Ecological Indicators 91 (2018) 241–248

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases

Jorge M. Lobo[a,*], Joaquín Hortal[a], José Luís Yela[b], Andrés Millán[c], David Sánchez-Fernández[d], Emilio García-Roselló[e], Jacinto González-Dacosta[e], Juergen Heine[e], Luís González-Vilas[f], Castor Guisande[f]

[a] Department of Biogeography and Global Change, Museo Nacional de Ciencias Naturales, CSIC, Madrid, Spain
[b] Facultad de Ciencias Ambientales y Bioquímica, Universidad de Castilla-La Mancha, Campus Tecnológico de la Fábrica de Armas, 4507 Toledo, Spain
[c] Departamento de Ecología e Hidrología, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
[d] Instituto de Ciencias Ambientales (ICAM), Universidad de Castilla-La Mancha, Campus Tecnológico de la Fábrica de Armas, 45071 Toledo, Spain
[e] Departamento de Informática, Edificio Fundición, Universidad de Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Spain
[f] Facultad de Ciencias de Mar, Universidad de Vigo, Lagoas-Marcosende, 36200 Vigo, Spain

ARTICLE INFO

Keywords:
Spatial bias
Data limitations
Database records
Geographic distribution
Survey completeness
Wallacean shortfall

ABSTRACT

Biodiversity databases are typically incomplete and biased. We identify their three main limitations for characterizing the geographic distributions of species: unknown levels of survey effort, unknown absences of a species from a region, and unknown level of repeated occurrence of a species in different samples collected at the same location. These limitations hinder our ability to distinguish between the actual absence of a species at a given location and its (erroneous) apparent absence as consequence of inadequate surveys. Good practice in biodiversity research requires knowledge of the number, location and degree of completeness of relatively well-surveyed inventories within territorial units. We herein present *KnowBR*, an application designed to simultaneously estimate the completeness of species inventories across an unlimited number of spatial units and different geographical extents, resolutions and unit expanses from any biodiversity database. We use the number of database records gathered in a territorial unit as a surrogate of survey effort, assuming that such number correlates positively with the probability of recording a species within such area. Consequently, *KnowBR* uses a "record-by-species" matrix to estimate the relationship between the accumulated number of species and the number of database records to characterize the degree of completeness of the surveys. The final slope of the species accumulation curves and completeness percentages are used to discriminate and map well-surveyed territorial units according to user criteria. The capacity and possibilities of *KnowBR* are demonstrated through two examples derived from data of varying geographic extent and numbers of records. Further, we identify the main advances that would improve the current functionality of *KnowBR*.

1. Introduction

Current development of information technology and biodiversity informatics allows storing, retrieving, sharing, filtering and manipulating massive datasets such as those on species distributions (Bisby, 2000; Godfray, 2002; Soberón and Peterson, 2004; Graham et al., 2004; Guralnick et al., 2007). Global initiatives such as the Global Biodiversity Information Facility (GBIF) provide support for these big data (Saarenmaa and Nielsen, 2002) that can provide critical information for large-scale environmental questions (Hampton et al., 2013). However, even these comprehensively compiled databases suffer from a number of problems and shortfalls (Hortal et al., 2015). In fact, available data on the geographical distribution of biodiversity is limited and, often, inaccurate (Rocchini et al., 2011; Ladle and Hortal, 2013), so our knowledge on species distributions is typically incomplete (the so-called Wallacean shortfall; Lomolino, 2004; Whittaker et al., 2005). Consequently, rather than providing accurate descriptions of species geographic ranges, the extant databases are typically characterized by incompleteness and biases (e.g., Dennis and Hardy, 1999; Soberón et al., 2000; Zaniewski et al., 2002; Anderson, 2003; Martínez-Meyer, 2005; Dennis et al., 2006; Lobo et al., 2007; Hortal et al., 2008; Stropp et al., 2016).

Three limitations of the information from biodiversity databases are particularly important when characterizing the geographic

* Corresponding author at: Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (MNCN-CSIC), c/José Gutiérrez Abascal 2, 28006 Madrid, Spain.
E-mail address: mcnj117@mncn.csic.es (J.M. Lobo).

https://doi.org/10.1016/j.ecolind.2018.03.077
Received 18 September 2017; Received in revised form 26 March 2018; Accepted 26 March 2018

distributions of species:

1. *Unknown survey effort*, a lack of knowledge of the effort devoted to survey each territorial unit that is due to most occurrence records lacking any associated measure of the effort carried out to obtain them.
2. *Unknown absences*, as almost all the available information involves only species occurrences (i.e., the localities in which a species has been collected), without any indication of the likelihood that a species is actually absent from the localities where it was not collected (whether these have been surveyed or not).
3. *Unknown recurrence*, which results from the incomplete compilation of species occurrences in many biodiversity databases, as multiple records of the same species in the same site or territorial unit are considered redundant and not reported (Hortal et al., 2007). This prevents teasing apart occasional records from the continued presence of the species in an area.

These three limitations are mutually interrelated, so only when all known occurrences are comprehensively compiled it is possible to estimate survey effort with some reliability, thereby helping to differentiate the absence of evidence from the evidence of absence. Therefore, a biodiversity database that compiles exhaustively all available information on the identity and distribution of a group of species would enable both identifying well-surveyed areas (e.g. Hortal and Lobo, 2005) and obtaining estimates of the repeated occurrence and/or the probability of absence of particular species (e.g. Guillera-Arroita et al., 2010).

An important consequence of data limitations for biogeographical and conservation analyses is the impossibility of distinguishing whether the apparent lack of occurrence of a target species in a given location reflects its actual absence or is the result of insufficient survey effort. As a result, maps of observed species richness are often suspiciously similar to maps of the number of records per territorial unit (Hortal et al., 2007). Species Distribution Models (SDMs) are commonly used to offset such data incompleteness. Briefly, SDMs relate the available occurrence data with a number of environmental variables (often via sophisticated modelling techniques). The model created during this training phase is then projected into the geographical space to predict the probable, albeit unknown, distribution of species (Guisan and Zimmermann, 2000). Such predicted distribution, whether potential or realized, is often larger than the range documented by occurrence data (Soberón and Nakamura, 2009). Most SDM techniques rely on absence data to limit the geographical response of the species, so they are particularly sensitive to the *unknown absences* limitation. However, common usage of SDMs promotes an almost-universal use of random pseudo-absences (a.k.a. background absences) to include absences into the training data used to derive the predictive function. This practice comes from the classic procedure followed in Resource-Selection Functions (Johnson, 1980). Use of background absence data is, however, inadequate for estimating the probability of occurrence of a species (Hastie and Fithiam, 2013), because it only reflects the intensity of the collection process that led to the data used to train the model (Aarts et al., 2012). Hence, complex SDM algorithms calibrated with data containing background absences yield poor and inconsistent predictions, a fact that often passes unnoticed due to the use of inadequate evaluation methods (Hijmans, 2012).

Employing statistical shortcuts on data with unknown levels of error and bias can generate unreliable results. Consequently, good practice in biodiversity informatics requires knowledge about the number, location and degree of completeness of surveys for those territorial units that have been, at least relatively, well inventoried. Such knowledge would facilitate identifying localities where the lack of records for a target species can be reliably assumed to correspond to its actual absence. Nonetheless, it can be used to guide the location of future surveys and/or determine uncertain or ignorance areas in which biodiversity data are insufficiently consistent (Hortal and Lobo, 2005; Ladle and Hortal, 2013; Hortal et al., 2015; Ruete, 2015; Meyer et al., 2015; Meyer et al., 2016).

The effects of uneven levels of sampling effort have been traditionally addressed through species richness estimators and species accumulation curves (Soberón and Llorente, 1993; Colwell and Coddington, 1994; Hortal and Lobo, 2005). This is done under the assumption that they allow comparing the values of species richness and other aspects of biodiversity between sites surveyed with different levels of effort. Indeed, Chao and Jost (2012) and Colwell et al. (2012) recently demonstrated that it is more appropriate to compare estimated species richness values between sites showing similar rates of species accumulation with survey effort than between sites surveyed with the same intensity. That is, estimates can be reliably compared when the slopes of the relationships between observed number of species and the amount of survey effort are similar (i.e., standardizing by survey coverage *sensu* Chao and Jost, 2012). This implies that estimating survey coverage is crucial when we aim to identify those locations with probable reliable inventories.

Despite the widely recognized importance of evaluating data quality and completeness as a preliminary step in any biodiversity study, this process is often neglected. Arguably, this is in part because such evaluation process is highly time-consuming, it requires the use of several software applications and/or *R* packages, and repeating the same process for each one of the territorial units or sites considered (or, in general, for any type of spatial unit). Here we present *KnowBR*, a freely available *R* package to estimate the survey completeness of species inventories across an unlimited number of territorial units or sites simultaneously. Starting with any biodiversity database, *KnowBR* calculates the survey coverage per spatial unit as the final slope of the relationship between the number of collected species and the number of database records, which is used as a surrogate of the survey effort. *KnowBR* calculate the accumulation curve in each spatial unit according to the *exact* estimator of Ugland et al. (2003) (default estimator), as well performing 200 permutations of the observed data (*random* estimator) to obtain a smoothed accumulation curve that is subsequently adjusted to four different asymptotic accumulation functions. These functions allow to obtain a completeness percentage (the percentage representing the observed number of species against the predicted one) that also may be used to estimate the territorial units with probable complete inventories.

With *KnowBR* we aim to provide a tool to assess the levels of survey completeness across a territory, rather than an application for comparing species richness between sites by the use of the analytic rarefaction and extrapolation techniques developed by Chao and Jost (2012) and Colwell et al. (2012). *KnowBR* therefore estimates the degree of completeness of the inventories of all the territorial units within a given territory and, through that, allows identifying those spatial units that can be considered well surveyed (herein, *WSsus*) at a given resolution and extent, according to the information gathered in any biodiversity database. *KnowBR* allows performing all these time-consuming analyses in a very simple way, and simultaneously for a large number of spatial units both regular (*cell* option) and irregular (*polygon* option).

2. Installation and data entry

KnowBR can be used as a regular *R* add-on package in both Linux and Mac OS by installing the file KnowBR.tar.gz (package source), as well as in RGUI for Windows by installing the file KnowBR.zip (Windows binaries). Both files are available on CRAN (Development Core Team R, 2016) and also at the web site http://www.ipez.es/RWizard, in the download section. However, *KnowBR* can also be used as a regular application as a plug-in of RWizard, an easy-to-use graphical user interface for the *R* environment (Guisande et al., 2014). RWizard is an open-source interface under GNU General Public License