Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2015) **24**, 335–347

Global Ecology and Biogeography
A Journal of Macroecology

RESEARCH PAPER

Can we derive macroecological patterns from primary Global Biodiversity Information Facility data?

Emilio García-Roselló[1], Cástor Guisande[2], Ana Manjarrés-Hernández[3], Jacinto González-Dacosta[1], Juergen Heine[1], Patricia Pelayo-Villamil[4], Luis González-Vilas[2], Richard P. Vari[5], Antonio Vaamonde[6], Carlos Granado-Lorencio[7] and Jorge M. Lobo[8]

[1]Departamento de Informática, Universidad de Vigo, Campus Lagoas-Marcosende s/n, 36310 Vigo, Spain, [2]Facultad de Ciencias del Mar, Universidad de Vigo, Campus Lagoas-Marcosende s/n, 36310 Vigo, Spain, [3]Instituto Amazónico de Investigaciones (IMANI), Universidad Nacional de Colombia, Km 2 vía Tarapacá, Leticia, Colombia, [4]Grupo de Ictiología, Universidad de Antioquia, A.A. 1226, Medellín, Colombia, [5]Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, PO Box 37012, MRC 159, Washington, DC, USA, [6]Departamento de Estadística e Investigación Operativa Facultad de CCEE y Empresariales, Universidad de Vigo, Torrecedeira 105, 36208 Vigo, Spain, [7]Departamento de Biología Vegetal y Ecología, Facultad de Biología, Universidad de Sevilla, Avenida de Reina Mercedes s/n, 41012 Sevilla, Spain, [8]Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (CSIC), c/ José Gutiérrez Abascal 2, 28006 Madrid, Spain

*Correspondence: Cástor Guisande, Facultad de Ciencias del Mar, Universidad de Vigo, Campus Lagoas-Marcosende s/n, 36310 Vigo, Spain.
E-mail: castor@uvigo.es

ABSTRACT

Aim To determine whether the method used to build distributional maps from raw data influences the representation of two principal macroecological patterns: the latitudinal gradient in species richness and the latitudinal variation in range sizes (Rapoport's rule).

Location World-wide.

Methods All available distribution data from the Global Biodiversity Information Facility (GBIF) for those fish species that are members of orders of fishes with only marine representatives in each order were extracted and cleaned so as to compare four different procedures: point-to-grid (GBIF maps), range maps applying an α-shape [GBIF-extent of occurrence (EOO) maps], the MaxEnt method of species distribution modelling (GBIF-MaxEnt maps) and the MaxEnt method but restricted to the area delimited by the α-shape (GBIF-MaxEnt-restricted maps).

Results The location of hotspots and the latitudinal gradient in species richness or range sizes are relatively similar in the four procedures. GBIF-EOO maps and most GBIF-MaxEnt-maps provide overestimations of species richness when compared with those present in a priori well-surveyed cells. GBIF-EOO maps seem to provide more reasonable world macroecological patterns. MaxEnt can erroneously predict the presence of species in environmentally similar cells of another hemisphere or in other regions that lie outside the range of the species. Limiting this overpredictive capacity, as in the case of GBIF-MaxEnt-restricted maps, seems to mimic the frequency of observations derived from a simple point-to-grid procedure, with the utility of this procedure consequently being limited.

Main conclusions In studies of macroecological patterns at a global scale, the simple α-shape method seems to be a more parsimonious option for extrapolating species distributions from primary data than are distribution models performed indiscriminately and automatically with MaxEnt. GBIF data may be used in macroecological patterns if original data are cleaned, autocorrelation is corrected and species richness figures do not constitute obvious underestimations. Efforts therefore should focus on improving the number and quality of records that can serve as the source of primary data in macroecological studies.

Keywords
Distribution models, GBIF, macroecological patterns, marine fishes, point-to-grid, range maps, Rapoport' rule.

DOI: 10.1111/geb.12260
http://wileyonlinelibrary.com/journal/geb

INTRODUCTION

Estimations of species distributions from range-wide occurrences using sources such as the geographical distribution records available at the Global Biodiversity Information Facility (GBIF; http://www.gbif.org) are necessary for both basic and applied purposes. Although the non-systematic character of these databases means that they are prone to harbour errors and biases (Yesson *et al.*, 2007; Mesibov, 2013), they nonetheless constitute the most important initiative to date aiming to compile at a global scale the colossal amount of dispersed information about biodiversity. As such, these databases are our main source of massive, comprehensive and freely available data on the distribution of species world-wide.

Two main approaches have been devised to overcome the shortcomings of these primary data: (1) generating polygon range maps by joining line segments connecting each pair of points in order to obtain continuous areas, and (2) developing distribution models using environmental predictors. The first purely geographical procedure is based on the assumption that areas of distribution should be composed of sets of connected populations. The second method assumes that environmental variables are the main determinants of the distribution of species and that using these variables we may predict the range of the species when actual ranges are unknown. Some studies have attempted to compare the effectiveness of these two procedures (Graham & Hijmans, 2006; Amboni & Laffan, 2012; Pineda & Lobo, 2012; Vasconcelos *et al.*, 2012). Conclusions as to the comparative effectiveness of the alternative procedures are hindered by the absence of independent reliable data capable of evaluating the obtained patterns (but see Pineda & Lobo, 2012). These studies show that distribution models (e.g. MaxEnt; see below) frequently generate higher overpredictions of species richness, a result that also appears when richness gradients are generated by summing up the range maps in a grid system (Hurlbert & Jetz, 2007; Jetz *et al.*, 2008; Bombi *et al.*, 2011; Cantú-Salazar & Gaston, 2013). Furthermore, the cited studies also demonstrate that both methods may provide similar species richness representations when the size of the grid cells is greater (i.e. low levels of resolution; Graham & Hijmans, 2006; Amboni & Laffan, 2012; Pineda & Lobo, 2012).

Primary data, range maps and distribution models may yield either relatively divergent or similar macroecological patterns. In the case of divergence, we speculate that these methods of generating distributions provide different representations, one of which best approximates reality. In this situation, the observed geographical patterns would depend predominantly on the geographical (range maps) or ecological (distribution models) rules by which the estimates of species distribution were constructed. Alternatively, if similar macroecological patterns can be obtained via different procedures for building species distributions then there would be no obvious advantage in using a particular procedure in so far as distribution patterns would be observed independent of the method used to represent them.

We herein focus on two of the most universal macroecological patterns: the latitudinal gradient in species richness (Pianka, 1966) and the latitudinal variation in range sizes, also known as Rapoport's rule (Rapoport, 1982; Stevens, 1989).

We estimate the magnitude and covariation of these two macroecological patterns in the case of species of marine fishes world-wide using data derived from four different methods of mapping species distributions: (1) point observations of species occurrences downloaded from the GBIF (GBIF maps); (2) extent of occurrence (EOO) maps obtained by applying the α-shape method to these species occurrences (GBIF-EOO maps); (3) maps obtained through the widely used MaxEnt method of species distribution modelling (GBIF-MaxEnt maps); and (4) maps derived from applying MaxEnt to the area delimited by the α-shape (GBIF-MaxEnt-restricted maps).Our aim was to determine whether the method used to build distributional maps from raw data influences the representation of the two selected macroecological patterns – latitudinal gradients in species richness and Rapoport's rule.

MATERIAL AND METHODS

Species

Analyses were based on all species of marine fishes currently recognized as valid (see Eschmeyer, 2013) that meet two criteria: (1) being members of those orders of fishes with only marine representatives; and (2) being available in IPez (http://www.ipez.es/; Guisande *et al.*, 2010; see Appendix S1 in Supporting Information). A total of 1835 species distributed across 23 orders met both criteria (see Appendix S1).

GBIF maps

Using the facilities available in the free ModestR application (see http://www.ipez.es/ModestR/; García-Roselló *et al.*, 2013), we imported all available distributional data of the selected species from the GBIF portal (http://www.gbif.org/; accessed May 2013). ModestR allows this importation for all the species at the same time by including a file with the species names following a simple taxonomic classification (Pelayo-Villamil *et al.*, 2012; García-Roselló *et al.*, 2013). We additionally used the cleaning facilities of ModestR (García-Roselló *et al.*, 2014) to minimize the errors frequently appearing in such massive databases (Yesson *et al.*, 2007; Mesibov, 2013). Using these cleaning facilities, we removed: (1) duplicates (multiple records of a species from a particular location); (2) those records with geographical coordinates of 0° longitude and 0° latitude which are likely to reflect default entries in cases of lack of data; (3) records in which the values of longitude and latitude are identical and probably represent erroneous repetitive data entry; (4) erroneous synonyms and obviously misidentified species in the light of known distributional ranges; and (5) those locations that fall outside the habitat of the species (marine waters) and are thus invalid. These import and cleaning procedures allowed us to obtain valid distributional information for 1678 of the 1835 species that are members of those orders with only marine representatives and which have data in GBIF (see Appendix S1). This yielded a total of 372,337 records, or 226.2 ± 918.5 records per species (mean ± SD). These primary data were used to build richness maps simply by overlaying individual species point occurrence data via a point-to-grid procedure (1° cell resolution).

GBIF-EOO maps

The EOO is defined as the area contained within the shortest continuous imaginary boundary that can be drawn to encompass all the known, inferred or projected sites representing the occurrence of a taxon (IUCN, 2013). We also estimate EOO for each species with ModestR, selecting an α-shape procedure. Both convex hulls and α-shapes can be used to generate range maps from a finite set of observed occurrences (Pateiro-Lopez & Rodriguez-Casal, 2011; CGAL, 2013), but the α-shape minimizes EOO overestimations by incorporating discontinuities in species distributions (Burgman & Fox, 2003). α-Shape is a generalization of the convex hull concept (Edelsbrunner *et al.*, 1983) that uses a single parameter (α) to construct a geometric shape from a set of points. When α approaches zero, the generated shape is near to the original point set, whereas when α increases we are able to obtain a range map similar to the typical convex hull. In our study we examine different α values to determine which is most appropriate.

GBIF-MaxEnt maps

We modelled the distribution of all species using the recommended (Elith *et al.*, 2006) and widely used MaxEnt application with default options (version 3.3.3; see Phillips *et al.*, 2006) in which 10,000 background absences are randomly selected across the entire world. To achieve this goal, the occurrence records obtained from GBIF were related to the Bio-ORACLE global dataset of environmental factors consisting of 23 geophysical, biotic and climate raster variables (Tyberghein *et al.*, 2012; see http://www.oracle.ugent.be).The suitability values thus obtained were imported in ModestR with continuous values transformed into binary values by applying a threshold to subsequently overlay all the individual models in order to obtain a species richness representation (see below).

GBIF-MaxEnt-restricted maps

To minimize commission errors in MaxEnt predictions for those environmentally similar areas far from the observed species range of species, we implemented a new facility in ModestR 2.0 which constrains the MaxEnt-derived maps to the EOO area of each species estimated by means of an α-shape procedure with $\alpha = 6$ (see 'Performance comparisons' under Results).Thus, background absences are not randomly selected across the entire world but rather within the area delimited by the α-shape procedure. For a detailed description of the method and further details see the step-by-step tutorial 3 'How to integrate and use environmental data in ModestR' or the ModestR tutorial. Both pdf files are available from the documentation section of the web site http://www.ipez.es/ModestR.

Exploring species richness

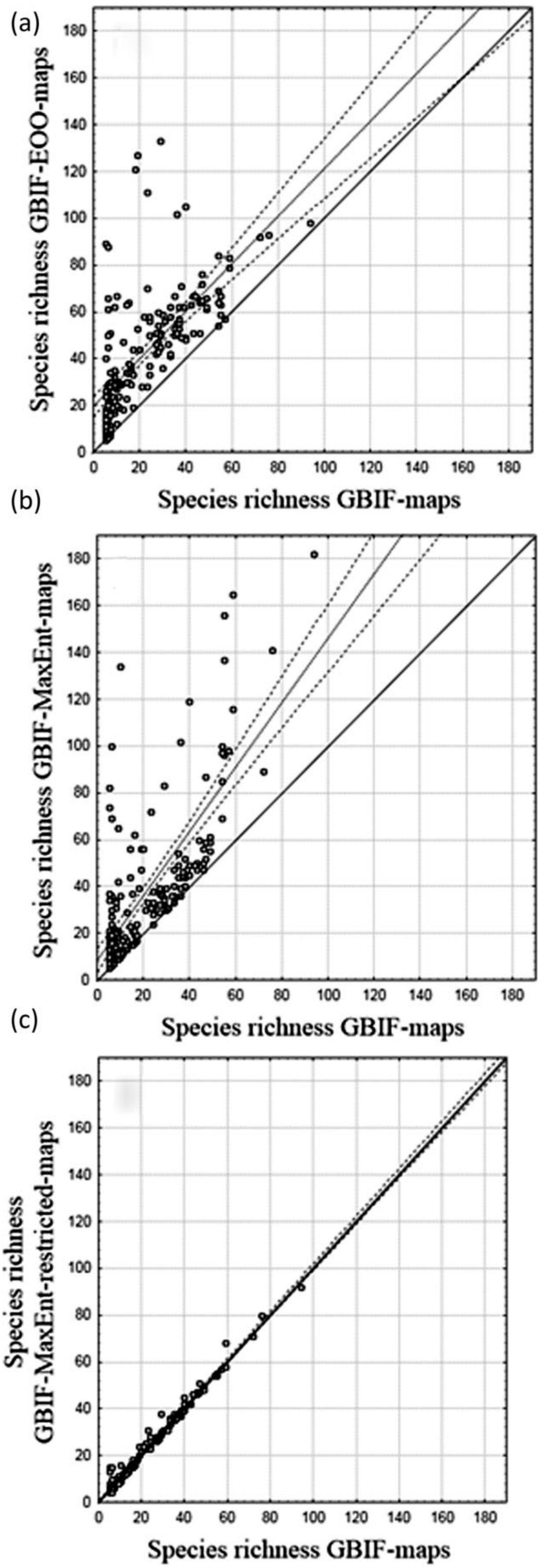
Several α-values (from 3 to 8) were considered when building GBIF-EOO maps, as were different thresholds at intervals of 0.05 to transform continuous MaxEnt suitability values into binary presence–absence values. In the case of MaxEnt, we also used the lowest predicted values associated with an observed presence as a threshold (Pearson *et al.*, 2004). Each one of the thus generated species richness maps resulting from the overlay of individual estimations (at 1° cell resolution) was related with the species richness values derived from nonparametric and species accumulation extrapolations. To do this, we used all the downloaded records ($n = 371{,}903$) to estimate the total number of species in cells of 1° size using the two different procedures available in the vegan R package (Oksanen *et al.*, 2013): (1) the first-order jackknife estimator, and (2) accumulation curves. Only those 1° cells with twice as many records as species ($n = 6338$) were considered in these estimations. The first-order jackknife estimator is a recommended nonparametric, incidence-based extrapolation method (Hortal *et al.*, 2006). It uses the frequencies of species in each cell to generate an extrapolated species richness value(S_p) taking into account the observed number of species (S_0) and the number of species occurring at only one site within the cell (a_1):

$$S_p = S_0 + a_1 \frac{N-1}{N}$$

where N is the number of sites in the cell. The accumulation-based estimation is based on the classical Clench function (Clench, 1979) for predicting an 'upper-limit' asymptote (the total number of species when the survey effort is infinite). Different methods can be used to sort the available records in the accumulation curve. In our study, we selected the 'random' procedure with 100 permutations without replacement, and also the 'exact' and 'Coleman' procedures (Oksanen *et al.*, 2013). The extrapolated species richness values for each 1° cell are the mean values obtained under these three accumulation methods. Although the use of these extrapolation methods can be prone to underestimate the 'real' species richness, we assumed herein that the provided values constitute the best available approach to identify grid cells with relatively reliable inventories and actual species richness values.

The completeness values generated by the first-order jackknife estimator and accumulation curves (the percentage of observed species against predicted ones) are used to discriminate the most probable well-surveyed 1° cells (WSCs) in which the GBIF data may provide relatively accurate inventories. To do this, we selected all cells with more than five species and completeness values of 95% or more in either of the two extrapolation methods ($n = 211$ grid cells). Acting on the assumption that the observed species richness values in these cells are reliable, we used them to select the best correlated α and threshold values for GBIF-EOO maps and GBIF-MaxEnt maps, respectively. Species richness values in WSCs were also used to compare the relative performance of the four procedures selected to estimate species range maps. To achieve this,the species richness values coming

Figure 1 Relationships between species richness values of the most probable well-surveyed 1° cells obtained by overlaying Global Biodiversity Information Facility (GBIF) individual species point occurrence data (GBIF maps) and those generated by (a) MaxEnt predictions (GBIF-MaxEnt maps; threshold = 0.75), (b) α-shape convex hulls [GBIF-EOO (extent of occurrence) maps; α-value = 6] and (c) GBIF-MaxEnt-restricted maps (α-shape = 6, threshold = 0.75). Linear regression lines (dots), 95% confidence bands (dashes) and equality lines (continuous lines) are shown.



from GBIF maps, GBIF-EOO maps, GBIF-MaxEnt maps and GBIF-MaxEnt-restricted maps were correlated (Pearson correlation coefficient *r*) against the values obtained by the two extrapolation methods (first-order jackknife estimator and accumulation curves). Those methods producing a higher correlation were considered to demonstrate a better performance.

Finally, we use the species richness values in WSCs to examine the differences in the distances among the centroids of the species when calculated from the distributions obtained with the four methods. The centroid was estimated as the average longitudinal and latitudinal position of all the 1° cells in which each species is present. The distance between the four estimated centroids was calculated using the haversine formula (Sinnott, 1984) due to the capacity of that formula to deal with points on a sphere.

Exploring Rapoport's rule

Area of occupancy (AOO) is defined as the area occupied by a taxon within its extent of occurrence (IUCN, 2013). We estimated AOO for each species with ModestR using rasterized maps with a resolution of $1' \times 1'$ cells, in which the occupied area was estimated using the following equation:

$$1.852 \times \frac{12{,}756.2\pi}{21{,}600}\cos\left(\text{latitude} \times \frac{\pi}{180}\right).$$

The value 1.852 is a nautical mile expressed in km, 12,756.2 is twice the radius of the earth in km and finally 21,600 in the value used to transform $1'$. The latitude in the equation is the central value of the individual pixels considered. The equation was applied to all the pixels occupied by the species and all the obtained values were summed to calculate the AOO.We also calculate the latitudinal range of each species (LR) as the difference between the maximum and minimum latitudes.

Steven's method (Stevens, 1989), Pagel's method (Pagel *et al.*, 1991), the mid-point method (Rohde, 1992) and the cross-species method (Letcher & Harvey, 1994) are among the procedures frequently employed to represent the latitudinal gradient in range size as well as to evaluate Rapoport's rule. Each of these methods has it advantages and disadvantages (Ruggiero & Werenkraut, 2007). In this study, we used the commonly applied Steven's method (Stevens, 1989) which shows the latitudinal variation in species range sizes for all species recorded within each 5° latitudinal band.

Using AOO and LR estimations from the distributional representations provided by the three considered procedures, we examined the latitudinal variation in their averaged values. We also

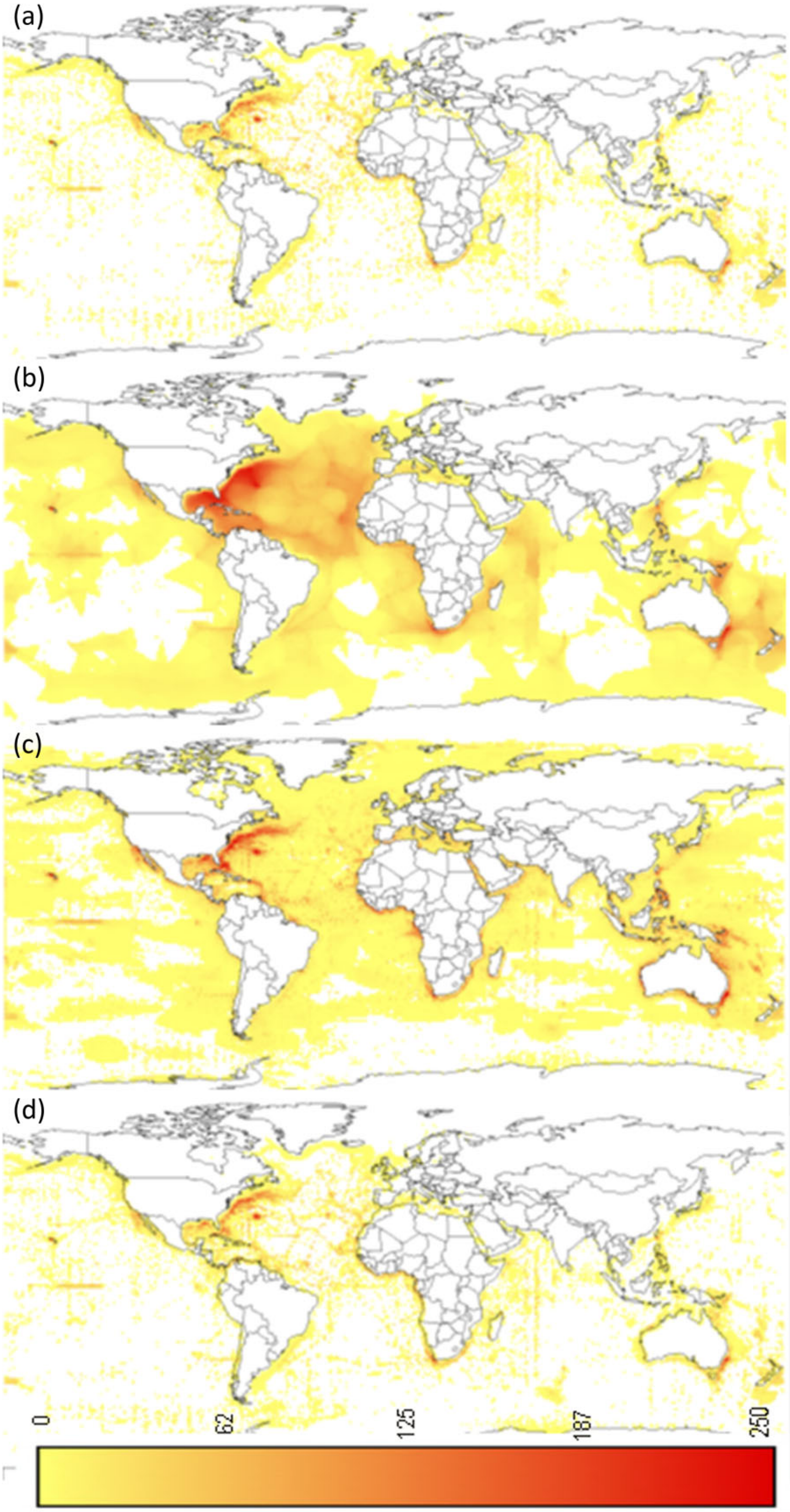


Figure 2 World variation in species richness of marine fish species according to (a) Global Biodiversity Information Facility (GBIF) maps, (b) GBIF-EOO (extent of occurrence) maps (α-value = 6), (c) GBIF-MaxEnt maps (fixed threshold = 0.75), and (d) GBIF-MaxEnt-restricted maps (α-shape = 6, threshold = 0.75) at 1° resolution.

examined the relationships of the AOO and LR estimations with the species richness values provided by these same procedures.

Statistical analyses

All statistical analyses were carried out with R (R Development Core Team, 2013). The Kolmogorov–Smirnov test with the Lilliefors correction was used to test for normality of residuals and was performed with the package *nortest* (Gross, 2013). The Breusch–Pagan test was used to test for homoscedasticity of residuals and performed with the package *lmtest* (Zeileis & Hothorn, 2002; Hothorn *et al*., 2013). Standardized residuals of linear regression models did not have a normal distribution (Kolmogorov–Smirnov, $P < 0.001$) and homoscedasticity was not present in the residuals (Breusch–Pagan test, $P < 0.001$), even when using transformations. Consequently we use a nonparametric statistic, the Theil–Sen estimator, to avoid potential problems with normality, homogeneity of variances and/or homoscedasticity. The Theil–Sen estimator, which is considered to be a robust linear regression (Hollander *et al*., 2013), was calculated using the package *rtk* (Marchetto, 2013).

Residuals from a regression must be independent and Steven's method is susceptible to problems of autocorrelation (Rohde, 1992). Although autocorrelation is not considered a problem in nonparametric methods (Hollander *et al*., 2013), autocorrelation might have the same effect on Theil–Sen regression slopes as it has on the least-squares regression slopes, i.e. the estimator variance increases thereby making it less accurate. Both species richness and AOO and LR values can be influenced by spatial autocorrelation since their values by latitudinal band are related to those of neighbouring bands. Using *MRFinder* – one of the applications of ModestR (García-Roselló *et al*., 2013) – we thus first obtained the list of species in latitudinal bands of 5°. Subsequently, we applied the Theil–Sen estimator to the thus obtained relationships between species richness and AOO or LR. In order to avoid any potential problems associated with autocorrelation, we also applied the Theil–Sen estimator to a subset of the original dataset in which no duplicate species appear among latitudinal bands, thereby minimizing the occurrence of spatial autocorrelation. When a species was present in more than one latitudinal band, we randomly selected one of the bands, because autocorrelation is a consequence of the presence of the same species in several bands. We used the Durbin–Watson statistic (Durbin & Watson, 1951) to detect the presence of autocorrelation.

RESULTS

Performance comparison

Species richness values in the WSCs derived from the first-order jackknife estimator and accumulation curves are inevitably highly correlated with the values of GBIF maps ($r = 0.991$ and $r = 0.923$, respectively; $P < 0.001$). In the case of the GBIF-MaxEnt maps, the best relationship is obtained at a fixed probability threshold of 0.75 ($r = 0.715$ for the first-order jackknife estimator and $r = 0.796$ for the accumulation curve values; $P < 0.001$). Applying the lowest predicted values associated with an observed presence as a threshold generates significantly lower correlation values ($r = 0.51$ and $r = 0.56$ in the case of GBIF-MaxEnt maps, and $r = 0.69$ and $r = 0.65$ in the case of GBIF-MaxEnt-restricted maps; $P < 0.001$) and clearly overpredicted maps (see Fig. S1). In the case of GBIF-EOO maps, the best correlations are obtained at an α-value of 6 ($r = 0.670$ for the first-order jackknife estimator and $r = 0.640$ for the accumulation curve values; $P < 0.001$). Hence, we subsequently used an α-value of 6 and a fixed threshold probability of 0.75. Using these criteria in GBIF-MaxEnt-restricted maps, we obtained the best correlations with the species richness derived from such maps for the first-order jackknife estimator ($r = 0.984$; $P < 0.001$) and accumulation curve values ($r = 0.918$; $P < 0.001$; see Fig. 1). Considering that the species richness values coming from GBIF maps in the WSCs can be considered reliable, a very high correlation also exists between these values and those provided by GBIF-MaxEnt-restricted maps ($r = 0.994$, $P < 0.001$).

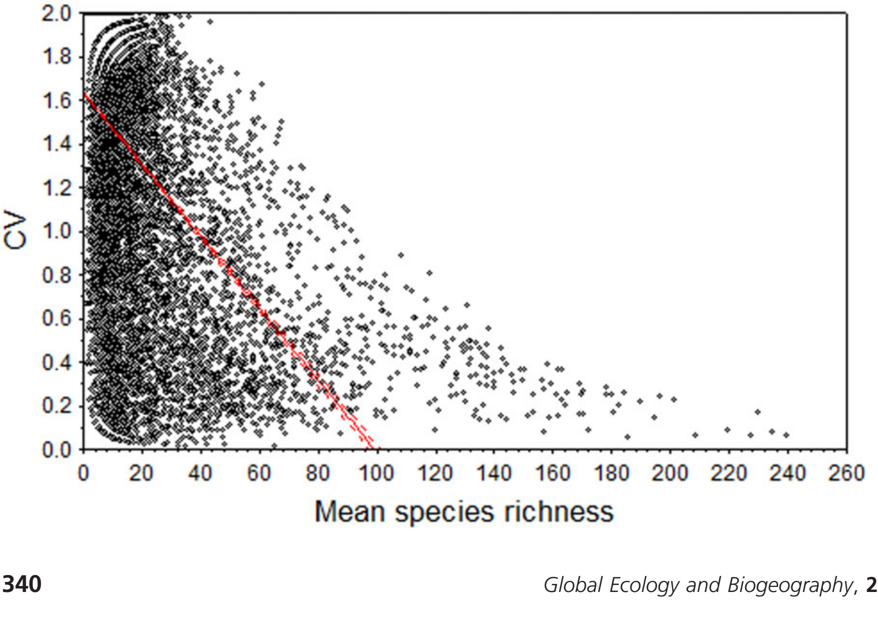


Figure 3 Relationship between coefficient of variation (CV) and mean species richness derived from the four procedures used shown in Fig. 2. The red line indicates linear regression ± 95% confidence interval (broken lines).

However, correlations are alternatively moderate between the species richness values provided by GBIF maps and GBIF-EOO maps ($r = 0.685$, $P < 0.001$) or GBIF maps and GBIF-MaxEnt maps ($r = 0.746$; $P < 0.001$). Both the GBIF-EOO maps and GBIF-MaxEnt maps seem to overpredict species richness values (Fig. 1). GBIF-MaxEnt maps seem to especially overpredict species richness in the richest cells, whereas GBIF-EOO maps overpredict species richness across the entire range of observed values (i.e. equally in cells of high versus low species richness).

Species richness

The obtained world-wide variations in species richness values under the four methods (GBIF maps, GBIF-EOO maps, GBIF-MaxEnt maps and GBIF-MaxEnt-restricted maps) show relatively similar but differentially contrasted geographical patterns (Fig. 2). The negative relationship between the coefficient of variation of these four estimations and the averaged species richness per cell ($r = -0.45$, $P < 0.001$) suggests, however, that the higher discrepancy in the estimations of the four methods occurs in the least species rich cells and that 'hotspots' are similarly depicted independent of the procedure (Fig. 3). The correlation values (r) between the world variation in species richness obtained with GBIF maps, GBIF-EOO maps and GBIF-MaxEnt maps are all statistically significant ($P < 0.001$; $n = 33{,}125$), oscillating from 0.633 in the case GBIF maps versus GBIF-EOO maps to 0.772 for GBIF-maps versus GBIF-MaxEnt maps. GBIF maps demonstrate a less contrasting pattern than GBIF-MaxEnt maps or GBIF-EOO maps, having a lower mean species richness per cell (mean ± 95% CI 2.05 ± 0.08) than in the case with these two other procedures (6.99 ± 0.17 and 11.55 ± 0.23, respectively). However, GBIF-MaxEnt-restricted maps not only provide mean species richness values similar to those of GBIF maps (2.24 ± 0.09) but also a very similar geographical pattern (Fig. 2). The correlation between the species richness values derived from both procedures is surprisingly high ($r = 0.988$). Thus, while GBIF-EOO maps and GBIF-MaxEnt maps can expand the distribution area of the species, thereby increasing the observed richness of the cells (Fig. 4a,b), GBIF-MaxEnt-restricted maps provide species richness values that follow observed values (Fig. 4c).

Latitudinal variation in the number of marine species increased overall towards the equator, with a relatively similar pattern apparent in the data derived from all four procedures (Fig. 5a). Species richness seems to be higher at latitudes between 25° and 35°, both north and south. However, the number of species per latitudinal band is higher in GBIF-MaxEnt maps. This pattern is most pronounced in the richest latitudinal bands but is also apparent at higher latitudes in the least species-rich bands (Fig. 5a).

Finally, the pairwise comparisons of the distances from the centroid of species distributions among the three methods clearly indicate that GBIF maps, GBIF-EOO maps and GBIF-MaxEnt-restricted maps all generate species distributions with very close centroids (Fig. 5b).

Rapoport's rule

Both AOO and LR decrease toward the equator under all four mapping procedures (Fig. 6), with this pattern being more

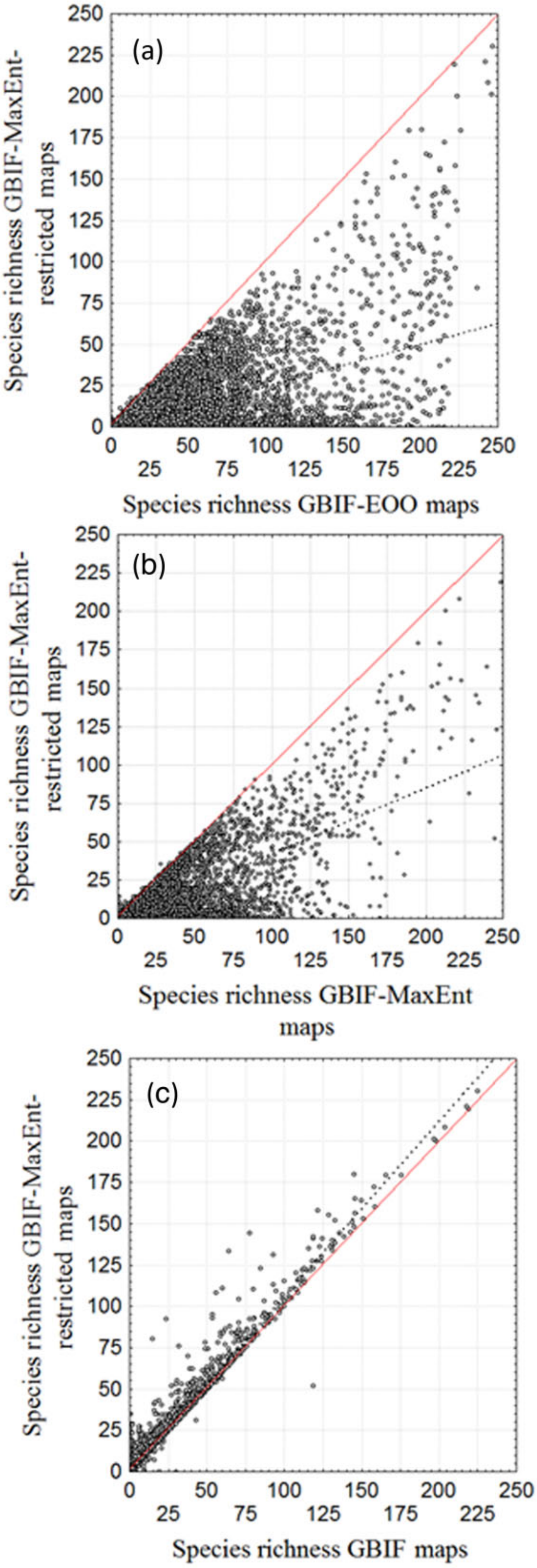


Figure 4 Relationships between world species richness of marine fish species at 1° cells derived from the Global Biodiversity Information Facility (GBIF)-MaxEnt-restricted procedure and those generated by (a) GBIF-EOO (extent of occurrence) maps, (b) GBIF-MaxEnt maps, and (c) GBIF maps.

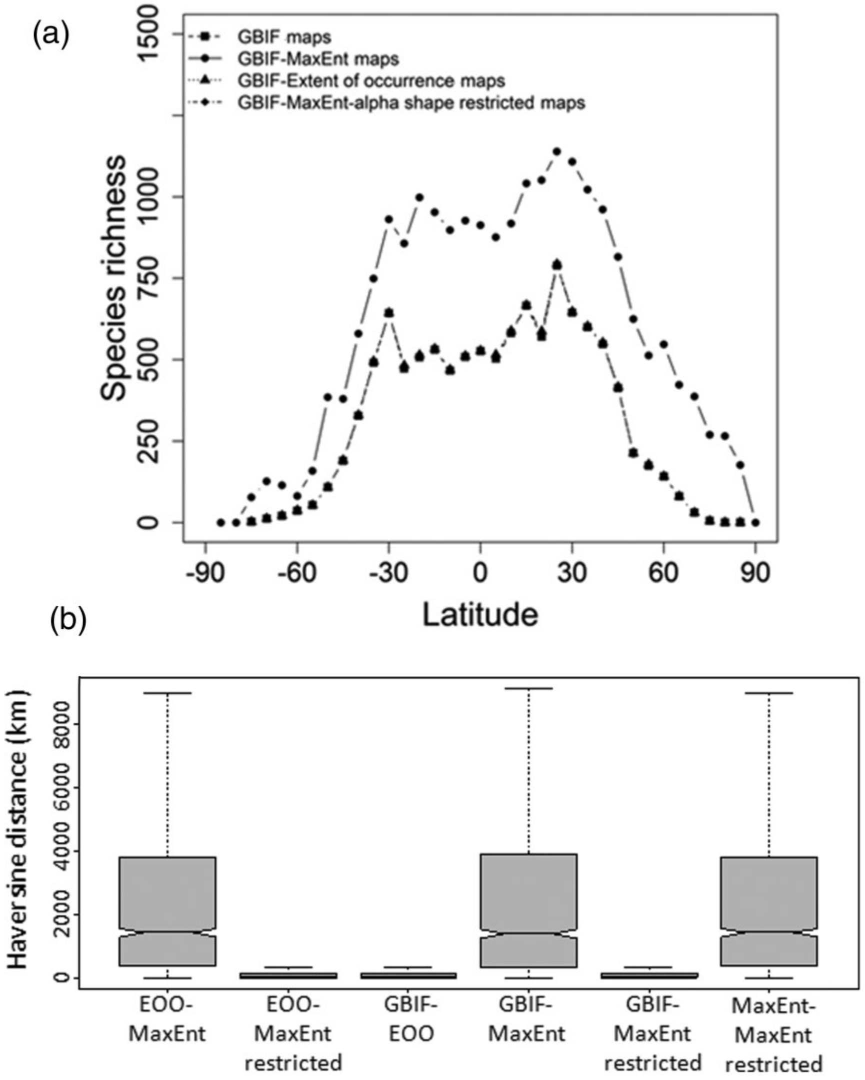


Figure 5 (a) Latitudinal gradient of species richness in latitudinal bands of 5°×5° using Global Biodiversity Information Facility (GBIF) maps, GBIF-EOO (extent of occurrence) maps (α-value = 6), GBIF-MaxEnt maps (fixed threshold probability = 0.75) and GBIF-MaxEnt-restricted-maps (α-value = 6, threshold = 0.75). (b) Median, lower-upper quartiles, and maximum-minimum haversine distances between centroid of species distributions compared pairwise among four considered mapping procedures: GBIF-EOO maps versus GBIF-MaxEnt maps (EOO-Maxent), GBIF-maps versus GBIF-EOO maps (GBIF-EOO), GBIF maps versus GBIF-MaxEnt maps (GBIF-MaxEnt), GBIF-EOO maps versus GBIF-MaxEnt-restricted maps (EOO-MaxEnt restricted), GBIF maps versus GBIF-MaxEnt-restricted maps (GBIF-MaxEnt restricted) and GBIF-MaxEnt maps versus GBIF-MaxEnt-restricted maps (MaxEnt-MaxEnt restricted). Multiple comparisons of mean ranks provide statistically significant differences ($P < 0.001$).

apparent in the case of GBIF-EOO maps and even more so in the case of GBIF-MaxEnt maps which notably increase the mean latitudinal range size and the mean area of occupancy of the species. This is particularly the case in the Southern Hemisphere (see the geographical representation of the changes in the mean latitudinal range size of the species in Fig. S2). In the case of AOO, GBIF maps do consequently not yield a symmetrical pattern in the Northern and Southern Hemispheres. GBIF-EOO maps and GBIF-MaxEnt-restricted maps, but particularly GBIF-MaxEnt maps, greatly amplify the occurrence of more widely distributed species below 40° latitude in the Southern Hemisphere.

The Theil–Sen slope estimators always demonstrated a statistically significant negative relationship between species richness and the two utilized measures of latitudinal variation in range sizes, thereby corroborating Rapoport's rule (Table 1a). However, the Durbin–Watson statistic indicates that the relationships derived from GBIF maps are highly and positively autocorrelated. When this autocorrelation is minimized to avoid the use of duplicate species among latitudinal bands (Table 1b) the Durbin–Watson statistic ranged between 1.3 and 2 in all regressions, which indicates a range of moderate to no autocorrelation. In this case, GBIF maps, GBIF-EOO maps and GBIF-MaxEnt-restricted maps also provide a statistically significant negative relationship between species richness and AOO, but this is not the case for GBIF-MaxEnt maps. However, the relationship between species richness and latitudinal range was not statistically significant in any of the four methods for mapping species ranges (Table 1b).

DISCUSSION

Are there differences among marine fishes across the world in the macroecological patterns derived from primary data, range maps and both types of MaxEnt distribution models used in this study? Notwithstanding differences in species richness, the GBIF maps, GBIF-EOO maps and GBIF-MaxEnt-restricted maps do not seem to provide excessively discordant global representations, but rather show similar latitudinal variation in the case of both species richness patterns and range sizes. Other than when we correct for autocorrelation, the distribution information obtained from GBIF-MaxEnt maps shows similar patterns.

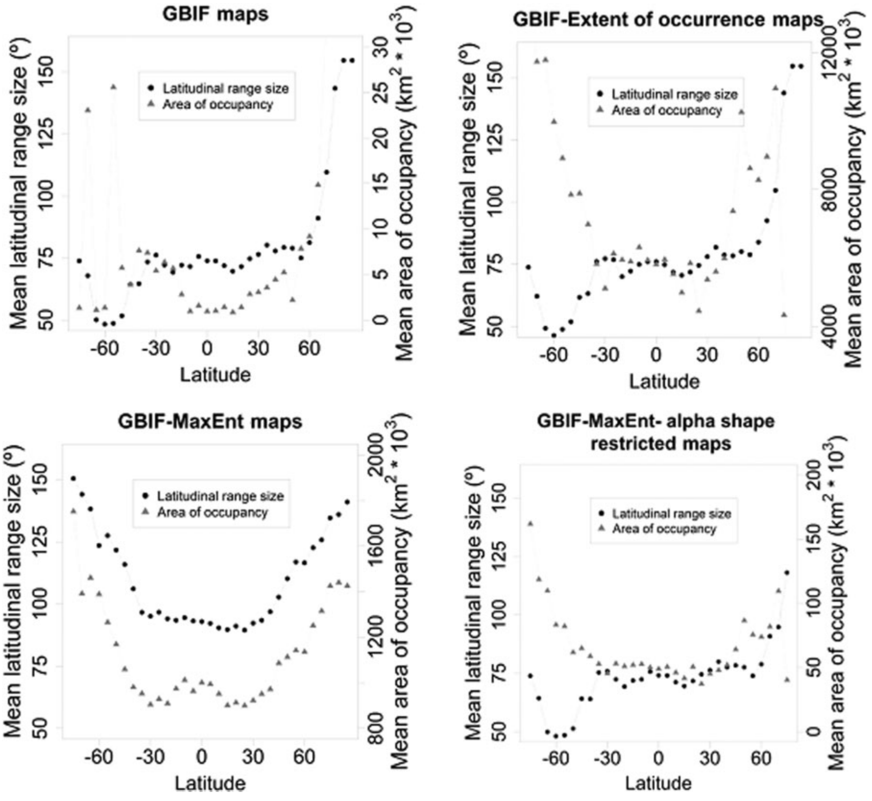


Figure 6 Mean latitudinal area of occupancy (in 10^3 km^2) and mean latitudinal range size (in degrees) of species in latitudinal bands of 5°×5° using Global Biodiversity Information Facility (GBIF) maps, GBIF-EOO (extent of occurrence) maps, GBIF-MaxEnt maps and GBIF-MaxEnt-restricted maps. Higher latitudes (90° to 80° both north and south) were not included in analysis due to small numbers of species in those regions.

Table 1 Results using the Theil–Sen slope estimator for relationships between species richness (*S*) and average area of occupancy (AOO) or latitudinal range (LR) of species in 5° latitudinal bands when calculated by three considered procedures for estimating species ranges: Global Biodiversity Information Facility (GBIF) maps, GBIF-EOO (extent of occurrence) maps (α-value = 6), GBIF-MaxEnt maps (fixed threshold probability = 0.75) and GBIF-MaxEnt-[α-shape]-restricted maps. Both all original data (a) and a subset of the original dataset (b) in which no duplicate species appear among latitudinal bands used.

	Relationship	Method	*P*-value	Slope	DW
(a)		GBIF-maps	0.002	−14.26	0.02
	S/AOO	GBIF-EOO maps	<0.001	−6740	1.76
		GBIF-MaxEnt maps	<0.001	−409	1.77
		GBIF-MaxEnt-restricted maps	<0.001	−25.6	0.25
		GBIF maps	<0.001	−0.03	0.15
	S/LR	GBIF-EOO maps	<0.001	−0.03	1.61
		GBIF-MaxEnt maps	<0.001	−0.05	1.71
		GBIF-MaxEnt-restricted maps	<0.001	−0.03	0.37
(b)		GBIF maps	0.012	−3.42	1.98
	S/AOO	GBIF-EOO maps	0.001	−7007	1.87
		GBIF-MaxEnt maps	0.272	−289	1.60
		GBIF-MaxEnt-restricted maps	0.001	−6956	1.86
		GBIF maps	0.252	−0.07	1.31
	S/LR	GBIF-EOO maps	0.252	−0.07	1.31
		GBIF-MaxEnt maps	0.498	−0.05	1.37
		GBIF-MaxEnt-restricted maps	0.267	−0.07	1.32

DW, Durbin–Watson statistic to measure autocorrelation (see Methods).

Our results indicate that the inclusion of species in localities from which they had not been recorded by the use of range maps or predicted by distribution models generally entails an increase in species richness values for species-poor cells. Extrapolations of individual species ranges, alternatively, do not appear to affect the geographical position of 'hotspots', patterns of global species richness and range size, or the existence of a negative relationship between the average range size of species and species richness. Inevitably, both MaxEnt distribution models and range maps generate higher richness values because these procedures will on occasion place species at localities where the species was not recorded in the GBIF records. Extrapolations based on species ranges seem to generate higher species richness values in our analyses (almost 65% higher) than do those produced by MaxEnt distribution models. Thus, including the species in spatially close, but environmentally unsuitable, cells generates on average higher species richness values than does including the species in environmentally similar localities independent of their proximity. Despite this fact, it is interesting that the number of species by latitudinal band is higher when MaxEnt distribution models are used. This apparently contradictory result is, in our opinion, a consequence of the overpredictive tendency of MaxEnt distribution models (Graham & Hijmans, 2006; Amboni & Laffan, 2012; Pineda & Lobo, 2012; Vasconcelos *et al.*, 2012). Species richness tends to inevitably increase by latitudinal band because the species occurring in a hemisphere or region can be predicted to be present in environmentally similar cells of the other hemisphere or other regions (see maps in Appendix S2). This effect is exemplified by the dramatic increase in the distance between the distribution centroids of species and also in the increase of the mean latitudinal range size and the mean area of occupancy of species when MaxEnt predictions are considered. As the marine environs in the Southern Hemisphere are much more expansive than those in the Northern Hemisphere, we observed that the highest levels of extrapolations appear in the southern zone. The lack of a significant negative relationship between species richness and AOO values derived from MaxEnt distribution models after minimizing the spatial autocorrelation also supports the supposition of the overpredictive tendency of MaxEnt. Even when a species is present in more than one latitudinal band and we randomly selected one of those bands, those methods, nonetheless, sometimes included a species in distant, but climatically suitable, regions. This inevitably tended to include the species in 'atypical' bands; thus, blurring the pattern.

In our study we used MaxEnt default options, but we are aware that alternative settings selected according to biologically based decisions can often be more appropriate (Merow *et al.*, 2013). This implies a need to apply models individually for each species, thereby specifically tuning the model settings (Anderson & Gonzalez, 2011) as well as selecting the most appropriate extent to which the model is calibrated (Giovanelli *et al.*, 2010). Such recommendations are corroborated by the results of our analysis, but run contrary to the standard practice of the automatic indiscriminate use of MaxEnt. Following these recommendations could reduce the overestimations of our MaxEnt predictions, thus yielding an even greater similarity in the macroecological patterns provided by the models and range maps. This question has been examined herein by implementing MaxEnt models individually for each species via a protocol directed towards avoiding overpredictions in distant but environmentally similar localities. In the so-called GBIF-MaxEnt-restricted maps, the models are constrained to the area limited by the α-shape range maps so that extrapolations beyond observations are restricted and extrapolations within the distribution area are limited. Interestingly, the results provided by this approach mimic the GBIF observations herein (Fig. 4c), with the correlation between observed species richness and the values derived from these restricted MaxEnt models being 0.99 in the case of both well-surveyed cells and global estimations. This noteworthy result is one of the most important caveats for the procedures used to estimate the distribution of species from presence observations. Recent studies demonstrate that use-availability designs in which presence observations are modelled against a set of points reflecting the general environmental conditions of the territory under consideration (background absences) reflect only the density of the incorporated observations (Aarts *et al.*, 2012) rather than the species occurrence probability (Hastie & Fithian, 2013). Thus, GBIF-MaxEnt-restricted maps tend to mirror the frequency of the presence observations used in the modelling process.

Distribution estimates of species based on GBIF primary data are probably biased (Yesson *et al.*, 2007; Mesibov, 2013; Guisande *et al.* 2013) and cannot provide reliable predictions of the extent of ranges (Beck *et al.*, 2013). Our general lack of a 'gold standard' to evaluate this bias hinders an estimation of the accuracy provided by range maps and distribution models carried out using raw data. If the macroecological patterns generated by using these four procedures for building species ranges are basically coincident, such coincidence may be due to the simple nature of these macroecological patterns which can be detected independently of the method used to obtain species distributions. Alternatively, this coincidence can also be due to the fact that the errors and biases in the raw data are propagated across all of these extrapolations in so far as these models and the resultant range maps are only capable of describing the patterns present in the raw data. If presence–background absence models attempt to represent the density of observations, it is inevitable that the obtained macroecological patterns derived from the models will not differ from those generated by simple procedures using raw data. Taking into account the important but neglected conceptual implications of the studies mentioned above as well as our results, we thus question the usefulness of using sophisticated modelling methods. This is especially the case when the degree of sampling bias in the available presence information is unknown and the lack of reliable inventories consequently prevents the use of 'true' absence information that is a prerequisite to estimating occurrence probabilities.

Are these extrapolations credible and reasonable? What is the advantage of using sophisticated methods when simple ones may provide similar or even better results? Our discrimination of the probable well-surveyed cells permits us to propose that both range maps and distribution models overpredict 'real' distributions – a conclusion arrived at in previous analyses (Graham & Hijmans, 2006; Jetz *et al.*, 2008; Bombi *et al.*, 2011; Amboni & Laffan, 2012; Pineda & Lobo, 2012; Vasconcelos *et al.*, 2012; Cantú-Salazar & Gaston, 2013). We also propose that the most mathematically sophisticated techniques do not seem to provide better estimates than do maps using simple point-to-grid methods. Simple point-to-grid procedures using primary data may provide underestimates, primarily at higher resolutions (Hurlbert & White, 2005; Graham & Hijmans, 2006; Hurlbert & Jetz, 2007; Hawkins *et al.*, 2008), but may be useful for studying macroecological patterns once the original data are cleaned, autocorrelation is corrected, and species richness is not notably underestimated. Therefore, efforts should be focused on improving the number and quality of distribution records for use as primary data in macroecological studies. Range maps extrapolated with an α-shape method, and to a lesser extent MaxEnt maps, may be an alternative when primary data are significantly underestimated, but efforts should be focused on correcting overpredictions and obtaining more representative distributional data.

ACKNOWLEDGEMENTS

We thank ENDESA for technical and financial support.

REFERENCES

Aarts, G., Fieberg, J. & Matthiopoulos, J. (2012) Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution*, **3**, 177–187.

Amboni, M.P.M. & Laffan, S.W. (2012) The effect of species geographical distribution estimation methods on richness and phylogenetic diversity estimates. *International Journal of Geographical Information Science*, **26**, 2097–2109.

Anderson, R.P. & Gonzalez, I. (2011) Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with MaxEnt. *Ecological Modelling*, **222**, 2796–2811.

Beck, J., Ballesteros-Mejia, L., Nagel, P. & Kitching, I.J. (2013) Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions*, **19**, 1043–1050.

Bombi, P., Luiselli, L. & D'Amen, M. (2011) When the method for mapping species matters: defining priority areas for conservation of African freshwater turtles. *Diversity and Distributions*, **17**, 581–592.

Burgman, M.A. & Fox, J.C. (2003) Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation*, **6**, 19–28.

Cantú-Salazar, L. & Gaston, K.J. (2013) Species richness and representation in protected areas of the western hemisphere: discrepancies between checklists and range maps. *Diversity and Distributions*, **19**, 782–793.

CGAL (2013) Computational geometry algorithms library. Available at: http://www.cgal.org (accessed 16 June 2013).

Clench, H. (1979) How to make regional list of butterflies: some thoughts. *Journal of Lepidopterists' Society*, **33**, 216–231.

Durbin, J. & Watson, G.S. (1951) Testing for serial correlation in least squares regression. *Biometrika*, **38**, 159–171.

Edelsbrunner, H., Kirkpatrick, D.G. & Seidel, R. (1983) On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, **29**, 551–559.

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distribution from occurrence data. *Ecography*, **29**, 129–151.

Eschmeyer, W.N. (2013) Catalog of fishes: genera, species, references. Available at: http://research.calacademy.org/research/ichthyology/catalog/fishcatmain.asp (accessed 20 June 2013).

García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrás-Hernández, A., Vaamonde, A. & Granado-Lorencio, C. (2013) ModestR: a software tool for managing and analyzing species distribution map databases. *Ecography*, **36**, 1202–1207.

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

Giovanelli, J.G.R., de Siqueirac, M.F., Haddadb, C.F.B. & Alexandrino, J. (2010) Modeling a spatially restricted distribution in the Neotropics: How the size of calibration area affects the performance of five presence-only methods. *cological Modelling*, **221**, 215–224.

Graham, C.H. & Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578–587.

Gross, J. (2013) *Five omnibus tests for the composite hypothesis of normality*. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest (accessed 30 June 2013).

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, L.F., Duque, S. & Salmerón, F. (2010) IPez: an expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, **102**, 240–247.

Guisande, C., Patti, B., Vaamonde, A., Manjarrés-Hernández, A., Pelayo-Villamil, P., García-Roselló, E., González-Dacosta, J., Heine, J. & Granado-Lorencio, C. (2013) Factors affecting species richness of marine elasmobranchs. *Biodiversity and Conservation*, **22**, 1703–1714.

Hastie, T. & Fithian, W. (2013) Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 864–867.

Hawkins, B.A., Rueda, M. & Rodríguez, M.A. (2008) What do range maps and surveys tell us about diversity patterns? *Folia Geobotanica*, **43**, 345–355.

Hollander, M., Wolfe, D.A. & Chicken, E. (2013) *Nonparametric statistical methods*, 3rd edn. Wiley, New York.

Hortal, J., Borges, P.A. & Gaspar, C. (2006) Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology*, **75**, 274–287.

Hothorn, T., Zeileis, A., Farebrother, R.W., Cummins, C., Millo, G. & Mitchell, D. (2013) *Testing linear regression models*. R package version 0.9-32. Available at: http://CRAN.R-project.org/package=lmtest (accessed 30 June 2013).

Hurlbert, A.H. & Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences USA*, **104**, 13384–13389.

Hurlbert, A.H. & White, E.P. (2005) Disparity between range map- and survey-based analyses of species richness: patterns, processes and implications. *Ecology Letters*, **8**, 319–327.

IUCN (2013) *Guidelines for using the IUCN Red List categories and criteria. Version 10*. IUCN Standards and Petitions Subcommittee. http://www.iucnredlist.org/documents/RedListGuidelines.pdf (accessed 16 June 2013).

Jetz, W., Sekercioglu, C.H. & Watson, J.E.M. (2008) Ecological correlates and conservation implications of overestimating species geographic ranges. *Conservation Biology*, **22**, 110–119.

Letcher, A. & Harvey, P. (1994) Variation in geographical range size among mammals of the Palearctic. *The American Naturalist*, **144**, 30–42.

Marchetto, A. (2013) *Mann–Kendall test, seasonal and regional Kendall tests*. R package version 1.2. Available at: http://CRAN.R-project.org/package=rkt (accessed September 2013).

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Mesibov, R. (2013) A specialist's audit of aggregated occurrence records. *ZooKeys*, **293**, 1–18.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.M. & Wagner, H. (2013) *Community ecology package*. R package version 2.0-9. Available at: http://CRAN.R-project.org/package=vegan (accessed September 2013).

Pagel, M., May, R. & Collie, A. (1991) Ecological aspects of the geographical distribution and diversity of mammalian species. *The American Naturalist*, **137**, 791–815.

Pateiro-Lopez, B. & Rodriguez-Casal, A. (2011) *Alphahull: generalization of the convex hull of a sample of points in the plane*. Available at: http://cran.r-project.org/web/packages/alphahull (accessed 16 June 2013).

Pearson, R.G., Dawson, T.P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.

Pelayo-Villamil, P., Guisande, C., González-Vilas, L., Carvajal-Quintero, J.D., Jiménez-Segura, L.F., García-Roselló, E., Heine, J., González-Dacosta, J., Manjarrés-Hernández, A., Vaamonde, A. & Granado-Lorencio, C. (2012) ModestR: Una herramienta infromática para el estudio de los ecosistemas acuáticos de Colombia. *Actualidades Biológicas*, **34**, 225–239.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modeling*, **190**, 231–259.

Pianka, E.R. (1966) Latitudinal gradients in species diversity: a review of concepts. *The American Naturalist*, **100**, 33–46.

Pineda, E. & Lobo, J.M. (2012) The performance of range maps and species distribution models representing the geographic variation of species richness at different resolutions. *Global Ecology and Biogeography*, **21**, 935–944.

R Development Core Team (2013) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Rapoport, E.H. (1982) *Areography: geographic strategies of species*. Pergamon, Oxford.

Rohde, K. (1992) Latitudinal gradients in species diversity: the search for the primary cause. *Oikos*, **65**, 514–527.

Ruggiero, A. & Werenkraut, V. (2007) One-dimensional analyses of Rapoport's rule reviewed through meta-analysis. *Global Ecology and Biogeography*, **16**, 401–414.

Sinnott, R.W. (1984) Virtues of the haversine. *Sky and Telescope*, **68**, 159.

Stevens, G.C. (1989) The latitudinal gradient in geographical range: how so many species coexist in the tropics. *The American Naturalist*, **133**, 240–256.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F. & de Clerck, O. (2012) Bio-ORACLE: a global environmental dataset for marine species distribution modeling. *Global Ecology and Biogeography*, **21**, 272–281.

Vasconcelos, T.S., Rodríguez, M.Á. & Hawkins, B.A. (2012) Species distribution modeling as a macroecological tool: a case study using New World amphibians. *Ecography*, **35**, 539–548.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Grey, W.A., White, R.J., Jones, A.C., Bisby, F.A. & Culham, A. (2007) How global is the Global Biodiversity Information Facility? *PLoS ONE*, **2**, 11, e1124.

Zeileis, A. & Hothorn, T. (2002) Diagnostic checking in regression relationships. *R News*, **2**, 7–10. Available at: http://CRAN.R-project.org/doc/Rnews/ (accessed 30 June 2013).

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Figure S1. World distribution of species richness of marine fish species after applying lowest predicted values associated with observed presence as threshold predicted by MaxEnt (upper map) and constraining MaxEnt derived maps to the extent of occurrence area of each species estimated by means of the α-shape procedure (lower map).

Figure S2. Mean latitudinal range size (°) of species in cells of 5′ using Global Biodiversity Information Facility (GBIF) maps, GBIF-extent of occurrence maps, GBIF-MaxEnt maps and GBIF-MaxEnt-restricted maps.

Appendix S1. Taxonomy of studied species including the number of database records for each species. Lack of information on number of records means no map is available for that species.

Appendix S2. Distribution maps of *Cetichthys indagator*, *Beryx mollis*, *Centroberyx druzhinini*, *Myripristis earlei*, *M. seychellensis*, *M. tiki*, *Sargocentron macrosquamis*, *S. xantherythrum*, *Aulotrachichthys argyrophanus*, *A. sajademalensis*, *Gephyroberyx japonicus* and *Paratrachichthys fernandezianus* showing predictions obtained from MaxEnt and records available in the Global Biodiversity Information Facility.

BIOSKETCH

Emilio García-Roselló is a university lecturer at the Department of Computer Sciences of the University of Vigo. His research interests include component-oriented software engineering, scientific software design and software reusability. He applies these interests to the design of software for ecological data processing and biodiversity management.

Editor: José Alexandre Diniz-Filho