Zooplankton-Cyanobacteria project, random forest calculations

Laura Uusitalo; laura.uusitalo@ymparisto.fi

2018-2020

Introduction

This paper documents the R code and the analyses done for assessing the effect of cyanobacteria on zooplankton based on field sampling data. The work is lead by Sanna Suikkanen in SYKE Marine Research Centre.

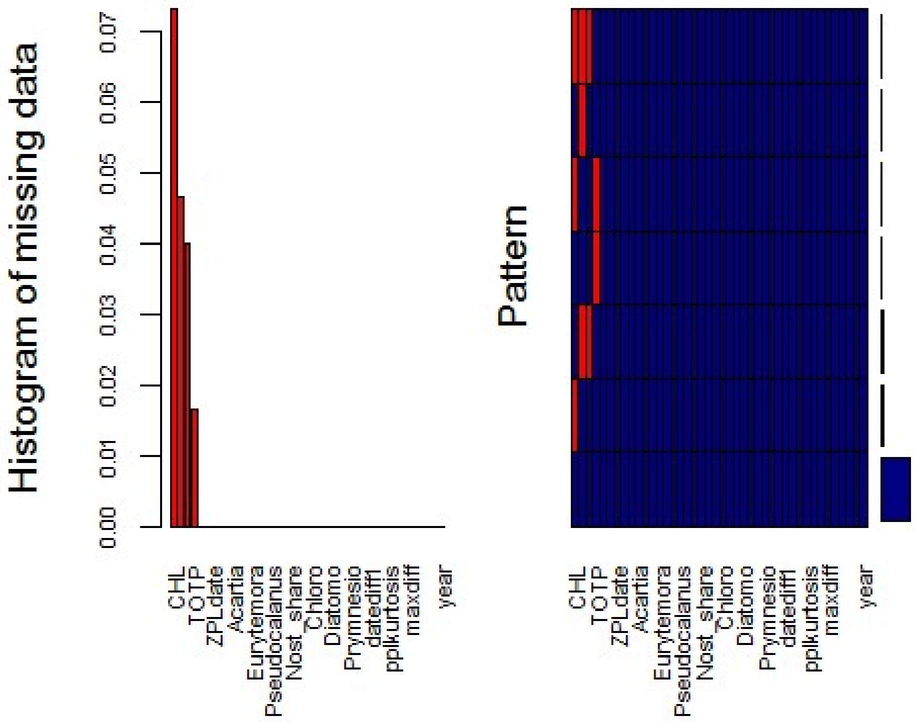
Prepping the data & packages

```
setwd("D:/Users/uusitalol/RStudio/ZplCyanoCalculations_Sept2015")

#packages
library(mice)
library(caret)
library(VIM)
library(missForest)
library(ggplot2)

#this file is created in the CombineZplPplWithPhyschem.R script:
load("dataforRF.RData") # object name  = dat

#Set seed
seed <- 42
set.seed(seed)
```

Visualize & impute missing values

Visualize the missing values. See https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/.

```
aggr(dat, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, 
labels=names(dat), cex.axis=.7, gap=3, ylab=c("Histogram of missing 
data","Pattern"))

## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```



```
##
##  Variables sorted by number of missings:
##        Variable      Count
##             CHL 0.07308970
##             SAL 0.04651163
##            TEMP 0.03986711
##            TOTP 0.01661130
##         Station 0.00000000
##         PPLdate 0.00000000
##         ZPLdate 0.00000000
##         ZPLsize 0.00000000
##         NFratio 0.00000000
##         Acartia 0.00000000
##         Bosmina 0.00000000
##      Cercopagis 0.00000000
##      Eurytemora 0.00000000
##     Limnocalanus 0.00000000
##        Nauplius 0.00000000
##   Pseudocalanus 0.00000000
##       Rotatoria 0.00000000
##         Nost_bm 0.00000000
##      Nost_share 0.00000000
##        OtherC_bm 0.00000000
##    OtherC_share 0.00000000
##          Chloro 0.00000000
##          Chryso 0.00000000
##          Crypto 0.00000000
##         Diatomo 0.00000000
##            Dino 0.00000000
##         Eugleno 0.00000000
##       Prymnesio 0.00000000
##           Mx_bm 0.00000000
##         envDate 0.00000000
##       datediff1 0.00000000
##       datediff2 0.00000000
##     zplkurtosis 0.00000000
##     pplkurtosis 0.00000000
##             Lat 0.00000000
##            Long 0.00000000
##         maxdiff 0.00000000
##         diffsum 0.00000000
##             nas 0.00000000
##   nonCyanoTotBM 0.00000000
##  nonCyanoAutoBM 0.00000000
##            year 0.00000000
```

Impute missing values using missing forest method.

```
dat.mF<-missForest(dat[, colSums(is.na(dat))>0], maxiter=50)

##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
##   missForest iteration 5 in progress...done!

dat.mF$OOBerror

##     NRMSE
## 0.2187928

#replace
datI<-dat
datI$TEMP<-dat.mF$ximp$TEMP
datI$SAL<-dat.mF$ximp$SAL
datI$TOTP<-dat.mF$ximp$TOTP
datI$CHL<-dat.mF$ximp$CHL
```

For each class variable (the variable we're explaining), make a separate data set. In these data sets use only a subset of the features (predictors, explaining variables):

For zooplankton variables (mean size, kurtosis, nauplius:female ratio, taxa biomasses) we use: - Nost_bm - Nost_share - OtherC_bm - OtherC_share - Chloro - Chryso - Crypto - Diatomo - Dino - Eugleno - Prymnesio - Mx_bm - TEMP - SAL - CHL - TOTP -pplkurtosis - Lat - Long - nonCyanoTotBM - nonCyanoAutoBM - year

For phytoplankton variables (total biomass, autotroph biomass, mixotroph biomass, kurtosis), we use

the physico-chemical variables (excluding chl-a), year, and sampling station coordinates as features. Phytoplankton class variables (the targets we are predicting) are measured without cyanobacteria, and the biomasses of Nostocales and other cyanobacteria are used as features for these classes. - ZPLsize - NFratio - Acartia - Bosmina - Cercopagis - Eurytemora - Limnocalanus - Nauplius - Pseudocalanus - Rotatoria - TEMP - SAL - TOTP - zplkurtosis - Lat - Long - year - Nost_bm - OtherC_bm

```
#all colnames in varI
#c("Station", "PPLdate", "ZPLdate",  "ZPLsize", "NFratio", "Acartia",
"Bosmina", "Cercopagis", "Eurytemora",      "Limnocalanus", "Nauplius",
"Pseudocalanus", "Rotatoria", "Nost_bm", "Nost_share", "OtherC_bm",
"OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo", "Dino",
"Eugleno", "Prymnesio", "Mx_bm", "envDate", "TEMP", "SAL", "CHL",
"TOTP", "datediff1", "datediff2", "zplkurtosis", "pplkurtosis", "Lat",
"Long", "maxdiff", "diffsum", "nas", "nonCyanoTotBM", "nonCyanoAutoBM",
"year")

#Zooplankton classes; each has 22 explaining variables

dat.ms <- datI[, c("ZPLsize", "Nost_bm", "Nost_share", "OtherC_bm",
"OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo", "Dino",
"Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.nf <- datI[, c("NFratio", "Nost_bm", "Nost_share", "OtherC_bm",
"OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo", "Dino",
"Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.nauplius <- datI[, c("Nauplius", "Nost_bm", "Nost_share", "OtherC_bm",
"OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo", "Dino",
"Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.zplkurtosis <- datI[, c("zplkurtosis", "Nost_bm", "Nost_share",
"OtherC_bm", "OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo",
"Dino", "Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.Acartia <- datI[, c("Acartia", "Nost_bm", "Nost_share", "OtherC_bm",
"OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo", "Dino",
"Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.Bosmina <-datI[, c("Bosmina", "Nost_bm", "Nost_share", "OtherC_bm",
"OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo", "Dino",
"Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.Eurytemora <- datI[, c("Eurytemora", "Nost_bm", "Nost_share",
"OtherC_bm", "OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo",
"Dino", "Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.Pseudocalanus <- datI[, c("Pseudocalanus", "Nost_bm", "Nost_share",
"OtherC_bm", "OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo",
"Dino", "Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

dat.Rotatoria <- datI[, c("Rotatoria", "Nost_bm", "Nost_share",
"OtherC_bm", "OtherC_share",    "Chloro", "Chryso", "Crypto", "Diatomo",
"Dino", "Eugleno", "Prymnesio", "Mx_bm", "TEMP", "SAL", "CHL",    "TOTP",
"pplkurtosis", "Lat",  "Long","nonCyanoTotBM", "nonCyanoAutoBM", "year")]

#phytoplankton classes; ; each has 19 explaining variables

dat.auto <- datI[, c("nonCyanoAutoBM", "ZPLsize", "NFratio", "Acartia",
"Bosmina", "Cercopagis", "Eurytemora",      "Limnocalanus", "Nauplius",
"Pseudocalanus", "Rotatoria", "TEMP", "SAL", "TOTP", "zplkurtosis", "Lat",
"Long",  "year", "Nost_bm", "OtherC_bm")]

dat.mixo <-  datI[, c("Mx_bm", "ZPLsize", "NFratio", "Acartia", "Bosmina",
"Cercopagis", "Eurytemora",      "Limnocalanus", "Nauplius",
"Pseudocalanus", "Rotatoria", "TEMP", "SAL", "TOTP", "zplkurtosis", "Lat",
"Long",  "year", "Nost_bm", "OtherC_bm")]

dat.tot <-  datI[, c("nonCyanoTotBM", "ZPLsize", "NFratio", "Acartia",
"Bosmina", "Cercopagis", "Eurytemora",      "Limnocalanus", "Nauplius",
"Pseudocalanus", "Rotatoria", "TEMP", "SAL", "TOTP", "zplkurtosis", "Lat",
"Long",  "year", "Nost_bm", "OtherC_bm")]

dat.pplkurtosis<- datI[, c("pplkurtosis", "ZPLsize", "NFratio", "Acartia",
"Bosmina", "Cercopagis", "Eurytemora",      "Limnocalanus", "Nauplius",
"Pseudocalanus", "Rotatoria", "TEMP", "SAL", "TOTP", "zplkurtosis", "Lat",
"Long",  "year", "Nost_bm", "OtherC_bm")]
```

Finally, make random forest regression models for all explained variables. We use out-of-bag validation and root mean squared error (RMSE) to evaluate the model fit. The optimal number of predictors for each target variable was selected using the root mean square error (RMSE) of the repeated cross-validation.

Below, the optimal number of predictors can be found in the result text on row that says "## The final value used for the model was mtry =". The RMSE with different numbers of predictors has also been plotted for each target variable. Scatter plot of the target variable has been plotted against all the predictors that were chosen for the final model.

Random regression forests do a multiple nonlinear regression. A nice explanation can be found here: https://www.quora.com/How-does-random-forest-work-for-regression-1

```
#########
#RFCV cross-validation (check what happens to classification accuracy when
new explanatory variables are included)
# and cross-validated Random Forest regression

#settings
#graphics settings
theme1 <- trellis.par.get()
theme1$plot.symbol$col = rgb(.2, .2, .2, .4)
theme1$plot.symbol$pch = 16
theme1$plot.line$col = rgb(1, 0, 0, .7)
theme1$plot.line$lwd <- 2
trellis.par.set(theme1)
```

```
#rf settings
#seed has been set above
control <- trainControl(method="oob")
mtryZ <- c(1:22)
mtryP <- c(1:19)
tunegridZ <- expand.grid(.mtry=mtryZ)
tunegridP <- expand.grid(.mtry=mtryP)
```

Create the function to run the random forest

```
runRF <- function(dd, rf) {
  respvar <- colnames(dd)[1]
  colnames(dd)[1] <- "Class"
  rf <- train(Class ~ ., data=dd, method="rf", tuneGrid=tunegrid,
trControl=control, ntree = 100000,  importance=TRUE)
  print(rf)
  plot(rf)
  v <- varImp(rf, scale = TRUE)
  print(v)

  imp <- rownames(v$importance)[order(v$importance, decreasing=TRUE)]
  n <- rf$bestTune$mtry
  ddtmp <- ddtmp<-dd[,1, drop=F]


  return(rf)
}
```

Run the random forests for each target variable:

```
#set tunegrid for zpl class variables

tunegrid <- tunegridZ
```

Zpl mean size

```
dd<-dat.ms
rf<-runRF(dd)

## Random Forest
##
## 301 samples
##  22 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE         Rsquared
##    1    0.006332072  0.13950996
##    2    0.006320739  0.14258724
##    3    0.006333700  0.13906746
##    4    0.006345738  0.13579163
##    5    0.006356951  0.13273489
##    6    0.006377713  0.12706045
##    7    0.006382629  0.12571434
##    8    0.006400183  0.12089849
##    9    0.006416172  0.11650067
##   10    0.006435999  0.11103192
##   11    0.006448163  0.10766855
##   12    0.006464632  0.10310464
##   13    0.006477042  0.09965784
##   14    0.006496043  0.09436757
##   15    0.006514660  0.08916917
##   16    0.006533764  0.08381939
##   17    0.006546025  0.08037753
##   18    0.006565396  0.07492698
##   19    0.006591858  0.06745486
##   20    0.006610520  0.06216699
##   21    0.006629686  0.05672104
##   22    0.006655476  0.04936803
## 
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
## rf variable importance
## 
##   only 20 most important variables shown (out of 22)
## 
##                 Overall
## SAL             100.000
## year             81.767
## Prymnesio        75.433
## Lat              59.580
## Nost_share       54.682
## Eugleno          50.153
## nonCyanoAutoBM   49.991
## nonCyanoTotBM    49.121
## Long             48.946
## Chryso           46.409
## OtherC_bm        46.060
## Diatomo          38.251
## CHL              35.947
## Mx_bm            35.128
## Dino             34.206
## OtherC_share     33.404
## Nost_bm          31.064
## TEMP             22.934
## TOTP             14.665
## Crypto            6.966
```

```
 save(rf, file = "ms_rf.RData")
```

Zooplankton kurtosis

```
dd<-dat.zplkurtosis
rf<-runRF(dd)
```

```
## Random Forest 
## 
```

```
## 301 samples
##  22 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    3.628376  0.07428713
##    2    3.619148  0.07898984
##    3    3.614363  0.08142367
##    4    3.610617  0.08332697
##    5    3.608721  0.08428965
##    6    3.608138  0.08458551
##    7    3.608167  0.08457056
##    8    3.608855  0.08422164
##    9    3.606413  0.08546024
##   10    3.608093  0.08460825
##   11    3.610685  0.08329243
##   12    3.611529  0.08286361
##   13    3.611232  0.08301471
##   14    3.611722  0.08276565
##   15    3.614680  0.08126262
##   16    3.613476  0.08187483
##   17    3.615229  0.08098383
##   18    3.615756  0.08071578
##   19    3.618276  0.07943397
##   20    3.618051  0.07954813
##   21    3.620831  0.07813321
##   22    3.618656  0.07924057
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 9.
## rf variable importance
##
##   only 20 most important variables shown (out of 22)
##
##                Overall
## year           100.000
## SAL             67.826
## Long            50.002
## Crypto          48.900
## nonCyanoAutoBM  46.303
## Chryso          44.141
## TOTP            40.510
## TEMP            39.925
## Lat             39.031
## nonCyanoTotBM   37.078
## Diatomo         31.157
## Nost_share      28.719
## Prymnesio       28.523
## Eugleno         23.899
## Nost_bm         21.740
## Dino            19.841
## Mx_bm           17.921
## CHL              9.756
## pplkurtosis      8.121
## OtherC_bm        7.568
```

```
 save(rf, file = "zplkurtosis_rf.RData")
```

zpl Nauplius to female ratio

```
dd<-dat.nf
rf<-runRF(dd)
```

```
## Random Forest
##
## 301 samples
##  22 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    4.829862   0.1052675980
##    2    4.823786   0.1075175247
##    3    4.838688   0.1019948539
##    4    4.854173   0.0962377605
##    5    4.867948   0.0911012131
##    6    4.886294   0.0842374219
##    7    4.904612   0.0773586526
##    8    4.919982   0.0715668474
##    9    4.935217   0.0658081115
##   10    4.950112   0.0601605498
##   11    4.967209   0.0536570083
##   12    4.983722   0.0473547073
##   13    5.006224   0.0387326395
##   14    5.019275   0.0337140795
##   15    5.032979   0.0284306648
##   16    5.045886   0.0234410075
##   17    5.061302   0.0174648836
##   18    5.079631   0.0103357138
##   19    5.089924   0.0063209132
##   20    5.104537   0.0006068652
##   21    5.120063  -0.0054820035
##   22    5.131506  -0.0099813429
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
## rf variable importance
##
##   only 20 most important variables shown (out of 22)
##
##                 Overall
## Lat              100.00
## Crypto            93.22
## SAL               84.93
## TOTP              74.17
## nonCyanoAutoBM   73.54
## nonCyanoTotBM    73.00
## Long             55.66
## year             55.64
## Diatomo          53.10
## Prymnesio        49.56
## OtherC_bm        45.03
## Chloro           40.83
## Dino             39.96
## Chryso           38.07
## Mx_bm            37.06
## Nost_bm          35.53
## CHL              33.75
## OtherC_share     30.81
## Nost_share       25.96
## Eugleno          23.69
```

```
save(rf, file = "nf_rf.RData")
```

zpl Nauplius

```
dd<-dat.nauplius
rf<-runRF(dd)
```

```
## Random Forest
##
## 301 samples
##  22 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    3.571398  0.1822910
##    2    3.514860  0.2079762
##    3    3.493128  0.2177399
##    4    3.483797  0.2219132
##    5    3.479243  0.2239463
##    6    3.481293  0.2230312
##    7    3.481642  0.2228758
##    8    3.483561  0.2220188
##    9    3.489922  0.2191751
##   10    3.493084  0.2177592
##   11    3.499804  0.2147469
##   12    3.507990  0.2110690
##   13    3.514137  0.2083019
##   14    3.522376  0.2045849
##   15    3.529730  0.2012604
##   16    3.536460  0.1982114
##   17    3.544046  0.1947680
##   18    3.553861  0.1903020
##   19    3.559808  0.1875898
##   20    3.567557  0.1840488
##   21    3.575257  0.1805229
##    22    3.584617  0.1762265
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 5.
## rf variable importance
##
##   only 20 most important variables shown (out of 22)
##
##                  Overall
## Lat              100.00
## Long              99.44
## Prymnesio         66.39
## Crypto            57.96
## SAL               57.65
## year              55.73
## nonCyanoTotBM     49.61
## Mx_bm             48.26
## TOTP              48.09
## CHL               46.53
## nonCyanoAutoBM    42.26
## Chryso            38.81
## Nost_bm           35.22
## OtherC_share      29.38
## Nost_share        26.14
## Dino              25.45
## TEMP              22.88
## OtherC_bm         22.74
## Chloro            22.15
## Diatomo           17.35
```

```
save(rf, file = "nauplius_rf.RData")
```

Acartia BM

```
dd<-dat.Acartia
rf<-runRF(dd)
```

```
## Random Forest
##
## 301 samples
##  22 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    35.83415  0.3566353
##    2    35.09674  0.3828414
##    3    34.80345  0.3931131
##    4    34.66255  0.3980171
##    5    34.63487  0.3989781
##    6    34.65164  0.3983959
##    7    34.69328  0.3969493
##    8    34.77084  0.3942501
##    9    34.86350  0.3910173
##   10    34.89837  0.3897983
##   11    34.99491  0.3864178
##   12    35.05588  0.3842778
##   13    35.12705  0.3817753
##   14    35.24964  0.3774526
##   15    35.27456  0.3765721
##   16    35.37758  0.3729252
##   17    35.45193  0.3702868
##   18    35.55345  0.3666750
##   19    35.61693  0.3644113
##   20    35.66684  0.3626288
##   21    35.73579  0.3601621
##   22    35.83374  0.3566500
## 
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 5.
## rf variable importance
## 
##   only 20 most important variables shown (out of 22)
## 
##                Overall
## Long            100.00
## TOTP             83.72
## SAL              78.77
## Lat              72.86
## Crypto           72.66
## nonCyanoTotBM    61.34
## nonCyanoAutoBM   59.50
## year             52.10
## TEMP             44.35
## Dino             42.59
## Prymnesio        41.12
## Mx_bm            40.40
## CHL              40.38
## Chloro           35.60
## Nost_bm          32.24
## pplkurtosis      32.21
## OtherC_bm        31.04
## OtherC_share     29.00
## Chryso           26.26
## Nost_share       24.61
```

```
 save(rf, file = "Acartia_rf.RData")
```

Bosmina BM

```
dd<-dat.Bosmina
rf<-runRF(dd)
```

```
## Random Forest 
## 
## 301 samples
##  22 predictor
```

```
## 
## No pre-processing
## Resampling results across tuning parameters:
## 
##   mtry  RMSE       Rsquared 
##    1    100.10003  0.1712118
##    2     99.49589  0.1811857
##    3     99.39232  0.1828895
##    4     99.41974  0.1824386
##    5     99.46781  0.1816477
##    6     99.55895  0.1801475
##    7     99.67240  0.1782778
##    8     99.91413  0.1742872
##    9    100.05759  0.1719144
##   10    100.17774  0.1699245
##   11    100.29609  0.1679621
##   12    100.51426  0.1643382
##   13    100.57201  0.1633778
##   14    100.87328  0.1583579
##   15    101.06382  0.1551754
##   16    101.19825  0.1529264
##   17    101.29747  0.1512646
##   18    101.53534  0.1472737
##   19    101.80612  0.1427196
##   20    101.95033  0.1402891
##   21    102.23623  0.1354605
##   22    102.25754  0.1351001
## 
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 3.
## rf variable importance
## 
##   only 20 most important variables shown (out of 22)
## 
##                Overall
## OtherC_bm      100.000
## TEMP            97.966
## Long            93.325
## OtherC_share    90.314
## year            68.140
## Dino            56.757
## nonCyanoTotBM   56.756
## TOTP            44.451
## Crypto          42.819
## Nost_bm         41.755
## CHL             41.668
## Nost_share      38.582
## nonCyanoAutoBM  38.241
## Prymnesio       36.942
## Lat             36.858
## Diatomo         30.886
## Mx_bm           30.334
## Chryso          25.531
## Chloro           8.824
## SAL              1.855

 save(rf, file = "Bosmina_rf.RData")
```

Eurytemora BM

```
dd<-dat.Eurytemora
rf<-runRF(dd)

## Random Forest 
## 
## 301 samples
##  22 predictor
## 
## No pre-processing
## Resampling results across tuning parameters:
## 
##   mtry  RMSE      Rsquared 
##    1    74.21702  0.2991151
##    2    72.01438  0.3400999
##    3    70.87740  0.3607727
##    4    70.31620  0.3708553
##    5    70.09766  0.3747600
##    6    69.89392  0.3783893
##    7    69.92629  0.3778134
##    8    69.92003  0.3779248
##    9    70.04927  0.3756228
##   10    70.29722  0.3711948
##   11    70.41871  0.3690195
##   12    70.68286  0.3642770
##   13    70.92338  0.3599431
##   14    71.08707  0.3569852
##   15    71.37867  0.3516991
##   16    71.66004  0.3465778
##   17    71.90951  0.3420205
##   18    72.22259  0.3362785
##   19    72.53421  0.3305386
##   20    72.84322  0.3248223
##   21    73.07180  0.3205784
##   22    73.39334  0.3145858
## 
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 6.
## rf variable importance
## 
##   only 20 most important variables shown (out of 22)
## 
##                 Overall
## Long           100.000
## Prymnesio       88.473
## Lat             47.368
## Mx_bm           45.818
## Crypto          37.317
## SAL             29.733
## TOTP            22.892
## year            22.157
## OtherC_bm       20.300
## Nost_share      19.892
## CHL             13.305
## nonCyanoTotBM   12.491
## Nost_bm         11.488
## nonCyanoAutoBM  11.147
## Dino             9.248
## Chloro           6.638
## Diatomo          5.872
## OtherC_share     4.031
## pplkurtosis      3.895
## Eugleno          2.248
```

```
save(rf, file = "Eurytemora_rf.RData")
```

Pseudocalanus BM

```
dd<-dat.Pseudocalanus
rf<-runRF(dd)
```

```
## Random Forest 
## 
## 301 samples
##  22 predictor
## 
## No pre-processing
## Resampling results across tuning parameters:
## 
##   mtry  RMSE      Rsquared 
##    1    15.33209  0.4044395
##    2    14.48881  0.4681505
##    3    14.13424  0.4938632
##    4    13.94158  0.5075669
##    5    13.84802  0.5141543
##    6    13.79069  0.5181683
##    7    13.75059  0.5209663
##    8    13.72394  0.5228218
##    9    13.71547  0.5234102
##   10    13.70502  0.5241360
##   11    13.72294  0.5228910
##   12    13.71863  0.5231907
##   13    13.74892  0.5210826
##   14    13.76457  0.5199919
##   15    13.78444  0.5186054
##   16    13.80688  0.5170362
##   17    13.82536  0.5157426
##   18    13.87403  0.5123274
##   19    13.89961  0.5105274
##   20    13.93580  0.5079755
##   21    13.97442  0.5052442
##   22    14.01251  0.5025433
```

```
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 10.
## rf variable importance
##
##   only 20 most important variables shown (out of 22)
##
##                Overall
## Lat            100.000
## SAL             36.020
## Long            27.569
## year            23.925
## Chryso          16.990
## Dino            13.851
## Mx_bm           13.602
## TOTP            11.376
## OtherC_share     9.750
## Prymnesio        8.710
## OtherC_bm        8.071
## nonCyanoAutoBM   7.890
## Crypto           7.755
## TEMP             7.712
## Chloro           7.306
## nonCyanoTotBM    7.157
## CHL              7.025
## Diatomo          4.165
## Eugleno          3.829
## Nost_bm          3.409
```

```
save(rf, file = "Pseudocalanus_rf.RData")
```

Rotatoria BM

```
dd<-dat.Rotatoria
rf<-runRF(dd)
```

```
## Random Forest
##
## 301 samples
##  22 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    18.86283  0.128563554
##    2    19.03624  0.112467047
##    3    19.17901  0.099104351
##    4    19.28855  0.088783942
##    5    19.36469  0.081576427
##    6    19.44807  0.073649832
##    7    19.52602  0.066209373
##    8    19.56701  0.062285207
##    9    19.65358  0.053969516
##   10    19.74028   0.045604301
##   11    19.81133   0.038721486
##   12    19.87407   0.032623549
##   13    19.93726   0.026462240
##   14    20.02853   0.017528347
##   15    20.10775   0.009740901
##   16    20.17843   0.002766743
##   17    20.23020  -0.002357000
##   18    20.34343  -0.013608364
##   19    20.36977  -0.016235382
##   20    20.41657  -0.020910115
##   21    20.48509  -0.027774713
##   22    20.53762  -0.033051732
## 
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 1.
## rf variable importance
## 
##   only 20 most important variables shown (out of 22)
## 
##                 Overall
## CHL              100.00
## nonCyanoTotBM     87.94
## nonCyanoAutoBM    83.85
## Diatomo           73.15
## pplkurtosis       66.56
## year              60.66
## OtherC_bm         59.97
## Nost_bm           59.57
## Nost_share        59.50
## OtherC_share      55.97
## Crypto            55.57
## TOTP              55.13
## Dino              50.89
## Long              50.26
## Chloro            49.27
## SAL               37.35
## Mx_bm             35.26
## Prymnesio         30.26
## Eugleno           27.00
## Lat               26.49
```

```
 save(rf, file = "Rotatoria_rf.RData")
```

Phytoplankton kurtosis

```
dd<-dat.pplkurtosis
rf<-runRF(dd)
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid
## mtry: reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid
## mtry: reset to within valid range
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid
## mtry: reset to within valid range
```

```
## Random Forest
##
## 301 samples
##  19 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    1.500287  -0.03414989
##    2    1.506322  -0.04248744
##    3    1.509811  -0.04732205
##    4    1.512443  -0.05097596
##    5    1.514794  -0.05424645
##    6    1.516631  -0.05680551
##    7    1.517495  -0.05800989
##    8    1.519902  -0.06136841
##    9    1.521130  -0.06308441
##   10    1.522391  -0.06484685
##   11    1.523676  -0.06664543
##   12    1.524821  -0.06824907
##   13    1.524858  -0.06830146
##   14    1.526377  -0.07043090
##   15    1.526426  -0.07049985
##   16    1.527254  -0.07166154
##   17    1.528182  -0.07296462
##   18    1.528808  -0.07384357
##   19    1.529020  -0.07414082
##   20    1.528927  -0.07401098
##   21    1.528831  -0.07387615
##   22    1.528894  -0.07396433
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 1.
## rf variable importance
##
##               Overall
## year          100.000
## Rotatoria      98.961
## TEMP           76.569
## Eurytemora     71.904
## NFratio        71.687
## OtherC_bm      69.883
## Lat            58.080
## Long           57.994
## Bosmina        54.482
## Nost_bm        41.021
## Nauplius       39.959
## Acartia        37.329
## Pseudocalanus  34.560
## ZPLsize        33.306
## TOTP           30.338
## Limnocalanus   24.243
## Cercopagis     23.390
## SAL             5.145
## zplkurtosis     0.000
```

```
 save(rf, file = "pplkurtosis_rf.RData")

#set tunegrid for ppl class variables

tunegrid <- tunegridP
```

Phytoplankton autotroph biomass (without cyanobacteria)

```
dd<-dat.auto
rf<-runRF(dd)
```

```
## Random Forest
##
## 301 samples
##  19 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    606.3715  0.1173349
##    2    600.9165  0.1331447
##    3    597.8912  0.1418511
##    4    595.8526  0.1476929
##    5    593.4436  0.1545706
##    6    591.9980  0.1586846
##    7    591.2268  0.1608752
##    8    589.4387  0.1659432
##    9    588.5242  0.1685293
##   10    588.1234  0.1696613
##   11    587.7050  0.1708423
##   12    587.6384  0.1710302
##   13    587.0406  0.1727160
##   14    586.4205  0.1744627
##   15    585.9690  0.1757335
##   16    585.4115  0.1773011
##   17    584.7373  0.1791949
##   18    584.7650  0.1791173
##   19    584.8501  0.1788783
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 17.
## rf variable importance
##
##               Overall
## TEMP          100.000
## year           57.746
## TOTP          52.382
## Pseudocalanus 48.647
## Nauplius      46.097
## OtherC_bm     39.820
## SAL           28.652
## ZPLsize       19.796
## Rotatoria     19.321
## Limnocalanus  18.846
## NFratio       17.354
## Lat           16.091
## Acartia       14.063
## Bosmina       11.911
## Long          11.280
## Nost_bm       10.705
## Eurytemora     7.901
## zplkurtosis    3.726
## Cercopagis     0.000
```

```
save(rf, file = "auto_rf.RData")
```

phytoplankton mixotroph biomass

```
dd<-dat.mixo
rf<-runRF(dd)
```

```
## Random Forest
##
## 301 samples
##  19 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    157.3353   0.153403553
##    2    156.9357   0.157699052
##    3    157.7702   0.148716740
##    4    158.7294   0.138335156
##    5    159.7251   0.127489935
##    6    160.9188   0.114400813
##    7    161.8052   0.104617388
##    8    162.5374   0.096494974
##    9    163.4917   0.085854574
##   10    164.3229   0.076536166
##   11    165.1720   0.066967639
##   12    165.7487   0.060441176
##   13    166.6401   0.050307880
##   14    167.6335   0.038951031
##   15    168.4251   0.029853530
##   16    169.0252   0.022927463
##   17    170.0239   0.011346601
##   18    170.9406   0.000657336
##   19    171.8612  -0.010135375
##
```

```
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
## rf variable importance
##
##                Overall
## SAL             100.00
## Eurytemora       87.64
## Nauplius         80.17
## NFratio          70.94
## Lat              66.06
## year             59.90
## Acartia          52.49
## Limnocalanus     51.97
## ZPLsize          51.39
## Pseudocalanus    50.91
## Long             49.89
## OtherC_bm        49.00
## Rotatoria        42.85
## TEMP             34.84
## Cercopagis       18.76
## TOTP             16.30
## zplkurtosis       9.29
## Nost_bm           4.65
## Bosmina           0.00
```

```
save(rf, file = "mixo_rf.RData")
```

Phytoplankton total biomass (without cyanobacteria)

```
dd<-dat.tot
rf<-runRF(dd)
```

```
## Random Forest
##
## 301 samples
##  19 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    619.0501  0.1795308
##    2    612.4997  0.1968023
##    3    608.9145  0.2061777
##    4    608.1400  0.2081957
##    5    606.1684  0.2133216
##    6    604.9199  0.2165586
##    7    604.8061  0.2168535
##    8    604.6566  0.2172405
##    9    604.5780  0.2174442
##   10    605.0146  0.2163135
##   11    604.3394  0.2180616
##   12    604.6297  0.2173102
##   13    603.9808  0.2189895
##   14    604.1453  0.2185638
##   15    604.8283  0.2167961
##   16    604.5344  0.2175569
##   17    604.9910  0.2163745
##   18    604.4116  0.2178749
##   19    605.0638  0.2161860
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 13.
## rf variable importance
##
##               Overall
## TEMP          100.000
## TOTP           65.523
## year           59.230
## Pseudocalanus  55.072
## Nauplius       45.255
## SAL            41.317
## OtherC_bm      36.309
## Lat            28.456
## Acartia        27.355
## NFratio        26.437
## Limnocalanus   21.678
## ZPLsize        19.744
## Rotatoria      17.185
## Nost_bm        11.121
## Long           10.468
## Bosmina         8.520
## Eurytemora      7.528
## zplkurtosis     4.254
## Cercopagis      0.000
```

```
save(rf, file = "totppl_rf.RData")
```