How can Machine Learning aid in understanding copepod dynamics in a changing North Sea?

Petersen Kaj[1], Semmouri Ilias[1], Lazendic Srdan[2], Asselman Jana[1] and Janssen Colin[1]

[1] Blue Growth Research Lab, Ghent University, Wetenschapspark 1, 8400 Ostend
E-mail: kaj.petersen@outlook.com
[2] Foundations Lab: Clifford Research Group, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium

Zooplankton, including copepods, are fundamental to marine ecosystems, serving as a crucial link in the transfer of energy from primary producers (phytoplankton) to higher trophic levels. Additionally, they play a significant role in the biological carbon pump, facilitating the sequestration of atmospheric carbon dioxide to the deep ocean. Their sensitivity to environmental changes makes copepods reliable indicators of ecosystem health and climate-induced shifts, underscoring the importance of studying their dynamics to better understand and mitigate the impacts of climate change on marine biodiversity. Planktonic organisms are notoriously understudied due to the challenges in sampling their vast size range, which spans over six orders of magnitude, from less than a micron to several meters. Identifying specimens is both time-consuming and requires expertise. Despite these difficulties, understanding the responses of plankton communities to climate change is crucial, as these changes directly impact valuable fisheries and, ultimately, global food security.

Therefore, the aim of this research is twofold: (1) to gain insights into copepod population dynamics under changing environmental conditions and (2) to assess the applicability and limitations of machine learning (ML) techniques to analyze and predict copepod dynamics in the Belgian part of the North Sea (BPNS). The data for this study were primarily sourced from the long-term monitoring efforts from the LifeWatch Observatory in the BPNS, in addition to data collected through our own efforts. As such, this research integrates key environmental variables, including sea surface temperature, salinity, nutrient concentrations, and chlorophyll levels, which are known to influence copepod populations and their dynamics.

We developed a robust data preprocessing pipeline to address the inherent challenges of ecological data. This included handling missing values, removing outliers, scaling, and augmenting data to mitigate sparsity and improve model performance. The study also explored complexity reduction techniques to focus on the most relevant features. Various machine learning models were implemented and evaluated for performance, ranging from traditional approaches such as linear regression and decision trees to advanced methods like Random Forest, Gradient Boosting and neural networks. These models were evaluated for their ability to integrate and learn from multidimensional data, providing insights into the relationships between environmental stressors and copepod dynamics.

Hyperparameter tuning and cross-validation techniques were employed on all models to optimize model performance and ensure generalizability. Ensemble learning approaches were also explored to combine the strengths of the top-performing individual models, enhancing robustness and predictive accuracy. The performance of these models was assessed using metrics such as mean absolute error (MAE) and coefficient of determination (R^2), ensuring a comprehensive evaluation of their predictive capabilities. The best-performing model was an ensemble combining the strengths of the RandomForestRegressor, XGBRegressor, and CatBoostRegressor, achieving an average accuracy of 70% The findings demonstrate the potential of ML techniques to uncover patterns and relationships within complex ecological data. Key environmental predictors, such as temperature and salinity, emerged as significant drivers of copepod abundance, highlighting their sensitivity to climate-induced changes. However, the study also underscores the limitations of ML in ecological contexts, particularly when faced with sparse, noisy, or incomplete datasets. The results emphasize the need for integrating domain expertise and expanding datasets through field validation and collaboration with marine biologists to improve the reliability and applicability of these models.

By focusing on the ecological significance of zooplankton and their responses to environmental stressors, this study bridges computational methodologies and marine ecological research. Beyond copepod dynamics, the work contributes to a growing body of research demonstrating the utility of artificial intelligence in ecological modelling and conservation, offering a foundation for future studies on the impacts of climate change on marine biodiversity and ecosystem services.

Keywords
Machine Learning; Marine Ecosystems; Copepod Dynamics; Belgian Part Of The North Sea; Regression