Ecography

**ECOG-04563**
García-Roselló, E., Guisande, C., González-Vilas, L., González-Dacosta, J., Heine, J., Pérez-Costas, E. and Lobo, J. M. 2019. A simple method to estimate the probable distribution of species. – Ecography doi: 10.1111/ecog.04563

## Supplementary material

**Appendix 1**

The most appropriate percentage of contribution of predictors *(*%*C*) using the Instability Index (Guisande *et al*. 2017), the best smoothing value of the kernel density function (*Sm*) for the accomplishment of the compounded environmental layer (CEL), and the most appropriate percentage of tolerance (%*T*) to expand maximum and minimum found environmental values of the predictors in the presence cells were selected to include them in NOO as default options. The same three virtual species mentioned in this study were used for this purpose (five repetitions). Different contribution percentages may affect the predictions by including variables not directly related to the distribution of the species. The values selected for %*C* were 70%, 80% and 90%. *Sm* values oscillated from 1 to 11, while %*T* values varied from 0 to 5. Performance metrics (sensitivity, specificity and AUC) were calculated comparing predictions against "true" presence-absence maps of the virtual species. Continuous predictions were transformed into binary ones by using the minimum training presence threshold. The values of these three metrics were related to the three factors (%*C*, *Sm*, and %*T*) by means of General Linear Models using a 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) and percentage of presences (%*P*, four levels) were also included in these analyses as supplementary factors. In total, 180 cases were used to examine the influence of the percentage of contribution of predictors (3 *Prev* x 4 %*P* x 3 %*C* x 5 repetitions), 660 cases were used to examine the influence of smoothing values (3 *Prev* x 4 %*P* x 11 *Sm* x 5 repetitions), and 360 cases were used to assess the influence of the tolerance to expand maximum and minimum found environmental values (3 *Prev* x 4 %*P* x 6 %*T* x 5 repetitions). All two-way and three-way interactions between the target factor (%*C*, *Sm* or %*T*) and the two supplementary factors (*Prev* and %*P*) were included in the model to

assess whether the best option to be selected in the NOO procedure could depend on the geographic range and/or environmental tolerance of the species and the quantity of the used data.

The percentage of contribution of predictors (%*C*) using the Instability Index do not significantly affect AUC or specificity values, but it does affect sensitivity ($F_{(1,144)}$ = 4.9; $p$= 0.009) although the explanatory capacity barely reaches 1%. The two-way and three-way interactions affecting *%C* are not statistically significant in the case of AUC or specificity. However, the three-way interaction including %*C*, *Prev* and %*P* is statistically significant ($F_{(1,144)}$ = 2.3; $p$ = 0.01) indicating that the rate of success in predicting presences is similar at medium or high prevalence values. However, an 80% of predictor contribution is preferable when the species have a low prevalence and the number of used presences increases (very low = 16; low = 164). (see Fig. A1).

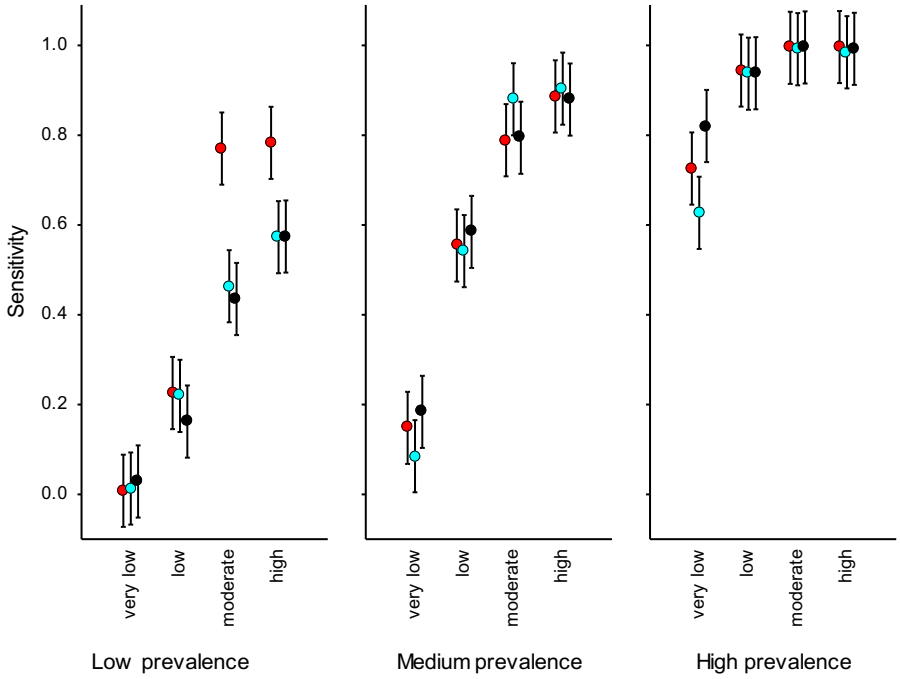



**Figure A1**.- Effect of the percentage of contribution of predictors *(%C)* using the Instability Index (Guisande *et al*. 2017) on sensitivity metrics (± 95% CI) according to the four levels of the percentage of presences used in the model training and the three levels of the species prevalence. Blue circles = 70%; red circles = 80%; black circles = 90%.

The variation in smoothing values of the kernel density function (*Sm*) directed to carry out the compounded environmental layer significantly influence the three performance metrics: AUC ($F_{(10, 528)} = 2.1$; $p = 0.02$; $R^2 = 1.0\%$), sensitivity ($F_{(10, 528)} = 251.4$; $p < 0.0001$; $R^2 = 18.5\%$), and specificity ($F_{(10, 528)} = 17.9$; $p < 0.0001$; $R^2 = 2.9\%$). In the case of AUC, only the interaction *Prev* x *Sm* is statistically significant ($F_{(20, 528)} = 2.0$; $p = 0.006$), indicating that the increase in the smoothing factor negatively effects AUC values when the species prevalence is high (Fig. 2).

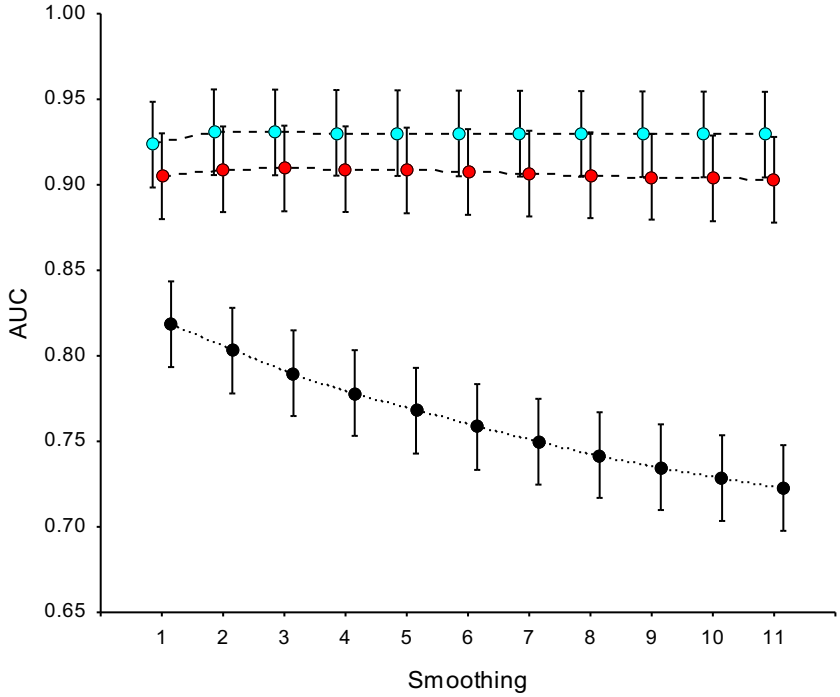
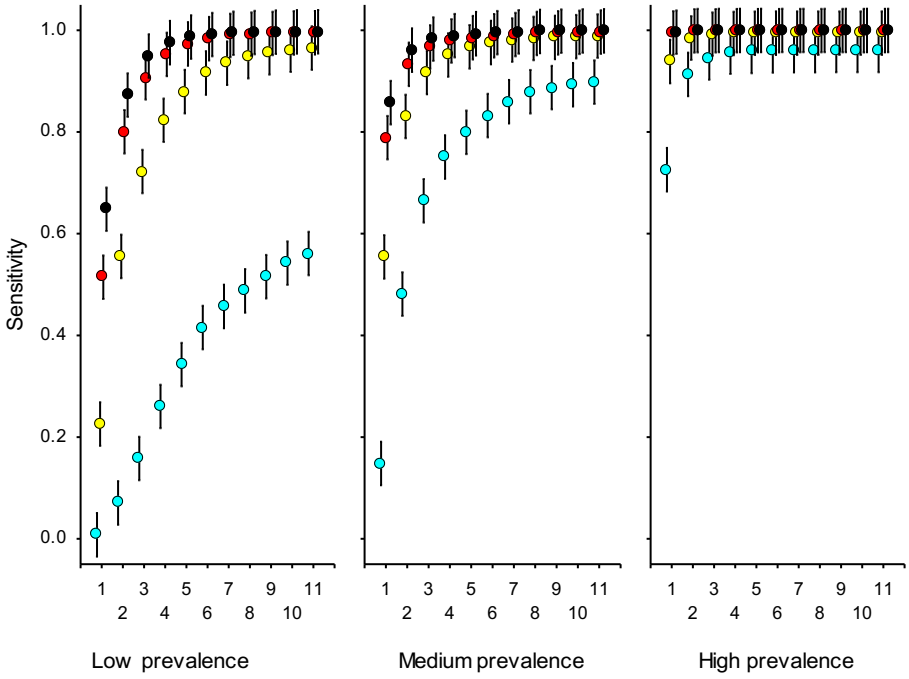



**Figure A2**.- Effect of the smoothing values of the kernel density function (*Sm*) directed to carry out the compounded environmental layer on AUC metrics (± 95% CI) values depending of the prevalence of the species (blue circles = low prevalence; red circles = medium prevalence; black circles = high prevalence).

On the contrary, sensitivity is positively affected by the increase in the smoothing value, so that the rate of success in presence data is maximum when the smoothing attains a value of six. However the statistical significant *Sm* x *Prev* x *%P* three-way interaction ($F_{(60, 528)} = 6.1$; $p < 0.0001$) suggests that low smoothing values negatively influence success in predicting presence mainly when the percentage of presence used and the

prevalence of the species are low (Fig. 3). Specificity follows the same pattern as AUC; success in predicting absence diminishes with the increase in the smoothing value. However, only the *Prev* x *Sm* interaction ($F_{(20, 528)} = 4.5$; $p < 0.0001$) clarified that this pattern appears mainly when the prevalence of the species is higher (Fig. 4).

**Figure A3**.- Effect of the smoothing values of the kernel density function (*Sm*) directed to carry out the compounded environmental layer on sensitivity (± 95% CI) according to




the four levels of the percentage of presences used in the model training and the three levels of the species prevalence. Black circles = 10% of total presences; red circles = 5%; yellow circles = 1%; blue circles = 0.1%.

These results indicate that the smoothing factor influenced model predictions depending of the prevalence of the species, or better said, the ratio between the extent of occurrence and the whole extent of the studied region (the relative occurrence area, ROA; Jiménez-Valverde *et al.* 2008). Thus, we recommend using a low smoothing value when the species has a high prevalence since then the success in predicting absence is improved (less commission errors). However, such smoothing values may reduce success in predicting presence when the species have a low prevalence and a low number of presence data are used to train the model. When the species is geographically and environmentally

restricted it is easy to achieve high success rates in absence predictions, and therefore it is desirable to have a relatively high smoothing value to facilitate the correct prediction of presence. As consequence, we recommend a smoothing value of 1 for most situations, except when a very small number of presence data points are available, and the prevalence of the species is low (i.e., when the data used in NOO poorly represent the distribution of the species). In these cases, a smoothing value of 5-6 would be recommended to avoid many omission errors (i.e., predict as absences localities with true presence).

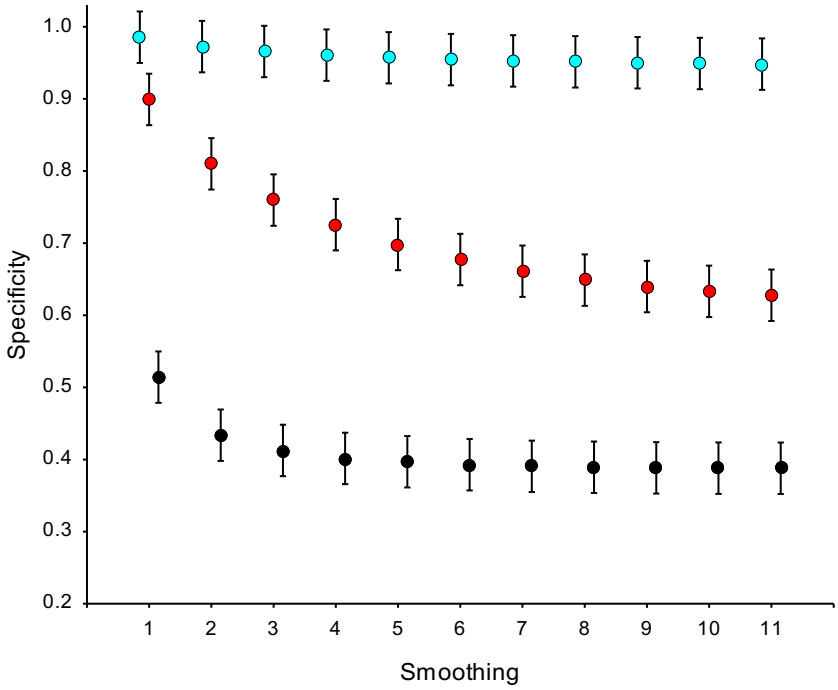



**Figure A4**.- Effect of the smoothing values of the kernel density function (*Sm*) directed to carry out the compounded environmental layer on specificity (± 95% CI) according to the three levels of the species prevalence. Black circles = high prevalence; red circles = medium prevalence; blue circles = low prevalence.

Finally, the percentage of tolerance (%*T*) to expand maximum and minimum environmental values found in the predictors of the presence cells do not show any statistically significant influence in the three performance metrics (interactions are not significant either) (Fig. 5).

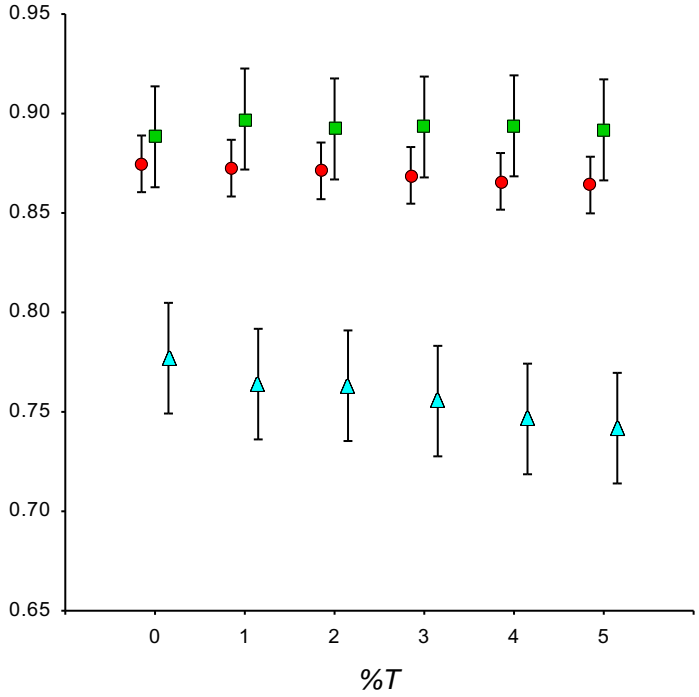



**Figure A5**.- Effect of the tolerance (%*T*) to expand maximum and minimum values of the predictors in the presence cells on AUC (red circles), sensitivity (green squares) and specificity (blue triangles) metrics (± 95% CI). These values are controlled for the effect of species prevalence and number of presences used in model training.

## REFERENCES

Guisande, C., García-Roselló, E., Heine, J., González-Dacosta, J., González Vilas, L., García Pérez, B.J. *et al.* (2017). SPEDInstabR: An algorithm based on a fluctuation index for selecting predictors in species distribution modelling. *Ecol. Inform.*, **37**, 18-23.

Jiménez-Valverde, A., Lobo, J. M. & Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. *Divers. Distrib.*, **14**, 885-890.