WI_chloride_randomforest

Hilary Dugan

3/4/2019

Tidy data

```
dat <- datin %>% dplyr::filter(Chloride < 10000 & Chloride >=0) %>%
  dplyr::mutate(Chloride = ifelse(Chloride == 0, 0.0001, Chloride)) %>%
  dplyr::filter(!(coastdist < 4 | Chloride > 1000)) %>%
  dplyr::filter(state_zoneid == 'State_9' | state_zoneid == 'State_14') %>% # Wisconsin and Minnesota
  dplyr::filter(Date > as.Date('1990-01-01')) %>%
  dplyr::mutate(rdDist_WIMN = pmin(rdDist_WI,rdDist_MN)) %>%
  dplyr::mutate(iws_forest = iws_nlcd2011_pct_41 + iws_nlcd2011_pct_42 +iws_nlcd2011_pct_43) %>%
  dplyr::mutate(iws_develop = iws_nlcd2011_pct_24 + iws_nlcd2011_pct_23) %>%
  dplyr::group_by(lagoslakeid) %>%
  dplyr::summarise_if(is.numeric,funs(mean)) %>%
  dplyr::left_join(distinct(dplyr::select(datin,lagoslakeid,lakeconnection,gnis_name,state_zoneid)))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## please use list() instead
##
## # Before:
## funs(name = f(.)
##
## # After:
## list(name = ~f(.))
## This warning is displayed once per session.
```

```
## Joining, by = "lagoslakeid"
```

```
#adding connectivity dummies
oneHot=function(x) {
  conn = factor(x)
  apply(data.frame(model.matrix(~conn+0)),2,FUN=factor)
}

dat_conn_dummy<-oneHot(dat$lakeconnection)

dat<-cbind(dat,dat_conn_dummy)

log01 <- function(x){log(x + 0.001)} # log of columns
dat_rf <- dat %>%
  mutate_at(vars(Chloride,iws_ha:rdDist_WIMN,iws_forest,iws_develop),log01) %>%
  filter(!is.na(iws_nlcd2011_pct_22))
```

Random Forest model

```
rf_cov <- dat_rf %>% dplyr::select(nhd_lat, nhd_long,
                                   iws_forest,
                                   iws_nlcd2011_pct_81,
```

```
                                    iws_nlcd2011_pct_82,
                                    iws_nlcd2011_pct_21,
                                    iws_nlcd2011_pct_22,
                                    iws_develop,
                                    iws_roaddensity_density_mperha,
                                    rdDist_Interstate,rdDist_WIMN,connDR_LakeStream:connIsolated)

##grid search to select random forest hyperparameters
control <- trainControl(method="oob", number=10, search="random")
rf_random <- train(y=dat_rf$Chloride, x = rf_cov, method="rf",tuneLength=15, trControl=control)
#best r2 model has mtry=4, but I like maximizing mtry to better interpret interactions in forestFloor p
#small grid search for the right sampsize. Lower sampsize decorrelates trees a bit, which is important
sampsize<-c(100,200,500,1000,length(dat_rf$Chloride))
samp_df<-data.frame()

for(i in 1:length(sampsize) ){
rf_sampsize<-randomForest(y=dat_rf$Chloride,
                        x = rf_cov,
                        keep.inbag = T,
                        importance = T,
                        ntree=500,
                        sampsize = sampsize[i],
                       # mtry=4,
                        mtry=length(rf_cov) )
samp_df<-rbind(samp_df,
               data.frame(sampsize=sampsize[i],
                          r2=rf_sampsize$rsq[length(rf_sampsize$rsq)],
                          mse=rf_sampsize$mse[length(rf_sampsize$mse)])
               )
}
samp_df
```

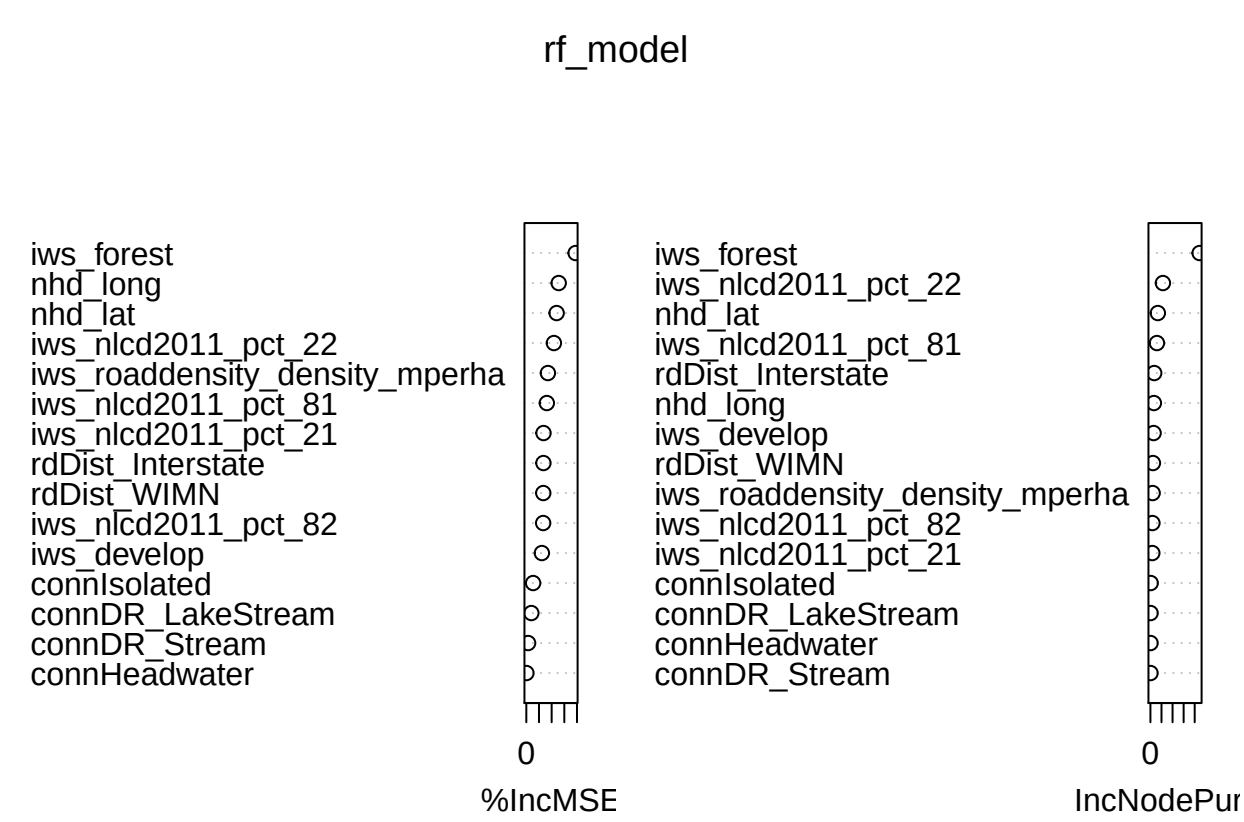
```
##   sampsize        r2       mse
## 1      100 0.7983260 0.5042481
## 2      200 0.8132195 0.4670097
## 3      500 0.8229908 0.4425783
## 4     1000 0.8273348 0.4317169
## 5     1936 0.8280910 0.4298263
```

```
#sampsize around the max is pretty good, though it's not the worst idea to reduce this value to decorre

rf_model<-randomForest(y=dat_rf$Chloride,
                        x = rf_cov,
                        keep.inbag = T,
                        importance = T,
                        ntree=500,
                        sampsize = 1739,
                       # mtry=4,
                        mtry=length(rf_cov) )

rf_model
```

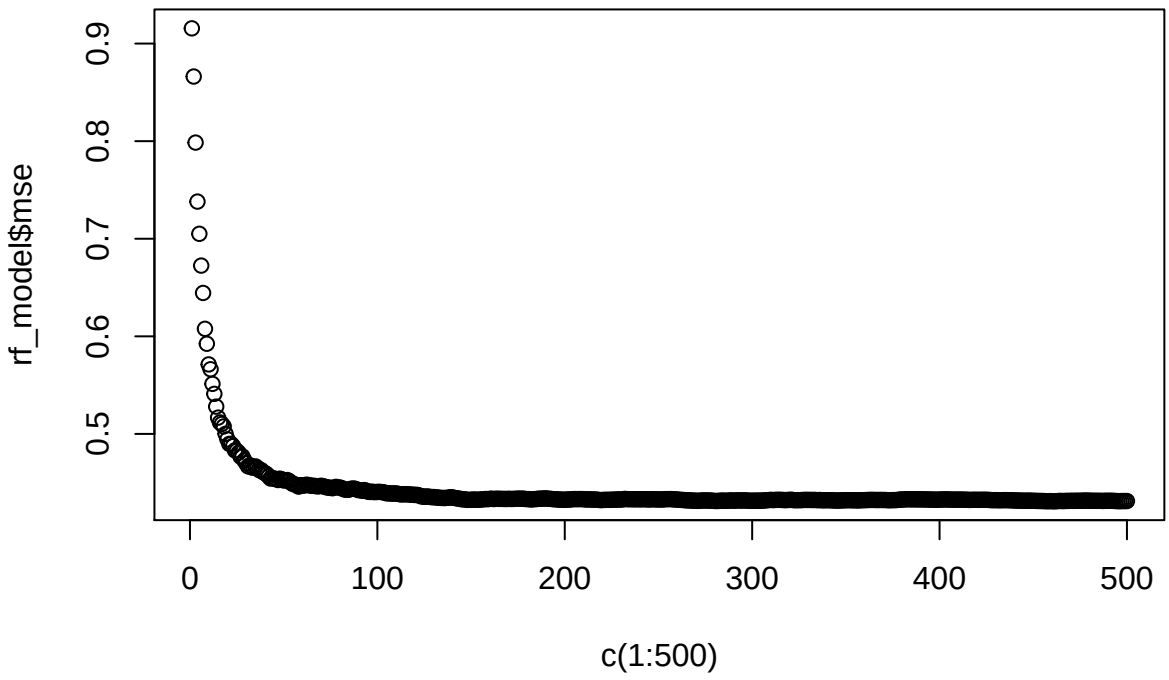
```
##
## Call:
##  randomForest(x = rf_cov, y = dat_rf$Chloride, ntree = 500, mtry = length(rf_cov),      sampsize = 1
```

```
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 15
##
##           Mean of squared residuals: 0.4312293
##                     % Var explained: 82.75
```

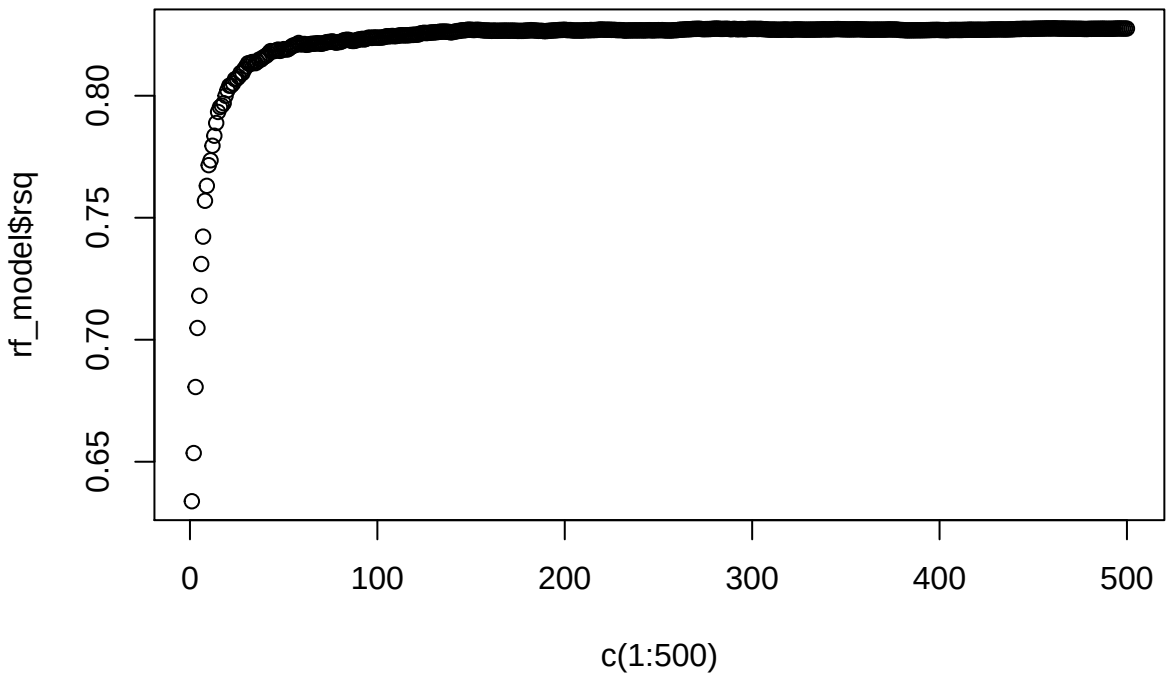
```
varImpPlot(rf_model)
```




```
#you'll notice that your interpretation of the forestfloor plot will change as you adjust mtry and samp

#determining # of trees to stabilize error
  plot(rf_model$mse~c(1:500))
```

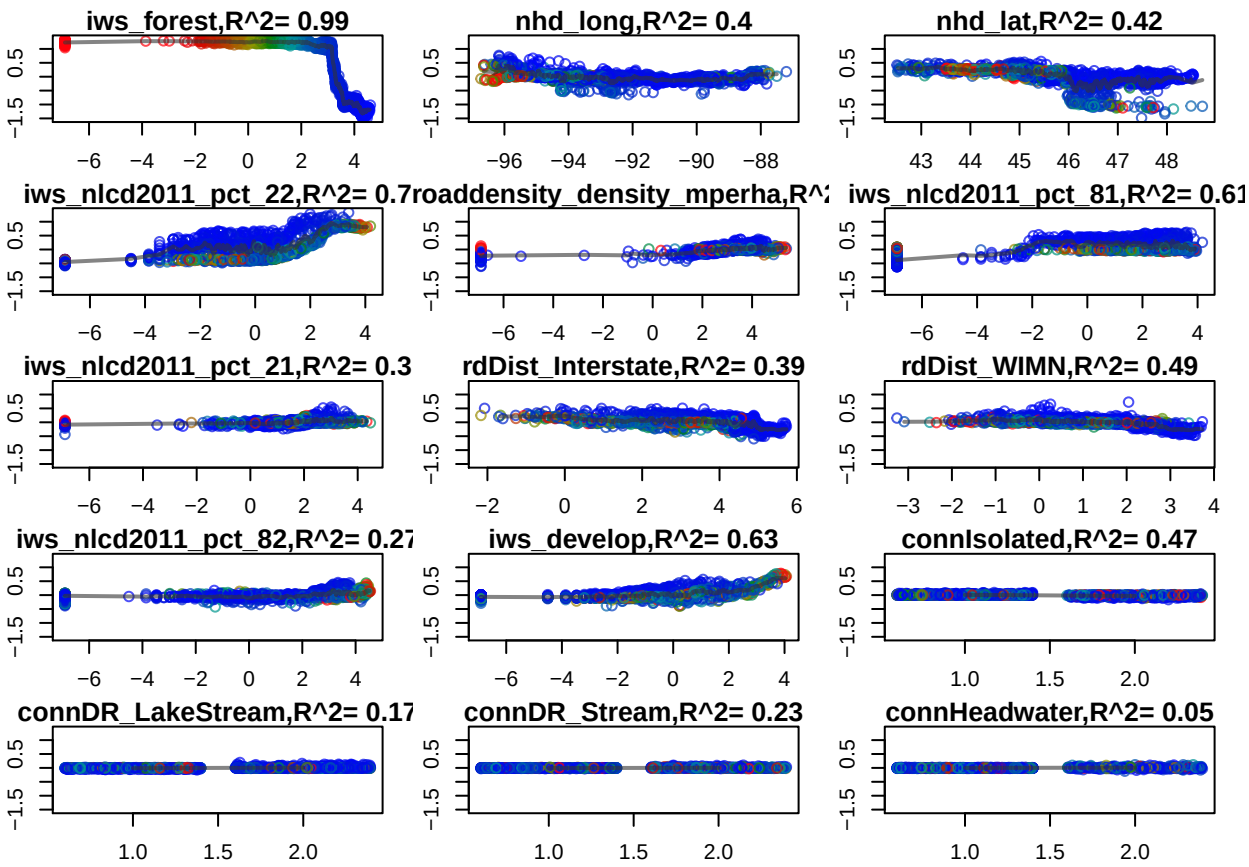



```
plot(rf_model$rsq~c(1:500))
```




```
#pretty good by 100 trees


ff_model<-forestFloor(rf_model,X = rf_cov,y= rf_model$y )
Col = fcol(ff_model, 1, orderByImportance=T)
plot(ff_model, col=Col, orderByImportance=T)
```

```
## [1] "compute goodness-of-fit with leave-one-out k-nearest neighbor(guassian weighting), kknn package
```




```
dat_rf = dat_rf %>% mutate(predicted = rf_model$predicted)
library(lme4)
```

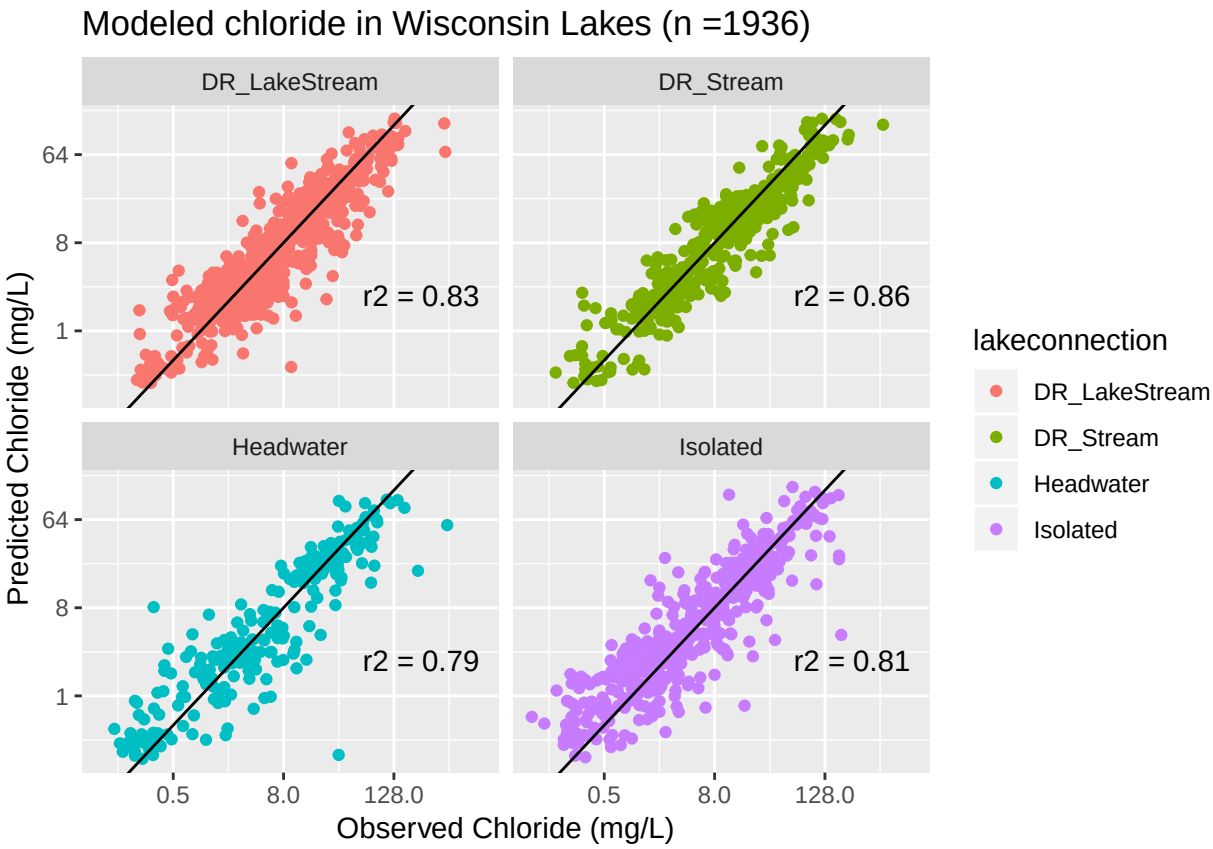
```
## Warning: package 'lme4' was built under R version 3.5.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
fits <- lmList(predicted ~ Chloride | lakeconnection, data=dat_rf)
fits1 = data.frame(r2 = paste0('r2 = ',round(summary(fits)$r.squared,2)), lakeconnection = unique(fits@g
                   Chloride = rep(7,4),
                   predicted = rep(0.1,4))

p = ggplot(dat_rf, aes(x = exp(Chloride), y = exp(predicted), color = lakeconnection)) + geom_point() +
  #scale_colour_viridis_d() +
  facet_wrap(~lakeconnection) +
  ylab('Predicted Chloride (mg/L)') + xlab('Observed Chloride (mg/L)') +
  labs(title = paste0('Modeled chloride in Wisconsin Lakes (n =',nrow(dat_rf),')')) +
  scale_y_continuous(trans = log2_trans()) + scale_x_continuous(trans = log2_trans()) +
  geom_text(data = fits1, aes(label = r2),hjust = 1,vjust = -1, color = 'black')
p
```




```
#plotting residuals over space
library(tigris)
```

```
## To enable
## caching of data, set `options(tigris_use_cache = TRUE)` in your R script or .Rprofile.
```

```
##
## Attaching package: 'tigris'
```

```
## The following object is masked from 'package:graphics':
##
##     plot
```

```
states <- states(cb = TRUE)
```

```
##
  |
  |                                                                      |   0%
  |
  |=                                                                     |   1%
  |
  |=                                                                     |   2%
  |
  |==                                                                    |   3%
  |
  |===                                                                   |   4%
  |
  |===                                                                   |   5%
```

```
|
|====                                                   |   6%
|
|====                                                   |   7%
|
|=====                                                  |   7%
|
|=====                                                  |   8%
|
|======                                                 |   9%
|
|======                                                 |  10%
|
|=======                                                |  10%
|
|=======                                                |  11%
|
|=======                                                |  12%
|
|========                                               |  12%
|
|========                                               |  13%
|
|=========                                              |  14%
|
|=========                                              |  15%
|
|==========                                             |  15%
|
|==========                                             |  16%
|
|===========                                            |  16%
|
|===========                                            |  17%
|
|============                                           |  18%
|
|============                                           |  19%
|
|=============                                          |  19%
|
|=============                                          |  20%
|
|==============                                         |  21%
|
|==============                                         |  22%
|
|===============                                        |  23%
|
|===============                                        |  24%
|
|================                                       |  24%
|
|================                                       |  25%
```

```
|
|=================                                                     |  25%
|
|==================                                                    |  26%
|
|==================                                                    |  27%
|
|===================                                                   |  27%
|
|===================                                                   |  28%
|
|====================                                                  |  29%
|
|====================                                                  |  30%
|
|=====================                                                 |  30%
|
|=====================                                                 |  31%
|
|======================                                                |  32%
|
|======================                                                |  32%
|
|======================                                                |  33%
|
|=======================                                               |  33%
|
|=======================                                               |  34%
|
|========================                                              |  35%
|
|========================                                              |  36%
|
|========================                                              |  36%
|
|=========================                                             |  37%
|
|=========================                                             |  38%
|
|=========================                                             |  38%
|
|==========================                                            |  39%
|
|==========================                                            |  39%
|
|==========================                                            |  40%
|
|===========================                                           |  41%
|
|===========================                                           |  41%
|
|============================                                          |  42%
|
|============================                                          |  43%
```

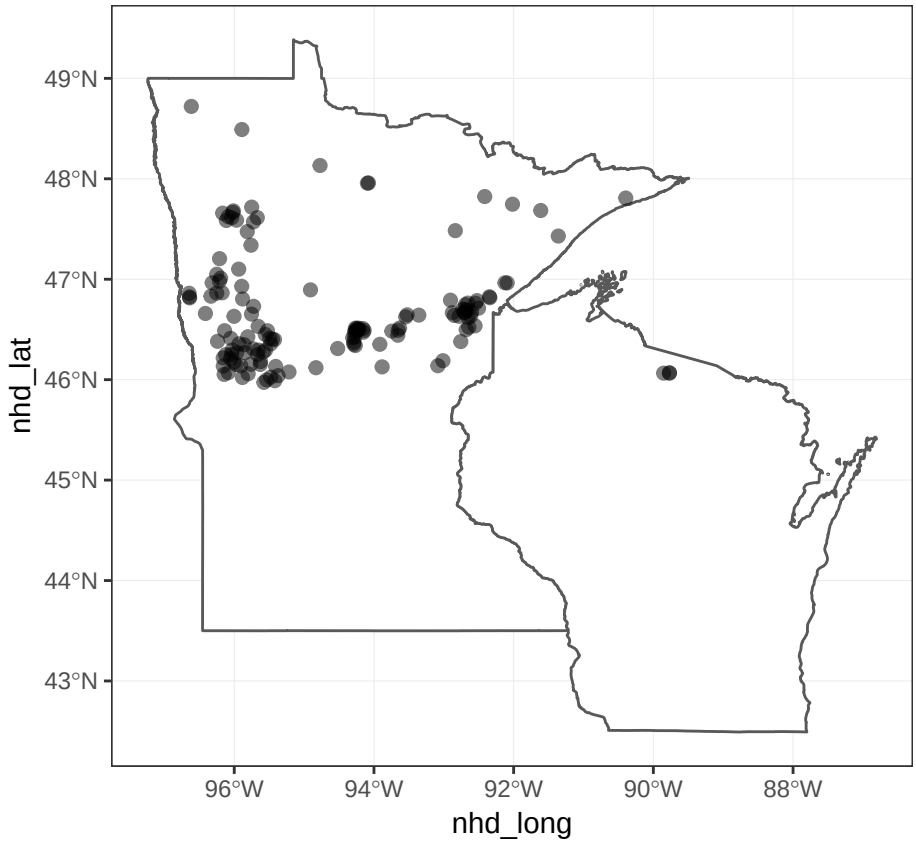
```
|
|============================                                    |  44%
|
|============================                                    |  44%
|
|============================                                    |  45%
|
|============================                                    |  46%
|
|============================                                    |  47%
|
|============================                                    |  47%
|
|============================                                    |  48%
|
|============================                                    |  49%
|
|============================                                    |  50%
|
|============================                                    |  50%
|
|============================                                    |  51%
|
|============================                                    |  52%
|
|============================                                    |  53%
|
|============================                                    |  53%
|
|============================                                    |  54%
|
|============================                                    |  55%
|
|============================                                    |  56%
|
|============================                                    |  56%
|
|============================                                    |  57%
|
|============================                                    |  58%
|
|============================                                    |  59%
|
|============================                                    |  59%
|
|============================                                    |  60%
|
|============================                                    |  61%
|
|============================                                    |  62%
|
|============================                                    |  63%
|
|============================                                    |  64%
```

```
|
|=======================================                               |  64%
|
|=======================================                               |  65%
|
|========================================                              |  66%
|
|========================================                              |  67%
|
|=========================================                             |  67%
|
|=========================================                             |  68%
|
|==========================================                            |  69%
|
|==========================================                            |  70%
|
|==========================================                            |  70%
|
|==========================================                            |  71%
|
|===========================================                           |  72%
|
|===========================================                           |  73%
|
|===========================================                           |  73%
|
|===========================================                           |  74%
|
|============================================                          |  75%
|
|============================================                          |  76%
|
|============================================                          |  76%
|
|============================================                          |  77%
|
|============================================                          |  78%
|
|=============================================                         |  78%
|
|=============================================                         |  79%
|
|=============================================                         |  79%
|
|=============================================                         |  80%
|
|==============================================                        |  81%
|
|==============================================                        |  81%
|
|==============================================                        |  82%
|
|==============================================                        |  82%
```

```
  |
  |=================================================             |  83%
  |
  |=================================================             |  84%
  |
  |==================================================            |  84%
  |
  |==================================================            |  85%
  |
  |===================================================           |  86%
  |
  |===================================================           |  87%
  |
  |====================================================          |  87%
  |
  |====================================================          |  88%
  |
  |=====================================================         |  88%
  |
  |=====================================================         |  89%
  |
  |=====================================================         |  90%
  |
  |======================================================        |  90%
  |
  |======================================================        |  91%
  |
  |=======================================================       |  92%
  |
  |=======================================================       |  93%
  |
  |========================================================      |  93%
  |
  |========================================================      |  94%
  |
  |========================================================      |  95%
  |
  |=========================================================     |  95%
  |
  |=========================================================     |  96%
  |
  |==========================================================    |  96%
  |
  |==========================================================    |  97%
  |
  |===========================================================   |  98%
  |
  |============================================================  |  99%
  |
  |======================================================================| 100%
```

```r
states_sf<- st_as_sf(states)
dat_rf$residuals<-dat_rf$predicted-dat_rf$Chloride
```

```
r_map<-ggplot(data=dat_rf) +
  geom_sf(data=states_sf[states_sf$NAME %in% c("Wisconsin", "Minnesota"),], fill="white")+
  geom_point(aes(x=nhd_long, y=nhd_lat, col=abs(residuals), size = abs(residuals)), alpha=.5 )+
  scale_color_viridis_c(option="magma")+
  theme_bw()


#plotting the location of the weird points in the nhd_lat forest floor plot
ggplot()+
  geom_sf(data=states_sf[states_sf$NAME %in% c("Wisconsin", "Minnesota"),], fill="white")+
  geom_point(data=dat_rf[ff_model$FCmatrix[,"nhd_lat"]< (-.5),],aes(x=nhd_long, y=nhd_lat), size=2, alpl
  theme_bw()
```



Prediction for LAGOS

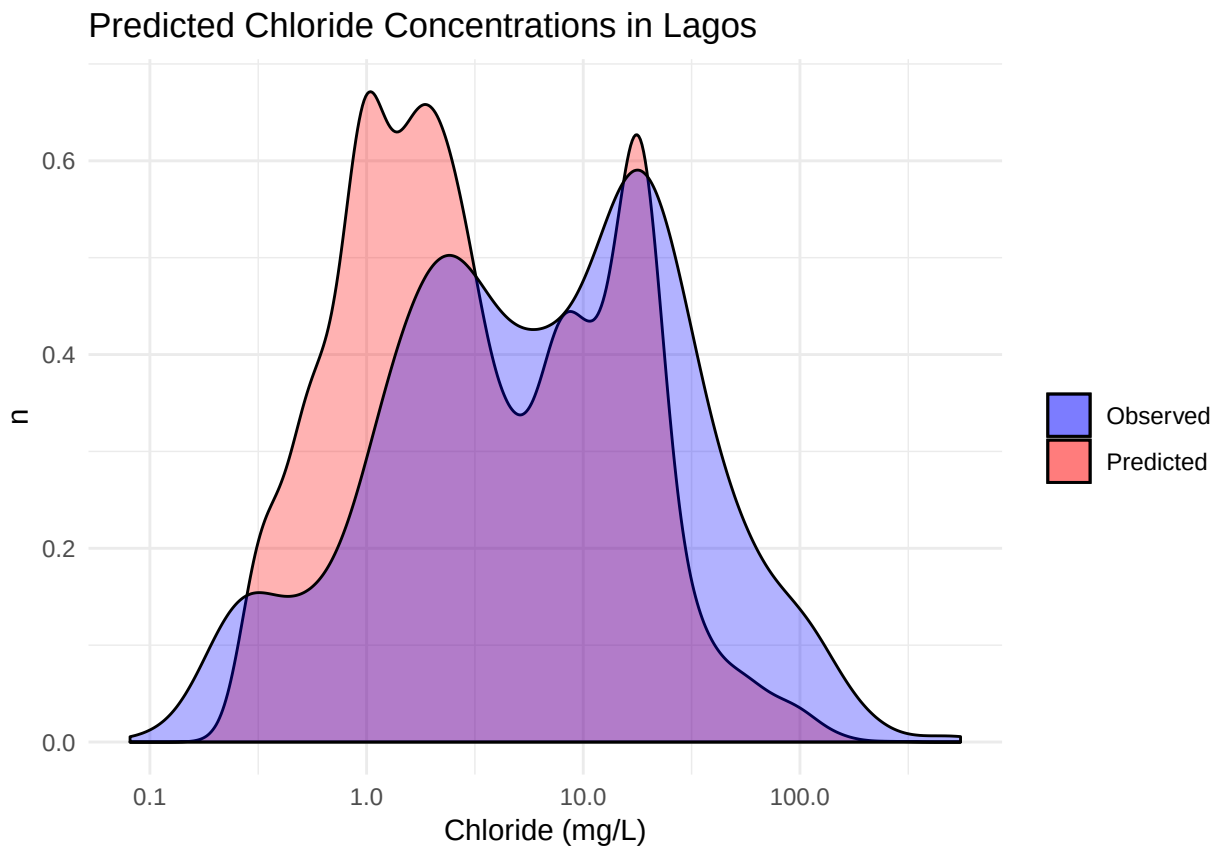
```
datPred2 = read_csv('data/LAGOS_allLakes.csv') %>%
  dplyr::filter(state_zoneid == 'State_9' | state_zoneid == 'State_14') %>% # Wisconsin and Minnesota
  dplyr::mutate(rdDist_WIMN = pmin(rdDist_WI,rdDist_MN)) %>%
  dplyr::mutate(iws_forest = iws_nlcd2011_pct_41 + iws_nlcd2011_pct_42 +iws_nlcd2011_pct_43) %>%
  dplyr::mutate(iws_develop = iws_nlcd2011_pct_24 + iws_nlcd2011_pct_23)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   nhdid = col_character(),
##   gnis_name = col_character(),
##   state_zoneid = col_character(),
```

```
##   lakeconnection = col_character(),
##   latewisconsinglaciation_glacial = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
datPred2 <- datPred2 %>%
  mutate_at(vars(iws_nlcd2011_ha_0:iws_develop),log01) %>%
  filter(!is.na(iws_nlcd2011_pct_22))

dat_conn_dummy<-oneHot(datPred2$lakeconnection)
datPred2<-cbind(datPred2,dat_conn_dummy)
datPred2<- datPred2 %>% select(nhd_lat,nhd_long,
                               iws_forest,
                               iws_nlcd2011_pct_81,
                               iws_nlcd2011_pct_82,
                               iws_nlcd2011_pct_21,
                               iws_nlcd2011_pct_22,
                               iws_develop,
                          iws_roaddensity_density_mperha,
                          rdDist_Interstate,rdDist_WIMN,connDR_LakeStream:connIsolated)

preds <- predict(rf_model, newdata = datPred2)
datPred2$clPred = preds

# Plot prediction histogram
ggplot() +
  geom_density(data = datPred2, aes(x = exp(clPred), fill = "r"), alpha = 0.3) +
  geom_density(data = dat_rf, aes(x = exp(Chloride), fill = "b"), alpha = 0.3) +
  scale_colour_manual(name ="", values = c("r" = "red", "b" = "blue"), labels=c("b" = "Observed", "r" =
  scale_fill_manual(name ="", values = c("r" = "red", "b" = "blue"), labels=c("b" = "Observed", "r" = "
  scale_x_continuous(trans='log10') +
  labs(x = "Chloride (mg/L)", y = "n") +
  ggtitle("Predicted Chloride Concentrations in Lagos") +
  theme_minimal()
```


Predicted Chloride Concentrations in Lagos
0.6
0.4
0.2
0.0
n
0.1
1.0
10.0
100.0
Chloride (mg/L)
Observed
Predicted