Bioinformatics, 2016, 1–2
doi: 10.1093/bioinformatics/btw333
Advance Access Publication Date: 10 June 2016
Applications Note

OXFORD

Databases and ontologies

bdvis: visualizing biodiversity data in R

Vijay Barve[1,*] and Javier Otegui[2]

[1]Florida Museum of Natural History, University of Florida, Gainesville, FL, USA and [2]University of Colorado Museum of Natural History, Boulder, CO, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on December 2, 2015; revised on May 19, 2016; accepted on May 23, 2016

Abstract

Summary: Biodiversity studies are relying increasingly on primary biodiversity records (PBRs) for modelling and analysis. Because biodiversity data are frequently 'harvested'—i.e. not collected by the researcher for that particular study, but obtained from data aggregators such as the Global Biodiversity Information Facility—researchers need to be aware of strengths and weaknesses of their data before they venture into further analysis. R is becoming a lingua franca of data exploration and analysis. Here, we describe an R package, bdvis, which facilitates efforts to understand the gaps and strengths of PBR data with quick and useful visualization functions.

Availability and Implementation: The full code of the R package bdvis, along with instructions on how to install and use it, is available via CRAN – The Comprehensive R Archive Network (http://cran.r-project.org/web/packages/bdvis/index.html) and in the corresponding author's main GitHub repository: http://www.github.com/vijaybarve/bdvis. The source code is licensed under CC0

Contact: vijay.barve@gmail.com

1 Introduction

Biodiversity studies are in focus because of the perceived risk of mass extinction due to rapid environmental changes in recent years. Most studies rely on primary biodiversity records (PBR) (Andrew *et al.*, 2012, de la Torre *et al.*, 2012, Ramírez-Bastida *et al.*, 2008), which are basically records of species' occurrences in a specific place at a specific time. PBR are being used to study almost every aspect of human endeavor, from basic needs like food and shelter to science and politics (Chapman and Speers, 2005). Publications citing data served by the Global Biodiversity Information Facility (GBIF), currently the most preeminent network of PBR institutions, cover diverse areas like invasive alien species, climate change effects, conservation, human health, agriculture, etc. (http://www.gbif.org/mendeley), which illustrates broad relevance of PBRs.

Informatics tools are becoming essential in biodiversity science for improved management, exploration, discovery, analysis and presentation of biological and ecological information (Soberón and Peterson, 2004), challenges that are collectively referred to as biodiversity informatics. This is a relatively young, but rapidly growing, field whose aim is to leverage current computational techniques and information technologies to solve biodiversity problems. The solutions to many of the key challenges rely on availability of sets of large and good enough information.

More and more PBRs are being made available through aggregators or networks like GBIF and VertNet at global scale; on regional scales portals like BioCASE (biocase.org) and Indian Biodiversity Portal (indiabiodiversity.org) are actively serving PBR. GBIF currently serves more than 560 million PBRs. Major citizen science initiatives like eBird (ebird.org) and iNaturalist (inaturalist.org) have joined the venture, and have greatly fueled the growth of GBIF in recent years. However, due to the distributed nature of these huge data aggregators, spatial, taxonomic and temporal gaps may arise when collating the different sources they comprise. The package helps in identifying the gaps.

Visualizing data is a powerful technique in the biodiversity informatics domain, useful to quickly identify the strengths and weaknesses of a dataset, especially in terms of geo-spatial, temporal and taxonomic gaps (Otegui *et al.*, 2013a). These assessments help (i) data rights holders, to efficiently invest in improvement of the quality of their dataset, and (ii) users, to better understand the existing gaps in the data (Otegui and Ariño, 2012).

The R language (http://www.r-project.org/) is rapidly becoming the preferred tool for all kinds of data analysis. The package ecosystem

supported by R is very effective in making reusable functions available to users. R has numerous packages that serve an increasing range of purposes, several of which are useful for various biodiversity informatics-related tasks like the packages rinat (http://cran.r-project.org/package=rinat), rgbif (http://cran.r-project.org/package=rgbif), or dismo (http://cran.r-project.org/package=dismo). However, there is a lack of integrative tools for performing gap analysis on biodiversity data. In this paper, we briefly introduce the bdvis package, a tool that aims to bridge that gap.

2 Package description

The package's functions may be classified broadly as follows:

1. Helper functions to convert data to the correct format to be used in `bdvis`, and to enrich an initial dataset with additional data (like higher taxonomy and grid identifiers).
2. Summary tables
3. Geographic, temporal and taxonomic visualizations
4. Other miscellaneous graphs and charts

The data need to be in a format that the package understands for it to work. The functions under (1) help to achieve that. They change the name and format of some required fields (namely scientific name, date collected, latitude and longitude) so that visualization functions work seamlessly, and calculate extra fields for some of the visualizations. Executing these functions is a recommended first step when using the package. There is a wrapper function (`format_bdvis`) for common formats in biodiversity studies, such as data extracted using `rgbif` package (see Section 3).

After this initial step, the package is ready to handle the dataset. For the sake of simplicity, each visualization is created calling a single function, with parameters to customize the output. The resulting outputs are R standard graphics, which can be exported to jpg or tiff images in the regular way.

3 Example

To illustrate the usage of the package and a few of the visualizations it can produce, we applied the functions in the package over a set of 925 194 records of the genus Icterus, a group of New World birds, extracted from GBIF using the `rgbif` package.

```
require(bdvis); require(rgbif)
# Download Icterus records using rgbif
icterus=occ_search(scientificName="Icterus",
limit = 10000)
# Fix the field names and add extra values
icterus=format_bdvis(icterus$data,"rgbif")
# Create gridded species density map
mapgrid(icterus, ptype="species")
# Create temporal polar plot
tempolar(icterus, color="blue", plottype="r")
```

The resulting plots are shown in Figure 1. The mapgrid plot (left) reveals a latitudinal gradient of species richness, being higher in the southernmost part of the map, as well as a knowledge gap in the central-northern border of Mexico. The temporal plot (right) shows that most of the records have been sampled during May and early June, but there are some spikes in September and on January first. This last feature is most likely a quality issue in the records, derived from using a non-adequate value for storing 'unknown' information (see Otegui *et al.*, 2013b).

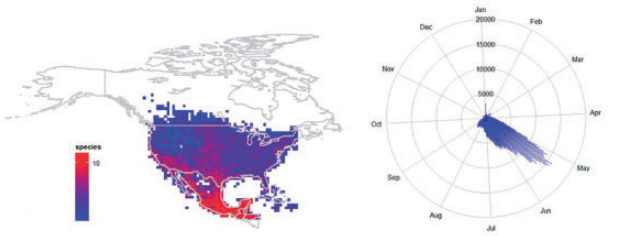

Fig. 1. Two visualizations of the Icterus dataset created with the **bdvis** package. *Left*: gridded map of species density for North America. *Right*: Temporal distribution of all records in the collection

With this simple example, we show the potential of the most basic visualizations produced with the `bdvis` package in detecting overall patterns as well as artifacts in a readily available set of records.

Acknowledgements

We are thankful to Google Inc. for the Google Summer of Code initiative, which brought the authors together to work on this package. We also thank the R Project for Statistical Computing for their support. For comments and early guidance on package development, we thank Scott Chamberlain, Carl Boettiger, Karthik Ram and Handley Wickham. We also thank A. Townsend Peterson, Jorge Soberón and Robert Guralnick, for guidance during development of the package, and Narayani Barve and Andrés Lira-Noriega for testing the package and offering suggestions on user interface. Toshita Barve offered helpful suggestions on the manuscript.

Funding

This work has been supported by the Google Summer of Code 2013 and 2014.

Conflict of Interest: none declared.

References

Andrew,M.E. *et al.* (2012) Beta-diversity gradients of butterflies along productivity axes. *Global Ecol. Biogeogr.*, **21**, 352–364.

Chapman,A.D. and Speers,L. (2005) *Uses of Primary species- Occurrence Data, Version 1.0*. Copenhagen.

De la Torre,L. *et al.* (2012) A biodiversity informatics approach to ethnobotany: Meta-analysis of plant use patterns in Ecuador. *Ecol. Soc.*, **17**,

Otegui,J. and Ariño,A.H. (2012) BIDDSAT: visualizing the content of biodiversity data publishers in the global biodiversity information facility network. *Bioinformatics*, **28**, 2207–2208.

Otegui,J. *et al.* (2013a) Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PLoS ONE*, **8**, e55144.

Otegui,J. *et al.* (2013b) On the dates of the GBIF mobilised primary biodiversity data records. *Biodivers. Inf.*, **8**, 173–184.

Ramírez-Bastida,P. *et al.* (2008) Aquatic bird distributions in Mexico: designing conservation approaches quantitatively. *Biodivers. Conserv.*, **17**, 2525–2558.

Soberón,J. and Peterson,A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.*, **359**, 689–698.