Selecting relevant predictors for presence-only species distribution modelling. A case study from the marine environment missing documented absences and uncertain distributions

S. Bosch[a,b], L. Tyberghein[b] and O. De Clerck[a]

[a] Phycology Research Group, Ghent University
Ghent, Belgium
samuel.bosch@ugent.be
olivier.declerck@ugent.be

[b] Flanders Marine Institute (VLIZ)
Ostend, Belgium
lennert.tyberghein@vliz.be

Keywords: benchmark dataset; SDM; marine; model selection.

Abstract: Ideally, datasets for species distribution modelling contain data covering the entire distribution of the species, with evenly sampled records, confirmed absences and auxiliary data (e.g. physiological experiments) allowing informed decisions on predictor selection. Unfortunately, these criteria are only rarely met in real world datasets. This is even more the case for marine organisms for which distributions are too often only scantly characterized and absences generally not recorded. In order to evaluate predictor selection (see Barbet-Massin & Jetz 2014) we compiled a dataset of marine species suitable for benchmarking species distribution modelling, MarineSPEED. We selected 530 well studied and well identifiable species from all major marine taxonomic groups with different range sizes and from different ecoregions. Over three million distribution records where compiled from public sources (e.g. OBIS, GBIF, EMODNET, Reef Life Survey) and linked to environmental and biological data enabling refinement of the analytical results by species group, ecology, geography or life history characteristics.
Using this dataset, predictor selection was performed under different variations of numbers of predictor variables and sampling bias correction. Distributions were modelled with combinations of 3, 4, 7 and 8 environmental variables (calcite, chlorophyll a, diffuse attenuation, nitrate, photosynthetically active radiation, pH, phosphate, salinity, silicate, sea surface temperature, bathymetry and distance to shore) selected from Bio-ORACLE and MARSPEC. Sampling bias correction was performed using spatial thinning and a target group background. Performance of the models was evaluated using random as well as spatial cross-validation (Hijmans 2012). From the results we can conclude that the model setup has a significant impact on predictor selection. The MarineSPEED dataset can serve to evaluate the performance of modelling techniques aimed at predicting distributions of species under current and future climatic scenarios and under a wide array of parameter settings.

References

Barbet-Massin, M., & Jetz, W. (2014). A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. *Diversity and Distributions*, 20(11), 1285–1295.
Hijmans, R. J. (2012). Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, 93(3), 679–688.