



Using citizen science data to estimate climatic niches and species distributions

Patrícia Tiago^{a,b,*}, Henrique M. Pereira^{b,c,d}, César Capinha^{b,e}

^aCE3C, Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências da Universidade de Lisboa, 1749-016 Lisbon, Portugal

^bCátedra Infraestruturas de Portugal-Biodiversidade, CIBIO/InBIO, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

^cGerman Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

^dInstitute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, 06108 Halle, Saale, Germany

^eGlobal Health and Tropical Medicine (GHTM), Instituto de Higiene e Medicina Tropical (IHMT), Universidade Nova de Lisboa (UNL), Rua da Junqueira 100, 1349-008 Lisboa, Portugal

Received 20 August 2016; received in revised form 31 March 2017; accepted 1 April 2017
Available online 8 April 2017

Abstract

Opportunistic citizen data documenting species observations – i.e. observations collected by citizens in a non-standardized way – is becoming increasingly available. In the absence of scientific observations, this data may be a viable alternative for a number of research questions. Here we test the ability of opportunistic species records to provide predictions of the realized distribution of species and if species attributes can act as indicators of the reliability and completeness of these data. We use data for 39 reptile and amphibian species across mainland Portugal as a case study. We collected distribution data from two independent sources: a national citizen-science project and a scientific. We measure and compare the climatic niche width of the species as represented by each of the two data sources. Generalized linear mixed models (GLMM) were used to relate a set of response variables describing the species' morphology, life-history, communication, type of locomotion, habitat and geographic distribution, to observed differences in niche widths. We also performed species distribution models (SDMs) for each of the two types of data using generalized additive models. We found that 12 species had more than 50% of their climate niche covered by citizen science data. Results from GLMMs suggested that the number of grid cells in which a species occurs and its use of forest habitat were positively related to the comprehensiveness of the sampling of climatic niches by citizen science data. Variation in the p of SDMs for both types of data (as measured by the true skill statistic; TSS) was highly similar but SDMs from citizen science data had an overall lower performance. Nevertheless a few species achieved good predictions (TSS > 0.6) using these data. We conclude that species observations in citizen science projects can provide accurate predictions of species

*Corresponding author at: CE3C, Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências da Universidade de Lisboa, 1749-016 Lisbon, Portugal.

E-mail address: patricia.tiago@gmail.com (P. Tiago).

realized distributions, however, efforts should be made to identify the conditions under which these data are more likely to provide reliable representations of the species niches.

© 2017 The Authors. Published by Elsevier GmbH on behalf of Gesellschaft für Ökologie. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Citizen science; Opportunistic species observations; Climatic niche; Species distribution models

Introduction

Currently there is a large interest in citizen science – i.e. the engagement of non-professionals in scientific research (Miller-Rushing, Primack, & Bonney 2012) – marked by a strong increase on citizen science programs (Dickinson et al. 2012). The scope of these programs is wide, covering research areas such as conservation biology and biodiversity monitoring, which are using citizen science programs to collect large amounts of species distribution data allowing to fill existing gaps of information (e.g. Pereira & Cooper 2006; Danielsen et al. 2009; Danielsen, Burgess, Jensen, & Pirhofer-Walzl 2010; Pereira et al. 2010; Szabo, Vesk, Baxter, & Possingham 2010; Dickinson et al. 2012). These citizen science projects can collect data in a similar way to scientific surveys, i.e. following strict protocols. In these cases, the main difference between the two types of survey concerns the general lack of experience of the citizen scientists which can lead to taxonomic misidentifications, reducing data accuracy. Alternatively, citizen science surveys of species distributions can take place through the collection of opportunistic data, data collected by non-standardized methods, with no sampling design and no systematic protocol (Dickinson, Zuckerberg, & Bonter 2010). These later programs can cover wide spatial extents and often provide a large number of records (Chandler et al. 2012, 2017). These opportunistic records have the same problem of lack of taxonomic expertise of the participants and can, in many cases, be spatially and temporally biased (Beck, Ballesteros-Mejia, Nagel, & Kitching 2013; Higa et al. 2014). Bias in species observations provided by citizen-science programs may hinder the usefulness of these records in ecological research. Importantly, the sampling effort of opportunistic records is generally not known but it can vary widely over time (Dickinson et al. 2010; Snäll, Kindvall, Nilsson, & Pärt 2011) and across space (Dickinson et al. 2010; van Strien, van Swaay, & Termaat 2013) and between and within taxonomic groups (Kéry et al. 2010; Snäll et al. 2011).

Species distribution models (SDMs) – also known as ecological niche models or habitat selection models – are now widely used in ecological and evolutionary research (Elith & Leathwick 2009; Kozak, Graham, & Wiens 2008). These models relate data on species distributions with spatial environmental data in order to estimate locations where the species could occur (Elith & Leathwick 2009). The questions they allow to address are wide-ranging and include how climate change may modify biodiversity patterns (Thuiller et al. 2008), where invasive species may become established

(Capinha & Anastácio 2011), where the hotspots of endangered species are located (Godown & Peterson 2000), which areas should be prioritized for conservation (Chen & Peterson 2002) or which locations are suitable for species translocations or cultivation (Jovanovic, Arnold, & Booth 2000; Cunningham, Anderson, & Walsh 2002). SDMs rely on two types of data, species distribution data and environmental data. While the latter is now widely available at high spatial resolution and for wide spatial extents (Kozak et al. 2008), mainly due to large-scale mapping and modelling projects (e.g. Higa et al. 2014; Levinsky et al. 2013), the geographic distribution of many species still remains poorly known (Scheffers, Joppa, Pimm, & Laurance 2012). In this context, it is relevant to understand if species observation records coming from citizen science projects are useful for inferring species distributions, and if so, under which conditions these records are more or less reliable.

In this work we assess whether opportunistic citizen science databases are viable data sources to model species distributions and test if species attributes can indicate the reliability and completeness of the opportunistic distribution data. We use amphibians and reptiles records from the BioDiversity4All database (www.biodiversity4all.org), a country-wide citizen science project in Portugal. We use herptiles (i.e. reptiles and amphibians) because many of these species tend to be cryptic and pass unnoticed and also because the prejudice associated with this group can affect the observations recorded in a citizen science project, as several of these species are feared and despised by many people (Price & Dorcas 2011). These characteristics contribute to a distributional data shortfall – as opposed, for instance, to a few other conspicuous and ‘attractive’ groups such as birds for which distributional data is more abundant. We use opportunistic citizen science records to measure the climatic niche width of 39 herptiles. We then compare these niche widths to the ones obtained using records from an, independent, long-term, scientific atlas and test for species traits and characteristics of the species distributions as indicators of the differences found. Finally, we also build projections of species distribution models based on each of the two distinct sources of species records and compare their predictive performances.

Materials and methods

In this work, we perform three main analyses to assess the merits of opportunistic citizen science records of species observations. In the first analysis, we measure the climatic

niche width of the species, as represented by citizen science observations and as represented by an atlas using scientific observations. In the second analysis, we statistically relate the differences of widths we found between these two data sources with a large set of variables describing properties of the species. This aims to test if species properties can serve as predictors of the sampling of climatic niches by citizen science records. Finally, in our third analysis, we perform species distribution models for each type of species observation data and compare the predictive performances of each. We describe each of these analyses in detail below.

Species data

We used citizen science data for reptiles and amphibians recorded in mainland Portugal. These data were retrieved from the BioDiversity4All web portal (www.biodiversity4all.org), a Portuguese project that is similar to citizen science biodiversity databases elsewhere such as iNaturalist (<http://www.inaturalist.org/>) or iSpot (<http://www.ispot.org/>). We used only species occurrences that provided GPS recordings of geographical coordinates, from 1990 until December 2013 and resampled these records to a unique record per 10×10 km grid cell. Species occurring in less than three distinct grid cells were not considered because of insufficient records to calculate niche breadths (see below). In total we used data for 39 species (15 amphibians and 24 reptiles) (see Appendix A in Supplementary material: Table 1), representing 91% of all herptile species known for mainland Portugal. The data set had a mean number of grid cells per species of 23.95 ± 20.27 (SD).

Simultaneously, we also collected 10×10 km grid cell distribution data of these species from the atlas of Portuguese herptofauna (Loureiro et al. 2008; hereafter referred to as “atlas”). This is the most updated inventory of the herptiles species distributions available for Portugal and, unlike opportunistic citizen science databases, this atlas provides distribution data collected from previous scientific publications, systematic surveying in sampling stations and ad hoc observations from experts (Crespo & Oliveira 1989; Malkmus 1995; Godinho et al. 1999; Malkmus 2004; Loureiro et al. 2008).

Environmental data

We collected five spatial variables for representing climatic variation across mainland Portugal: annual mean temperature; temperature seasonality; maximum temperature of warmest month; annual precipitation and precipitation seasonality. These variables were collected from the Worldclim database (Hijmans, Cameron, Parra, Jones, & Jarvis 2005) at a spatial resolution of 5 arc minutes (c. 9 km at the mid-latitude of Portugal). We selected these variables because temperature and humidity are known to be highly influen-

tial in determining the distribution of herptiles, affecting their physiology and key aspects of biology and phenology (Girardello, Griggio, Whittingham, & Rushton 2010). Moreover, mainland Portugal has a strong latitudinal climatic gradient, ranging from semi-aridity in regions of the south to moist Atlantic climate in the north, allowing for the filtering effect of climate to be highly influential in the distribution of the studied species (Márquez, Real, Olivero, & Estrada 2011). We performed a principal component analysis (PCA) to remove multi-collinearity among the 5 climatic variables and retained the two first components, which together accounted for 89% of the total variance in the original data (see Appendix A in Supplementary material: Table 2).

Measuring sampling of climatic niche from citizen science data

For each species we calculated two distinct climatic niches, one using the species occurrence records found in the citizen science project (opportunistic data), and the other using the occurrences found in the atlas: for each of the data sources, we projected the occurrences of each species into the climatic space defined by the two first principal components obtained from the PCA. Next, we used a convex hull to delimit the climatic space occupied by each species. The width of the climatic niche of each species corresponded to the area enclosed by each convex hull. Finally, for each species we calculated the percentage of the total climatic niche width measured using occurrences from the atlas that was captured by the climatic niche calculated for the citizen science data – i.e. the area of the intersected polygons divided by the area of the scientific atlas polygon, multiplied by 100. By considering only the intersected polygon to represent the climatic niched sampled by citizen science, we exclude any portion falling outside the niche space that is sampled by the atlas, as these cases are likely to have its origin on observation errors (see Discussion). A value of 100% corresponds to a situation where the citizen science data fully captures the climatic niche as represented by the atlas.

Species traits and characteristics of species distributions

In order to potentially explain interspecific differences in the climatic niche sampling from citizen science data, we collected a set of explanatory variables describing species morphology, life-history, communication mode, type of locomotion, habitat and geographic distribution. Data for these variables were collected from Trochet et al. (2014), Encyclopedia of Life (<http://www.eol.org>), Reptile database (<http://reptile-database.org/>), Reptile Trait Database (<http://scales.ckff.si/scaletool>), HerpNet (<http://www.herpnet.org>), Naturdata (<http://naturdata.org>) and Loureiro et al. (2008). A total of 27 variables were

considered (see Appendix A in Supplementary material: Table 3).

Relating variation of climatic niche and species traits

We used generalized linear mixed models (GLMM) (Breslow & Clayton 1993) to relate the variation in the percentage of climatic niche sampled by citizen science (response variable) with species traits (the explanatory variables). The advantage of using GLMMs instead of simpler generalized linear models is that the former allow to incorporate information on the clustering of the sampling units, hence providing estimates of standard errors of model coefficients corrected for non-independence among observations. Here, to account for the phylogenetic non-independence of species (i.e., species that are phylogenetically closer should provide more similar observations than those that are further away), we built GLMMs using family (the taxonomic clustering unit used here) as random effect, while the remaining variables entered the models as fixed effects (Cassey, Blackburn, Sol, Duncan, & Lockwood 2004). The response variable, which varies between 0 and 100% of climate niche overlap, was arcsine transformed before analyses to meet assumptions of normality. The models assumed a Gaussian distribution of errors and used an identity link function (Breslow & Clayton 1993). To avoid problems of multi-collinearity among the explanatory variables, we calculated variance inflation factors (VIFs) for each. We then sequentially dropped the explanatory variable with the largest VIF until all values were below 5—a threshold below which collinearity is not a problem (Bowerman & O'Connell 1990). This procedure led to the removal of 5 variables (see Appendix A in Supplementary material: Table 3). Using the remaining variables, we built GLMMs for all possible combinations of fixed effects. All models used restricted maximum likelihood estimation and Akaike's information criterion for small samples size (AICc) was used for inter-model ranking. We obtained final model coefficients from averaging the set of 'best' models, instead of selecting a single best model. The best ranked models were considered to be those holding a cumulative AICc weight of at least 0.95 (Karanth, Gopalaswamy, DeFries, & Ballal 2012). We used the lme4 package (Bates, Maechler, & Dai 2009) for R 3.1 (R Development Core Team 2014) to implement the GLMMs, while model selection and averaging were conducted using the MuMIn 1.10.0 package (Barton 2014).

Species distribution models

We also built species distribution models (SDMs) for all species that had 20 or more unique grid cells in the citizen science data set. Species having a lower number of records were discarded because of insufficient variability to include in a model with two explanatory variables (i.e., the two first components of a PCA; see below) (Vittinghoff & McCulloch

2007). This corresponded to a total of 20 species (9 amphibians and 11 reptiles). SDMs for these species were made independently for the citizen science data set and the atlas data set. All models were made using generalized additive models (GAM) (Guisan, Edwards, & Hastie 2002) which contrasted the climatic conditions where the species was observed (as represented by the two main axes of the PCA, see above) with a random sample of climatic conditions in the grid cells where it was not recorded (i.e. 'pseudo-absences') (Capinha, Anastácio, & Tenedório 2012). The GAM models were implemented using the raster2 package of R using default values (see Appendix A in Supplementary material: Text 1). A total of 10 replicate predictions were made for each species, each using a new random set of pseudo-absences. Final prediction corresponded to the average of the 10 replicate predictions.

We evaluated the performance of SDMs using citizen science observations by comparing the predictions with the species presences and absences found in the atlas. The atlas-based SDMs were evaluated by means of 5-fold cross-validation. Model evaluation was assessed using the true skills statistics—TSS (Allouche, Tsoar, & Kadmon 2006). This metric accounts for both sensitivity (i.e. proportion of species presences correctly predicted) and specificity (i.e. proportion of pseudo-absences correctly predicted) of the predictions. This metric ranges from −1 to 1 and requires the definition of a probability threshold value for continuous predictions (Freeman & Moisen 2008), so we performed its calculation along all possible thresholds using sequential increments of 0.01. We retained the maximum value obtained across all thresholds. TSS values >0.6 indicate good predictive performance, 0.2–0.6 fair to moderate, and <0.2 poor (Capinha, Rocha, & Sousa 2014).

Results

The proportion of the climatic niche width that was captured by citizen science data differed markedly among species (Figs. 1A and B; 2). Of the 39 species analysed, 12 had more than 50% of their known climate space covered by citizen science data (Fig. 1A; B) (see Appendix A in Supplementary material: Table 4). One amphibian (*Hyla meridionalis*) and three reptiles (*Tarentola mauritanica*, *Timon lepidus* and *Psammmodromus algirus*) had niche coverages higher than 70%. The anurans *Pelophylax perezi*, *Bufo bufo* and *Hyla arborea* achieved niche coverages higher than 60% while one salamander (*Salamandra salamandra*) and four reptiles (*Psammmodromus hispanicus*, *Rinechis scalaris*, *Natrix maura* and *Podarcis hispanicus*) had coverages higher than 50%. The species with lowest coverages were two snakes: *Coronella austriaca* (5.9%) and *Vipera latastei* (6.0%) and two lizards: *Chalcides bedriagai* (1.2%) and *Hemidactylus turcicus* (6.9%). Eight of the 39 species had portions of the climatic niche sampled by citizen science that were not represented by observations of the atlas data set. For six of these

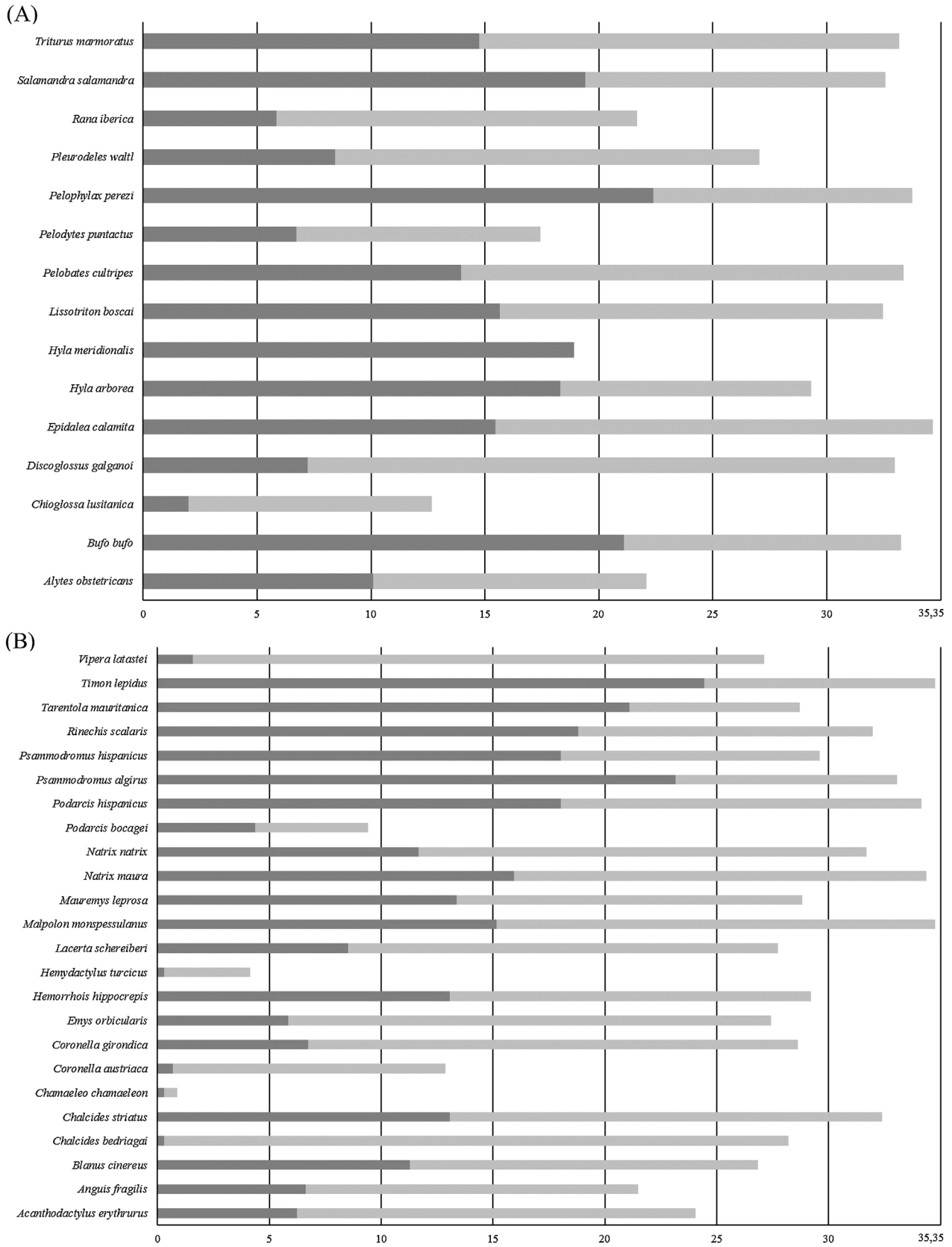


Fig. 1. Width of the climatic niche of (A) amphibians and (B) reptiles as measured from atlas records (light grey), citizen science records (dark grey) and the total climatic space existing in the study area (mainland Portugal) (x-axis).

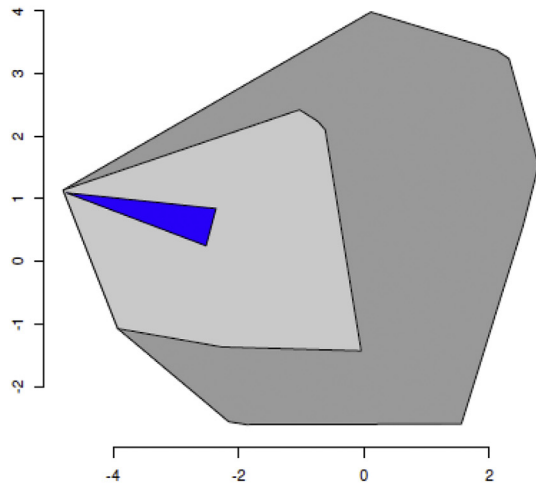
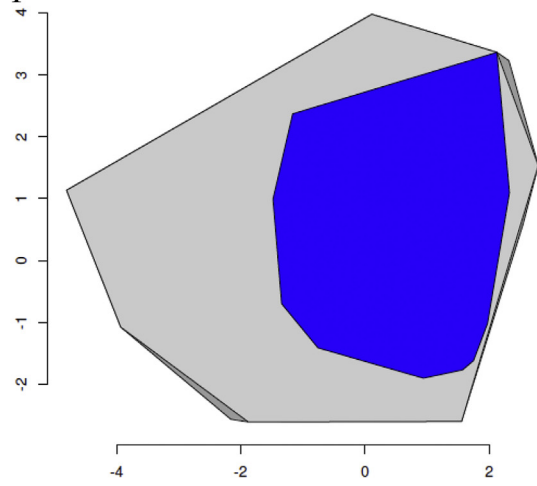
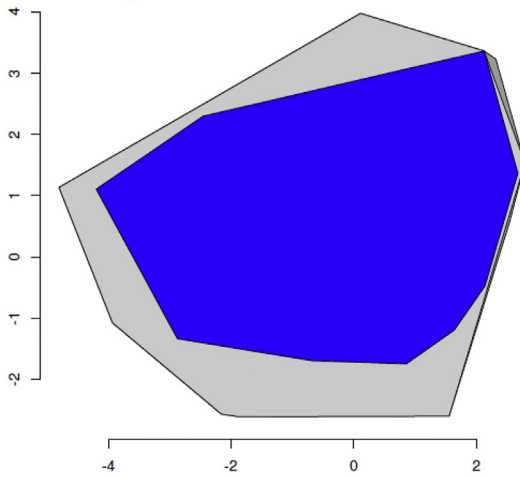
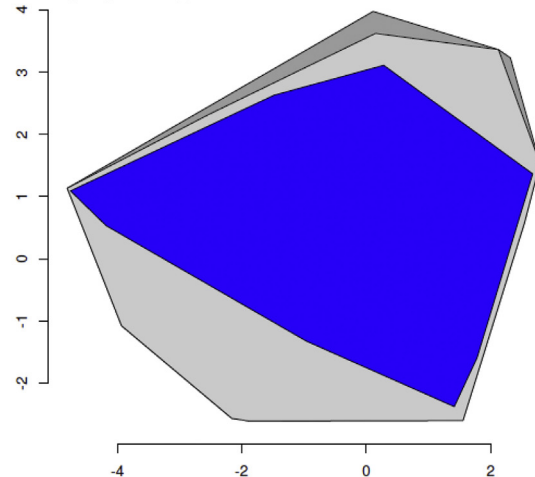
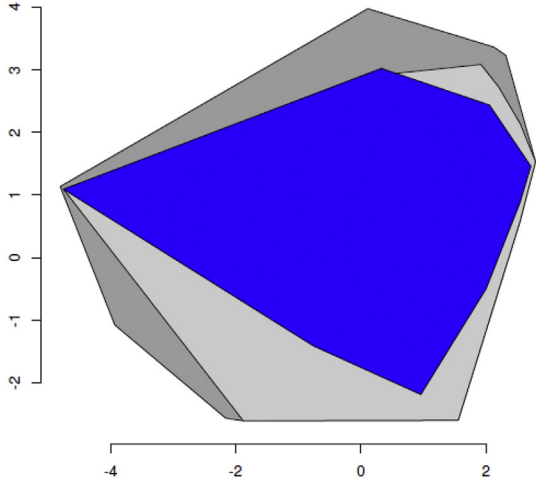
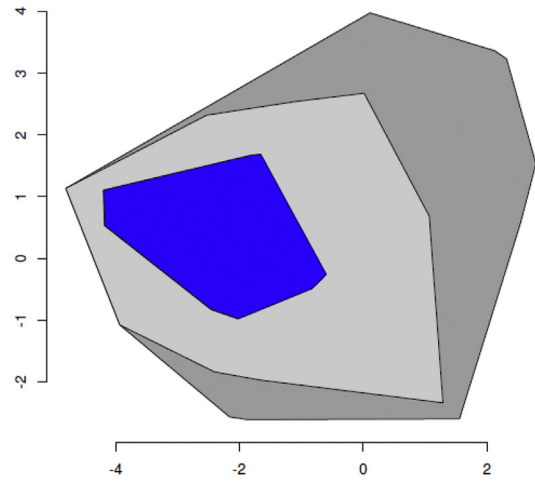
Coronella austriaca*Epidalea calamita**Timon lepidus**Pelophylax perezi**Tarentola mauritanica**Rana ibérica*

Fig. 2. Representative examples of species' climatic niche widths based on opportunistic data (blue) and scientific atlas data (light grey). The total climatic space existing in the study area (mainland Portugal) is also shown (dark grey). Polygons refer to the convex envelope of the corresponding species observations projected into a two-dimensional climatic space. The climatic space is defined by the two main axes of a principal components analyses summarizing overall climatic variation in the study area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1. Model averages of generalized linear mixed models (GLMM) relating the variation in the percentage of climatic niche sampled by citizen science to species traits (GridAtlas—number of grid atlas cells presence, H.Forest—forest habitat, StD.Altitude—standard deviation of the altitude, Climber—mode of displacement climber, H.Shrubland—shrubland habitat, H.Sand—sand habitat, H.Grassland—grassland habitat, Latmax—maximum latitude) (level of significance * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$). A total of 39 herptile species that occur in mainland Portugal is used in the models.

Variables	Relative variable importance	Model-averaged coefficients	Std. Error	$P_r (> z)$
GridAtlas	1.00	6.217e – 04	1.325e – 04	<0.001***
H.Forest	0.99	2.198e – 01	7.266e – 02	0.01**
StD.Altitude	0.51	–5.264e – 04	6.433e – 04	0.42
Climber	0.36	4.957e – 02	8.138e – 02	0.55
H.Shrubland	0.31	3.926e – 02	7.325e – 02	0.60
H.Sand	0.16	–1.570e – 02	4.992e – 02	0.76
H.Grassland	0.12	–8.510e – 03	3.392e – 02	0.80
Latmax	0.10	–2.499e – 03	1.790e – 02	0.89
Intercept		5.823e – 01	8.301e – 01	0.49

species this ‘outside region’ corresponded to about 5% or less of all niche breadth measured by citizen science (*Alytes obstetricans*, *Blanus cinereus*, *H. arborea*, *P. hispanicus*, *P. hispanicus* and *T. mauritanica*). For two species, this region, however, corresponded to about 35% of the total breadth of the niche (*H. meridionalis* and *Podarcis bocagei*).

Based on averaged model coefficients of GLMMs relating variation in the percentages of climatic niche sampled by citizen science with species traits, we found that the two variables that had higher relative importance were the number of grid cells occupied by the species and the use of forest habitat; both showing a positive association with niche coverage (Table 1). The remaining explanatory variables either had negligible relevance in the models or were not selected by any model (Table 1).

We built SDMs for 20 species based on citizen science records and validated the predictions using the scientific atlas. The performance of the models varied strongly among species (Fig. 3). The performance was good ($TSS \geq 0.6$) for three species: *H. meridionalis*, *Lacerta schreiberi* and *Rana iberica*, fair ($0.6 > TSS \geq 0.2$) for nine species: *Hemorrhoids hippocrepis*, *H. arborea*, *Lissotriton boscai*, *Mauremys lep-rosa*, *P. perezi*, *P. algirus*, *S. salamandra*, *T. mauritanica* and *Triturus marmoratus*, while models for eight species: *B. bufo*, *Epidalea calamita*, *Malpolon monspessulanus*, *N. maura*, *P. hispanicus*, *P. hispanicus*, *R. scalaris* and *T. lepidus* performed poorly ($TSS < 0.2$) (Fig. 3) (see Appendix A in Supplementary material: Table 5). We also found that the percentage of niche covered by citizen science is not associated with the predictive performance of the SDMs (Pearson correlation coefficient = -0.38 , $P = 0.096$).

We also built SDMs using scientific species occurrences which were validated by means of a 5-fold cross-validation. The variation of the predictive performance of these models and of those from citizen science was very similar (Pearson correlation coefficient = 0.9 , $P < 0.05$) but the latter had, on average, a lower performance (mean difference in $TSS = -0.151$, SD: 0.101) (Fig. 3).

Discussion

In this work, we assessed whether species records from opportunistic citizen science databases were able to provide useful estimates of (1) the climatic niches of the species and of (2) their potential distributions, as drawn from species distribution models. We further aimed to identify how the species’ characteristics may act as indicators of the reliability and completeness of opportunistic distribution data.

Our results showed that the sampling of the species climatic niches based on citizen science records can vary greatly. This is not unexpected because, as in most biological groups, herptile species differ greatly in terms of elusiveness and secludeness, inherently leading to variation in detectability and ease of identification in the natural environment (Mazerolle et al. 2007). However, we also found that some species attributes may serve as indicators of the usefulness of the opportunistic sampling. The number of grid cells where the species occurs was the variable with higher explanatory power, suggesting that climatic niches are better sampled for species that have wide ranges. This could be because these are, in general, common species, including for instance *P. perezi*, *T. lepidus*, *B. bufo* which are also relatively well known by the citizens. The other indicator we found good support for was if a species uses forest habitats, with those using them having a significantly better sampling of their climatic niche. The reason for this relationship is not clear. One possible explanation might be that this type of habitat is preferred by citizen scientists for making observations, but this requires further investigation.

We also identified some species having portions of the climatic niche represented by citizen science data that were not found in the climatic niches using the atlas data set. This was found for a small number of species (eight) and only for two of them this ‘novel’ portion of the niche represented more than 6% of its total niche breadth. Two non-mutually exclusive reasons can account for the niche expansions observed. The first is that the additional niche breadth is an ‘artifact’ resulting

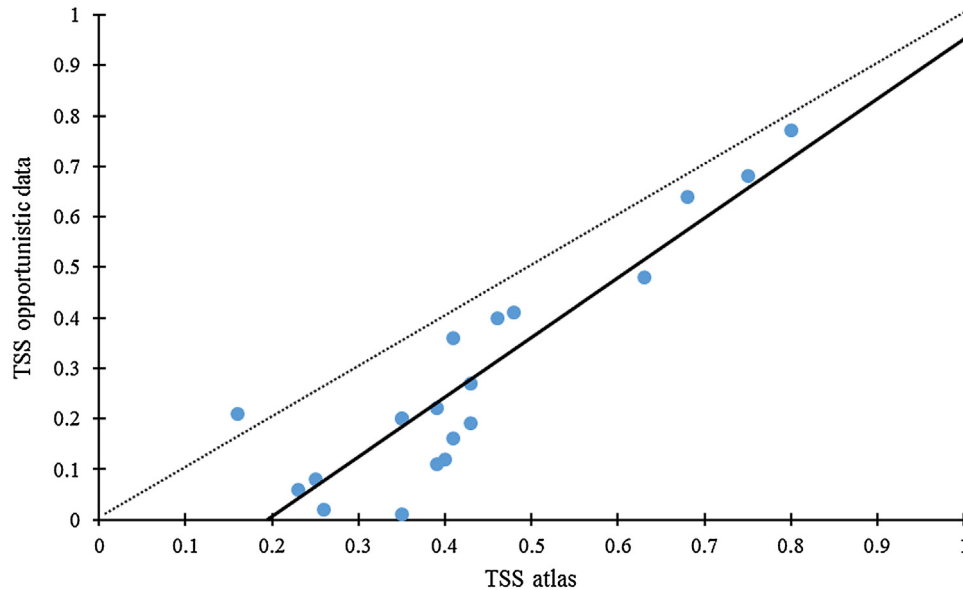


Fig. 3. Comparison of the performance of species distribution models based on opportunistic data (y-axis) and on scientific atlas data (x-axis). The models were performed using a total of 20 species of herptiles occurring in mainland in Portugal. Each dot represents one species. Species distribution models were made using generalized additive models (GAM) and the predictive performance of each model was measured by means of the true skills statistics (TSS). The predictive performances achieved by the two types of data are strongly correlated (Pearson correlation coefficient = 0.9, $P < 0.05$). The dotted line indicates the relationship expected if performances from both types of predictions matched perfectly. The straight line corresponds to the relationship that was indeed observed.

from misidentifications of the species, leading citizen scientists to record the species in places where it is in fact absent. The second reason is that it could indeed represent a niche expansion, resulting from true observations of the species in areas not recorded in the atlas data set. While it is hard to disentangle the relative role of the two possible causes, we expect the former to certainly have some expression, given the uncertain reliability of citizen science observations. Nevertheless, a role for citizen science observations of the species in new locations should also be considered as found elsewhere (Dickinson et al. 2010). Future work should consider the validation of citizen science observations, by experts, in those new areas.

Regarding the performance of potential distribution models, we also found a high interspecific variability, but importantly, several species were well modeled using citizen science occurrence records (3 species with good predictive performance and 9 with fair predictive performance). Moreover, variation in predictive performances among species was highly similar between models using citizen science data and those using data from a scientific database, despite a lower average performance of the former. This strongly suggests that opportunistic citizen science databases of species observations can represent a viable alternative to scientific records when these are not available, and the challenge might be to combine different data sources to achieve better results.

Interestingly, we found that SDMs for species having their climatic niches well sampled by citizen science were not necessarily the ones having higher predictive performance. This finding highlights that having good citizen science data does

not guarantee having good SDMs. This is because the performance of SDMs is contingent on a number of other factors besides the comprehensiveness of species distribution data. We found that the performance of SDMs (both opportunistic and atlas-based) had a strong negative relationship with the total climatic niche of the species, which suggests that the performance of models for the study area is strongly determined by the degree of climatic specialization of the species. In other words, generalist species are less well predicted, as found elsewhere (Evangelista et al. 2008).

Other factors can further contribute to reducing SDM performance when one uses opportunistic species occurrences. One of such factors is that the opportunistic databases do not provide records of species absences which are highly desirable to improve SDMs. Hence our models using pseudo-absences (i.e. a sampling of all climatic conditions available) may be biasing the species potential distributions (VanDerWall, Shoo, Graham, & Williams 2009) and it would be of great interest to have citizen science records of species absences as well. This is possible when observers use species check-lists (Sullivan et al. 2009). For studies in countries such as the Netherlands, where opportunistic data is numerous and widespread, geographical bias presents limited relevance, since for many species the data collection is already very good (van Strien et al. 2013). With fewer data or when particular areas or habitat types remain under-sampled, post-stratification and weighting of sites according to their share in the statistical population under study may be a useful option (van Strien et al. 2013).

Some studies demonstrated how widespread, inexpensive count data, collected by citizen scientists can be analyzed to reveal important information about the habitat preference and population dynamics of broadly dispersed and difficult-to-observe species (Thorson, Scheverell, Semmens, & Pattengill-Semmens 2014). Even though studies emphasized problems associated with opportunistic data, like being generated with uneven sampling effort over time (Botts, Erasmus, & Alexander 2012; Maes, Vanreusel, Jacobs, Berwaerts, & Van Dyck 2012; Prendergast, Wood, Lawton, & Eversham 1993), others gave evidence that opportunistically-gathered data has a great potential to make meaningful contributions to biodiversity science and policy-making (Schmeller et al. 2009; Tulloch, Possingham, Joseph, Szabo, & Martin 2013; Isaac et al. 2014). Knowledge of species' distributions is of particularly importance for ecology, evolution and conservation science. Opportunistic citizen science databases can provide scientists with a vast set of occurrence records, yet, understanding the benefits and limitations of this information is fundamental to achieve better results in modelling species distributions.

Acknowledgements

PT acknowledges support from the Portuguese Foundation for Science and Technology (FCT/MCTES) (SFRH/BD/89543/2012). CC acknowledges support from the Portuguese Foundation for Science and Technology (FCT/MCTES) and POPH/FSE (EC) (SFRH/BPD/84422/2012 and GHTM—UID/Multi/04413/2013). We wish to thank the three anonymous peer reviewers, who helped us to greatly improve the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.baae.2017.04.001>.

References

- Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43, 1223–1232.
- Barton, K. (2014). *MuMIn: multi-model inference*. R Package Version 1.10.0. Available from: <http://lme4.r-forge.r-project.org>.
- Bates, D., Maechler, M., & Dai, B. (2009). *lme4: linear mixed-effects models using Eigen and S4 classes*. R package version 0.999375-31. <http://cran.r-project.org/web/packages/lme4/>
- 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.
- Bowerman, B. L., & O'Connell, R. T. (1990). *Linear statistical models: An applied approach* (2nd edn.). Belmont, CA: Duxbury Press.
- Botts, E. A., Erasmus, B. F. N., & Alexander, G. J. (2012). Methods to detect species range size change from biological atlas data: A comparison using the South African frog atlas project. *Biological Conservation*, 146(1), 72–80. <http://dx.doi.org/10.1016/j.biocon.2011.10.035>
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.
- Capinha, C., & Anastácio, P. (2011). Assessing the environmental requirements of invaders using ensembles of distribution models. *Diversity and Distributions*, 17, 13–24.
- Capinha, C., Anastácio, P., & Tenedório, J. A. (2012). Predicting the impact of climate change on the invasive decapods of the Iberian inland waters: An assessment of reliability. *Biological Invasions*, 14(8), 1737–1751.
- Capinha, C., Rocha, J., & Sousa, C. A. (2014). Macroclimate determines the global range limit of *Aedes aegypti*. *EcoHealth*, 11(3), 420–428.
- Cassey, P., Blackburn, T. M., Sol, D., Duncan, R. P., & Lockwood, J. L. (2004). Global patterns of introduction effort and establishment success in birds. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(Suppl. 6), S405–S408.
- Chandler, M., Bebb, D. P., Castro, S., Lowman, M. D., Muoria, P., Ogue, N., & Rubenstein, D. I. (2012). International citizen science: Making the local global. *Frontiers in Ecology and the Environment*, 10(6), 328–331.
- Chandler, M., See, L., Buesching, C. D., Cousins, J. A., Gillies, C., Kays, R. W., ... & Tiago, P. (2017). Involving citizen scientists in biodiversity observation. In M. Walters, & R. J. Scholes (Eds.), *The GEO handbook on biodiversity observation networks* (pp. 211–237). Cham: Springer International Publishing.
- Chen, G., & Peterson, T. (2002). Prioritization of areas in China for the conservation of endangered birds using modelled geographical distributions. *Bird Conservation International*, 12, 197–209. <http://dx.doi.org/10.1017/s0959270902002125>
- Crespo, E. G., & Oliveira, M. E. (1989). *Atlas da distribuição dos anfíbios e répteis de Portugal continental*. Lisboa: Serviço Nacional de Parques e Reservas e Conservação da Natureza.
- Cunningham, D. C., Anderson, E. R., & Walsh, K. B. (2002). Ecology and biogeography of *Cassia brewsteri*: Assessment of potential sites for cultivation. *Australian Journal of Experimental Agriculture*, 42, 1071–1080.
- Danielsen, F., Burgess, N. D., Balmford, A., Donald, P. F., Funder, M., Jones, J. P., & Yonten, D. (2009). Local participation in natural resource monitoring: A characterization of approaches. *Conservation Biology*, 23, 31–42.
- Danielsen, F., Burgess, N. D., Jensen, P. M., & Pirhofer-Walzl, K. (2010). Environmental monitoring: The scale and speed of implementation varies according to the degree of people's involvement. *Journal of Applied Ecology*, 47, 1166–1168.
- Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Phillips, T., & Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, 10(6), 291–297.
- Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: Challenges and bene-

- fits. *Annual Review of Ecology Evolution and Systematics*, 41, 149–172.
- Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, 40(1), 677–697.
- Evangelista, P. H., Kumar, S., Stohlgren, T. J., Jamevich, C. S., Crall, A. W., Norman, J. B., & Barnett, D. T. (2008). Modelling invasion for a habitat generalist and a specialist plant species. *Diversity and Distributions*, 14, 808–817.
- Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, 217, 48–58.
- Girardello, M., Griggio, M., Whittingham, M. J., & Rushton, S. P. (2010). Models of climate associations and distributions of amphibians in Italy. *Ecological Research*, 25, 103–111.
- Godinho, R., Teixeira, J., Rebelo, R., Segurado, P., Loureiro, A., Alvares, F., ... & Brito, J. C. (1999). Atlas of the continental portuguese herpetofauna: An assemblage of published and new data. *Revista Española de Herpetología*, 13, 61–82.
- Godown, M., & Peterson, T. (2000). Preliminary distribution analysis of US endangered bird species. *Biodiversity and Conservation*, 9, 1313–1322.
- Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, 157(2), 89–100.
- Higa, M., Yamaur, Y., Koizumi, I., Yabuhara, Y., Senzaki, M., & Ono, S. (2014). Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. *Diversity and Distributions*, 1–9.
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965–1978.
- Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., Roy, D. B. (2014). Extracting robust trends in species' distributions from unstructured opportunistic data: a comparison of methods. 10.1101/006999.
- Jovanovic, T., Arnold, R., & Booth, T. (2000). Determining the climatic suitability of *Eucalyptus dunnii* for plantations in Australia: China and Central and South America. *New Forests*, 19, 215–226.
- Karanth, K. K., Gopalaswamy, A. M., DeFries, R., & Ballal, N. (2012). Assessing patterns of conflict risk and compensation distribution around a Central Indian protected area. *PLoS One*, 7(12), e50433. <http://dx.doi.org/10.1016/j.biocon.2013.06.027>
- Kéry, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., & Zbinden, N. (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, 24, 1388–1397.
- Kozak, K. H., Graham, C. H., & Wiens, J. J. (2008). Integrating GIS-based environmental data into evolutionary biology. *Trends in Ecology and Evolution*, 23, 141–148.
- Levinsky, I., Araújo, M. B., Nogués-Bravo, D., Haywood, A. M., Valdes, P. J., & Rahbek, C. (2013). Climate envelope models suggest spatio-temporal co-occurrence of refugia of African birds and mammals. *Global Ecology and Biogeography*, 22, 351–363.
- Loureiro, A., Ferrand de Almeida, N., Carretero, M. A., & Paulo, O. S. (Eds.). (2008). *Atlas dos Anfíbios e Répteis de Portugal*. Lisboa: Instituto de Conservação da Natureza e da Biodiversidade.
- Maes, D., Vanreusel, W., Jacobs, I., Berwaerts, K., & Van Dyck, H. (2012). Applying IUCN red list criteria at a small regional level: A test case with butterflies in Flanders (north Belgium). *Biological Conservation*, 145(1), 258–266. <http://dx.doi.org/10.1016/j.biocon.2011.11.021>
- Malkmus, R. (1995). *Die Amphibien und Reptilien Portugals, Madeiras und der Azoren*. Wissenschaften, Magdeburg: Westarp.
- Malkmus, R. (2004). *Amphibians and reptiles of Portugal, Madeira and the Azores Archipelago*. Ruggell: A.R.G. Gantner Verlag K.G.
- Márquez, A. L., Real, R., Olivero, J., & Estrada, A. (2011). Combining climate with other influential factors for modelling the impact of climate change on species distribution. *Climate Change*, 108(1–2), 135–157.
- Mazerolle, M. J., Bailey, L. L., Kendall, W. L., Royle, J. A., Converse, S. J., & Nichols, J. D. (2007). Making great leaps forward: Accounting for detectability in herpetological field studies. *Journal of Herpetology*, 41, 672–689.
- Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, 10(6), 285–290.
- Pereira, H. M., Belnap, J., Brummitt, N., Collen, B., Ding, H., Gonzalez-Espinosa, M., ... & Vieira, C. (2010). Global biodiversity monitoring. *Frontiers in Ecology and the Environment*, 8, 459–460.
- Pereira, H. M., & Cooper, H. D. (2006). Towards the global monitoring of biodiversity change. *Trends in Ecology & Evolution*, 21, 123–129.
- Prendergast, J., Wood, S., Lawton, J., & Eversham, B. (1993). Correcting for variation in recording effort in analyses of diversity hotspots. *Biodiversity Letters*, 1(2), 39–53.
- Price, S. J., & Dorcas, M. E. (2011). The Carolina Herp Atlas: An online, citizen science approach to document amphibian and reptile occurrences. *Herpetological Conservation Biology*, 6(2), 287–296.
- Scheffers, B. R., Joppa, L. N., Pimm, S. L., & Laurance, W. (2012). What we know and don't know about Earth's missing biodiversity. *Trends in Ecology and Evolution*, 27(9), 501–510.
- Schmeller, D. S., Henry, P. Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., ... & Henle, K. (2009). Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology*, 23(2), 307–316. <http://dx.doi.org/10.1111/j.1523-1739.2008.01125>
- Snäll, T., Kindvall, O., Nilsson, J., & Pärt, T. (2011). Evaluating citizen-based presence data for bird monitoring. *Biological Conservation*, 144, 804–810.
- Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142, 2282–2292.
- Szabo, J. K., Vesk, P. A., Baxter, P. W. J., & Possingham, H. P. (2010). Regional avian species declines estimated from volunteer-collected long-term data using list length analysis. *Ecological Applications*, 20, 2157–2169.
- Thorson, J. T., Scheverell, M. D., Semmens, B. X., & Pattengill-Semmens, C. V. (2014). Demographic modeling of citizen science data informs habitat preferences and

- population dynamics of recovering fishes. *Ecology*, 95(12), 3251–3258.
- Thuiller, W., Albert, C., Araújo, M. B., Berry, P. M., Cabeza, M., Guisan, A., . . . & Zimmermann, N. E. (2008). Predicting global change impacts on plant species' distribution: Future challenges. *Perspectives in Plant Ecology Evolution and Systematics*, 9, 137–152.
- Trochet, A., Moulherat, S., Calvez, O., Stevens, V., Clobert, J., & Schmeller, D. (2014). A database of life-history traits of European amphibians. *Biodiversity Data Journal*, 2, e4123. <http://dx.doi.org/10.3897/BDJ.2e4123>
- Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J., & Martin, T. G. (2013). Realising the full potential of citizen science monitoring programs. *Biological Conservation*, 165, 128–138.
- VanDerWall, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modelling: How far you stray from what you know? *Ecological Modelling*, 220(4), 589–594.
- van Strien, A. J., van Swaay, C. A. M., & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, 50, 1450–1458.
- Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6), 710–718. <http://dx.doi.org/10.1093/aje/kwk052>

Available online at www.sciencedirect.com

ScienceDirect