Title:

**Authors:** Jason D. Selwyn1, INSERT PEOPLE, Steven V. Vollmer1

# Affiliations:

1: Department of Marine and Environmental Sciences, Northeastern University, Nahant, MA, USA.

\*Corresponding author. Email: [j.selwyn@northeastern.edu](mailto:j.selwyn@northeastern.edu)

# Keywords

# Abstract

# Introduction

# Methods

* Data collection

The incidence rate of WBD was modelled using a generalized linear mixed model with a logit link function using both site and sampling time and their interaction as independent effects with significance assessed using likelihood ratio tests. **SOMETHING ABOUT POST-HOCS USED**

* 16s sequencing
* ASV Identification

Samples were pruned to keep only those with more than 1,000 reads and to keep ASVs identified as bacteria after the removal of cyanobacteria, mitochondria, and chloroplast sequences. ASVs were further filtered to retain those found in at least 10% of samples and in all four sampling timepoints. Read counts of the remaining ASVs were normalized for variable sequencing depth using the trimmed mean of M-values method with singleton pairing implemented in edgeR (TMMwsp; Robinson et al. 2010, Robinson and Oshlack 2010). Normalized read counts were converted to log2 counts per million reads with all subsequent analyses being performed on these normalized and log-transformed counts per million.

To reduce model overfitting the dataset was split into training (75%) and testing (25%) sets with all model tuning being done using the training set and only using the test set to calculate the final model metrics (e.g., accuracy).

* Preprocessing
* Models tested
* Hyperparamter tuning – incl resampling scheme
  + 10-fold repeated 10-time cross validation
  + Hyperparamters assessed based on minimizing log-loss
    - Target both accuracy and confidence
  + Initial random grid of 50 followed by up to 200 bayesian tuning iterations – each model (stop after 25 iterations no improvement)
* Tuned model comparison
  + Model of balanced accuracy using final tuned models for best log-loss
  + Choose all models with a 95%+ probability of being within 1% accuracy of the top model
* Variable Importance
  + Approximate MC based Shapley values (Shapley 1953, Štrumbelj and Kononenko 2014, Greenwell 2023)
  + 100

# Results

# Discussion

# Literature Cited

Greenwell, B. 2023. fastshap: Fast Approximate Shapley Values.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Robinson, M. D., and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11:R25.

Shapley, L. S. 1953. A value for n-person games. Page 409 Contributions to the Theory of Games (AM-28). Princeton University Press.

Štrumbelj, E., and I. Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41:647–665.

# Tables