Results of the RF for the Baldeggersee, r\_cyanobac:

Reminder: RMSE = Root Mean Square Error, it’s the average distance between the observed data values and the predicted data values.

Rf kNN:

Random Forest

252 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 225, 228, 227, 226, 228, 228, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.8375849 0.3201781 0.7155878

2 0.8248001 0.3287178 0.6947317

3 0.8222334 0.3284018 0.6835093

4 0.8200375 0.3315143 0.6776683

5 0.8235572 0.3263142 0.6785250

6 0.8287992 0.3182082 0.6812018

7 0.8367738 0.3075877 0.6867616

8 0.8322914 0.3134726 0.6807855

9 0.8331984 0.3128726 0.6823827

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 4.

RF bag:

Random Forest

252 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 225, 227, 225, 226, 228, 227, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.8228185 0.3653759 0.7064596

2 0.8049806 0.3749507 0.6832815

3 0.8041872 0.3719491 0.6750324

4 0.8070140 0.3648914 0.6744125

5 0.8101063 0.3570404 0.6741890

6 0.8110961 0.3557159 0.6737145

7 0.8179060 0.3442088 0.6789170

8 0.8146538 0.3497658 0.6745041

9 0.8201167 0.3419189 0.6776574

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 3.

RF median:

252 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 225, 227, 228, 227, 228, 227, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.8556283 0.2982789 0.7432768

2 0.8371070 0.3250398 0.7220678

3 0.8405781 0.3163299 0.7200629

4 0.8352914 0.3258632 0.7135966

5 0.8418871 0.3123812 0.7181200

6 0.8341570 0.3268686 0.7082941

7 0.8379828 0.3192659 0.7104446

8 0.8392534 0.3165630 0.7120521

9 0.8343156 0.3260723 0.7092216

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 6.

Testing the models (with the testing data set): predictive accuracy

kNN:

RMSE Rsquared MAE

0.6459448 0.2113350 0.5170144

Bag:

RMSE Rsquared MAE

0.6463085 0.1956656 0.5154139

Median:

RMSE Rsquared MAE

0.7388746 0.1660573 0.6285694

A screenshot of a computer

Description automatically generated with medium confidence

Other metric for the importance of variables: RFE, or Recursive Feature Elimination

How recursive feature elimination works: building many models of a type of machine learning method on the training set, and iteratively re-calculating the most important variables. At the end, the variables that were found important most often, across different subset sizes, can be reported

kNN:

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables RMSE Rsquared MAE RMSESD RsquaredSD MAESD Selected

1 0.9884 0.1651 0.8011 0.18585 0.1853 0.14778

2 0.9138 0.2280 0.7511 0.16545 0.1837 0.14975

3 0.8824 0.2510 0.7359 0.14286 0.1850 0.13337

4 0.8451 0.3005 0.7055 0.11755 0.1651 0.11321

5 0.8323 0.3238 0.6966 0.12343 0.1779 0.10936

6 0.8286 0.3319 0.6897 0.11559 0.1587 0.10355

7 0.8296 0.3334 0.6959 0.11095 0.1629 0.10114

8 0.8278 0.3350 0.6982 0.09973 0.1445 0.09494

9 0.8168 0.3503 0.6874 0.10697 0.1535 0.10112 \*

The top 5 variables (out of 9):

PO4\_P, microzoo, NO3\_N, cyclopoida, temperature

Bagging:

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables RMSE Rsquared MAE RMSESD RsquaredSD MAESD Selected

1 0.9517 0.2070 0.7700 0.07382 0.07842 0.06102

2 0.8714 0.2760 0.7051 0.08161 0.11862 0.08000

3 0.8337 0.3394 0.6834 0.10709 0.17364 0.07403

4 0.8194 0.3451 0.6758 0.09681 0.15200 0.07319

5 0.8253 0.3363 0.6903 0.08972 0.14608 0.06893

6 0.8206 0.3439 0.6829 0.09393 0.15672 0.07359

7 0.8106 0.3628 0.6778 0.09109 0.16294 0.07196

8 0.8042 0.3697 0.6772 0.08471 0.14744 0.07063

9 0.8028 0.3693 0.6737 0.08399 0.14768 0.06850 \*

The top 5 variables (out of 9):

PO4\_P, microzoo, NO3\_N, cyclopoida, tot\_zoo

Median:

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables RMSE Rsquared MAE RMSESD RsquaredSD MAESD Selected

1 1.0003 0.1483 0.8257 0.12603 0.1505 0.10672

2 0.9189 0.2034 0.7642 0.10131 0.1362 0.10628

3 0.8775 0.2555 0.7260 0.10192 0.1485 0.09641

4 0.8524 0.2936 0.7184 0.10855 0.1751 0.09856

5 0.8531 0.2938 0.7254 0.08282 0.1407 0.07316

6 0.8563 0.2834 0.7234 0.08538 0.1491 0.07374

7 0.8483 0.2985 0.7290 0.07070 0.1389 0.06392

8 0.8307 0.3298 0.7199 0.07390 0.1474 0.06771

9 0.8219 0.3451 0.7085 0.07776 0.1512 0.06960 \*

The top 5 variables (out of 9):

PO4\_P, NO3\_N, microzoo, cyclopoida, temperature

3D plots:

These are the best I could get

Graphical user interface, application

Description automatically generated

How I interpret these:

* Tot cyano rate: temperature doesn’t really matter, and

After replacing the NA’s with 0’s and removing the pre-processing:

The problem with these is: we have NA’s, we can’t use the RFE or the defaultSummary to estimate which predictors are most important and how accurate the models are :(

Total cyanobacteria ratio:

Random Forest

252 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 126, 127, 126, 126, 126, 128, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.2454512 0.4255477 0.2142632

2 0.2385657 0.4327380 0.2080455

3 0.2353137 0.4478425 0.2026320

4 0.2372298 0.4374039 0.2041931

5 0.2374534 0.4362990 0.2037180

6 0.2389597 0.4330035 0.2029545

7 0.2390801 0.4325208 0.2037121

8 0.2391615 0.4341331 0.2040505

9 0.2394798 0.4347000 0.2026571

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 3.

This looks good: the RMSE is low and the Rsquared is relatively ok

Chroococcales ratio:

Random Forest

252 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 123, 122, 123, 123, 124, 124, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.06270451 0.11602479 0.03558361

2 0.06513661 0.08067648 0.03680954

3 0.06648727 0.08033754 0.03725645

4 0.06773351 0.05522578 0.03768353

5 0.06846731 0.06407608 0.03825901

6 0.06856840 0.05674641 0.03825793

7 0.06878824 0.04917271 0.03816811

8 0.06877093 0.05119256 0.03830352

9 0.06859591 0.06458977 0.03815712

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 1.

Nostocales ratio:

Random Forest

252 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 126, 127, 126, 126, 126, 128, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.02929441 0.16959469 0.01840473

2 0.03064408 0.12562889 0.01923149

3 0.03216108 0.09760908 0.02011746

4 0.03255329 0.10075446 0.02015389

5 0.03351961 0.08365591 0.02057040

6 0.03390325 0.09106855 0.02079831

7 0.03425616 0.08089050 0.02103863

8 0.03485300 0.08375825 0.02131864

9 0.03483314 0.08595525 0.02125865

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 1.

RMSE values are very low (which is good), but the Rsquared are way too low to be significant

Oscillatoriales ratio:

Random Forest

252 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 126, 127, 126, 126, 126, 128, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.2423230 0.4831974 0.2090199

2 0.2298791 0.5119247 0.1973528

3 0.2270671 0.5120184 0.1946831

4 0.2279061 0.5048736 0.1933277

5 0.2284840 0.5027030 0.1934762

6 0.2294378 0.4973633 0.1930310

7 0.2306338 0.4963178 0.1941442

8 0.2304789 0.4965828 0.1934791

9 0.2308619 0.4941142 0.1934560

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 3.

This model seems to be quite accurate