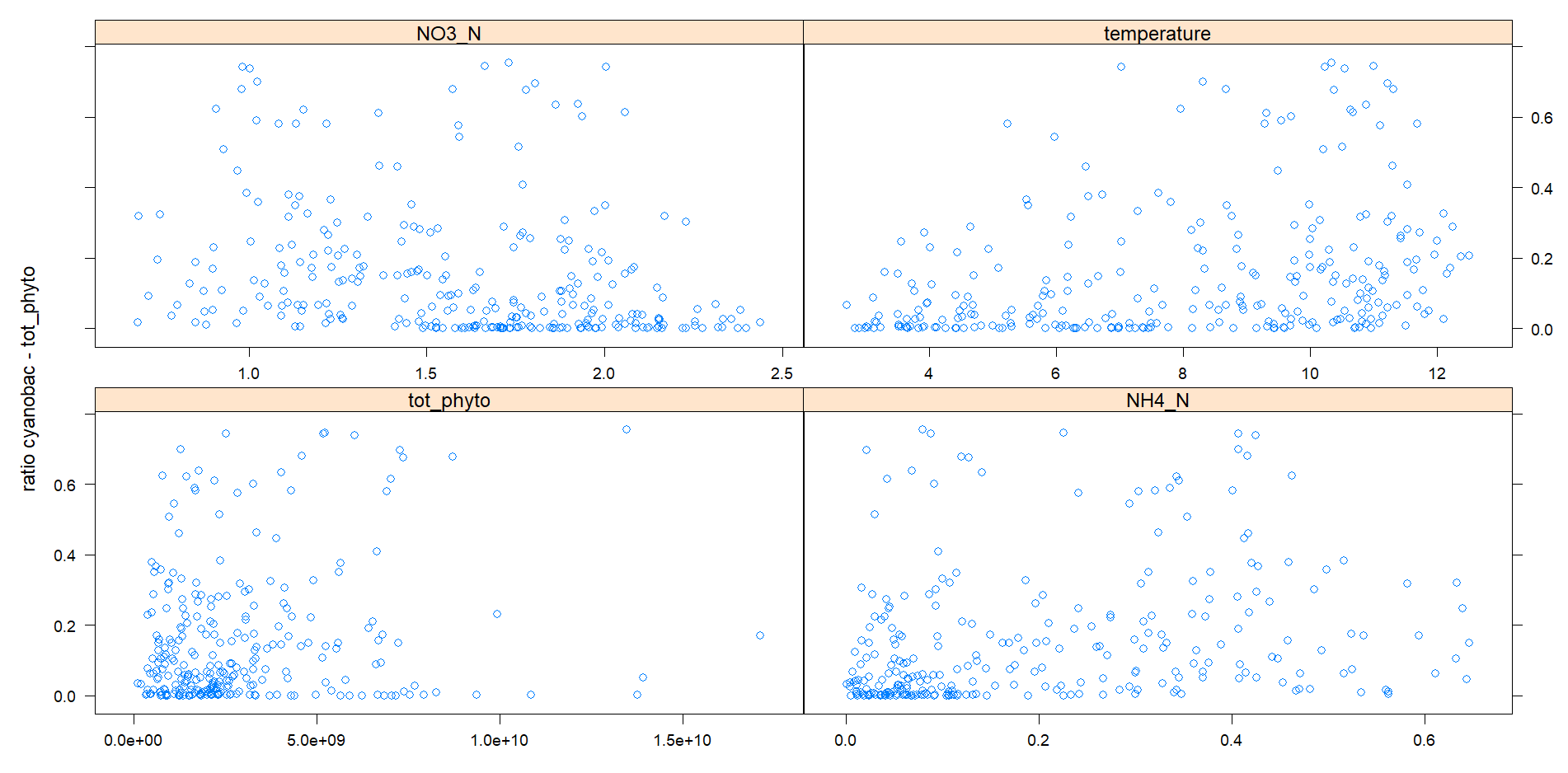
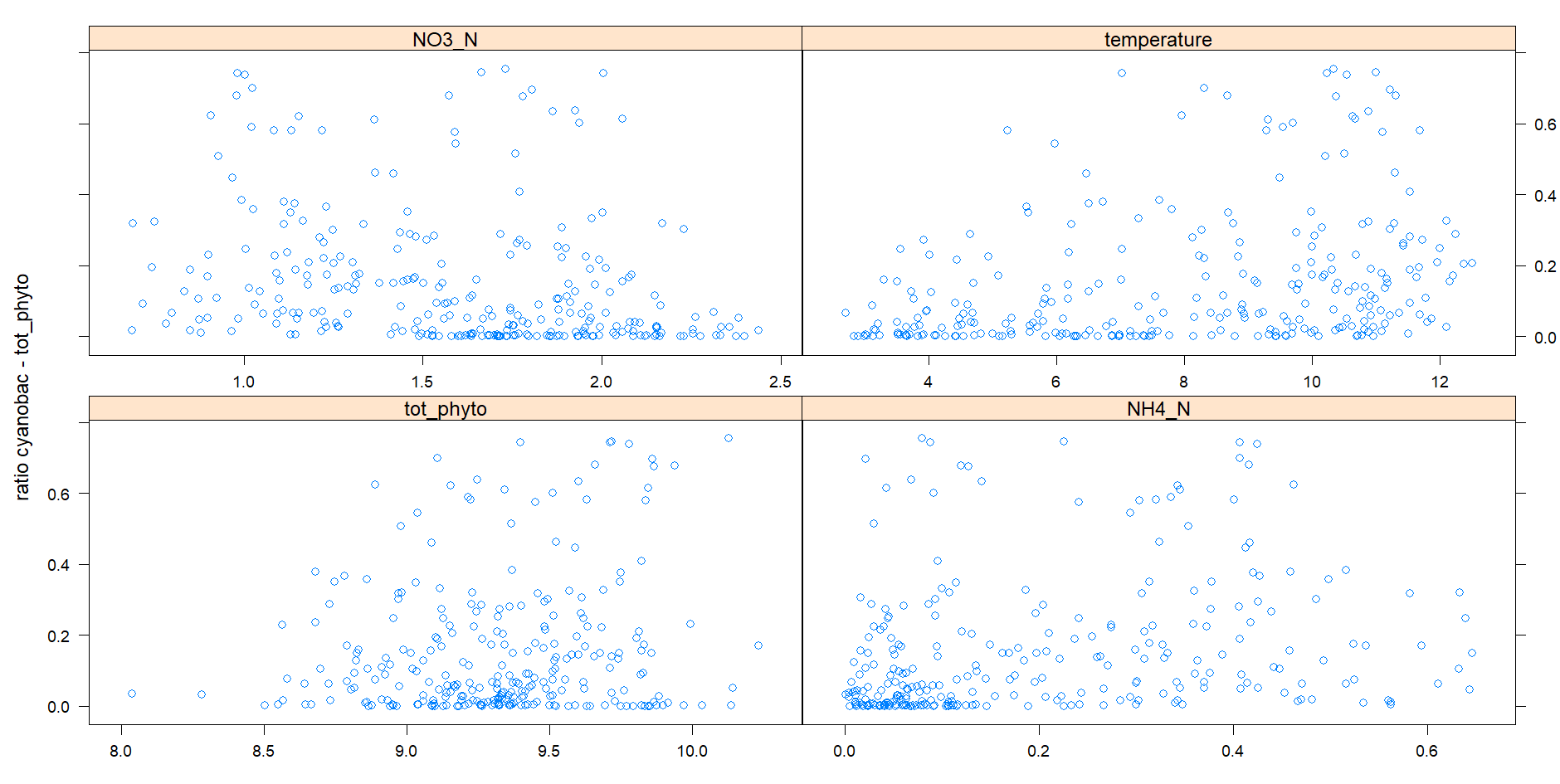
Here are 4 scatter plots describing the relationship between some predictors and the response variable, which is the ratio of total cyanobacteria biovolume to total phytoplankton biovolume





The top one is with normal values, and the bottom one uses log10-transformed values

No imputation was needed at all, since the chemistry and temperature data were always sampled 🡪 this is very good

Now the RF models:

For the ratio cynobac – total phyto:

Random Forest

196 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 175, 176, 177, 175, 175, 175, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.1537050 0.2709662 0.1128926

2 0.1557634 0.2492252 0.1142040

3 0.1567707 0.2439114 0.1146615

4 0.1563569 0.2498777 0.1144255

5 0.1586235 0.2364921 0.1153701

6 0.1587120 0.2366233 0.1153813

7 0.1608737 0.2190679 0.1165071

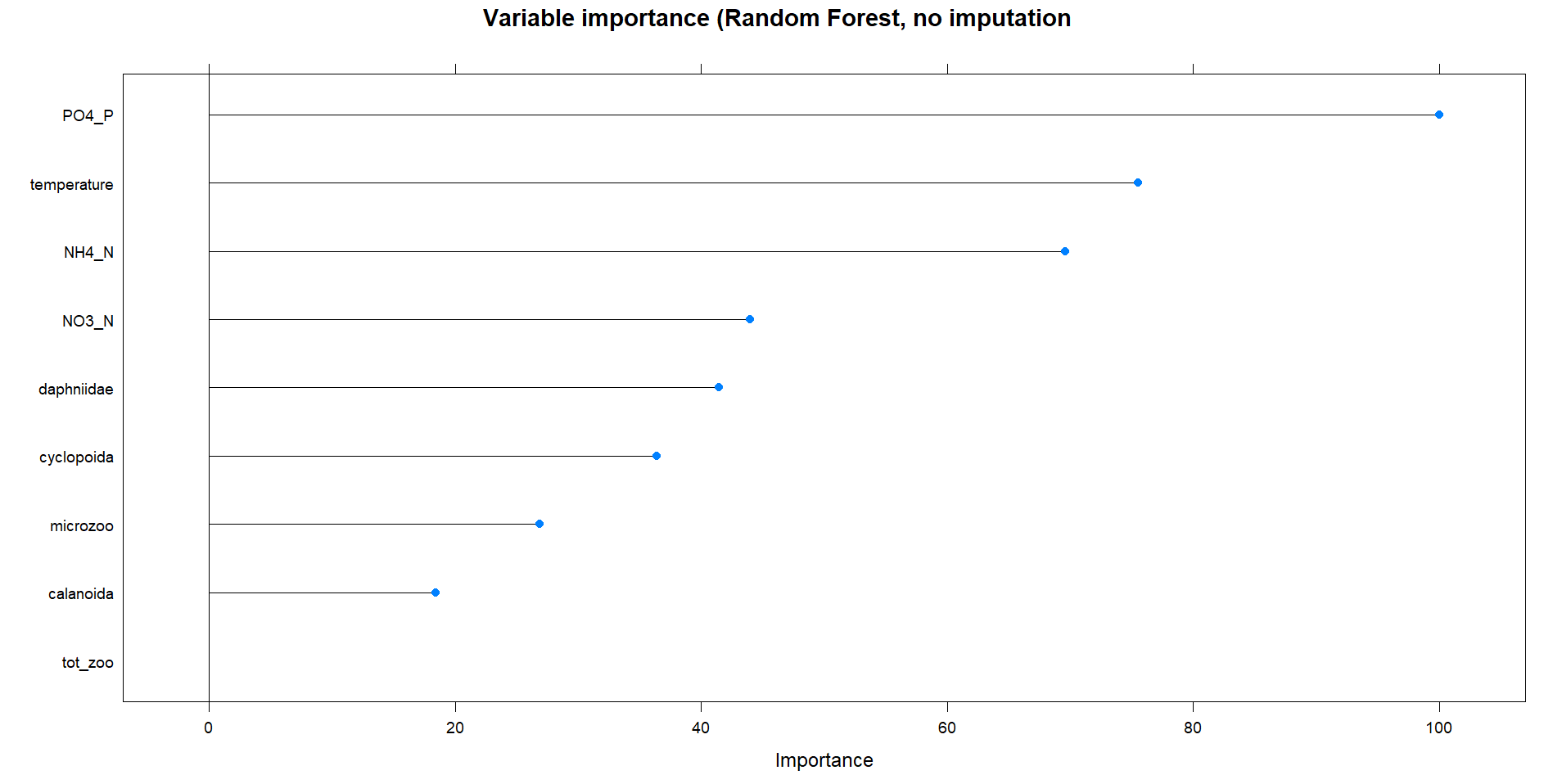
8 0.1603833 0.2224845 0.1161507

9 0.1597046 0.2278140 0.1158230

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

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

RMSE is good, but Rsquared is too low (and the fact that the best mtry is 1 shows the model is not good)



But the 3 top predictors seem to explain the variation pretty well (this plot looks very good)

For the Chroococcales

Random Forest

196 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 175, 175, 175, 175, 177, 176, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.1102638 0.1930612 0.07335769

2 0.1121106 0.1758560 0.07464866

3 0.1124273 0.1845097 0.07426064

4 0.1142749 0.1672897 0.07538757

5 0.1155896 0.1678063 0.07528329

6 0.1158568 0.1687681 0.07480972

7 0.1166633 0.1678166 0.07483966

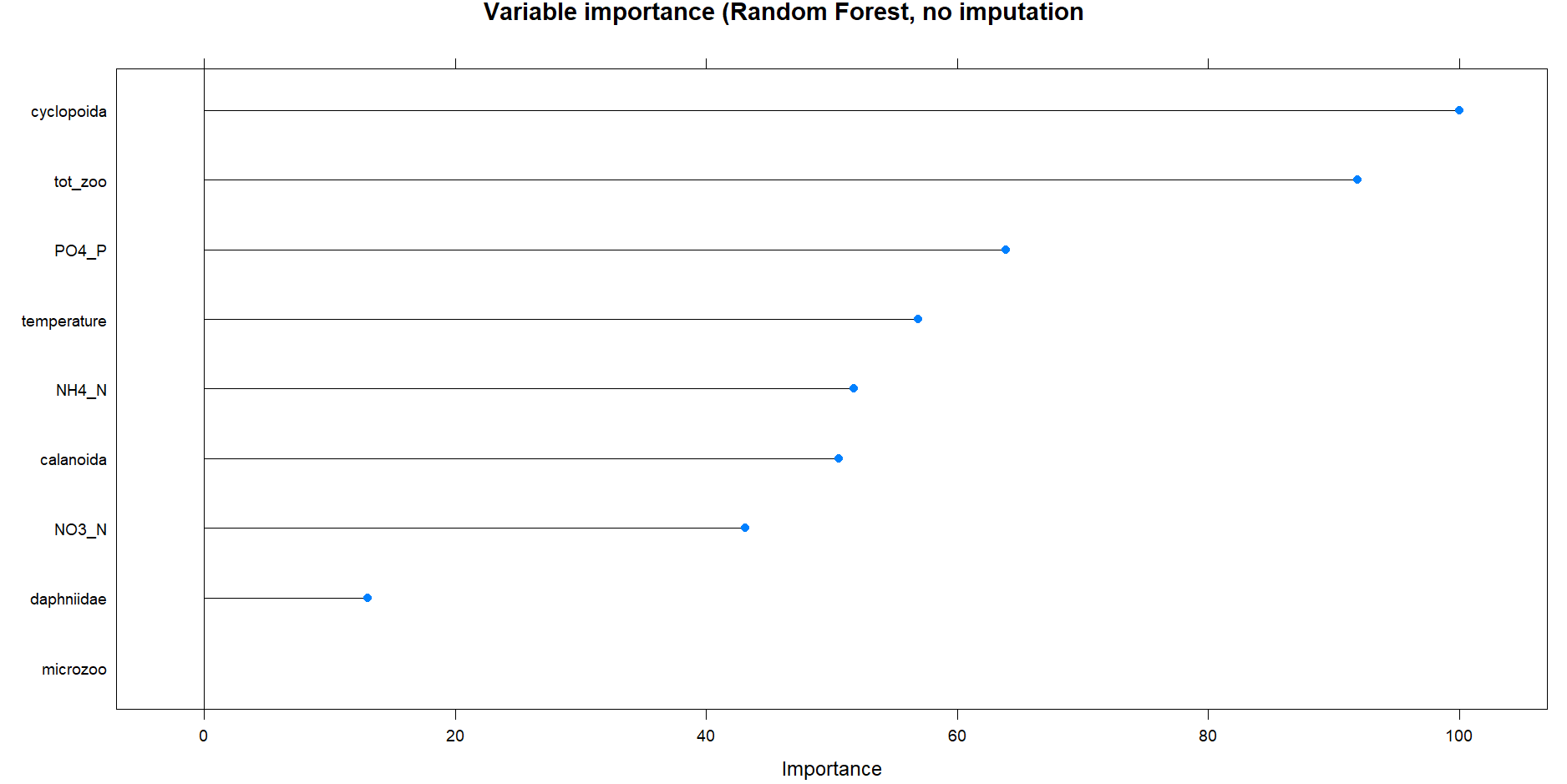
8 0.1172611 0.1698285 0.07487440

9 0.1185677 0.1668068 0.07516694

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

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

Even worse!



Random Forest

196 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 175, 175, 175, 175, 177, 176, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.1265553 0.04594297 0.08490169

2 0.1282732 0.04069518 0.08679766

3 0.1302195 0.03024106 0.08823907

4 0.1323934 0.02260685 0.08970667

5 0.1329103 0.02589359 0.08963520

6 0.1338362 0.02356605 0.09059027

7 0.1342507 0.02343591 0.09092614

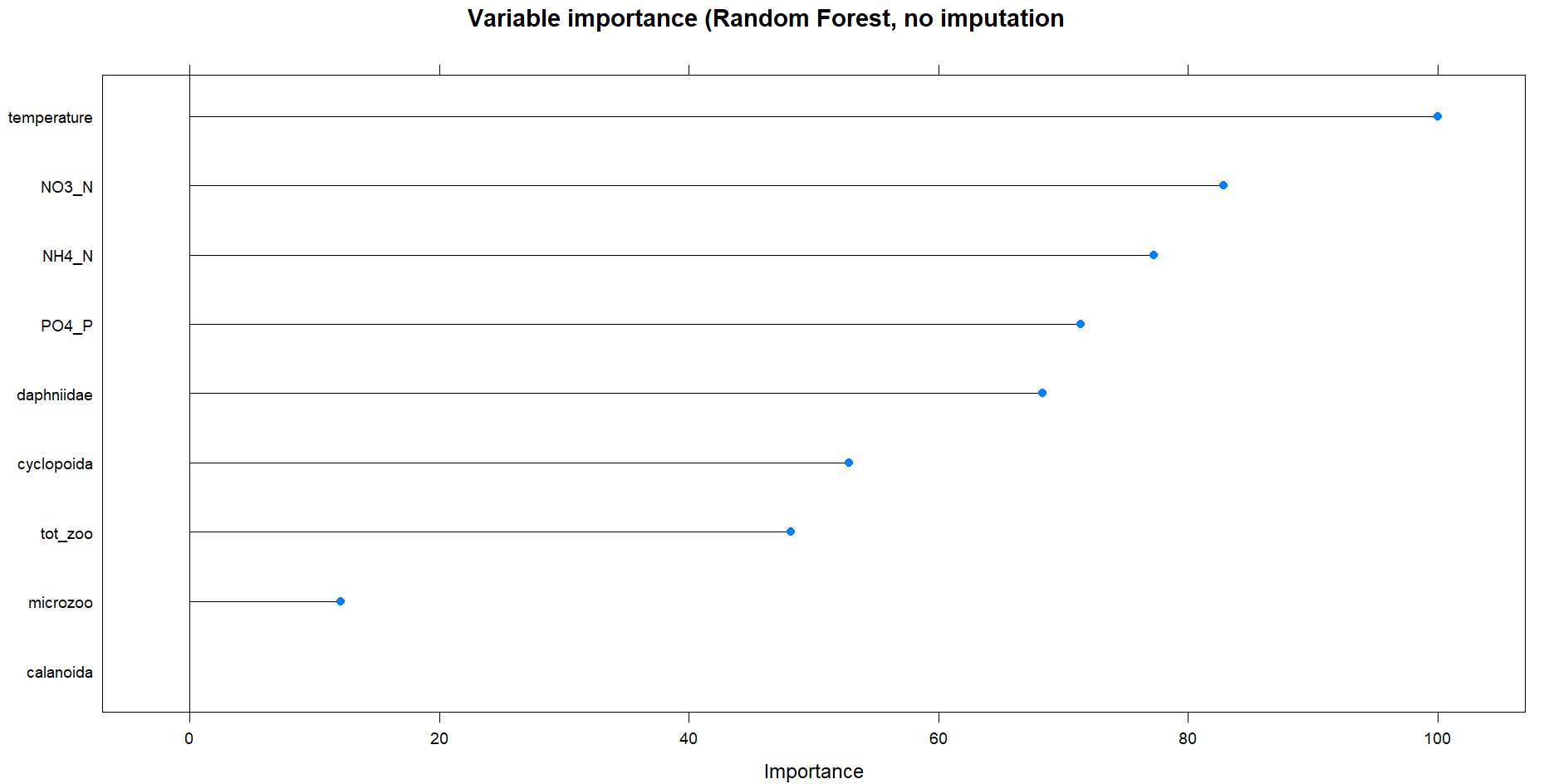
8 0.1349851 0.02339991 0.09111731

9 0.1369026 0.01459354 0.09212267

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

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

I didn’t know Raquared values could get so small…



Random Forest

194 samples

9 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 173, 174, 174, 174, 174, 174, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 0.01000985 0.1725504 0.005114620

2 0.01075885 0.1473236 0.005339951

3 0.01129438 0.1390015 0.005489862

4 0.01188606 0.1267989 0.005677605

5 0.01227031 0.1111826 0.005787667

6 0.01288410 0.1143450 0.005922550

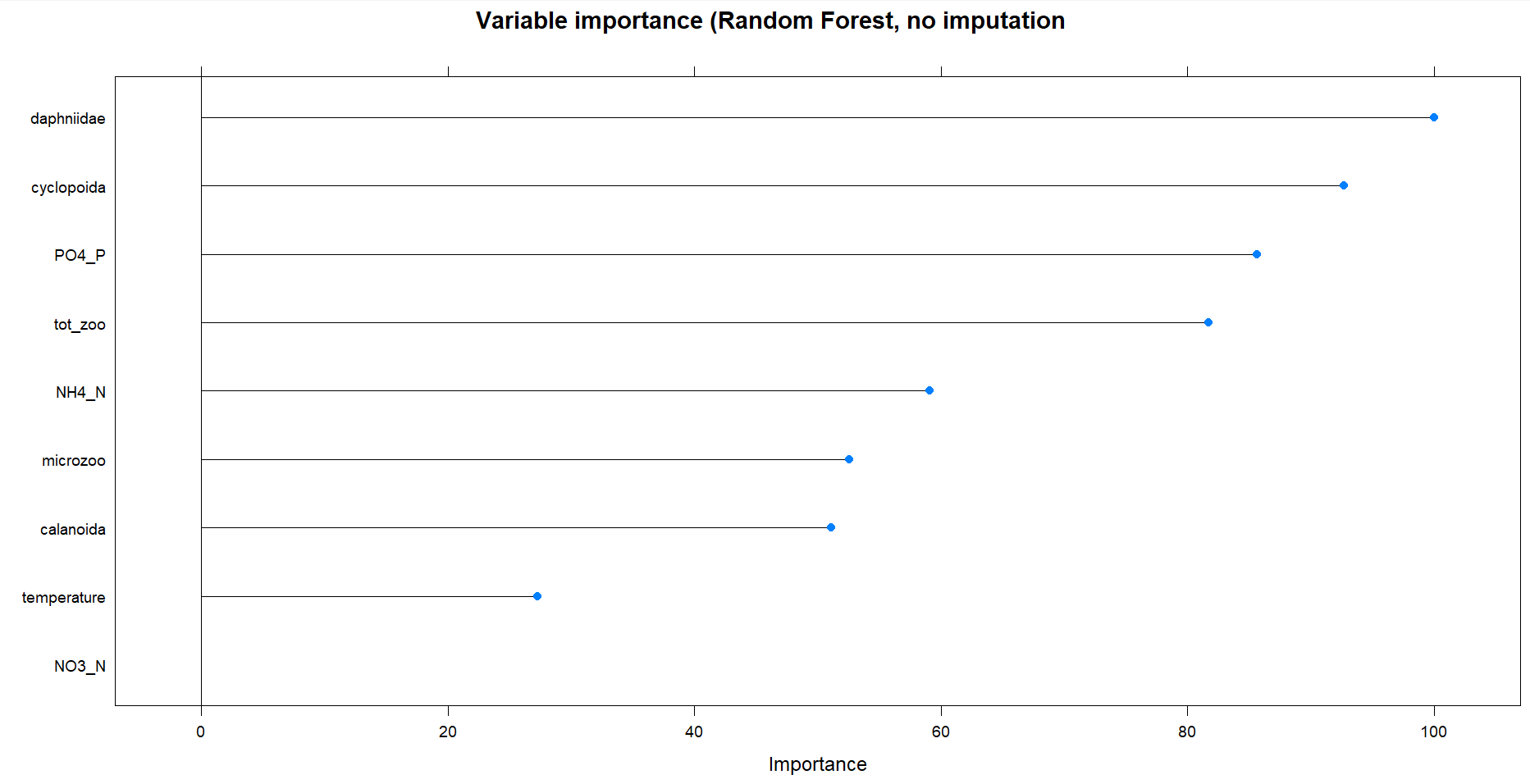
7 0.01303162 0.1141186 0.005944769

8 0.01347838 0.1079251 0.006017266

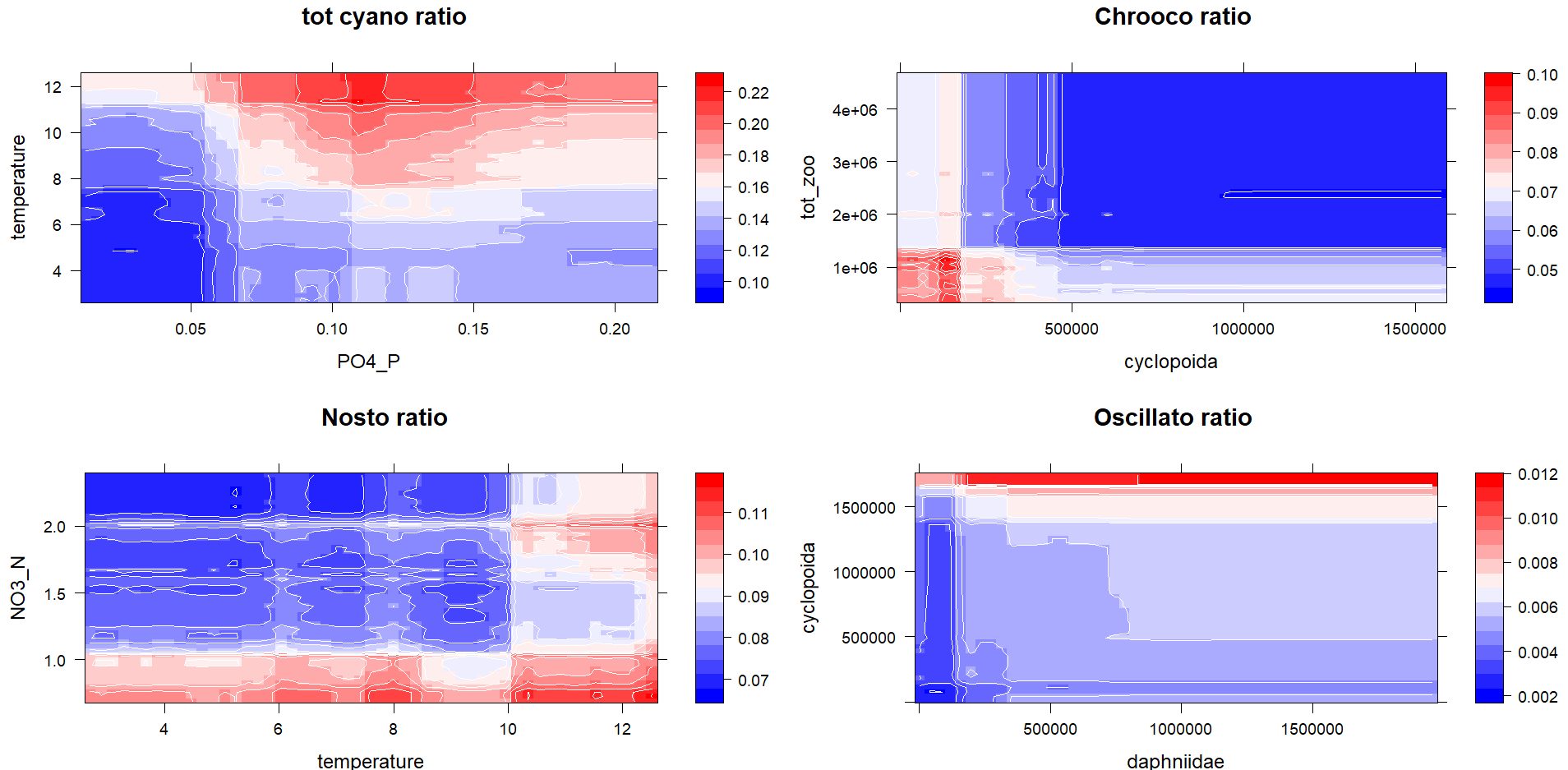
9 0.01397818 0.1139350 0.006127347

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

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



Now I know how to make 3D plots, so here are 8 of them:



The ones below only consider the 2 predictors temperature and phosphate

