

Estimation of species richness with stacked species distribution models (S-SDM)

Comparing two stacking methods (sum of binary outputs and mean probabilities) and the effect of varying species prevalences

Goal

The aim of the study is to estimate the species richness (SR) of four taxonomic species groups in Flanders (Belgium): butterflies, grasshoppers, amphibians, and dragonflies by stacking individual SDMs (S-SDM).

Data

The study area is the region of Flanders in the north of Belgium (fig. 1). The distribution data of the taxonomic species groups are collected from inventory databases containing the observations of a large number of monitoring studies and volunteer work. The spatial detail is 1 x 1km² UTM grid cells. No real absence data is available, therefore only a limited number of UTM grids is selected, known to be well inventoried for a given species group. In total 10 environmental variables are used for the SDM, including soil characteristics, land use type and climate.

Method

A generalized additive model (GAM) with four degrees of freedom and generalized linear model (GLM) (quadratic) are used for the species distribution modelling of the individual species. Each model runs three times with a different set of training and evaluation data (data split 75%) and the results are averaged.

A common method to calculate species richness (SR) is the approach of stacking the output of distribution models of individual species (S-SDM) (Guisan & Rahbek, 2011). We use two different methods of stacking SDMs: the first method is to sum of the calculated probabilities (PROB), the second method sums the binary output after a threshold is applied (TSSbin).

The mean relative prevalence of the species composition is calculated for each location. A large number of species richness maps were calculated, by generating 200 binomial trials from the probability maps. The mean relative prevalence is calculated as the average prevalence in the species composition divided by the mean species richness.

For the analysis we use R version 3.1.0 (R core team, 2014) and the R biomod2 package (Thuiller et al., 2014).

Results and discussion

The majority of the SDMs have reasonable to good values for area under the curve (AUC) evaluation criteria (fig. 2). The AUC's for dragonflies are lower, but still the majority of the models have scores above 0.7.

The calculation of SR based on probabilities (PROB) and the sum of the binary outputs (TSSbin) give quite different results. Both are slightly skewed, but the variance of SR for TSSbin method is much larger than for the PROB method (fig 4). The mean SR however are rather similar (mean 10.53 & 10.89; median 10.0 & 10.6).

The difference in spatial distribution is shown in fig. 3a, b and c. The SR map of the PROB method is much smoother than TSSbin. The differences between the two methods have clear spatial structure. Dubuis et al (2011) and Guisan & Rahbek (2011) find that the TSSbin method tends to estimate a higher SR compared to the S-SDM based on the probabilities. In this study we find only local over- and underestimation.

The effect of differences in prevalence on the SR map was analysed. The map of the relative species prevalence (fig 3d) shows that less common species (red color) occur mainly in the east and along the coast in the west. The threshold for the TSSbin method is known to be positively correlated with the prevalence (Allouche et al., 2006) which has some impact in de SR for rare and very common species. Figure 5 shows that the TSSbin tends to predict a higher SR for species compositions with relatively rare species and a lower SR for species composition with very common species.

Figure 6 shows the effect of the prevalence on the AUC values. Species with a low prevalence generally have a higher AUC. This can be explained as less common species generally require more specific environmental conditions which are easier to model. The results are consistent with other studies (Allouche et al., 2006). It can be expected that the estimation of SR map is more reliable in areas with a higher portion of rare species than in areas with very common species.

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.

Guisan, A., Rahbek, C., 2011. SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography* 38, 1433-1444.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.P., Guisan, A., 2011. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions* 17, 1122-1131.

BIOMOD library in R (Thuiller et al., 2009)

Thuiller, W., Georges, D., Engler, R. (2014). biomod2: Ensemble platform for species distribution modelling. R package version 3.1-41/r652. <http://R-Forge.R-project.org/projects/biomod/>

Fig 1: Flanders (Belgium)



Fig 3: Estimated SR with binary (TSSbin) (a) and probability data (PROB) (b), the difference (TSSbin - Prob) (c), and relative mean prevalence (d)

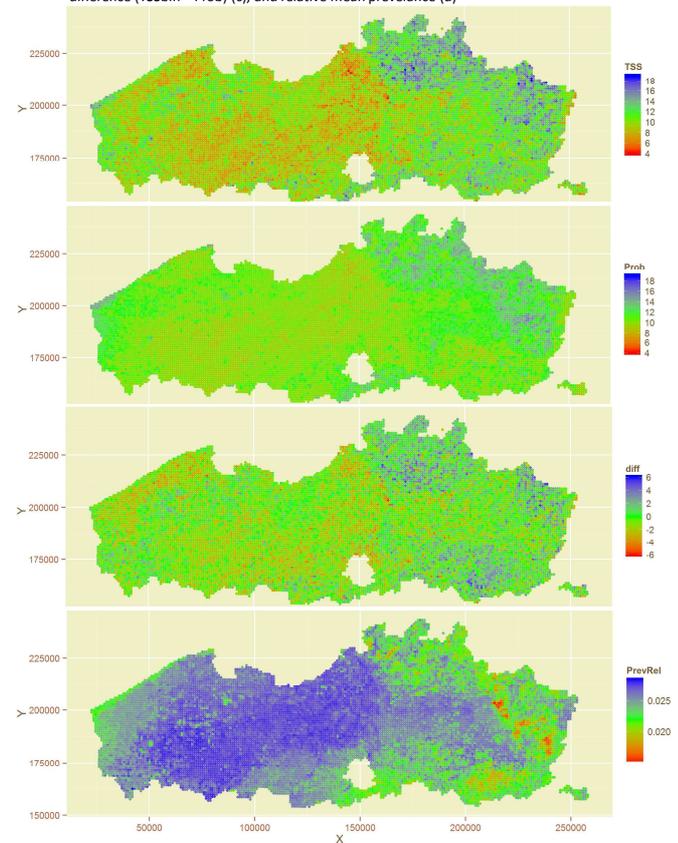


Fig 2: Boxplots of AUC/ROC for each species group

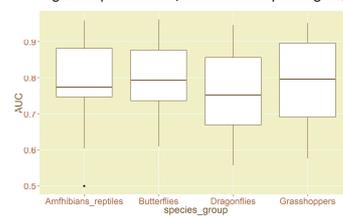


Fig 4: Distribution of estimated SR (TSSbin, Prob)

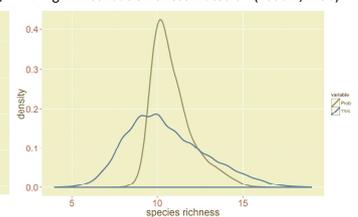


Fig 5: differences (TSSbin - Prob) versus prevalences

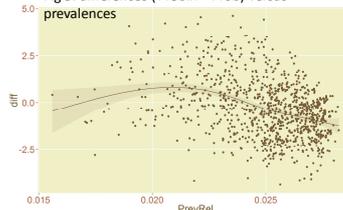
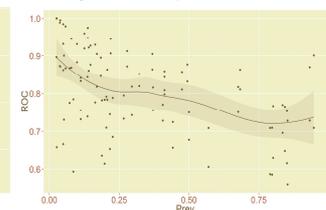


Fig 6: AUC versus prevalence.



Contact

*Toon Van Daele (toon.vandaele@inbo.be). Kliniekstraat 25, 1070 Brussels, Belgium.