DC5: Marine Biodiversity Prediction

Task description

The goal is to predict the abundance of multiple taxa (reef fish and plankton) and marine particulate matter across the ocean using socio-environmental predictors (e.g., physical and biogeochemical ocean parameters delivered by operational products, descriptors of the habitats and of socio-economic activities such as fishing pressure). Once tested and validated, using independent observations, the models will be used to extrapolate the spatiotemporal distribution of these taxa in unobserved parts of the ocean.

Training datasets

Response data: abundance per size class of ~700 reef fish species in 20k surveys, biovolume of ~10 morphotypes of particulate matter at 30k locations, biovolume of ~20 taxa of plankton at 50k locations.
Predictors: bathymetry (GEBCO), climatologies of environmental variables (World Ocean Atlas), standard resolution satellite data (CMEMS L4 ocean colour, CMEMS L4 physics), ocean reanalyses ([GLORYS]), high resolution satellite images (Sentinel-1 and 2), possibly high resolution biogeochemical model outputs (CMEMS PISCES)

The response variables have distributions with very long tails (rare occurrences of large values) which make usual regression metrics (e.g. R2) inappropriate. L1 or L2 loss on log transformed data will be used instead, for a given target.

To define the metric for all targets, micro and macro averages will be used, to also consider targets of low abundance.

Evaluation metrics / data

Data: For reef fish, data from 2006 to 2023 will be used to train the models while the 2024-2025 surveys will constitute independent evaluation datasets. For particulate matter and plankton, an evaluation dataset as independent of the training set as possible will be constructed by clustering observations in lat., lon., time space.

Baseline solutions

The baseline solutions are classical machine-learning based models (e.g. gradient boosted trees, random forest) fitted (i) on local observations (not at the seascape level), and (ii) per individual target (not in a multivariate way).

- code for Drago et al 2022: modelling of plankton biomass.
- code for Haute et al (in prep): modelling of fish abundance.

References

- Clements et al (2022)
- Drago et al (2022)
- Kaneko et al (2023)
- Schickele et al (2024)