A Marine Species Benchmark Dataset for Ecological Modelling

Bosch Samuel[1], Lennert Tyberghein[2] and Olivier De Clerck[1]

[1] Phycology Research Group, Ghent University, Krijgslaan 281-S8, 9000 Ghent, Belgium
E-mail: samuel.bosch@ugent.be

[2] Flanders Marine Institute (VLIZ), Wandelaarkaai 7, 8400 Oostende, Belgium

Species Distribution modelling (SDM) uses machine learning and statistical learning techniques to model ecological niches based on species distribution records and environmental data.. While most algorithms, methods and environmental data used for building SDMs are provided as R packages, ready to use software and data downloads, the species distribution records used in papers featuring new algorithms and methods generally use different datasets. Replication and comparison of SDM experiments are thus made unnecessarily difficult.

The usage of datasets from public repositories like the UCI Machine Learning Repository and the KEEL-dataset repository is a common practice in the machine learning and data mining community. This practice enables researchers to replicate experiments and compare new algorithms and methods with older ones without having to replicate the results of other researchers on their dataset. Usage of datasets from a public repository reduces the amount of tedious “technical work” and greatly eases the comparison of models with published results. So far however, benchmark datasets, are not common practice in ecological and biogeographical studies. For the marine environment a benchmark dataset to evaluate the performance of SDM algorithms is absent all together. The fact that commonly used SDM algorithms and methods were developed using terrestrial case studies only, makes the development of a marine benchmark dataset even more important.

To remedy this problem we compiled a list of marine species suitable for benchmarking SDM algorithms. We selected well studied and well identifiable species from all major taxonomic groups with different range sizes and from different ecoregions. With the above criteria in mind we looked at the availability of public species distribution records from public sources (e.g. OBIS, GBIF, EMODNET, Reef Life Survey) verifying whether sufficient records are available and checking if the distribution of the records didn’t contain significant gaps and errors. Species datasets are linked to environmental data and to biological and ecological properties (traits) enabling refinement of the analytical results by species group, ecology, geography or life history characteristics.

Here we present the first version of Marine SPEcies and Environment Dataset (MarineSPEED v1) containing over 300 species and over three million distribution records. The dataset can serve to evaluate the performance of modelling algorithms aimed at predicting distributions of species under current and future climatic scenarios under a wide array of parameter settings. Here we illustrate the potential of GLASER toward SDM predictor selection and sampling bias correction.

Keywords: sdm; enm; benchmark