ARPHA Conference Abstracts 4: e64661
doi: 10.3897/aca.4.e64661

Conference Abstract

Predicting classifications in marine biomonitoring with supervised machine learning: how much data is required?

Verena Dully‡, Tom Wilding§, Timo Mühlhaus‡, Thorsten Stoeck‡

‡ University of Kaiserslautern, Kaiserslautern, Germany
§ Scottish Marine Institute, Oban, United Kingdom

Corresponding author: Verena Dully (vdully@rhrk.uni-kl.de), Thorsten Stoeck (stoeck@rhrk.uni-kl.de)

Received: 19 Feb 2021 | Published: 04 Mar 2021

Citation: Dully V, Wilding T, Mühlhaus T, Stoeck T (2021) Predicting classifications in marine biomonitoring with supervised machine learning: how much data is required? ARPHA Conference Abstracts 4: e64661. https://doi.org/10.3897/aca.4.e64661

Abstract

Marine coastal ecosystems offer numerous ecosystem services and are therefore subject to a variety of stressors from anthropogenic activities. Environmental biomonitoring programs for effective management and conservation of coastal marine ecosystems are therefore crucial. Traditional monitoring has been based on macrofauna indices which are laborious and require expert knowledge. Recently, eDNA metabarcoding has become increasingly popular as it does not involve macrofauna species identification and is therefore cost and time inexpensive. Studies have shown that ecosystem monitoring based on eDNA metabarcoding is feasible and random forest (RF) algorithms can predict various biological indices, and therefore ecosystem health. To propose adequate designs for future eDNA metabarcoding-based marine coastal monitoring surveys, the aim of the study is to find out (1) What is the lower limit of reads for accurate RF predictions in coastal marine monitoring using microbial communities? (2) Is this limit the same for different monitoring targets? To achieve this goal, we exploited four different Illumina amplicon datasets obtained from bacterial communities in different costal environments. From these datasets, we predicted different objectives relevant for biomonitoring. For each dataset, those corresponding prediction objectives (labels) were predicted using amplicon sequence variants (ASVs) as features. After construction of RF models using all available sequences

of a dataset (full model, serving as benchmark for targeted prediction accuracy), we then successively down-sampled each dataset to lower sequence numbers. Prediction accuracies of the reduced models were then compared to the accuracies of the full models to assess the minimum number of features to obtain the targeted prediction accuracy. Our results show that there is no general answer to question (1) and that (2) the limit varies between different monitoring targets. We have identified the most informative criteria that are relevant to assess the sequencing depth required to predict a biomonitoring category using RF. This may guide future study designs and may help to estimate and control costs in applied routine DNA-based biomonitoring using RF to predict the biomonitoring target. In our contribution we will elucidate and discuss these criteria.

Keywords

Environmental biomonitoring, supervised machine learning, random forest classification, eDNA metabarcoding, sequence number

Presenting author

Verena Dully

Presented at

1st DNAQUA International Conference (March 9-11, 2021)