A simple method to estimate the probable distribution of species

Emilio García-Roselló[1], Cástor Guisande[2], Luis González-Vilas[2], Jacinto González-Dacosta[1], Juergen Heine[1], Elisa Pérez-Costas[2] and Jorge M. Lobo[3]

[1]Department of Computer Science, 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]Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (CSIC), c/ José Gutiérrez Abascal 2, 28006 Madrid, Spain

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

Decision date: 28-May-2019

ABSTRACT

Species distribution models (SDMs) are broadly used to predict species distributions from available presence data. However, SDMs results have been criticized for several reasons mainly related to two basic characteristics of most SDMs: 1) general lack of reliable species absence information, 2) the frequent use of an arbitrary geographical extent (GE) or accessible area of the species. These impediments have motivated us to generate a procedure called Niche of Occurrence (NOO). NOO provides the probable distribution of species (realized niche) relying solely on partial information about presence of species. It operates within a natural geographical extent delimited by available observations and avoids using misleading thresholds to obtain binary presence-absence estimations when the species prevalence is unknown. In this study the main characteristics of NOO are presented, comparing its performance with other recognized and more complex SDMs by using virtual species to avoid the omnipresent error sources of real data sets.

Keywords: species distribution models, biodiversity, ecological niche, biogeography, macroecology, ecological modelling

Introduction

The exponential increase in the processing power of computers combined with the free availability of digital environmental layers and primary data about the distribution of species has propagated an overgrowth of species distribution models (SDMs). More than one study is published each day about this subject (Hortal et al. 2012) without a consensus about the general procedures that must be followed (see Franklin 2010, Peterson et al. 2011or Guisan et al. 2017). The basic practice followed in SDMs consists of relating available site occupancy data on species against environmental predictors using a diverse array of algorithms and general modelling procedures. This simple practice is similar to the classic one relating different predictors to a response variable obtained after a factorial standardized design. However, the spatially explicit probable or potential distributions derived in the case of SDMs use as response variables data coming from unstandardized and unplanned surveys, generally carried out by biologists who have not coordinated their efforts (Sastre and Lobo 2009, Guralnick et al. 2018).

SDM results have been criticized for three main reasons: (1) the effects of biases and the quality of occurrence information about species (Graham and Hijmans 2006, Amboni and Laffan 2012); (2) the correlative character of the functions relating the response and predictor variables (Kumar et al. 2014, Peterson et al. 2015); and (3) the inconsistency of the validation methods used to estimate the accuracy of the obtained results (Lobo et al. 2008, Hijmans 2012). These drawbacks are related to two basic characteristics of the large majority of SDM exercises: the general lack of reliable absence information on the distribution of species (Lobo et al. 2018) and the frequent use of an arbitrary geographical extent (GE) in the process of model building (Acevedo et al. 2017, Cooper and Soberón 2018). As a measure of survey effort is frequently not available, spatial units with reliable inventories cannot be discriminated and, as consequence, reliable absence of species cannot be distinguished (i.e., a locality identified as well-surveyed would harbour reliable absence when the species does not appear in it).

This general lack of reliable absence data has led to the use in SDMs of pseudo or background absence data selected at random from the study area under consideration. This is the classic procedure followed in the use-availability approach (Resource Selection Functions, Johnson 1980), in which the presence of species is related against a sample of points (background absence data), which enable one to represent the available conditions in the selected study area in order to estimate the environmental preferences of species. This procedure remains the general rule in most SDMs despite the fact that it cannot estimate the probability of occurrence (Hastie and Fithian 2013) and tends to lead a geographical representation that reflects the intensity of the data used in the given modelling process (Aarts et al. 2012). Resource Selection Functions are, however, very useful in detecting the environmental preferences of species. Nevertheless, these preferences are heavily dependent on the use of GE and, therefore, on the diversity of the environmental conditions existing in the study area. Thus, as both the location and number of “false” absence data points are highly dependent on GE (or *M* according to Soberón and Peterson 2005 or Soberón 2010), it is fundamental to select the appropriate area to estimate the predictors that can delimit the distribution of the species (Chefaoui and Lobo 2008, VanDerWal et al. 2009, Barve et al. 2011, Acevedo et al. 2012, 2017, Cooper and Soberón 2018). A function relating presence versus background absence data may produce different results depending on the area at which these “false” absence points are selected (Lobo et al. 2010) and the capacity of the used modelling algorithm to generate complex relationships (Iturbide et al. 2018). In addition, a large and inadequate GE may generate misleading accuracy assessments because the rate of well-predicted absences (specificity) is inflated as a consequence of correct absence predictions in regions far from those in which the species has been observed (Lobo et al. 2008, Hijmans 2012, Somodi et al. 2017).

As SDMs generate continuous suitability or probability values, it is necessary to apply a threshold to separate presence from absence data in order to subsequently build a confusion matrix with presence-absence validation data to compute different measurements of discrimination performance. The best threshold to minimizing model prediction errors is one that minimizes the difference between sensitivity (the ratio of correctly predicted presences to the total number of presences) and specificity (the ratio of correctly predicted absences to the total number of absences) (Liu et al. 2005, Jiménez-Valverde and Lobo 2007). However, due to the general lack of reliable absence data and the use of pseudo or background absence data, neither this threshold nor species prevalence can be calculated (i.e., the frequency of the species over the entire study GE). As an alternative, 1-specificity or the fraction of absences predicted as present (commission error) is substituted by the fraction of the total study area predicted present (Phillips 2017) when discrimination performance metrics are calculated. The consequence of this strategy to circumvent the problem of the lack of absence data is that model representations correctly predict as many as possible presences in the smallest possible area; that is, a predicted area inevitably very similar in shape and extent to a simple density function of the available presence data used in the training process. In the absence of reliable absence data an alternative approach consists of using the lowest predicted value associated with an observed presence or minimum training presence threshold (MTPT) as a threshold to infer species presence; a criterion ensuring that all the used presences are predicted as suitable (Pearson et al. 2007).

In short, (1) reliable absence data are not available for most organisms and situations, (2) presence/background-absence relationships are extent-dependent, and (3) without previous knowledge about the prevalence of species it is difficult to find the best way to transform in binary the continuous output values provided by SDMs. All of these impediments have motivated us to generate a new and simple modelling procedure able to provide geographical representations about the probable distribution of species from partial data, without aiming for predictions to go beyond the accessible area. This procedure (hereafter called NOO or Niche of Occurrence) is based solely on information about species presence as BIOCLIM, ENFA or DOMAIN do (Carpenter et al. 1993, Hirzel et al. 2001, Booth 2007) but including in its implementation both the selection of the predictors and the delimitation of the geographical extent to be used. NOO aims to represent geographically the realized niche (*sensu* Soberón 2010) operating within a natural GE delimited by the available observations, and, in turn, does not offer binary presence-absence estimations based on misleading thresholds acting on continuous predictions when the prevalence of the species is unknown. In this study, the basic characteristics of this procedure are presented, comparing its performance with other recognized and more complex modelling methods by using virtual species to avoid the omnipresent error sources of real data sets.

Methods

The NOO approach

Estimating the distribution of species by means of the NOO approach starts with the delimitation of the Extent of Occurrence (EOO). In the Appendix there is a full description of the algorithm and a step-by-step tutorial has been provided at the website http://www.ipez.es/modestr/Manual_Tutorial.html (tutorial 20). According to IUCN (2017), EOO is the area contained within the shortest continuous imaginary boundary which can be drawn to encompass all the known, inferred or projected sites of present occurrence of a taxon. In our approach, the delimitation of EOO is a first step directed to select the geographical extent (GE) of the species. EOO can be delimited in ModestR (García-Roselló et al. 2103, 2014) using a convex hull, an α-shape or a Kernel density distribution (see Tutorial 20 in http://www.ipez.es/modestr/Manual_Tutorial.html). However, all these area delimitations are based on the geometry of the available occurrences and may not strictly reflect natural units determining the isolation and connectivity of local populations (i.e., the accessible area). Drainage basins constitute, on the contrary, natural spatially self-organized systems (Rodríguez-Iturbe et al. 2011), bounded topographically, and composed by hierarchical sub-basins which share a common history and constitute a network (O' Keefe et al. 2012). River basins may also be used in NOO with ModestR to delimit the GE for each species inhabiting terrestrial environments. Thus, the set of river basins with presence observations that, in turn, enable the connection of all the available occurrences (all the selected basins must be connected), is the default option in NOO to delimit the GE. The watershed information provided by the WaterBase project (www.waterbase.org) was used for this purpose which includes a hierarchical coding system to recognize river basins of different levels. This dataset of shapes with the information of water basins was introduced in ModestR (González-Vilas et al. 2106), and it is used in the estimation on NOO to select the minimum level of river basins with occurrences that generate a contiguous and connected area.

The following step in the NOO procedure consists of selecting the preferred predictor environmental variables within the GE able to account for the presence of the target species. This purpose is accomplished by first eliminating redundant predictors to subsequently select which of these explanatory variables better discriminate the conditions in the presence localities from those available in the GE. Redundancy among variables is assessed by sequentially deleting predictors using the variance inflation factor (VIF) (Fox and Weisberg 2011). VIF values higher than 30 (default) are considered as those indicating high collinearity among the explanatory variables, but this VIF value can be modified by the user at convenience. The selected predictors were subsequently submitted to a recently proposed Instability Index (Guisande 2016, Guisande et al. 2017) that does not require normalized data. Dividing each predictor into a number of intervals or bins decided by the user, the number of records in each bin was calculated considering separately the cells where the species occurs and those of the selected studied area. A peak of instability is observed when there are important differences in the predictor comparing the bins of presence with the corresponding ones of the study area. This index outperforms other methods proposed to identify the most appropriate environmental factors (Guisande et al. 2017, Fan et al. 2018). The explanatory variables with a higher percentage of contribution to the Instability Index are assumed to be those that most affect the distribution of the species in the accessible area or GE. In order to include only the variables with a higher contribution, it is possible to select an accumulated percentage of contribution (the default option is 80% but see Appendix in Supplementary Material) so that if 100% is selected all variables will be included.

Once the most important environmental variables affecting the distribution of each species in their accessible area have been identified, NOO generates a “compounded environmental layer” (CEL) by using polar coordinates (Guisande et al. 2011). Polar coordinates allow representing in a two-dimensional coordinate system the data coming from any number of different variables by assigning to each an angle and a distance from the centre according to their positive or negative value. The default order or angle of the variables is established in NOO by calculating the correlation matrix between them so that each variable will be followed by the one to which it is most highly correlated (this order can be modified by the user). The polar coordinates system is defined after standardizing the environmental variables from 1 to 2, and the CEL is finally generated by computing the values of all the selected environmental variables in each cell of the considered GE. These compounded layers can be displayed both geographically and in a two-dimensional polar coordinates system, but also can be stored for later use. The purpose of carrying out these CELs is to obtain in a single file all the information about the relevant predictors indicating the preferences of the species in the selected GE. Thus, selecting a CEL inherently implies selecting the values of the specific variables used to build the CEL at the desired extent.

Subsequently, the occurrence cells of the target species are projected onto the CEL and a kernel density estimation is calculated on this CEL reflecting the intensity of used presence data in the environmental space. The values of the environmental variables constituting the CEL, ranging from maximum to minimum density values attained in species presences (*Minimal Density at Presence* default option), are retained and projected geographically. As a consequence, all the cells with environmental conditions similar to those existing in the occurrence localities are discriminated along the previously selected GE. The final output can be a binary (*Distribution map* option) reflecting suitable and unsuitable cells, or a continuous map (*Density map* option) reflecting the quantity of presence observations in each set of environmental conditions within the GE. In this process, the user must select a smoothing factor and a tolerance value. The smoothing factor is the bandwidth or deviation of the smoothing kernel, so that greater values lead to a greater spread around presence samples (see Fig. 1). Tolerance allows for expanding the maximum and minimum found values of the used predictors in the presence cells. Previous tests using virtual species (see Appendix in Supplementary Material) indicate that a smoothing factor of 1 and a tolerance of 1% provide general good results and are used as default parameters.

Virtual species

The 19 bioclimatic variables of WorldClim at a resolution of five arc-minutes (Hijmans et al. 2005) were used to generate a virtual species in Europe (from 34º to 72º in latitude, and from -11º to 50º in longitude; n = 171,657 cells). Iteratively applying VIF on these variables and discarding all those with a VIF value higher than 5 as recommended (Hair et al. 2014) allowed us to ultimately retain four bioclimatic variables (mean diurnal range, mean temperature of wettest quarter, precipitation seasonality and precipitation of warmest quarter). These variables were subsequently used to generate the distribution of the virtual species by using the *virtualspecies* R package (Leroy et al. 2016), which allows one to produce a continuous suitability map according to a response function for each bioclimatic variable. Thus, three virtual species were generated differing in their prevalence values (0.1, 0.5 and 0.9), which are completely conditioned by the formerly selected bioclimatic variables (see Fig. 2). The final continuous suitability maps obtained in this way were scaled between 0 and 1 and converted in binary (1/0) using as thresholds 0.9, 0.7 and 0.5 values (see Table 1). Different prevalence values were used to examine if the predictive capacity of the different modelling procedures varies with environmental tolerance and the considered distribution range of the species.

SDMs on the virtual species

From each one of these three virtual species, four percentages of total presence (very low = 0.1%; low = 1%; medium = 5%; and high = 10% of total) were randomly selected and the process was repeated five times. Use of differing numbers of presence data aimed to examine the capacity of the different modelling methods to generate accurate predictions with differing amounts of data. In all cases the MTPT was used as the threshold to convert to binary the continuous predictions, thus simulating the common situation in which there is lack of absence data. Thus, a total of 60 sets of data were used (3 prevalence values x 4 percentages of presences x 5 repetitions). When required, an equal number of background absence cells was used in each model, corresponding to 5% (n = 8,583) of the total number of cells in the study region (n = 171,657 cells), thus simulating the common situation in which true absences are unknown (see Table 1).

The results provided by NOO using default options (see Appendix in Supplementary Material) were compared with those of seven different modelling techniques frequently used in SDMs. Thus, both profiling models using only information about presence (BIOCLIM and DOMAIN), and regression (GLM and GAM) or machine learning methods (MaxEnt, Boosted Regression Trees-BRT, and Support Vector Machines-SVM) requiring the use of background absences were computed. The *dismo*, *mgcv*, *glm,randomForest, gbm* and *e1071* R packages were used for this purpose. See Guisan et al. (2017) for a more detailed description of each of these methods. In total, 480 models were carried out (8 modelling techniques x 60 sets of data).

Evaluation

Evaluation of the predictive performance of SDMs is usually based on discrimination measures able to assess the difference between output-derived predictions and more or less independent presence-absence data (Fielding and Bell 1997). In our case, the confusion matrix and the performance metrics were calculated by comparing the modelled binary predictions against the entire presence-absence data of the virtual species. Four different metrics were used to estimate the performance of the predictions: sensitivity, specificity, true skill statistics (TSS) and AUC (see Fielding and Bell 1997). Of these, AUC is the only one that is not threshold dependent. AUC summarized model performance independently of the used threshold by plotting sensitivity against 1-specificity (false positive rate) over a number of thresholds (100 in this case), and the area under the receiver operating characteristic (ROC) curve, or AUC, was calculated. AUC is non-dependent on a single threshold and thereby considered a more reliable metric to compare results from different SDMs (but see Lobo et al. 2008 and Jiménez-Valverde 2014).

Statistic treatment of the data

The variation in the four performance metrics (sensitivity, specificity, TSS and AUC) was analysed by means of General Linear Models using an ANOVA design and type III sum of squares (i.e., estimating the partial effects of each factor while controlling for the effects of the remaining predictors). Species prevalence (*Prev*, three levels), percentage of presence data points (%*P*, four levels), and the modelling methods (M; eight levels) were the factors selected to assess the effects of the prevalence of the species, the quality of the used information (percentage of presences) and the modelling procedure. A full factorial model was carried out considering the three two-way interactions and the one three-way interaction among the three considered factors.

Results

The modelling method used is by far the factor with a higher capacity to explain variation in the different performance metrics when considered individually, accounting for from 23% to 50% of total variability (Table 2). Except for NOO, all of the remaining modelling methods offer high rates of success in predicting presence (sensitivity) (Fig. 3). Thus, selecting the lowest suitability value in a presence (MTPT) makes it possible to guess the total presence of the virtual species very well. However, the general success in predicting presence comes at the expense of a malfunction in predicting absence; model predictions generally are so wide that many absence areas are predicted as presence areas (commission errors). However, the NOO procedure is the only modelling method that does not follow this pattern since both commission and omission errors are relatively balanced (Fig. 3). Although some presence is erroneously predicted as absence, NOO includes a low number of erroneous absence predictions thereby providing a not too wide geographic representation.

The former general pattern varies according to the prevalence of the species and the percentage of presence used as indicated by the statistically significant three-way interaction *M* x *Prev* x *%P* (Table 2). Thus, NOO generates comparatively misleading predictions of the complete range of the virtual species when the species possesses a low prevalence and very few presence points are used to train the model (Fig. 4); i.e., when the data used in NOO poorly represent the distribution of the species. On the contrary, absences start to be comparatively well-predicted when the number of used presence data points increase and/or the prevalence of the species is medium or high (Fig. 4).

Discussion

We propose here a procedure especially suited to generate probable distributions of species when there is no reliable information about absence, the prevalence of species is unknown, the user is aware of the drawbacks generated by the use of background absence data, and there is no intent for predictions to go beyond the natural area in which the species has been observed (Jarnevich et al. 2015, Lobo 2016). Thus, our approach is especially suited in situations in which transferability is not a concern. Nevertheless, NOO would allow selecting in these cases (invasive species, climate change, etc.) the predictors in the accessible area and transferring the predictions to any desired region (see NOO manual in http://www.ipez.es/modestr/Manual_Tutorial.html). As many SDMs procedures already exist, the first question that may arise is whether a new method is necessary or at least relevant. By "procedure" we understand not only the modelling algorithm, but also other important characteristics or choices as the lack of "true" absence data, the refusal to use background absences, the need of a cautious selection of predictors, and the generation of predictions limited to the most accessible area. Among the different methodological choices that need to be made when we aim to predict the distribution of a species from partial data using a correlative approach, NOO is a simple method characterized by i) using only the available empirical evidence that exist in most cases (presence observations), ii) the mandatory selection of the predictors, and iii) the need to clearly delimit the accessible area over which the prediction will be carried out. The simplicity of NOO is shared with other applications as BIOCLIM, ENFA or DOMAIN (Carpenter et al. 1993, Hirzel et al. 2001, Booth 2007) but in this case the advantage of NOO is that the user can previously select the most convenient predictors within a reasonable and accessible area as an obligate prerequisite. Thus, the procedure proposed here aims to generate species distributions probably limited to the areas accessible to the species over relevant periods of its history, in agreement with the *M* concept and the terminology of the *BAM* diagram (Soberón and Peterson 2005, Cooper and Soberón 2018; Barve et al. 2011). As consequence, our proposed approach is particularly appropriate for offering distributional hypotheses about the distribution of terrestrial organisms not exhaustively surveyed as invertebrates in which species prevalence is generally completely unknown and do not exist reliable absence data (Guillera-Arroita et al. 2015).

Any SDM relies on a set of assumptions being inherently biased towards certain conditions. Therefore, different SDMs will exhibit better or worst performances depending on the context and quality of the data used. Any SDM procedure clearly providing better performance than existing ones in some contexts can be a useful contribution to the panoply of methods at the disposal of the scientific community. In this work we provide not only a description of a new approach to estimate species distributions, called NOO, but also a comparative analysis identifying its advantages and drawbacks regarding several of the most used SDM procedures. NOO allows obtaining relatively reliable predictions when the prevalence of the species is unknown and therefore it is impossible to estimate the most adequate threshold to transform continuous output values in binary ones. In these situations, NOO is able to provide a well-balanced rate of commission and omission errors, indicating its usefulness and superiority under certain conditions. However, NOO can provide wrong predictions when the target species has a low prevalence and the available training data are scarce. The delimitation of GE by using those river basins enabling the connection of all the available occurrences can provide excessively restricted accessible areas when few data are available, thus inflating prediction errors.

A supplementary advantage of NOO that can be claimed is its simplicity, which has been praised as a desirable quality of an SDM (Lobo 2016). Many of the most used

SDMs are based on machine learning techniques (maximum entropy, neural networks, SVM, etc.) working on presence-background absence data. Even if very powerful, these approaches are complex and black-box biased, a commonly reported drawback of this type of techniques (Ribeiro et al. 2016). In contrast to these black-box approaches, NOO has the advantage of relying on a simple and intuitive principle: the more a species presence is stronger under some environmental conditions, the more likely it can be also present under similar conditions within the same natural region in which it was observed.. Moreover, it is very easy to visualize this closeness in a polar coordinates graph such as that provided by the NOO implementation available in the ModestR software (see Appendix in Supplementary Material). This simplicity and understandability are clear advantages for the user when following the rationale of each step in the procedure and interpreting results.

To compute this closeness in the environmental space NOO uses a simple kernel density estimation on a polar coordinates system that provides a two-dimensional representation of a multidimensional environmental space. A true multidimensional approach using hypervolumes and multidimensional kernel density has been proposed (Blonder et al. 2014). But due to its multidimensional nature it is inherently more complex to use and interpret. Moreover, even if growing computational power makes this procedure feasible, it remains more expensive as soon the number of occurrences and dimensions increase. Further, it can be hypothesized that the superiority of this approach may not be significant when using a small number of dimensions, which can be done in most cases by previously selecting the most relevant variables predicting a species distribution.

Unlike most SDMs (with some exceptions, i.e., Blonder et al. 2014), NOO does not use pseudo or background absence data. This deliberate choice responds to a too often undervalued reality: the common lack of reliable absence data (Lobo et al. 2010). Despite the fact that most SDMs use pseudo or background absence data, it has been proven that this approach prevents one from accurately predicting the overall species occurrence probability (Hastie and Fithian 2013) and tends to provide misleadingly reliable results (Aarts et al. 2012). In addition to using SDMs that do not present this flaw, part of the solution to this problem will involve correctly assessing the survey efforts already done as those are needed to obtain more reliable presence and absence data (Lobo et al. 2018).

Another common problem of SDMs is the choice of a threshold needed to transform continuous probability or suitability values into the presence-absence binary variable needed for many further uses and performance measurements. An SDM may claim that it uses an ROC-based optimum threshold relying on pseudo or background absence data, supposed to minimize the difference between sensitivity and specificity. But to be congruent with avoiding the use of pseudo or background absence data, NOO does not apply this solution because this threshold is only reliable when real absence data are available to determine the ROC curve (Lobo et al. 2008, Jiménez-Valverde 2012), which is rarely the case. Therefore, NOO uses MTPT, which guarantees that all known presences are predicted as suitable (Pearson et al. 2007). However, this choice can be misleading as it requires an SDM that provides a good balance between sensitivity and specificity (i.e., between omission and commission errors), two qualities that are both desirable but frequently opposed, as increasing one of them can often only be done to the detriment of the other. A high threshold value will generate a SDM with a high sensitivity but a low specificity leading to higher commission errors, and vice versa. Therefore, we think that the balanced behaviour of NOO as shown in the comparative tests (see Appendix in Supplementary Material) is even more significant from this perspective. However, using MTPT prevents making easy adjustments of the threshold to adapt the model response to contexts in which omission/commission errors are not considered equally important and it would be desirable to specifically decrease one of them (Fielding 2002).

NOO also allows delimiting the GE, or accessible area, of the species using different techniques: convex hull, α-shape, Kernel density or minimal contiguous river basins where there are confirmed species occurrences. We think that this last one is a better approach, as it reflects natural accessible areas (Rodríguez-Iturbe et al. 2011), while the other options are based on simple geometrical procedures that do not take into account the terrain morphology that can limit or facilitate species dispersal. As far as we know, and despite its clear importance as a parameter of some widely used SDM (see for example Merrow et al., 2013), no other SDM currently includes this feature as an standard procedure, requiring in most cases the pre-processing of environmental data or the post-processing of SDM results (Cooper and Soberón 2018) in order to force the model to be circumscribed to a specific GE and avoid having unrealistic geographical representations of the realized niche. It can be argued that this use of GE is more difficult with marine species. Some studies showed that simple methods such as α-shape seem to provide better results than more complex approaches (García-Roselló et al. 2015*)*. Therefore, they may be used to define a GE under marine conditions, but more studies are needed to evaluate the performance of NOO in this context.

The tests (see Appendix in Supplementary Material) using virtual species allow comparing NOO with other SDM methods in conditions of nearly-perfect information (presence, absence, prevalence, and environmental data are all known). The results show that NOO performance is comparable to other widely used SDMs. It does not exhibit a high sensitivity, but it stands out particularly by its high specificity. That is, NOO predicts absences better than any other of the tested SDMs. It must be noted in this sense that all tests have been done using SDMs default parameters, a widespread practice that has been sometimes criticized as potentially inadequate in certain contexts, as it may lead to overestimation or overfitting in distribution predictions (Morales et al. 2017). Therefore, we think that this NOO specificity "by default" is a very valuable quality. Nonetheless, according to our tests about how the smoothing (bandwidth) parameter affects the NOO result, it seems clear that sensitivity may be increased if needed just by increasing bandwidth, but at the expense of specificity.

Acknowledgements –This study has been partially supported by the project BIOWEB (CGL2011-15622-E BOS) founded by Spanish Ministerio de Ciencia e Innovación, and by the projects BANDENCO (POII11-0277-5747) and IBERARTRO (SBPLY/17/180501/000492) founded by the Consejería de Educación, Ciencia y Cultura-Junta de Comunidades de Castilla-La Mancha.

References

Aarts, G. et al. 2012. Comparative interpretation of count, presence-absence and point methods for species distribution models. – Methods Ecol. Evol. 3**:**177-187

Acevedo, P. et al. 2012. Delimiting the geographical background in species distribution modelling. – J. Biogeogr. 39**:**1383–1390.

Acevedo, P. et al. 2017. Predictor weighting and geographical background delimitation: two synergetic sources of uncertainty when assessing species sensitivity to climate change. – Clim. Change 145**:** 131-143.

Amboni, M. P. and Laffan, S. W. 2012. The effect of species geographical distribution estimation methods on richness and phylogenetic diversity estimates. – Int. J. Geogr. Inf. Sci. 26**:** 2097–2109.

Barve, N. et al. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. – Ecol. Model. 222: 1810–1819.

Blonder, B. et al. 2014. The n-dimensional hypervolume. – Global Ecol. Biogeogr. 23: 595-609.

Booth, T. H. 2018. Why understanding the pioneering and continuing contributions of BIOCLIM to species distribution modelling is important. – Austral Ecol. 43: 852-860.

Carpenter, G. et al. 1993. DOMAIN: a flexible modelling ptrocedure for mapping potential distributions of plants and animals. – Biodiv. Conserv. 2: 667-680.

Chefaoui, R. and Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – Ecol. Mod. 210: 478-486.

Cooper, J. C. and Soberón, J. 2018. Creating individual accessible area hypotheses improves stacked species distribution model performance. – Global Ecol. Biogeogr. 27:156–165

Fan, J. Y. et al. 2018. What are the best predictors for invasive potential of weeds? Transferability evaluations of model predictions based on diverse environmental data sets for *Flaveria bidentis*. – Weed Res. 58: 141–149.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Fielding, A. H. 2002. What are the appropriate characteristics of an accuracy measure? – In: Scott, J. M et al. (eds.), Predicting Plant and Animal Occurences: Issues of Scale and Accuracy. Island Press, pp. 271-280.

Fox, J. and Weisberg, S. 2011. An R Companion to Applied Regression. – Sage.

Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. – Cambridge University Press.

García-Roselló, E. et al. 2013. ModestR: a software tool for managing and analyzing species distribution map databases. – Ecography 36:1202–1207.

García-Roselló, E. et al. 2014. Using ModestR to download, import and clean species distribution records. – Methods Ecol. Evol. 5: 703–713.

García-Roselló, E. et al. 2015. Can we derive macroecological patterns from primary Global Biodiversity Information Facility data?. – Global Ecol. Biogeogr. 24: 335-347.

González-Vilas, L. et al. 2016. Geospatial data of freshwater habitats for macroecological studies: an example with freshwater fishes. – Int. J. Geogr. Inf. Sci. 30: 126–141.

Graham, C. H. and Hijmans, R. J. 2006. A comparison of methods for mapping species ranges and species richness. – Global Ecol. Biogeogr. 15: 578-587.

Guillera-Arroita, G. et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. – Global Ecol. Biogeogr. 24: 276–292

Guisan, A. et al. 2017. Habitat suitability and distribution models. – Cambridge University Press.

Guisande, C. et al. 2011. Tratamiento de Datos Con R, Statistica y SPSS. – Díaz de Santos.

Guisande, C. 2016. Estimation of the relative importance of factors affecting species distribution based on stability concept. R package version 1.3. Available at: http://CRAN.R-project.org/package/SPEDInstabR (accessed 6 November 2016).

Guisande, C. et al. 2017. SPEDInstabR: An algorithm based on a fluctuation index for selecting predictors in species distribution modelling. – Ecol. Inform. 37: 18-23.

Guralnick, R. et al. 2018. Humboldt Core – toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. – Ecography 41: 713-725.

Hair, J. F. et al. 2014. Multivariate Data Analysis. – Pearson Education Limited.

Hastie, T. and Fithian, W. 2013. Inference from presence-only data; the ongoing controversy. Ecography 36: 864-867.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.

Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting and calibration with a null model bias. – Ecology 93: 679-688.

Hirzel A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – Ecol. Mod.145: 111–121.

Hortal, J. et al. 2012. Basic questions in Biogeography and the (lack of) simplicity of species distributions: Putting Species Distribution Models in the right place. – Nat. Conservacao 10: 108-118.

IUCN. 2017. Guidelines for Using the IUCN Red List Categories and Criteria. Version 13. IUCN Standards and Petitions Subcommitee. Available at: http://www.iucnredlist.org/documents/RedListGuidelines.pdf (accessed 6 March 2018).

Iturbide, M. et al. 2018. Background sampling and transferability of species distribution model ensembles under climate change. – Global Planet. Change 166: 19-29.

Jarnevich, C. S. et al. 2015. Caveats for correlative species distribution modeling. – Ecol. Inform. 29: 6–15.

Jiménez-Valverde, A. and Lobo, J.M. 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence. – Acta Oecol. 31: 361-369.

Jiménez-Valverde, A. et al. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. – Divers. Distrib. 14: 885-890.

Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. – Global Ecol. Biogeogr. 21: 498-507.

Jiménez-Valverde, A. 2014. Threshold-dependence as a desirable attribute for discrimination assessment: implications for the evaluation of species distribution models. – Biodivers. Conserv. 23: 369-385.

Johnson, D. H. 1980. The comparison of usage and availability measurements for evaluating resource preference. – Ecology 61: 65-71.

Kumar, S. et al. 2014. Evaluating correlative and mechanistic niche models for assessing the risk of pest establishment. – Ecosphere 5: 1-23.

Leroy, B. et al. 2016. Virtualspecies, an R package to generate virtual species distributions. – Ecography 39: 599-607

Liu, C. et al. 2005. Selecting thresholds of occurrence in the prediction of species distributions. – Ecography 28: 385–393.

Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.

Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.

Lobo J. M. 2016. The use of occurrence data to predict the effects of climate change on insects. – Curr. Opin. Insect. Sci. 17: 62–68.

Lobo, J. M. et al. 2018. KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. – Ecol. Indic. 91: 241-248.

Morales, N. S. et al. 2017. MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. – PeerJ 5:e3093.

O' Keefe, T. C. et al. 2012. Introduction to Watershed Ecology. – In: Environmental Protection Agency, Watershed Academy Web Documents, pp. 1-37.

Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102-117.

Peterson, A. T. et al. 2011. Ecological Niches and Geographic Distributions. – Princeton University Press.

Peterson, A. T. et al. 2015. Mechanistic and correlative models of ecological niches. – Eur. J. Ecol. 1: 28-38.

Phillips, S. J. 2017. A Brief Tutorial on Maxent. Available at: http://biodiversityinformatics.amnh.org/open_source/maxent/. (Accessed April, 10, 2018).

Ribeiro, M. T. et al. 2016. Model-Agnostic Interpretability of Machine Learning. – Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, New York.

Rodríguez-Iturbe, I. et al. 2011. Metabolic principles of river basin organization. – Proc. Natl. Acad. Sci. USA 108: 11751-11755.

Sastre, P. and Lobo, J. M. 2009. Taxonomist survey biases and the unveiling of biodiversity patterns. – Biol. Conserv. 142: 462-467.

Somodi, I. et al. 2017. Prevalence dependence in model goodness measures with special emphasis on true skill statistics. – Ecol. Evol. 7: 863–872.

Soberón, J. 2010. Niche and area of distribution modeling: a population ecology perspective. – Ecography 33: 159–167.

Soberón, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. – Biodiversity Informatics 2: 1–10.

VanDerWal, J. et al. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? – Ecol Model 220: 589–594.

Figure Legends

Figure 1. Representation of the differences in the density with smoothing values of 1, 2 and 3 (from left to right).

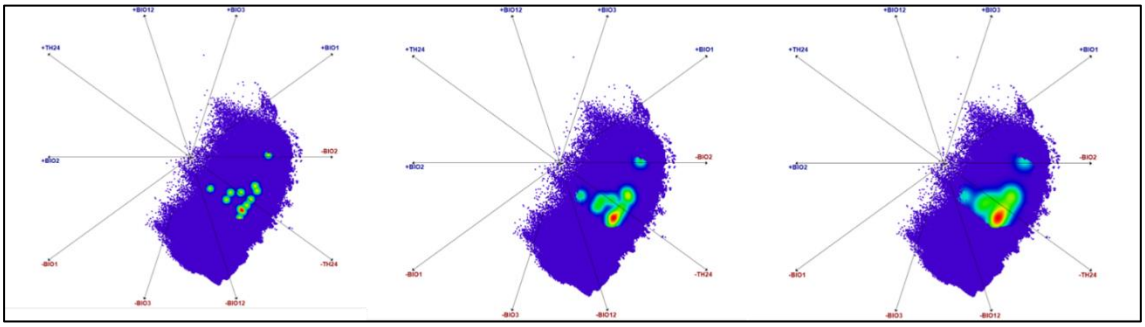

Figure 2. Continuous suitability values (left; from red-high to low-blue) of the virtual species and binary maps representing the presence (in red) and absence (in blue) of the species with a high prevalence (A; prevalence = 0.9), medium prevalence (B; prevalence = 0.5), and low prevalence (C; prevalence = 0.1).

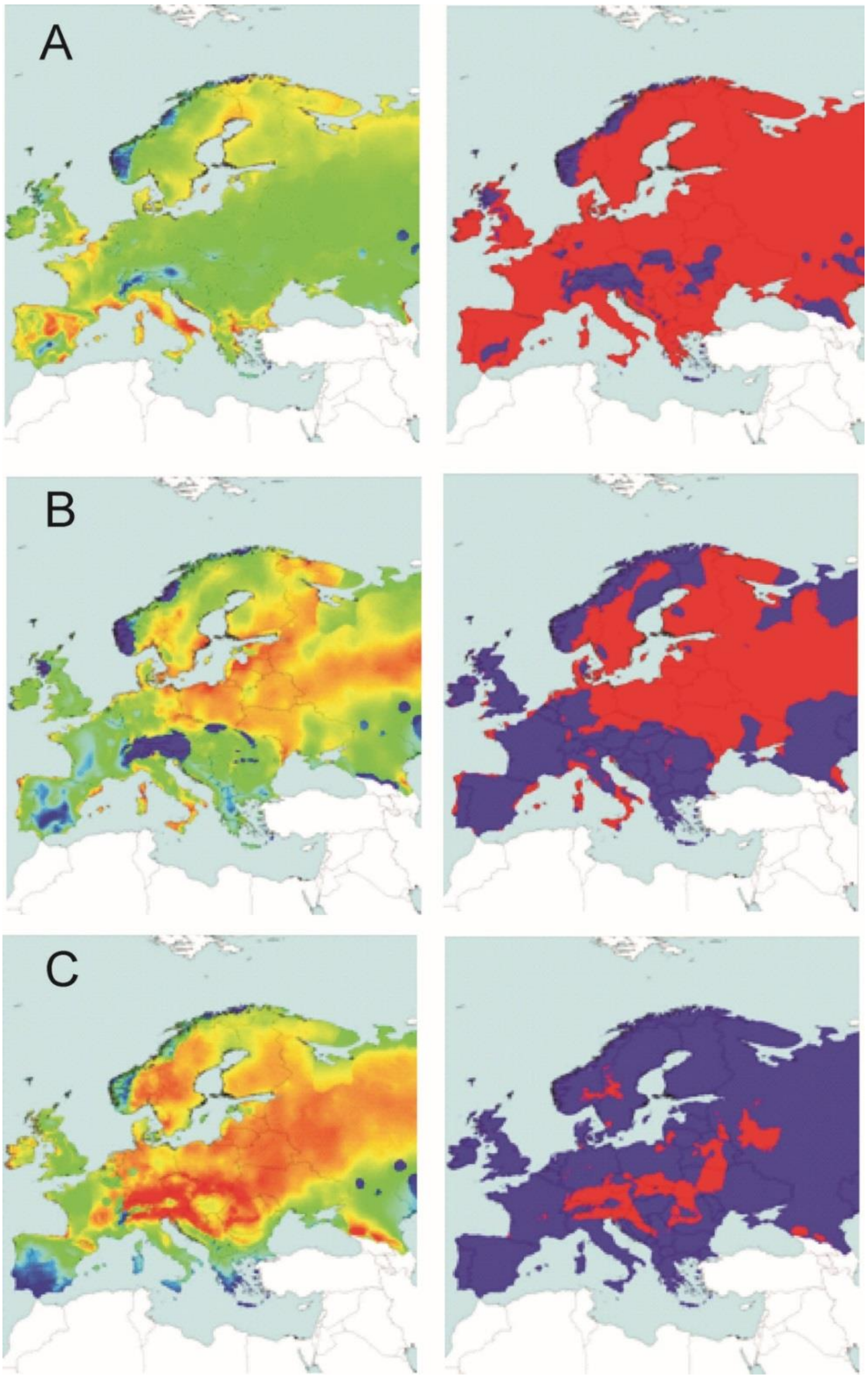

Figure 3. Variations in the four considered performance metrics (red circles = AUC; yellow squares = sensitivity; blue triangles = specificity; stars = TSS; ± 95% CI) among the eight different modelling techniques. The values are partial regressions representing the effects of each modelling technique controlling for the effects of the other two factors (species prevalence and number of presence data points used in model training). Performance metrics are calculated against the complete “true” presence-absence data of virtual species.

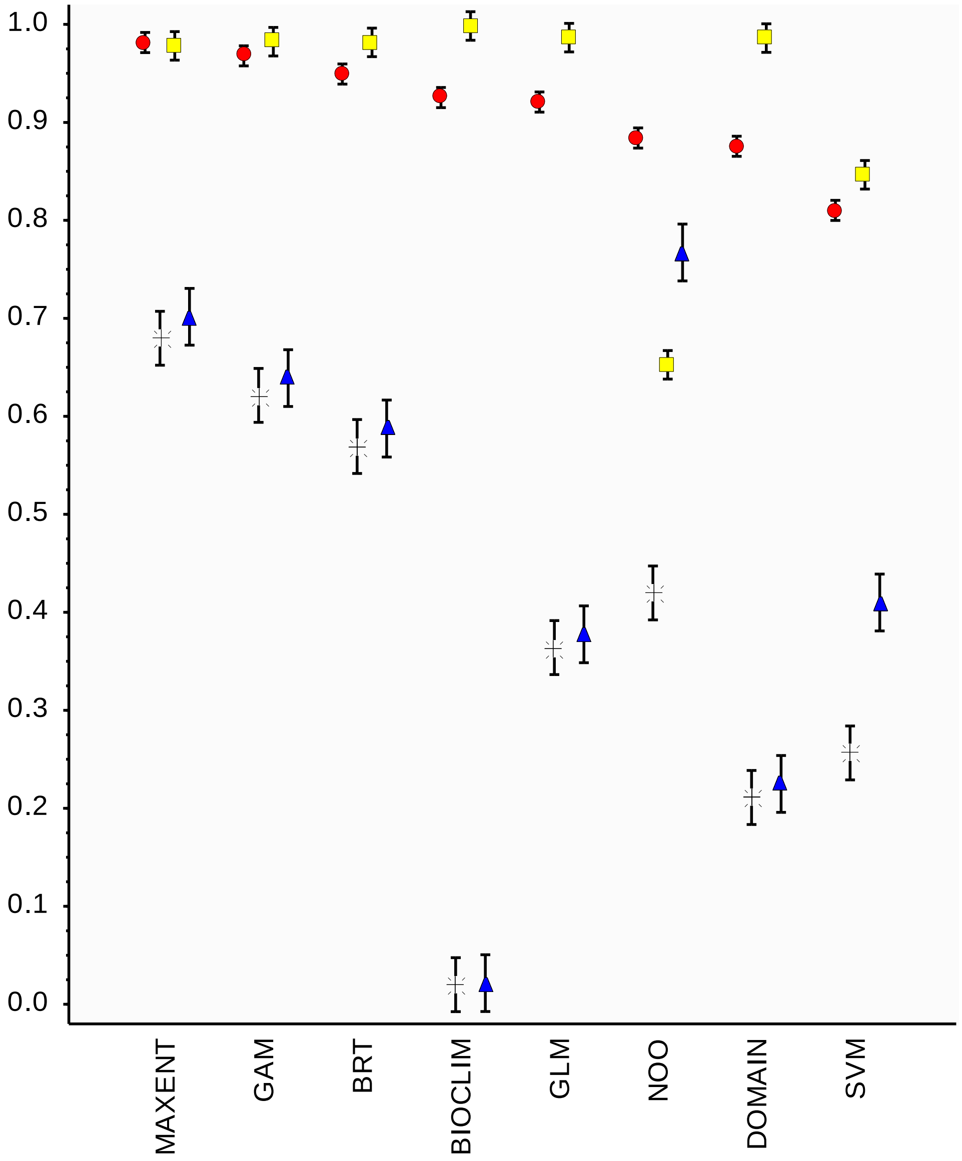

Figure 4. Effect of the modelling method on the four used performance metrics according to the percentage of presence data points used in model training (0.1%, 1%, 5% and 10%) and the prevalence of the virtual species (low = 0.1; medium = 0.5; and high = 0.9). AUC = red points; sensitivity = green points; specificity = yellow points; TSS = blue squares.

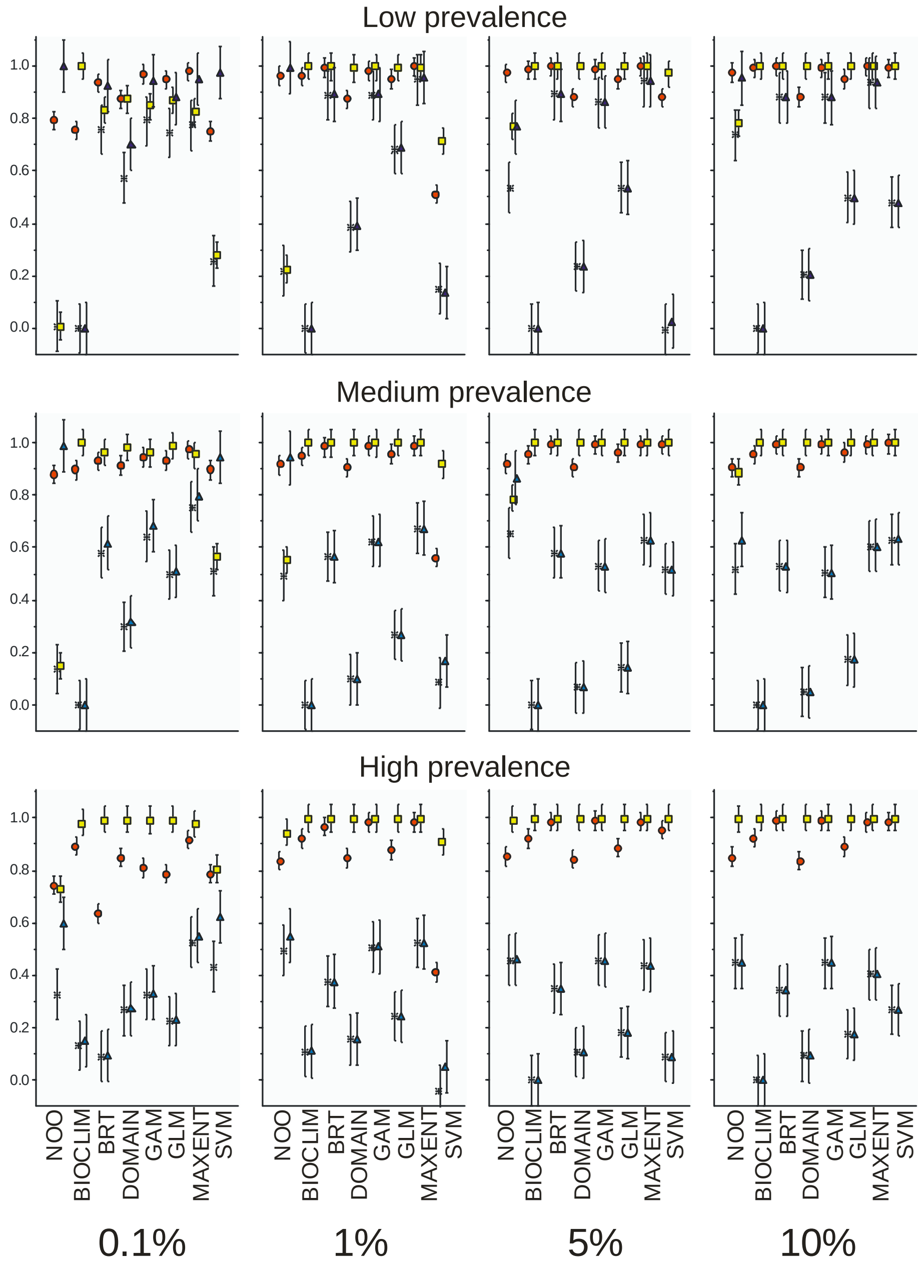

Table Legends

Table 1. Prevalence values (Prev) of the virtual species, threshold (Thres) used to convert to binary the continuous suitability values (from 0 to 1), and number of presence data points (NP) and absence data points (NA) of the generated virtual species. For each virtual species four percentages (%P) of the total number of presence points were selected and these numbers were used in the modelling process (NP_{MOD}). When required, the number of background absence data points ($NBA_{MOD;}$ n = 8,583) was defined as 5% of the total cells (n = 171,657).

Prev	Thres	NP	NA	%P	NP_{MOD}	NBA_{MOD}
0.9	0.5	153,512	18,145	0.10%	154	8,583
				1%	1,535	
				5%	7,676	
				10%	15,351	
0.5	0.7	86,791	84,866	0.10%	87	8,583
				1%	868	
				5%	4,340	
				10%	8,679	
0.1	0.9	16,372	155,285	0.10%	16	8,583
				1%	164	
				5%	819	
				10%	1,637	

Table 2. Statistically significant *F* values of the relationships between the four used performance metrics and the three considered factors related with the species prevalence (*Prev*), the percentage of the presence data points used in the model (*%P*), and the employing modelling method (*M*). Only those relationships statistically significant with a $p \leq 0.0001$ are shown. The percentage of explained variability is showed in brackets. General Linear Models with type III sum of squares are used, so that the explained variability of each factor is the one estimated as controlling for the effects of the remaining predictors.

	AUC	**Sensitivity**	**Specificity**	**TSS**
Prev	94.9 (5.4%)	133.9 (5.1%)	347.8 (16.5%)	183.2 (9.9%)
%P	144.0 (12.3%)	229.5 (13.2%)	63.4 (4.5%)	
M	115.3 (22.9%)	267.0 (35.7%)	300.2 (49.7%)	261.0 (49.2%)
M x Prev	7.0 (2.8%)	40.7 (10.9%)	27.1 (9.0%)	37.3 (14.1%)
M x %P	62.9 (37.5%)	48.8 (19.6%)	16.4 (8.2%)	21.0 (11.9%)
Prev x *%P*	10.2 (1.7%)	31.2 (3.6%)	6.0 (0.8%)	4.3 (0.7%)
Prev x *%P x M*	5.5 (6.6%)	35.6 (4.5%)	2.2 (2.2%)	3.4 (3.8%)