biodiversity, data quality

Introduction

Species distributions and phylogenetic trees constitute core data in biogeography, ecology and evolution. Ancestral range estimation and area-specific diversification rate analyses are two prominent examples (e.g. Meseguer et al., 2014; Antonelli et al., 2015). Most methods applied in this context depend on the classification of taxonomic occurrences (typically species or populations) into discrete geographic units, such as continents, biomes, mountain ranges, or user-defined operational areas. These methods include, among others, Lagrange (Ree & Smith 2008), BayesRates (Silvestro *et al.* 2011), GeoSSE/diversitree (Goldberg *et al.* 2011; FitzJohn 2012), BAT (Cardoso *et al.* 2015) and BioGeoBEARS (Matzke 2013). However, detailed information on species distributions is often available as GPS-based point locations, for example from collected specimens or field observations. A "manual" classification of point occurrences into operational areas using expert knowledge or graphical user interfaces, including GIS software, is feasible for small-scale studies. However, it becomes time consuming and error-prone as the number of species and areas increase, or when the spatial resolution for the classification increases (e.g., assigning species to environmentally and topographically complex regions). Manual data curation is particularly impractical for large-scale analyses using data from public sources, such as the Global Biodiversity Information Facility (GBIF, www.gbif.org), ebird (Sullivan *et al.* 2009) or SUPERSMART (Antonelli *et al.* 2014) which has led to a discussion on data-quality and reproducibility of large scale studies (e.g. Yang et al., 2013; Hjarding et al., 2014; Maldonado et al., 2015).

Here we present speciesgeocodeR, an R package to automatically clean, process and analyse species occurrence data and to code them into discrete areas. The core functionalities of the package are: 1) automated and reproducible data cleaning, 2) the rapid and reproducible classification of point occurrences into discrete operational areas, in an adequate format for subsequent phylogenetic analyses, including widely used nexus files as well as input files for the BioGeoBEARS package

(Matzke 2013) and 3) a publication-quality and rich visualization of the results. The package summarizes occurrence and species numbers per area, which allows the rapid exploration of even very large datasets, and provides additional functionalities such as: including elevation ranges and occurrence thresholds in the area classification, automatic downloading of GBIF data and WWF biomes and ecoregions (Olson *et al.* 2001), as well as calculating and visualizing coexistence patterns and species ranges. SpeciesgeocodeR is a tool for exploration, visualization and quality control of species distribution. The package is particularly suitable for (but not restricted to) the preparation of input files for use with phylogenetic trees, such as for biogeographic reconstructions or diversification rate analyses.

Description

SpeciesgeocodeR is written for the use in R (R Developement Core Team 2015). It can handle datasets of any size (from a few species to hundreds of thousands), and is particularly user-friendly, also for R-beginners. A comprehensive tutorial ("Data cleaning and exploration with speciesgeocodeR") is provided as part of the package (Supplementary file 1). SpeciesgeocodeR takes advantage of methods and functionalities developed in the sp (Pebesma & Bivand 2005), raster (Hijmans 2014b), maps (Becker *et al.* 2013) and maptools (Bivand & Lewin-Koh 2013) packages. Furthermore some particular functionalities use the mapdata (Becker *et al.* 2014), rgbif (Chamberlain *et al.* 2015) and geosphere (Hijmans 2014a) packages.

Workflow example and case study

We demonstrate a typical workflow with the major functions of the package on a case study of Lemur (*Lemuriformes*) distributions in Madagascan biomes. Figure 1 shows the results and

mentions the functions applied. For illustration we use occurrences from GBIF (http://data.gbif.org 2015) and a simplified version of the WWF biomes for Madagascar as example data. The occurrence dataset included global records, and also comprised individuals in captivity. As lemurs in the wild are restricted to Madagascar, we use these specimens to illustrate one function of the automated data cleaning. The data are distributed together with the package. The code to reproduce the analyses can be found in the tutorial (Supplementary file 1) together with additional examples.

DATA CLEANING AND DATA QUALITY

The quality of occurrence data downloaded from public databases has often been debated and, less often, tested (e.g. Yang et al., 2014; Maldonado et al., 2015; Meyer et al., 2015). With the *GeoClean* function speciesgeocodeR offers an automated and reproducible flagging of potentially problematic records (see tutorial for code examples). The function includes basic tests for coordinate validity (e.g. invalid coordinates, zero coordinates, equal latitude and longitude) and more complex tests accounting for common problems in large datasets (e.g. if occurrences fall within the right country borders, if occurrences have been assigned to the country centroid or country capital or to the GBIF headquarters). Depending on the "verbose" argument (determining the amount of output information), the result can be a single vector summarizing the results of all tests, or a "data.frame" table with results of each individual test. Figure 1A shows the results of automated cleaning of the example dataset. Of 627 records, 224 were flagged as problematic (red and blue on the map) and excluded from subsequent analyses.

OCCURRENCE CLASSIFICATION

The core function of the package, *SpGeoCod*, classifies species occurrences into operational areas and calculates summary tables. The standards input are sets of (1) species occurrence points and (2) geographical areas (polygons). Both can be provided as R objects of the classes "data.frame" and

"SpatialPolygons", or as tab delimited .txt and .shp files in the working directory. Alternatively, a vector of species names can be used as occurrence data. In this case the occurrences will be downloaded from GBIF. In a similar manner, the WWF ecoregion and biomes (Olson *et al.* 2001) can be directly included as areas via the *WwfLoad* function. *SpGeoCod* offers a set of options to customize the area classification. For example, it is possible to condition the presence of a species in a given area on a minimum number of occurrences (as percentage of total species occurrences), or to include user-defined elevation zones into the classification. If desired, the entire analysis can be run with a single line of code after the data is uploaded:

```
data(lemurs)
data(mdg_biomes)
outp <- SpGeoCod(lemurs, mdg_poly, areanames = "name")
```

The results of *SpGeoCod* are saved in an object of the S3 class "spgeoOUT", which can be explored using plot or summary methods. The *WriteOut* function exports the area classification in nexus format for the direct use with phylogenetic software or as a text file in the adequate format for the BioGeoBEARS package. Additionally the function can create a set of graphics and maps (Fig 1A-I) in pdf format, and summary tables in text format. As part of the typical output, figure 1B shows an overview map of all lemur occurrences in broadleaved tropical forests (the most species rich biome in Madagascar) and figure 1C shows the occurrence number per species in this biome. Similar statistics and maps are automatically created for each area and each species, and could e.g. be included as supplemental files in publications.

DATA EXPLORATION

SpeciesgeocodeR includes multiple additional functions to explore the data based on the structure of the "spgeoOUT" class (see function help files for details on calculations). Species numbers in each area can be visualized using *MapRichness* (Fig 1D). Grids with species and occurrence numbers can be calculated using *RichnessGrid* and visualized using *MapGrid* (Fig. 1E). Species ranges and range

sizes given the occurrence points can be calculated using *CalcRange* and visualized using *PlotHull* (Fig. 1F, mean range size per lemur species = 56,000 km^2, median = 12,000 km^2). As for Madagascan lemurs, occurrence data is often scarce and gridded diversity maps can be strongly influenced by sampling patterns (Fig 1E). To circumvent this problem *RangeRichness* can create richness maps based on the species ranges (Fig 1G). A coexistence matrix of the species in the given polygons can be calculated using *CoExClass* and visualized as a heat plot using *plot* (Fig 1H). Figure 1I shows an example of including user-defined elevation zones in the analysis, showing the species number per area and elevation range, produced with *SpGeoCod(elevation = T, threshold = c(500, 100, 1500)*. Elevation is automatically calculated from georeferences by linking them to the CGIAR-CSI SRTM 90m digital elevation data (http://srtm.csi.cgiar.org/) that cover the world at 90m resolution.

Comparison with other tools

SpeciesgeocodeR adds new functionalities to toolbox of recent, primarily macroecological packages in the R-environment (e.g. modestR, García-Roselló et al., 2014 and letsR, Vilela & Villalobos, 2015). Conceptually, it differs from other packages by its focus on data quality control and the link to evolutionary and biogeographic analyses. Some of the implemented differences between SpeciesgeocodeR and available tools include 1) automated and reproducible data-cleaning, 2) the easy and quick area classification of species into pre-defined polygons producing a format suitable for analyses with phylogenetic trees, 3) the possibility to calculate elevation profiles and code species into elevation zones, 4) allowing the user to specify minimum occurrence thresholds for area classifications, 5) the seamless inclusion of GBIF data and WWF biomes and ecoregions and 6) the calculation of coexistence patterns in the areas of interest. Finally, a particular strength of speciesgeocodeR is the easy analysis and visualization of even very large datasets (including millions of records) with just a few lines of code. In the future, speciesgeocodeR could be easily expanded to

carry out additional analyses, including implementations of ancestral state reconstructions and geographic mapping of sister clades.

Availability

SpeciesgeocodeR is available from CRAN since October 2015 (http://CRAN.R-project.org/package=speciesgeocodeR) under a GPL-3 license, and a developmental version is available from GitHub (https://github.com/azizka/Speciesgeocoder-Rpackage). Users are welcome to join a users' list (speciesgeocodeR@googlegroups) for receiving updates, reporting issues and providing feedback and suggestions. A python version of speciesgeocodeR is currently being developed (Töpel *et al.* 2014).

Acknowledgments

The authors would like to thank Mats Töpel, Maria Fernanda Calió and Ruud Scharn for discussions and ideas on this project, Esther Nieto and Diogo Provete for testing the package and Angela Cano, Gaëlle Bocksberger and Daniele Silvestro for helpful comments on the manuscript. Funding for this work was provided by the Swedish Research Council (B0569601), the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024), and a Wallenberg Academy Fellowship to A.A.

References

Antonelli, A., Condamine, F., Hettling, H., Nilsson, K., Nilsson, R., Oxelman, B., Sanderson, M., Sauquet, H., Scharn, R., Silvestro, D., Töpel, M. & Vos, R. (2014) SUPERSMART:Ecology and Evolution in the Era of Big Data. *PeerJ PrePrints*, **2:e501v1**, 1–17.

Antonelli, A., Zizka, A., Silvestro, D., Scharn, R., Cascales-Miñana, B. & Bacon, C.D. (2015) An engine for global plant diversity: highest evolutionary turnover and emigration in the American tropics. *Frontiers in Genetics*, **6**, 1–14.

177 Becker, R.A., Brownrigg, R. & Wilks, A.R. (2014) Mapdata: Extra Map Databases. http://cran.r–
178 project.org/package=mapdata.

179 Becker, R.A., Wilks, A.R., Brownrigg, R. & Minka, T.P. (2013) maps: Draw Geographical Maps.
180 http://cran.r–project.org/package=maps.

181 Bivand, R. & Lewin-Koh, N. (2013) Maptools: Tools for reading and handling spatial objects.
182 http://cran.r–project.org/package=maptools.

183 Cardoso, P., Rigal, F. & Carvalho, J.C. (2015) BAT - Biodiversity Assessment Tools, an R package for the
184 measurement and estimation of alpha and beta taxon, phylogenetic and functional diversity.
185 *Methods in Ecology and Evolution*, **6**, 232–236.

186 Chamberlain, S., Ram, K., Barve, V. & Mcglinn, D. (2015) Rgbif: Interface to the Global Biodiversity
187 Information Facility 'API'.

188 FitzJohn, R.G. (2012) Diversitree : comparative phylogenetic analyses of diversification in R. *Methods
189 in Ecology and Evolution*, **3**, 1084–1092.

190 García-Roselló, E., Guisande, C., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., González
191 Vilas, L., González-Dacosta, J., Vaamonde, A. & Granado-Lorencio, C. (2014) Using modestr to
192 download, import and clean species distribution records (D. Orme, Ed.). *Methods in Ecology and
193 Evolution*, **5**, 708–713.

194 Goldberg, E.E., Lancaster, L.T. & Ree, R.H. (2011) Phylogenetic inference of reciprocal effects
195 between geographic range evolution and diversification. *Systematic biology*, **60**, 451–65.

196 Hijmans, R.J. (2014a) geosphere: Spherical Trigonometry. http://cran.r–
197 project.org/package=geosphere.

198 Hijmans, R.J. (2014b) raster: Geographic data analysis and modeling. http://cran.r–
199 project.org/package=raster.

200 Hjarding, A., Tolley, K. a. & Burgess, N.D. (2014) Red List assessments of East African chameleons: a
201 case study of why we need experts. *Oryx*, 1–7.

202 http://data.gbif.org. (2015) Lemurs, 627 records, accessed on 12 May 2015, contributed by 16
203 datasets, publisher identifiers: 10.15468/ckcdpy (143 records), 10.15468/ab3s5x (74),
204 10.15468/lwu4fj (49), 10.15468/llxrqq (33), 10.15468/j0xw9i (30), 10.15468/zwtpo2 (29),
205 10.15468/uc1apo (.

206 Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J., Chilquillo, E., Rønsted, N. &
207 Antonelli, A. (2015) Estimating species diveristy and distribution in the era of Big Data: To what
208 extent can we trust public databases. *Global Ecology and Biogeography*, **24**, 973–984.

209 Matzke, N.J. (2013) Probabilistic historical biogeography: new models for founder-event speciation,
210 imperfect detection, and fossils allow improved accuracy and model-testing. *Frontiers of
211 Biogeography*, **5**, 242–248.

212 Meseguer, A.S., Lobo, J.M., Ree, R., Beerling, D.J. & Sanmartin, I. (2014) Integrating Fossils,
213 Phylogenies, and Niche Models into Biogeography to Reveal Ancient Evolutionary History: The

Case of Hypericum (Hypericaceae). *Systematic Biology*, **64**, 215–232.

Meyer, C., Kreft, H., Guralnick, R.P. & Jetz, W. (2015) Global priorities for an effective information basis of biodiversity distributions. *PeerJ PrePrints*, **3**.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D.N.. D., Powell, G.V.N., Underwood, E.C., D'amico, J. a., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P. & Kassem, and K.R. (2001) Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, **51**, 933.

Pebesma, E.J. & Bivand, R.S. (2005) Classes and methods for spatial data in R. *R News*, **5**, 9–13.

R Developement Core Team. (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.r–project.org/.

Ree, R.H. & Smith, S. a. (2008) Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic biology*, **57**, 4–14.

Silvestro, D., Schnitzler, J. & Zizka, G. (2011) A Bayesian framework to estimate diversification rates and their variation through time and space. *BMC evolutionary biology*, **11**, 311.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009) eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*, **142**, 2282–2292.

Töpel, M., Calio, M.F., Zizka, A., Scharn, R., Silvestro, D., Antonelli, A., Calió, M.F., Zizka, A., Scharn, R., Silvestro, D. & Antonelli, A. (2014) SpeciesGeoCoder: Fast categorisation of species occurrences for analyses of biodiversity, biogeography, ecology and evolution. *bioRxiv.10.1101/009274.10.1101/009274.*

Vilela, B. & Villalobos, F. (2015) letsR: a new R package for data handling and analysis in macroecology. *Methods in Ecology and Evolution*, n/a–n/a.

Yang, W., Ma, K. & Kreft, H. (2013) Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *Journal of Biogeography*, **40**, 1415–1426.

Yang, W., Ma, K. & Kreft, H. (2014) Identifying factors shaping the geography of floristic collections in China. *Global Ecology and Biogeography*, **23**, 1284–1292.

Supporting information

Additional Supporting Information may be found in the online version.

Appendix S1. SpeciesgeocodeR tutorial

245 **Figure legend**

246 **Figure 1.** The major functions of the speciesgeocodeR package (function name in *italics*)
247 demonstrated for lemurs. **A)** Results of the automated cleaning of geographic occurrence data
248 (*GeoClean*), red = records flagged by checks for coordinate validity, blue = records flagged by test for
249 country borders. **B)** Lemur species occurring in moist broadleaf forest, as an example of the summary
250 maps automatically produced by *SpGeoCod*. **C)** The number of occurrence records for each species in
251 Moist Broadleaf Forest (*SpGeoCod*). **D)** Lemur species richness in the biomes of Madagascar
252 (*MapRichness*). **E)** Species richness raster of lemur species in Madagascar (*RichnessGrid*, *MapGrid*). **F)**
253 Overlaid distribution Ranges of 26 Lemur species (*CalcRange*, *PlotHull*). **G)** Species richness raster
254 derived from the ranges shown in I) (*RangeRichness*). **H)** Heatplot of the coexistence matrix of 39
255 Lemur species in three Madagascan biomes. The colors indicate the percentage (per row) of shared
256 occurrences given the biomes (*CoExClass*). **I)** Species number in different elevation ranges in each
257 biome (*SpGeoCod(elevation = T, threshold = c(500,100,1500)*)

258