# List of figures

[**Figure 1** Freshwater lakes from southern Norway and Sweden sampled for the COMSAT project. Secchi depth is displayed to provide a general impression of the longitudinal gradient in the dataset. 2](#_Toc27597365)

[**Table 1** Subsets of the ASV and metadata tables. The columns to the right show the first 5 ASVs in decreasing order of abundance while the columns to the left show linearly independent metadata variables that can be treated as explanatory to bacterial community composition. 2](#_Toc27597366)

[**Figure 2** Pearson correlations among environmental metadata variables. Correlation values are displayed by color; the values printed in tiles indicate p-values for correlation significance. 4](#_Toc27597367)

[**Figure 3** Bray-Curtis community composition distances by DOM gradient. a) Bacterial community composition (BCC) Bray-Curtis distances by DOM (raw data). b) XGboost-Predicted BCC Bray distances by sites' DOM. c) OLS (SVD) regression-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid. d) NN-smoothed XGboost-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid. e) NN-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid. f) XGboost-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid. 5](#_Toc27597368)

[**Figure 4** Comparison of UPGMA hierarchical clusters of Random Forest-predicted environmental metadata vs. true values of the same variables. Branch length between two leaves or nodes is proportional to the Euclidean distance between said leaves or nodes. Lines between both trees link corresponding leaves on each. The correlation between both trees is 0.8389 and is significant with a p-value of 0.001. 6](#_Toc27597369)

[**Figure 5** Comparison of UPGMA hierarchical clusters of XGboost-predicted environmental metadata vs. true values of the same variables. Branch length between two leaves or nodes is proportional to the Euclidean distance between said leaves or nodes. Lines between both trees link corresponding leaves on each. The correlation between both trees is 0.7696 and is significant with a p-value of 0.001. 7](#_Toc27597370)

[**Figure 6** Comparison of UPGMA hierarchical clusters of XGboost-predicted environmental metadata vs. Random Forest predictions of the same variables. Branch length between two leaves or nodes is proportional to the Euclidean distance between said leaves or nodes. Lines between both trees link corresponding leaves on each. The correlation between both trees is 0.6584 and is significant with a p-value of 0.001. 8](#_Toc27597371)

# Introduction

## Chosen dataset

This study was performed on freshwater microbial ecology data generated by sampling 72 lakes from southern Norway and Sweden (fig. 1). It was part of a project designated COMSAT. The dataset comprises two tables. One consists of counts of amplicon sequence variants (ASV) for bacteria, while the other contains environmental metadata (table 1). Each observation in either table corresponds to a lake. Both tables are matched by observation and can thus be used as input and output for each other. Bacterial ASVs can be treated as a proxy for the abundance of bacterial species. These ASV counts can also be converted to binary with 0 equal to 0 and values above 0 set to 1 in order to study presence/absence patterns.



**Figure 1** Freshwater lakes from southern Norway and Sweden sampled for the COMSAT project. Secchi depth is displayed to provide a general impression of the longitudinal gradient in the dataset.

**Table 1** Subsets of the ASV and metadata tables. The columns to the right show the first 5 ASVs in decreasing order of abundance while the columns to the left show linearly independent metadata variables that can be treated as explanatory to bacterial community composition.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Site | ASV1 | ASV2 | ASV3 | ASV4 | ASV5 | … | Latitude | Longitude | Altitude | Area | Depth | Temperature | Secchi | O2 | CH4 | pH | TIC | SiO2 | KdPAR |
| 10000\_Hurdalsjøen | 18464 | 5231 | 6963 | 7563 | 9516 | … | 60.37648 | 11.04077 | 176 | 32.81 | 20.0 | 17.03 | 6.50 | 0.9044194 | 11.797343 | 6.870 | 0.82230 | 3312 | 0.62 |
| 10001\_Harestuvatnet | 15296 | 58728 | 30659 | 1614 | 17059 | … | 60.19323 | 10.71212 | 234 | 1.98 | 13.0 | 15.85 | 4.50 | 0.8468347 | 72.674567 | 7.365 | 4.05800 | 3783 | 0.89 |
| 170B\_Gjersjøen | 13356 | 52215 | 25810 | 1367 | 14586 | … | 59.78970 | 10.77485 | 40 | 2.64 | 22.0 | 19.65 | 3.30 | 0.8131012 | 52.953904 | 7.685 | 8.08500 | 3563 | 0.95 |
| 170\_Gjersjøen | 16227 | 53747 | 26456 | 2823 | 3119 | … | 59.78970 | 10.77485 | 40 | 2.64 | 22.0 | 19.65 | 3.30 | 0.8131012 | 52.953904 | 7.685 | 8.08500 | 3563 | 0.95 |
| 180\_Øgderen | 52862 | 4887 | 1361 | 14854 | 25616 | … | 59.71388 | 11.41303 | 133 | 12.66 | 9.5 | 18.61 | 1.10 | 0.8406025 | 85.639780 | 7.225 | 2.66800 | 1125 | 1.60 |
| 189\_Krøderen | 18830 | 53461 | 50015 | 12664 | 13253 | … | 60.13485 | 9.75860 | 133 | 43.91 | 14.0 | 15.44 | 2.80 | 0.8582522 | 29.100059 | 6.695 | 0.81360 | 2499 | 0.82 |
| 191\_Rødbyvatnet | 43828 | 7657 | 1836 | 34800 | 20517 | … | 59.58175 | 10.48715 | 118 | 1.16 | 10.0 | 18.55 | 2.10 | 0.8527711 | 260.596931 | 7.535 | 3.09200 | 2063 | 1.32 |
| 214\_Gjesåssjøen | 10532 | 588 | 9275 | 19181 | 5315 | … | 60.68167 | 11.99235 | 176 | 3.98 | 3.5 | 19.63 | 1.15 | 0.8360833 | 97.561306 | 7.070 | 1.73200 | 2924 | 2.27 |
| 2252\_Rotnessjøen | 14088 | 39265 | 35086 | 11061 | 7228 | … | 60.49690 | 12.34120 | 260 | 1.09 | 26.0 | 16.55 | 1.95 | 0.7350632 | 41.956068 | 6.635 | 0.77310 | 5559 | 1.08 |

## Research questions

It was chosen in this study to 1) identify patterns of variation in bacterial community composition along environmental gradients, 2) predict environmental metadata variables from bacterial community composition and 3) predict presence or absence of bacterial ASVs from environmental metadata. The first objective was pursued using pairs of dissolved organic matter water content as input with corresponding pairwise bacterial community Bray-Curtis distances as output. The second objective used the full bacterial community composition – ASV – table as input and single environmental metadata variables as outputs, yielding a separate model for each predicted metadata variable. The third objective used a subset of environmental variables as input to predict a subset of the bacterial community composition table converted to presence/absence values.

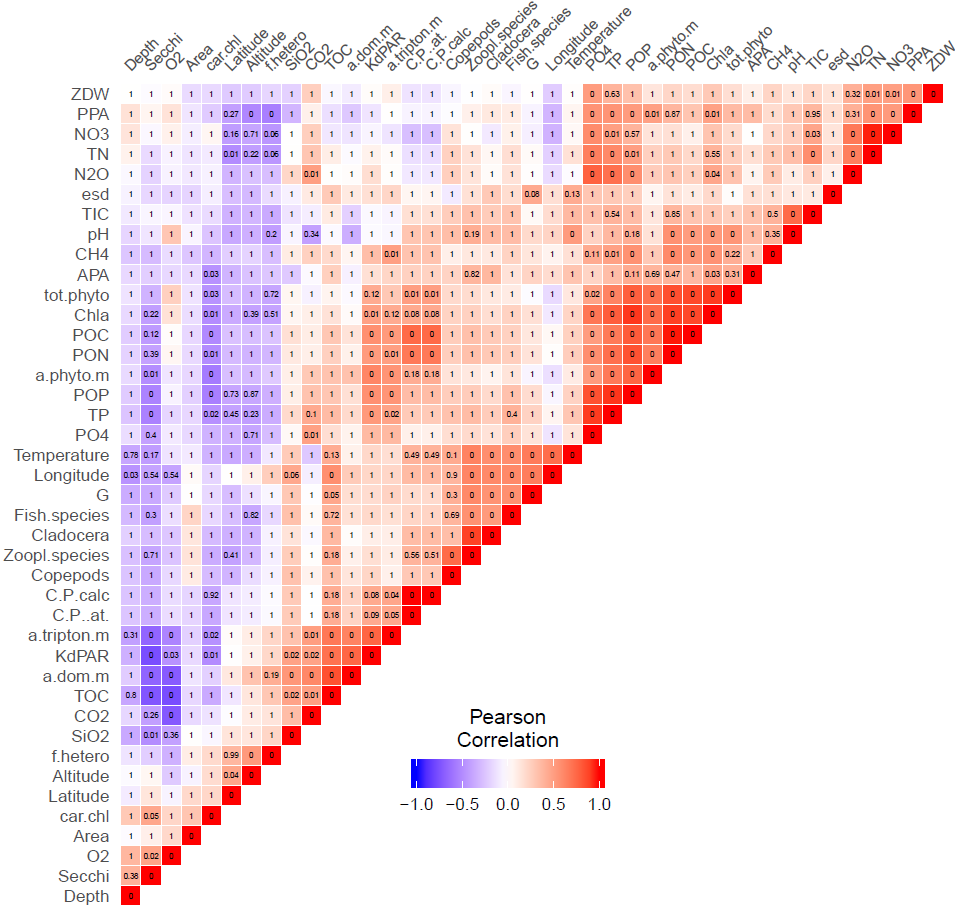
# Methods

## Data filtering, formatting and transformation

The COMSAT dataset contained only continuous variables. There were missing values in the environmental metadata, which were replaced by intrapolation using Multivariate Imputation by Chained Equations (Buuren & Groothuis-Oudshoorn 2011). This decision to keep observations with missing values is motivated by the very small number of observations in the full dataset compared to the number of descriptors which would make it even harder to train models successfully, were the number of observations to be reduced any further. The ASV table was scaled using ranging (set values to interval of 0-1) due to this transformation yielding somewhat better clustering in terms of ecological meaningfulness compared to subtraction of mean and division by standard deviation (Legendre & Legendre 2012). Environmental metadata were used with or without scaling, the former being performed by subtracting the mean and dividing by standard deviation.

## Dimensionality reduction

In order to extract patterns from the ASV data without using all predictors, pairwise distances between observations were used. Since the counts for most ASVs across most observations is 0, it was necessary to use an asymmetrical coefficient in order to avoid inflated similarity between observations. The Bray-Curtis distance was used accordingly. In the case of the environmental metadata, groups of linearly dependent variables had all but one variable kept, while a subsequent pruning was performed in the same manner where groups of variables presented variance inflation factors (VIF) above a threshold of 20 (ter Braak & Smilauer 2002). Correlations among environmental variables were visualized using a heatmap (fig. 2).



**Figure 2** Pearson correlations among environmental metadata variables. Correlation values are displayed by color; the values printed in tiles indicate p-values for correlation significance.

## Regression using XGboost

Regression using scikit-learn’s implementation of XGboost was used to predict bacterial community Bray-Curtis distances along the dissolved organic matter (DOM) gradient (“a.dom.m” in the environmental metadata table). The model was trained on data split into 80 % training and 20 % test sets. The same analysis was performed with own code for a feed forward neural network. After training the respective models, response data was generated for a meshgrid of values matching the minimum-maximum range of the DOM gradient. The idea here was to see how well XGboost performed for generating a model used for intrapolation. Intrapolated outputs were also generated with ordinary least-squares regression to appreciate how well an XGboost model can contain complex structures in the data where linear models cannot.

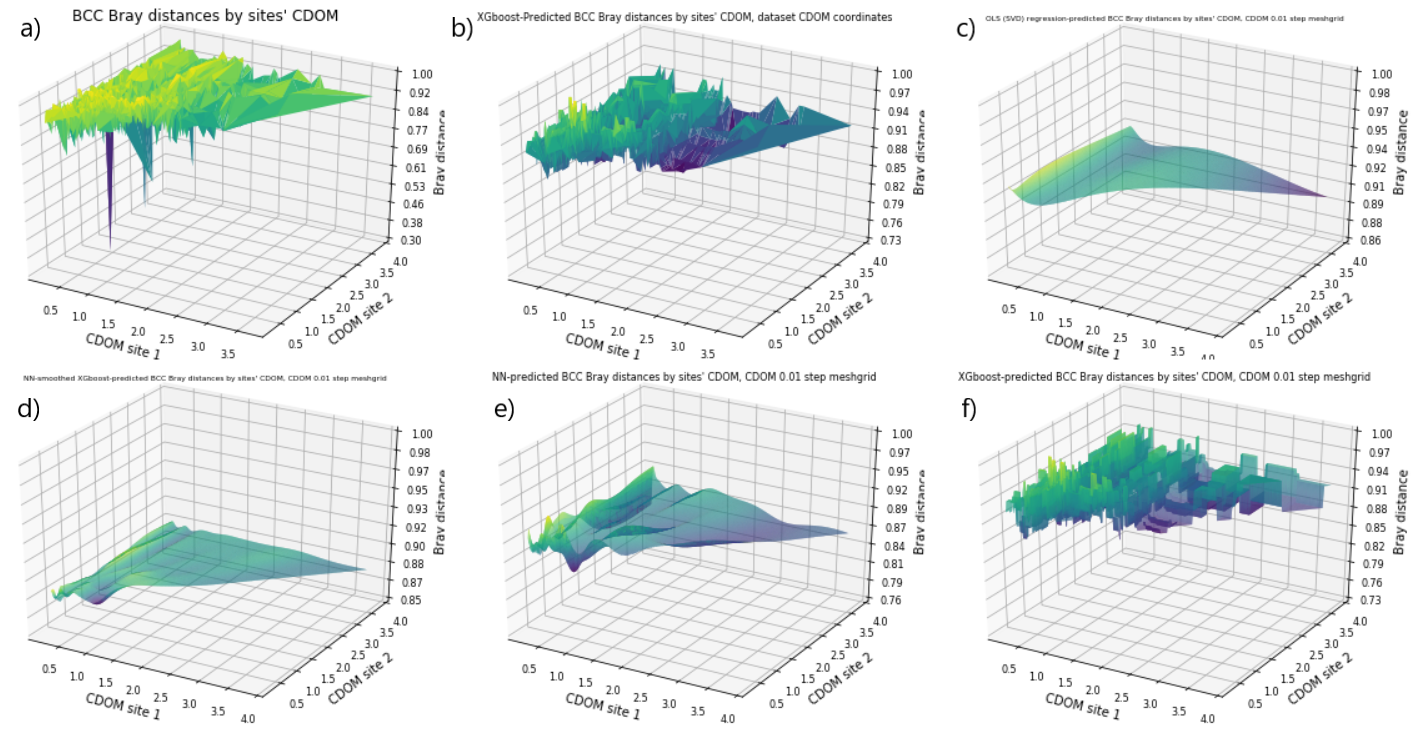
## Comparison of XGboost and Random Forest

A selection of non collinear environmental variables were predicted for an input of the 100 most abundant ASVs using scikit-learn’s XGboost and Random Forest libraries. The metadata tables were then scaled by subtracting the mean and dividing by standard deviation. Euclidean distance matrices were computed for true data as well as XGboost and Random Forest predictions, followed by Procrustes tests to determine whether the clusters of observations were significantly similar between predicted and true data. Unweighted pair group method with arithmetic mean (UPGMA) hierarchical clusterings of the distance matrices were visualized as Tanglegrams (Galili 2015).

# Results

## Regression using XGboost

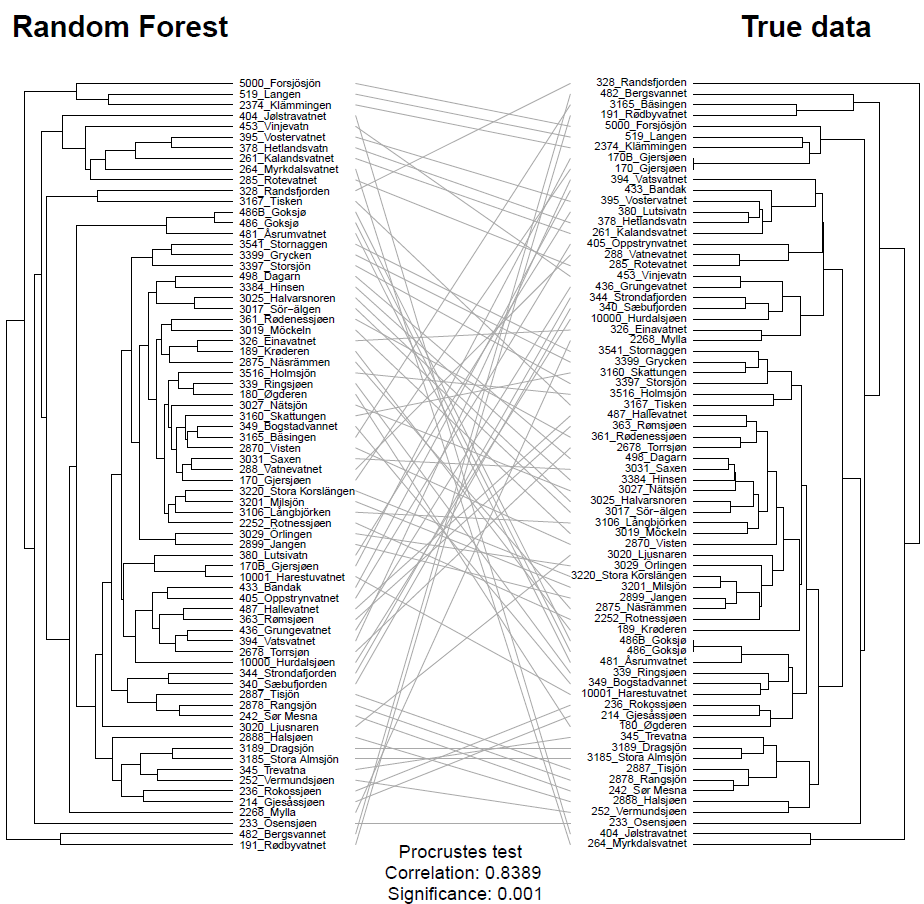
The patterns in both XGboost and feed forward neural network predicted values are similar. XGboost is however much faster when training the model. The linear model poorly captured the complex structures in data revealed by the XGboost and neural network regressions (fig. 3). While the predicted values from the XGboost model seem correct on the grid of values from the original dataset, the model returns some type of step function when predicting values for a meshgrid containing x,y values not present in the training dataset.



**Figure 3** Bray-Curtis community composition distances by DOM gradient. a) Bacterial community composition (BCC) Bray-Curtis distances by DOM (raw data). b) XGboost-Predicted BCC Bray distances by sites' DOM. c) OLS (SVD) regression-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid. d) NN-smoothed XGboost-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid. e) NN-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid. f) XGboost-predicted BCC Bray distances by sites' DOM, DOM 0.01 step meshgrid.

## Comparison of XGboost and Random Forest

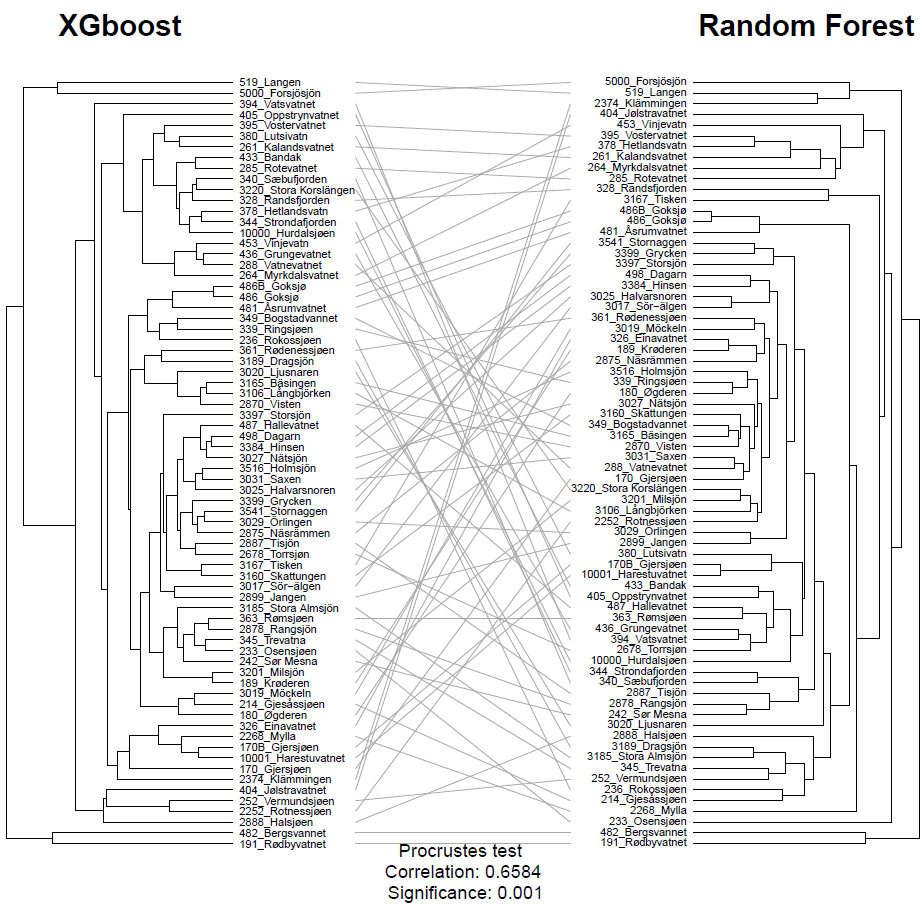
The tree topologies for true data, XGboost and Random Forest predictions are all significantly similar with p-values below 0.001. The correlation is highest between true data and Random Forest predictions, with a value of 0.84, while XGboost returned a correlation of 0.77. The correlation for XGboost and Random Forest was 0.66.



**Figure 4** Comparison of UPGMA hierarchical clusters of Random Forest-predicted environmental metadata vs. true values of the same variables. Branch length between two leaves or nodes is proportional to the Euclidean distance between said leaves or nodes. Lines between both trees link corresponding leaves on each. The correlation between both trees is 0.8389 and is significant with a p-value of 0.001.



**Figure 5** Comparison of UPGMA hierarchical clusters of XGboost-predicted environmental metadata vs. true values of the same variables. Branch length between two leaves or nodes is proportional to the Euclidean distance between said leaves or nodes. Lines between both trees link corresponding leaves on each. The correlation between both trees is 0.7696 and is significant with a p-value of 0.001.



**Figure 6** Comparison of UPGMA hierarchical clusters of XGboost-predicted environmental metadata vs. Random Forest predictions of the same variables. Branch length between two leaves or nodes is proportional to the Euclidean distance between said leaves or nodes. Lines between both trees link corresponding leaves on each. The correlation between both trees is 0.6584 and is significant with a p-value of 0.001.

# Discussion

## Regression using XGboost

Performing regression on the bacterial community composition distances by dissolved organic matter content was orders of magnitude faster than the own code for the feed-forward neural network. However, the usefulness of the resulting XGboost model falls somewhat short when an output is predicted for input values absent from the training data. Attempting the latter resulted in a step function pattern when the input was a meshgrid of DOM values found inside the interval of training DOM values. This problem did not occur for the FFNN when predicting the BCC distances for the same meshgrid of DOM values; the output was a smooth surface. Attempting to obtain a smooth surface of XGboost predictions by standard regression methods does not work either when structures in the data are as complex as is the case with this dataset. It is possible to train a feed-forward neural network on the XGboost predictions for training data in order to produce a smooth surface for a higher resolution meshgrid, although it comes at the cost of introducing possible artifacts and bias from slightly different results associated with distinct random seeds. Overall, the FFNN remains preferable when the input data does not follow regular steps along a gradient and when a model is trained for a subsequent intrapolation purpose.

## Comparison of XGboost and Random Forest

When using the same input to generate the same outputs, XGboost was much faster than Random Forest. An additional benefit of XGboost in the context of this study was the possibility to use the entire ASV table as input, whereas Random Forest could not complete the task in a reasonable time if much more than 100 predictors were used. Although XGboost could make use of more data in this case, the predictions from Random Forest turned out to be closest to reality. Overall, it can be said the main advantage of using XGboost over Random Forest is eliminating the need to select predictors, particularly when the number of predictors exceeds 100. Consequently, XGboost shows great potential for typical microbial ecology datasets such as the one analyzed in this study, since it is usually very difficult to select which predictors to exclude, especially when they may be part of a priori unforeseen but relevant structures in the data.

# References

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software, 45(3). https://doi.org/10.18637/jss.v045.i03

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics, 31(22), 3718–3720. https://doi.org/10.1093/bioinformatics/btv428

Legendre, P., & Legendre, L. F. (2012). Numerical ecology(Vol. 24). Elsevier

ter Braak, C. J. F. & P. Smilauer. 2002. CANOCO reference manual and CanoDraw for Windows user’s guide – Software for canonical community ordination (version 4.5). Microcomputer Power, Ithaca, New York. 500 pp.