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

Table 1
Format of the two CSV files that can be used as input data in *KnowBR*. In the two cases the matrices reflect the occurrence of a species in different localities. Note that the two first records correspond to the same locality and species because they represent different sampling events (e.g. collections made on different dates and/or by different collectors). The same value (count = 1) is given to each record independently of the number of collected specimens. Each database record is thus a pool of specimens from a single species with identical information in different fields as location, altitude, date of capture, type of habitat, food resource or collector.

A				B				
Species	Longitude	Latitude	Counts	Longitude	Latitude	spp1	spp2	spp3
spp1	−4.694	34.800	1	−4.694	34.800	1		
spp1	−4.641	34.741	1	−4.694	34.800	1		
spp1	−2.244	34.893	1	−2.244	34.893	1		
spp2	−1.443	35.098	1	−1.443	35.098		1	
spp3	−4.713	36.140	1	−4.713	36.140			1
spp3	−3.224	38.675	1	−3.224	38.675			1
spp3	−3.213	38.787	1	−3.213	38.787			1

that has been developed in C# on the Net platform. It is designed as an interface to facilitate the interaction with *R* (see video demonstration at http://www.ipez.es/RWizard). The only requirement for the installation of RWizard is that *R* and Net Framework 4.0 must be already installed, and then the file Setup.RWizard.V2.3.exe must be installed. Although *KnowBR* can be used without RWizard, this interface increases the ease of use. *KnowBR* can be installed into RWizard from http://www.ipez.es/RWizard, following the menu "download" → "RWizard applications" → "Install *KnowBR*". This way, *KnowBR* will be added as another RWizard applications (Guisande et al., 2015; Guisande, 2016a,b,c).

The primary data matrix used in *KnowBR* must be derived from an exhaustive database including all available georeferenced information. This implies that all the records of the study group have been gathered without discarding apparently redundant data resulting from the repeated occurrences of any target species in the same locality (see Hortal et al., 2007). Such redundancies can be divided among alternate database records that despite pertaining to the same species and having been gathered in the same place, differ in any other collection condition (i.e. date of capture, food source, collector, type of microhabitat, etc.). Here, any difference in the value of any database field yields a new record, regardless of the sex and/or number of individuals captured (see e.g., Lobo and Martín-Piera, 2002).

Most biodiversity data can originate from heterogeneous sources with different collecting methodologies. Because of this and due to the inability to obtain a universal sampling effort measure, *KnowBR* uses the number of database records as a surrogate of sampling effort (Hortal and Lobo, 2005; Soberón et al., 2007; Lobo, 2008). This practice assumes that the probability of recording a species as occurring in a given territorial unit correlates positively with the number of database records gathered for that unit. This assumption is not always true. For example, the accumulation of database records in a locality can be the consequence of surveys focused on the collection of one or a few species of particular interest within a large biological group (e.g., a diverse genus), rather than being due to the even survey of all the species from that group inhabiting the site. High numbers of database records gathered in such a biased fashion could falsely indicate that a particular territorial unit is well-surveyed for all of the species in the taxon of interest. In light of this potential problem, it is advisable to assess the degree of coverage of the original surveys, to ascertaining whether (and where) they were directed to collect all the potential species from the focal taxon or, on the contrary, the collection bias could lead to falsely identifying certain unevenly sampled territorial units as well-surveyed.

KnowBR allows implementing a standard approach for assessing the completeness of biodiversity data to represent the geographical distribution of the species diversity of any given biological group. This approach is particularly appropriate for poorly surveyed groups and/or regions lacking sufficient information to correct the unequal sampling efforts resulting from otherwise standardized survey protocols. When this rationale is applied, the typical "species-by-sites" matrix is replaced by a "records-by-species" matrix containing only four values: the name of the species, the longitude and latitude of each record in decimal degrees, and a count value (see Type A format in Table 1 and the description of data in the *KnowBR* PDF manual). The data can be also included in *KnowBR* as a CSV file in which each column represents the occurrence of a species and each file is a database record (see Type B format in Table 1 and the description of data in the *KnowBR* PDF manual, available at http://www.ipez.es/RWizard/ and https://CRAN.R-project.org/package=KnowBR).

3. Identifying well surveyed territories

KnowBR aggregate the occurrence data in the spatial units (cells or polygons). In the case of cells, a grid is built considering the resolution selected by the user and the minimum and maximum latitude and longitude of the database records. Subsequently, all the occurrences included within each cell were discriminated for subsequent analyses, being each cell identified by its central latitude and longitude. In the case of polygons, the algorithm used the function *in.out* of the package *mgcv* (Wood, 2018) in order to find the database records belonging to each polygon.

KnowBR uses the species accumulation curve that describes the relationship between the accumulated number of species and the surrogate of survey effort (database records) to characterize the rate of increase with survey effort in each spatial unit (Clench, 1979; Soberón and Llorente, 1993; Hortal and Lobo, 2005) and to determine *WSsus*. This accumulation curve is established in *KnowBR* both analytically (*exact* estimator) and by randomization (*random* estimator). The equation of Ugland et al. (2003) is used in the first case (default option) as provided in the *specaccum* function of the *vegan* R package (Oksanen et al., 2014). Alternatively, the user may perform 200 random permutations of the original data with replacement thus smoothing the species accumulation curve in order to avoid potential spurious effects resulting from the order of addition of the records. The user must take into account that the random procedure may generate slightly different final slopes depending on the selected random samples. The completeness of each territorial unit can be decided based on the final slope of these analytical or smoothed curves (the slope between the two last steps of the so-generated species accumulation curves). This final slope is thus calculated by *KnowBR* for each spatial unit. This final slope can be considered a measure of the degree of completeness of the inventory in each spatial unit (the final rate of increase in the number of species attained with the survey effort developed so far; see Hortal and Lobo, 2005). Previously, the user has to decide the minimum ratio between the number of database records and the number of species necessary to proceed with the calculations. A particular slope cutoff value can be discretionarily chosen by the user to decide which spatial units are well-sampled enough so as to be flagged as *WSsus* (e.g. Hortal and Lobo, 2005).

In addition, the obtained species accumulation curves are adjusted

to four different species-accumulation functions with three or less parameters: the Michaelis-Menten equation used by Clench (1979) (*Clench* option), a negative exponential model (*N-exponential* option), a classic sigmoid saturation model (*saturation* option) and the rational function (*rational* option) which is used as default (see Soberón and Llorente, 1993; Flather, 1996; Mora et al., 2008). The extrapolated asymptotic values of all these curves can be used, according to user's selection, to estimate the probable number of species in each spatial unit when the number of records will tend to infinite, calculating subsequently the completeness (i.e. the percentage of species that has been inventoried). Again, a completeness value chosen by the user can be selected to decide the *WSsus*. Both final slopes and completeness values can be calculated simultaneously for an unlimited number of territorial units across the geographical extent defined by the user. These territorial units can be regular cells of any resolution (*cell* option) but also irregular polygons (*polygon* option) according to user preferences. RWizard include in the "Area" argument the possibility of select administrative spatial units (countries and/or provinces) or rivers basins of different levels in which to perform the calculations. Instead of using the polygons available in RWizard, the user may also include any shapefile containing the desired irregular polygons (e.g. protected areas, countries, etc) by means of the "shape" argument. For further details, see PDF manual of *KnowBR*_1.5 (https://CRAN.R-project.org/package=KnowBR).

Classical non-parametric estimators and extrapolated richness values available in other packages as *vegan* (Oksanen et al., 2014) are not used here for discriminating *WSsus*. When the data come from heterogeneous biodiversity data the use of these methods is undesirable. In all the formerly mentioned calculations each database record is treated as a sampling unit or independent observation. This means that for each record only one species is present. However, in the usual incidence and abundance based procedures (Colwell et al., 2012), the number of species in each observation is a random variable (i.e., cannot not be fixed to be one in advance). According to Robert K. Colwell and Anne Chao (personal communication), the use of non-parametric estimators is undesirable when the data is structured in the format of database records.

In order to propose key metrics able to identify *WSsus* the new function *SurveyQ* (Survey Quality) was included in the package *KnowBR* to identify and plot the well-sampled, moderated-sampled and poorly-sampled cells or polygons. Stated briefly (see PDF manual of KnowBR_1.5 for further details), using the file called "Estimators" obtained from *KnowBR* which includes the final slope values of the accumulation curves, the obtained completeness values and the ratio between the number of records and the observed species (R/S ratio), *SurveyQ* provides a map with the location of well-sampled, moderated-sampled and poorly-sampled spatial units. *SurveyQ* also provides a 3D graph showing the distribution of these three parameters in all the considered spatial units (cells or polygons). The default values used to distinguish well-sampled, moderated-sampled and poorly-sampled cells or polygons are: slope < 0.02, completeness $> 90\%$ and R/S ratio > 15 for well-sampled spatial units, and slope > 0.3, completeness $< 50\%$ and R/S ratio < 3 for poorly-sampled spatial units. In the case of polygons, a plot representing the values of the three parameters is depicted by using polar coordinates (Van Sickle, 2010).

4. Two practical examples

The first example is the assessment of the completeness of world-wide distributional data on bees. Taxonomy is based on the World Bee Checklist Project, downloaded from the webpage of the Integrated Taxonomic Information System (www.itis.gov; last accessed in December 2016). This checklist includes 19,508 species names of world-wide Apoidea (Insecta, Hymenoptera) (Ascher and Pickering, 2014), one of the most important insect groups from an applied point of view. Subsequently, we used the software *ModestR* (see García-Roselló et al., 2013, 2014) to download and clean all GBIF data available for Apoidea (downloaded 4 July 2014). In total, 137,809 records for 5,836 species names ($\approx 30\%$ of total species) have georeferenced distributional data in GBIF. By using *ModestR* we exported all these data directly to the format required in *KnowBR* at a resolution of one-degree grid cells. We then calculated the final slope of the species accumulation curve with the *random* estimator for all the considered cells in 67 min and 53 s for a computer with 8 Gb RAM and 3.40 GHz 2-nuclei processor. The results (Fig. 1) show that only c.18% of all terrestrial world one-degree cells (n = 2,776) have georeferenced data for bees, and that only 9% of all Apoidea species have more than 10 records. Approximately 14% of the cells with georeferenced data (i.e. 386 cells) have twice the number of records than recorded species, only 0.8% (128 cells) have a final slope less than or equal to 0.1 (i.e. one new species added to the inventory each 10 database records), and only 0.5% (74 cells) have a final slope lower than or equal to 0.01 (i.e. one new species per 100 records) (Fig. 1). These results exemplify the general scarceness, and large degree of bias, in the georeferenced information available among databases dealing with the biodiversity of many groups, even as economically important as bees. Notably, in this case survey effort has been distributed in an extremely uneven fashion, with most records gathered in western North America and central and northern Europe, while Australia is the region harboring a comparatively higher number of well-surveyed cells (Fig. 1).

The second example of the utility of KnowBR for assessing the completeness of distributional information at a regional extent is based on the use of BANDASCA (Lobo and Martín-Piera, 2002; Hortal and Lobo, 2011). This database currently contains 15,142 records for the 54 species of the dung beetle family Scarabaeidae (Coleoptera) on the Iberian Peninsula and adjoining islands. In this example, final slopes for four grid cells of varying spatial resolutions (5, 15, 30 and 60 min) were calculated in 1 min and 20 s. This regional database illustrates some of the caveats and difficulties of applying this kind approach to detect areas with reliable inventories at varying scales. Changing the spatial resolution in the analysis notably affects the proportion of cells that can be considered to be well-surveyed (Fig. 2; see also the green broken line in Fig. 3A). Only 2% of all Iberian cells (138 cells) can be considered well-surveyed at a 5-min resolution, with the consequent reduction in the probability of reflecting the whole spectrum of environmental conditions present in the Iberian Peninsula, compared to larger cell sizes. Further, the resolution of the analyses can also affect our capacity to reflect the "true" faunistic composition of the obtained inventories. Some well-surveyed cells detected at low resolution (60') contain a large number of well-surveyed cells at the highest resolution (5'), showing that their well-surveyed inventories at the "regional" scale are the consequence of a number of complete "local" inventories (Fig. 3B). In some extreme cases, however, a 60' cell would be considered well-surveyed even when none of the 5' cells it contains presents complete inventories. This probably arises because many heterogeneous and incomplete "local" inventories may be capable of providing an, apparently, reliable "regional" species inventory. It follows that a low number of reliable "local" inventories can give rise high completeness values at coarse resolutions. We suspect that the apparent completeness of these "regional" inventories may actually underrepresent their local variability, at least in environmentally heterogeneous areas. Thus, carrying out an exhaustive survey effort at a locality may imply that larger territorial units containing this locality also appear as well-sampled, whereas their internal heterogeneity has not been adequately surveyed. The results provided by this example illustrate the need for caution when comparing estimates of inventory completeness carried out at different resolutions. It is therefore advisable to conduct exploratory analyses at varying spatial scales to determine whether this kind of effect may be hampering the completeness of the inventories in coarse territorial units, at least in highly heterogeneous regions. What is the most appropriate resolution in each case? We do not believe there is an easy answer to this question. Be that as it may, the used resolution

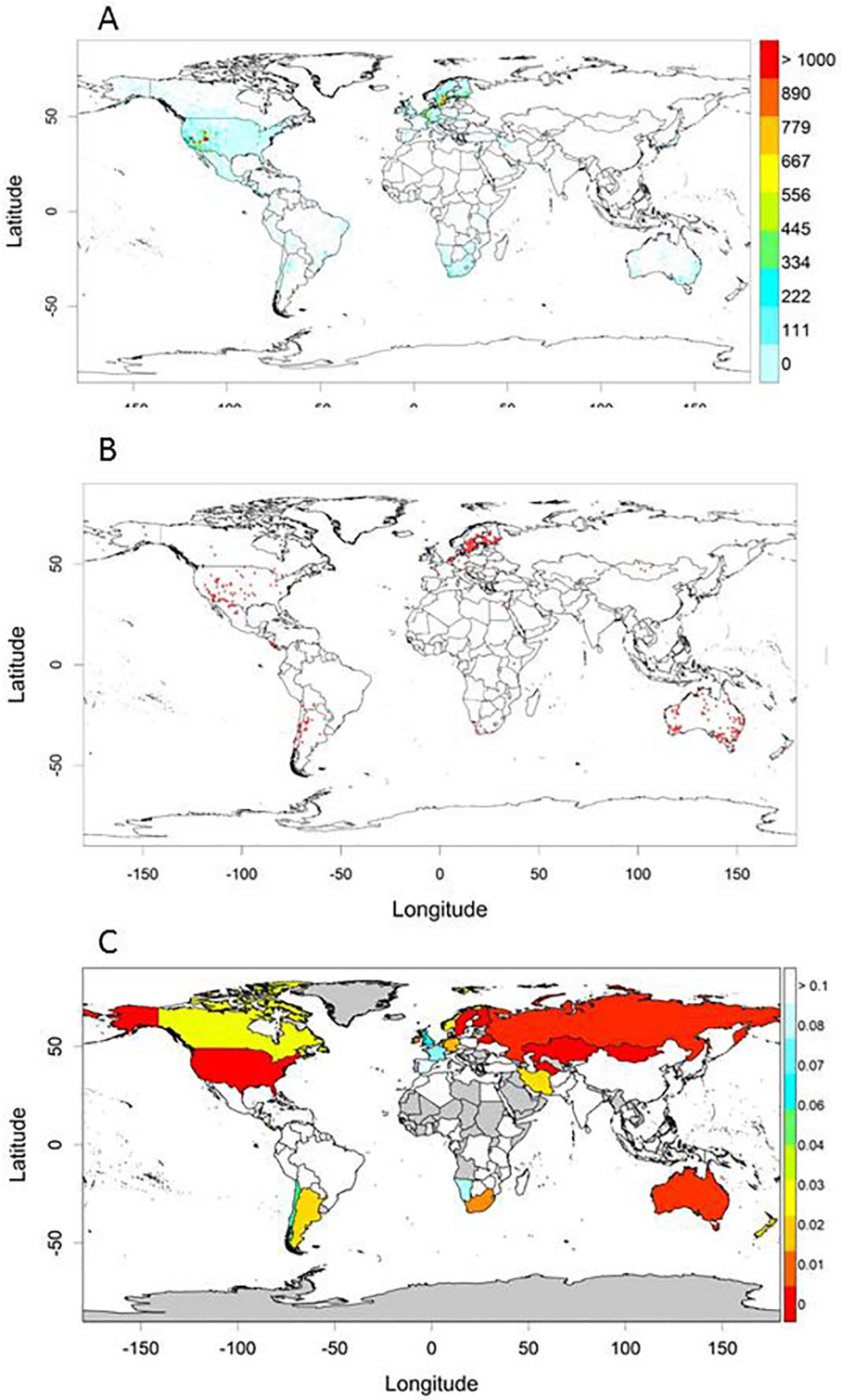

Fig. 1. (A) Number of database records included in GBIF for the Apoidea of the World (Ascher and Pickering, 2014) in one-degree grid cells. (B) Grid cells with final slopes of the relationship between the accumulated number of species and the number of database-records between 0 and 0.1. These cells are preliminarily qualified as well-sampled (see main text). (C) Final slopes of accumulated richness relationships for all world countries.

should be the one closest to the average home range of the considered species".

5. Future prospects

KnowBR has the potential to become a standard tool not only to assess survey completeness, but also to provide fair estimates of the distribution of biodiversity, and to study the survey process in detail. To increase the current functionality of *KnowBR,* future *R* applications should incorporate tools: i) to identify the spatial units that would be most appropriate for surveying to maximize the spatial and environmental coverage provided by the set of well-surveyed territorial units (such as Medina et al., 2013); ii) to identify reliable absences for any focal species and model species distributions using techniques that take

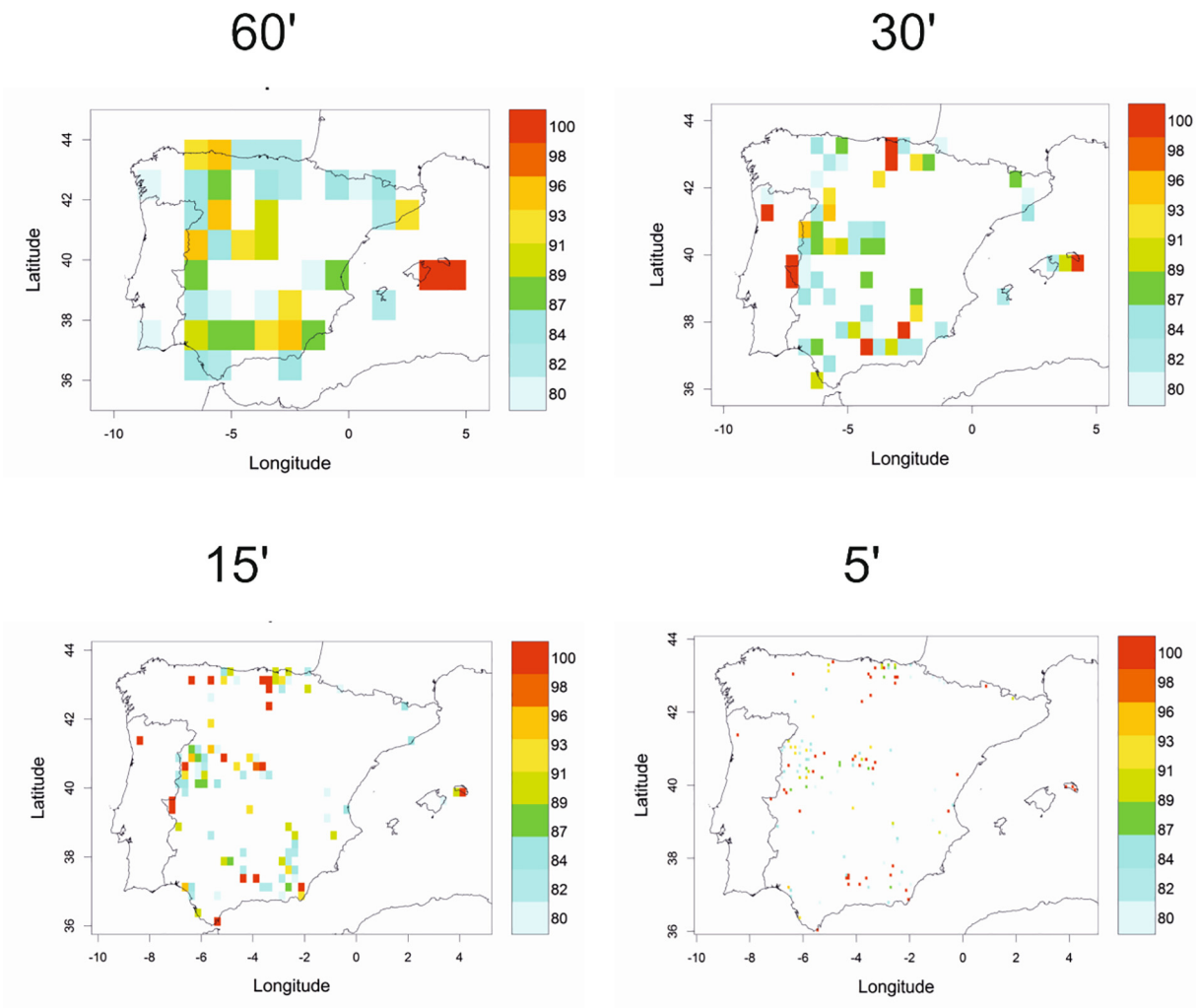

Fig. 2. Geographical distribution of the well-surveyed grid cells with reliable inventories for dung beetles (Coleoptera, Scarabaeidae) on the Iberian Peninsula and adjoining islands at multiple spatial resolutions. Here, well-surveyed grid cells are defined as those with completeness values ≥ 80% of the value predicted by the *exact* estimator (see text). The maps depict the results for four grid cell resolutions, 60, 30, 15 and 5 min, which correspond approximately to UTM cells of 14,400, 3,600, 900 and 100 km^2.

advantage of both occurrence and probabilistic absence data (Lobo et al., 2010); iii) to calculate the degree of uncertainty associated to the results of these models in poorly surveyed areas, based on their degree of completeness and the distance to well-surveyed areas (i.e. maps of ignorance *sensu* Rocchini et al., 2011); and iv) to describe the biases in the spatial distribution of sampling effort, and explore the factors behind these biases.

These improvements are aimed to allow developing the protocol for mapping biodiversity attributes described in Hortal et al. (2007) as the concatenated use of several *KnowBR* modules. If important gaps in species distribution information become apparent once well-surveyed territories have been identified, additional surveys will be required. These surveys should focus on covering as much of the spatial and environmental variability of the studied territory as possible while minimizing resource expenditure and efforts (Ferrier, 2002; Hortal and Lobo, 2005; Rocchini et al., 2011). Once well-surveyed areas provide enough coverage of the region, the future incorporation of species distribution modeling tools in *KnowBR* will allow filling in the gaps in the known distribution of species without resorting to the impractical task of exhaustively surveying the entire region. A practical consequence of the use of data from well-surveyed territories is that these models can provide reliable interpolations when these units represent the full range of environmental and spatial variation across the region of interest. On the contrary, when insufficient well surveyed units are available, the results of the models may extrapolate to unknown conditions or areas (Austin and Heyligers, 1989; Ferrier, 2002; Ferrier et al., 2002a,b), thereby diminishing the accuracy and reliability of model projections (Kadmon et al., 2004, Hortal and Lobo, 2011). To avoid these misleading extrapolations, it is critical to provide the tools necessary to calculate the environmental coverage provided by the data, to identify the territorial units that would likely result in better coverage once surveyed, and to generate predictions of species distributions that take into account the uncertainty in their projections to areas and domains with less environmental and spatial coverage.

In sum, *KnowBR* aims to improve the reliability of the results of biodiversity informatics applications, such as species distribution models. Good predictions require good biological data. The use of information from sites with inventories close to complete facilitates generating reliable predictions, and can also be used to design more efficient surveys that optimize data coverage for the analysis of biodiversity patterns. Good applications of biodiversity data will come only from assessing the reliability of data and accounting for its actual quality and accuracy. Of course, *KnowBR* does not allow overcoming the intrinsic limitations of the used data due to natural dynamics of ecological systems or survey difficulties. Rare, vagrants or difficult to survey species as well as collection bias may generate inadequate inventories (Dennis et al., 2006; Lobo et al., 2007; Hortal et al., 2008) that may add a supplementary uncertainty to the discrimination of well

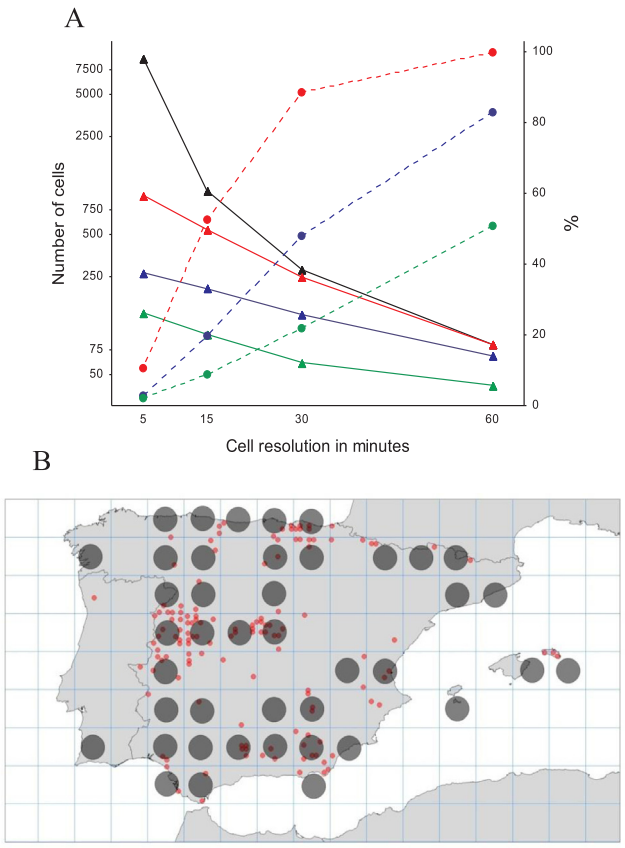

Fig. 3. (A) Variation in the number of grid cells (in logarithms) with reliable inventories for dung beetles in the Iberian Peninsula and adjoining islands (continuous lines and triangles) and their percentages (broken lines and circles) with regard to the total number of terrestrial grid cells (black continuous line). The numbers and percentages of the cells with at least one database record are represented in red, those cells with twice as many records as species are represented in blue, and cells with completeness values ≥ 80% according to the *exact* estimator of the accumulation curves are represented in green. (B) Comparison of well-surveyed cells at 60 (grey filled circles) and 5 min (red dots) resolution.

surveyed spatial units. If data weaknesses and shortfalls are known in advance, and the analyses take the associated limitations into account in a balanced approach, the conclusions obtained will increase in robustness and confidence. This kind of conclusions will go much farther in countering skeptics about the extent of the biodiversity crisis and/or the impacts of global change on biodiversity.

Acknowledgements

The development of *KnowBR* has been supported by the projects BIOWEB (CGL2011-15622-E BOS) and BANDENCO (POII11-0277-5747) founded by Spanish Ministerio de Ciencia e Innovación and the Consejería de Educación, Ciencia y Cultura-Junta de Comunidades de Castilla-La Mancha, respectively. DSF was supported by a post-doctoral contract funded by Universidad de Castilla-La Mancha and the European Social Fund (ESF).

References

Aarts, G., et al., 2012. Comparative interpretation of count, presence–absence and point methods for species distribution models. Meth. Ecol. Evol. 3, 177–187.
Anderson, R.P., 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. J. Biogeogr. 30, 591–605.
Ascher, J.S., Pickering, J., 2014. Discover Life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila). Available at http://www.discoverlife.org/mp/20q?guide=Apoidea_species.
Austin, M.P., Heyligers, P.C., 1989. Vegetation survey design for conservation: gradsect sampling of forests in north-eastern New South Wales. Biol. Conserv. 50, 13–32.
Bisby, F.A., 2000. The quiet revolution: biodiversity informatics and the internet. Science 289, 2309–2312.
Chao, A., Jost, L., 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93, 2533–2547.
Clench, H., 1979. How to make regional lists of butterflies: some thoughts. J. Lepid. Soc. 33, 216–231.
Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. Philos. Trans. R. Soc. B 345, 101–118.
Colwell, R.K., et al., 2012. Models and estimations linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J. Plant Ecol. 5, 3–21.
Dennis, R.L.H., Hardy, P.B., 1999. Targeting squares for survey: predicting species richness and incidence for a butterfly atlas. Glob. Ecol. Biogeogr. 8, 443–454.
Dennis, R.L.H., et al., 2006. The effects of visual apparency on bias in butterfly recording and monitoring. Biol. Conserv. 128, 486–492.
Development Core Team R, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
Ferrier, S., 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Syst. Biol. 51, 331–363.
Ferrier, S., et al., 2002a. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. Biodivers. Conserv. 11, 2275–2307.
Ferrier, S., et al., 2002b. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. Biodivers. Conserv. 11, 2309–2338.
Flather, C.H., 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. J. Biogeogr. 23, 155–168.
García-Roselló, E., et al., 2013. ModestR: a software tool for managing and analyzing species distribution map databases. Ecography 36, 102–1207.
García-Roselló, E., et al., 2014. Using ModestR to download, import and clean species distribution records. Meth. Ecol. Evol. 5, 708–713.
Godfray, C., 2002. Challenges for taxonomy. Nature 417, 17–19.
Graham, C.H., et al., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19, 497–503.
Guillera-Arroita, G., et al., 2010. Design of occupancy studies with imperfect detection. Meth. Ecol. Evol. 1, 131–139.
Guisande, C., et al., 2014. RWizard software, http://www.ipez.es/RWizard. University of Vigo. Spain.
Guisande, C., et al., 2015. FactorsR: an RWizard application for identifying the most likely causal factors in controlling species richness. Diversity 7, 385–396.
Guisande, C., 2016a. Niche estimation. R package version. 1:3 Available at < http://CRAN.Rproject.org/package=EnvNicheR > .
Guisande, C., et al., 2016b. SPEDInstabR: an algorithm based on a fluctuation index for selecting predictors in species distribution modeling. Ecol. Inform. 37, 18–23.
Guisande, C., 2016c. An algorithm for morphometric characters selection and statistical validation in morphological taxonomy. R package version. 1:1 Available at < http://CRAN.R-project.org/package=VARSEDIG > .
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Guralnick, R.P., et al., 2007. Towards a collaborative, global infrastructure for biodiversity assessment. Ecol. Lett. 10, 663–672.
Hampton, S.E., et al., 2013. Big data and the future of ecology. Front. Ecol. Environ. 11, 156–162.
Hastie, T., Fithian, W., 2013. Inference from presence-only data; the ongoing controversy. Ecography 36, 864–867.
Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93, 679–688.
Hortal, J., et al., 2008. Historical bias in biodiversity inventories affects the observed realized niche of the species. Oikos 117, 847–858.
Hortal, J., Lobo, J.M., 2005. An ED-based protocol for the optimal sampling of biodiversity. Biodivers. Conserv. 14, 2913–2947.
Hortal, J., Lobo, J.M., 2011. Can species richness patterns be interpolated from a limited number of well-known areas? Mapping diversity using GLM and kriging. Braz. J. Nat. Conserv. 9, 200–207.
Hortal, J., et al., 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). Conserv. Biol. 21, 853–863.
Hortal, J., et al., 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Ann. Rev. Ecol. Evol. Syst. 46, 523–549.
Johnson, D.H., 1980. The comparison of usage and availability measurements for evaluating resource preference. Ecology 61, 65–71.
Kadmon, R., et al., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14, 401–413.
Ladle, R., Hortal, J., 2013. Mapping species distributions: living with uncertainty. Front. Biogeogr. 5, 8–9.
Lobo, J.M., 2008. Database records as a surrogate for sampling effort provide higher species richness estimations. Biodivers. Conserv. 17, 873–881.
Lobo, J.M., Martín-Piera, F., 2002. Searching for a predictive model for Iberian dung beetle species richness based on spatial and environmental variables. Conserv. Biol. 16, 158–173.
Lobo, J.M., et al., 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? Divers. Distrib. 13, 772–780.
Lobo, J.M., et al., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33, 103–114.

Lomolino, M.V., 2004. Conservation biogeography. In: Lomolino, M.V., Heaney, L.R. (Eds.), Frontiers of Biogeography: New Directions in the Geography of Nature. Sinauer Associates, pp. 293–296.
Martínez-Meyer, E., 2005. Climate change and biodiversity: some considerations in forecasting shifts in species potential distributions. Biodivers. Inform. 2, 42–55.
Medina, N., et al., 2013. Designing bryophyte surveys for an optimal coverage of diversity gradients. Biodivers. Conserv. 22, 3121–3139.
Meyer, C., et al., 2015. Global priorities for an effective information basis of biodiversity distributions. Nat. Comm. 6 8221.
Meyer, C., et al., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19, 992–1006.
Mora, C., et al., 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. Proc. R. Soc. B 275, 149–155.
Oksanen, J., et al., 2014. Community Ecology Package. — R package version 2.0-10. Available at: http://CRAN.R-project.org/package = vegan.
Ruete, A., 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. Biodivers. Data J. 3, e5361.
Rocchini, D., et al., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Prog. Phys. Geogr. 35, 211–226.
Saarenmaa, H., Nielsen, E.S., 2002. Towards a global biological information infrastructure. Challenges, opportunities, synergies, and the role of entomology. European Environment Agency, Copenhagen.
Soberón, J., Llorente, B.J., 1993. The use of species accumulation functions for the prediction of species richness. Conserv. Biol. 7, 480–488.
Soberón, J., Nakamura, M., 2009. Niches and distributional areas: concepts, methods, and assumptions. Proc. Natl. Acad. Sci. U.S.A. 106, 19644–19650.
Soberón, J., Peterson, A.T., 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philos. Trans. R. Soc. Lond. 359, 689–698.
Soberón, J.M., et al., 2000. The use of specimen-label databases for conservation purposes: an example using Mexican papilionid and pierid butterflies. Biodivers. Conserv. 9, 1441–1466.
Soberón, J., et al., 2007. Assessing completeness of biodiversity databases at different spatial scales. Ecography 30, 152–160.
Stropp, J., et al., 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. Glob. Ecol. Biogeogr. 25, 1085–1096.
Ugland, K.I., et al., 2003. The species-accumulation curve and estimation of species richness. J. Anim. Ecol. 72, 888–897.
Van Sickle, J., 2010. Basic GIS Coordinates. CRC Press, Boca Raton, Florida.
Whittaker, R.J., et al., 2005. Conservation biogeography: assessment and prospect. Divers. Distrib. 11, 3–23.
Wood, S., 2018. Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version. 1.8-23 Available at < https://cran.r-project.org/web/packages/mgcv/index.html > .
Zaniewski, A.E., et al., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol. Model. 157, 261–280.