STA141A-Tree Models-ATW-CMD-Markdown

Andrew T. Weakley, Christina De Cesaris

12/15/2020

— Step 1: Data loading and procressing —

```
## --- Part a: Upload Metadata for samples ---
path_data<-file.path(getwd(),"data")
META_DATA<-as_tibble(read.csv(file.path(path_data,"IMPROVE_metadata.csv")))
## --- Filter samples from Korea and Canada ---
US_META<-META_DATA %>% filter(Country %nin% c("KR","CA"))


## --- Filter stats not in continental US ---
US_META<-META_DATA %>% filter(State %nin% c("HI","AK","VI"))

## --- Part b: Load samples data ---
DATA<-as_tibble(read.csv(file.path(path_data,"IMPROVE_2015_data_w_UNC_v2.csv")))

## --- Part c: Select samples from SW given site identifiers from SW_META table ("C
ode")
US_DATA_all<-as_tibble(DATA %>% filter(SiteCode %in% US_META$Code))
```

```
# Let's identify any samples that (grossly) violate PM2.5 mass balances
# PM2.5 (=Y) cannot be negative!
# Since there's some probability that PM2.5 is negative due to errors at low concen
tration, we may use PM2.5 uncertainties to remove samples that fall outside -3*PM2.
5_UNC.
# In this way, we don't risk censoring the data but do remove likely erroneous dat
a.
US_DATA_all<-US_DATA_all %>% dplyr::filter(PM2.5 > -3*PM2.5_UNC)
```

```
exclude<-c("PM10","POC","ammNO3","ammSO4","SOIL","SeaSalt","OC1","OC2","OC3","OC
4","EC1","EC2","EC3","fAbs_MDL","fAbs")
US_DATA_LRG<- US_DATA_all %>% dplyr::select(!contains(exclude) & !matches("_UNC") |
matches("PM2.5_UNC"))
any(is.na(US_DATA_LRG))
```

```
## [1] TRUE
```

```
US_DATA_LRG<-US_DATA_LRG[which(complete.cases(US_DATA_LRG)),]
any(is.na(US_DATA_LRG))
```

```
## [1] FALSE
```

```
set.seed(123)
## --- Instead of random partitioning, I will partition by first sorting samples by
SiteCode and DATE (already done) and place every other sample in the test set.
# --- This data has seasonality. Sorting by date therefore ensures seasonality is e
quivalent between datasets
n<-nrow(US_DATA_LRG)
ind_test<-seq(1,n,2)
US_DATA_LRG_test<-US_DATA_LRG[ind_test,]
US_DATA_LRG<-US_DATA_LRG[-ind_test,]
```

— Step 2: mclust for GMMs —

```
## --- Normalize US data by PM2.5 conc --
US_DATA_LRG_PM_norm<-US_DATA_LRG %>% dplyr::select(everything()/"PM2.5")
#rename_with()
```

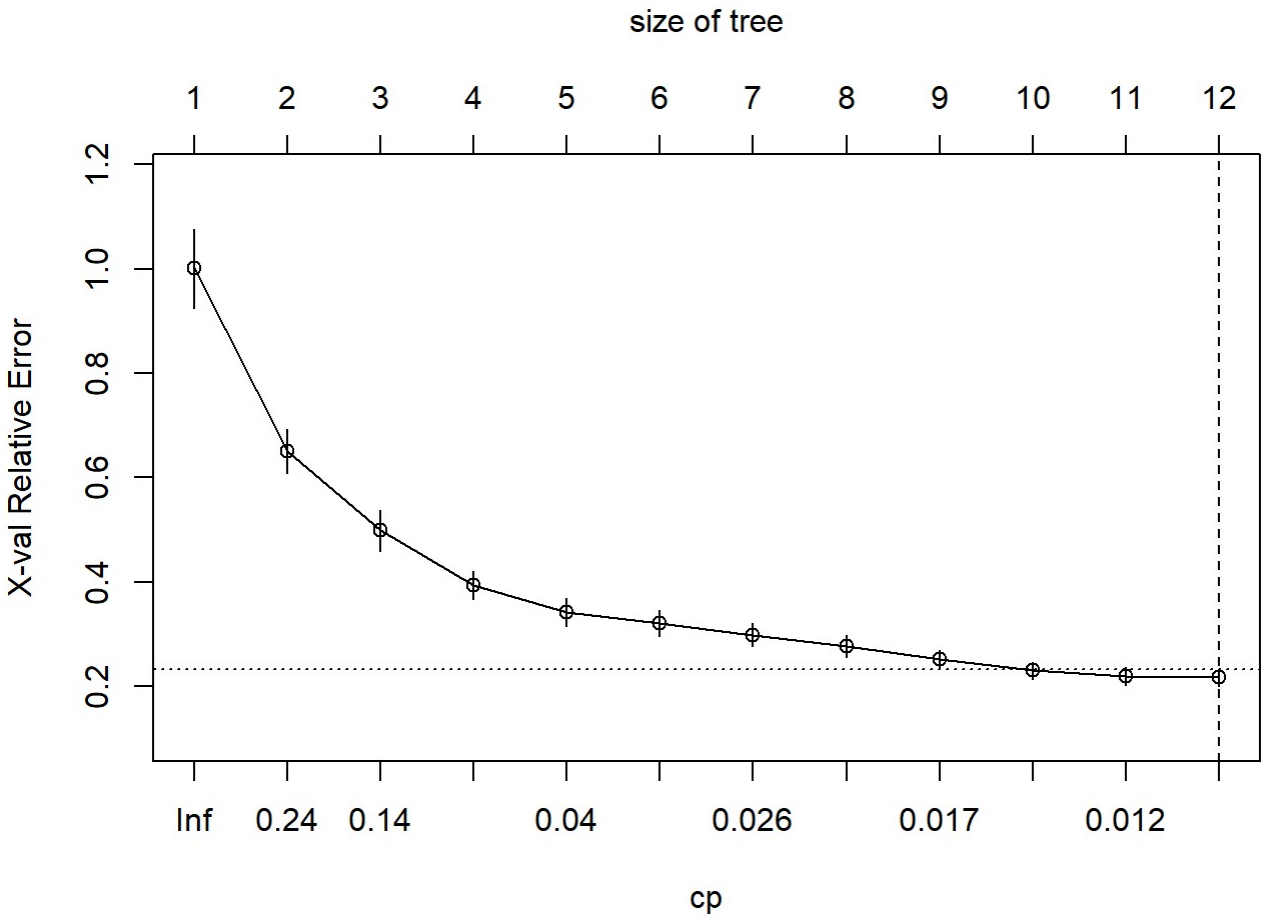
— Tree Regression —

1. initial fits

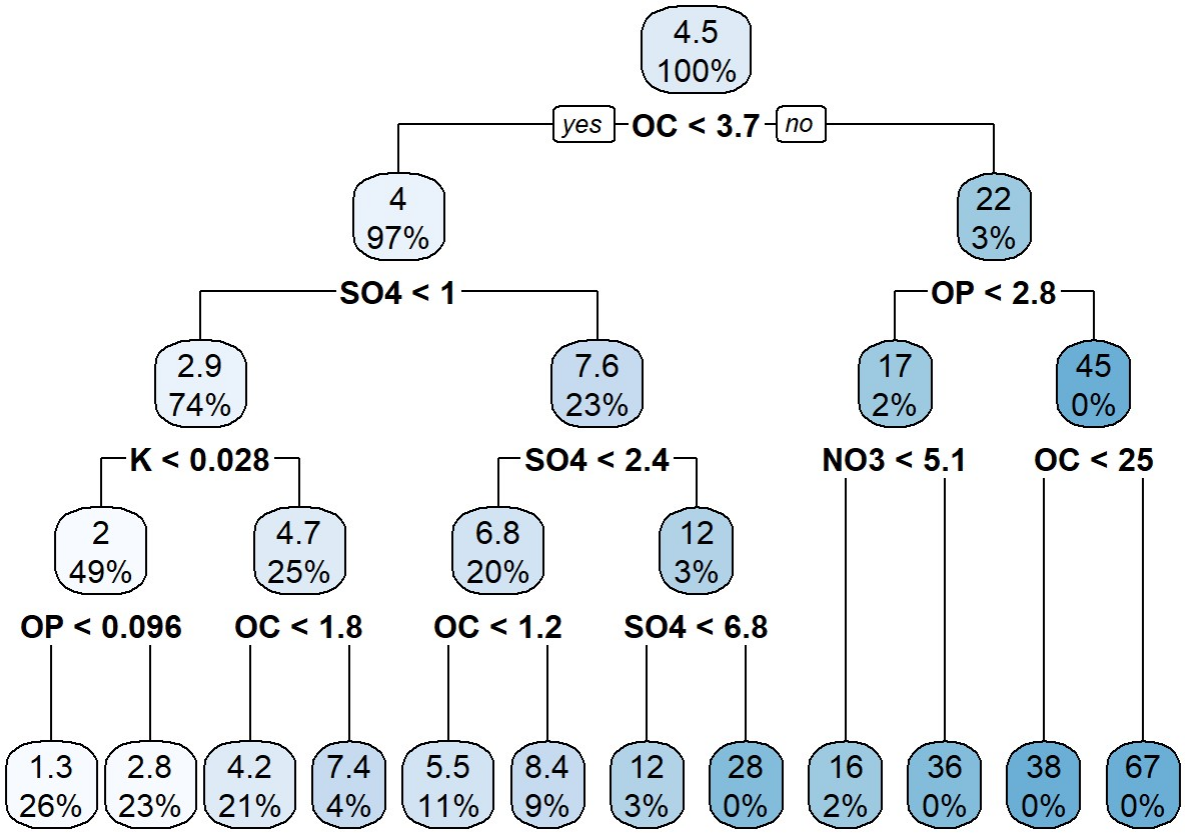
```
fit1 <- rpart(
    formula = PM2.5 ~ .-PM2.5_UNC,
    data    = US_DATA_LRG,
    method  = "anova",
    control = list(minsplit = 10, maxdepth = 20, xval = 10)
)
fit1
```

```
## n= 8647
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 8647 219000.9000  4.549602
##    2) OC< 3.693035 8390  90149.2100  4.017542
##      4) SO4< 1.0019 6378  26722.0200  2.890826
##        8) K< 0.02811 4237   6138.1860  1.992608
##         16) OP< 0.09625 2207   1385.3870  1.256258 *
##         17) OP>=0.09625 2030   2255.1420  2.793161 *
##        9) K>=0.02811 2141  10400.5200  4.668382
##         18) OC< 1.81129 1819   6084.1310  4.182025 *
##         19) OC>=1.81129 322   1455.4710  7.415849 *
##      5) SO4>=1.0019 2012  29663.6600  7.589211
##       10) SO4< 2.38635 1728  13878.1800  6.802017
##         20) OC< 1.16163 964   4246.3170  5.500812 *
##         21) OC>=1.16163 764   5940.2230  8.443853 *
##       11) SO4>=2.38635 284   8199.4220 12.378890
##         22) SO4< 6.76665 270   3368.5850 11.546900 *
##         23) SO4>=6.76665 14   1039.4700 28.424520 *
##    3) OC>=3.693035 257  48939.1500 21.919180
##      6) OP< 2.8095 214  12843.1400 17.344850
##       12) NO3< 5.11497 201   5742.5780 16.107340 *
##       13) NO3>=5.11497 13   2033.4390 36.478600 *
##      7) OP>=2.8095 43   9333.0300 44.684480
##       14) OC< 25.39485 33   2434.7920 38.052000 *
##       15) OC>=25.39485 10    656.0985 66.571640 *
```

```
#pairs(US_DATA_LRG[which(sapply(US_DATA_LRG, is.numeric))])
plotcp(fit1)

abline(v = 12, lty = "dashed")
```




```
rpart.plot(fit1)
```




```
summary(fit1)
```

```
## Call:
## rpart(formula = PM2.5 ~ . - PM2.5_UNC, data = US_DATA_LRG, method = "anova",
##     control = list(minsplit = 10, maxdepth = 20, xval = 10))
##   n= 8647
##
##            CP nsplit rel error    xerror       xstd
## 1  0.36489610      0 1.0000000 1.0003518 0.07605944
## 2  0.15417066      1 0.6351039 0.6501399 0.04173009
## 3  0.12220485      2 0.4809332 0.4985611 0.03998255
## 4  0.04649899      3 0.3587284 0.3941103 0.02711683
## 5  0.03463939      4 0.3122294 0.3421852 0.02628566
## 6  0.02850279      5 0.2775900 0.3213551 0.02532643
## 7  0.02313744      6 0.2490872 0.2992325 0.02255095
## 8  0.01731210      7 0.2259498 0.2773869 0.02106606
## 9  0.01685673      8 0.2086377 0.2520433 0.01790144
## 10 0.01306348      9 0.1917809 0.2305255 0.01659266
## 11 0.01140478     10 0.1787175 0.2197864 0.01640755
## 12 0.01000000     11 0.1673127 0.2171949 0.01646286
##
## Variable importance
##       OC       OP      SO4        S       EC SiteCode        K       SE
##       28       21       11       10        7        5        3        2
##       ZN      NO3     Date       MN       FE       SI       TI       AL
##        2        2        2        1        1        1        1        1
##       NI        V       PB
##        1        1        1
##
## Node number 1: 8647 observations,    complexity param=0.3648961
##   mean=4.549602, MSE=25.32681
##   left son=2 (8390 obs) right son=3 (257 obs)
##   Primary splits:
##       OC < 3.693035 to the left,  improve=0.3648961, (0 missing)
##       OP < 0.838775 to the left,  improve=0.3426816, (0 missing)
##       EC < 0.477345 to the left,  improve=0.3353746, (0 missing)
##       K  < 0.078975 to the left,  improve=0.3108213, (0 missing)
##       ZN < 0.003445 to the left,  improve=0.2197968, (0 missing)
##   Surrogate splits:
##       OP  < 0.815255 to the left,  agree=0.986, adj=0.545, (0 split)
##       EC  < 0.8572   to the left,  agree=0.978, adj=0.257, (0 split)
##       PB  < 0.03269  to the left,  agree=0.971, adj=0.027, (0 split)
##       K   < 0.70079  to the left,  agree=0.971, adj=0.027, (0 split)
##       NO3 < 10.94156 to the left,  agree=0.971, adj=0.027, (0 split)
##
## Node number 2: 8390 observations,    complexity param=0.1541707
##   mean=4.017542, MSE=10.74484
##   left son=4 (6378 obs) right son=5 (2012 obs)
##   Primary splits:
##       SO4 < 1.0019   to the left,  improve=0.3745293, (0 missing)
##       S   < 0.35026  to the left,  improve=0.3738907, (0 missing)
##       K   < 0.032105 to the left,  improve=0.3698196, (0 missing)
```

```
##       OC  < 1.06415  to the left,  improve=0.3622052, (0 missing)
##       OP  < 0.237865 to the left,  improve=0.3482584, (0 missing)
##   Surrogate splits:
##       S        < 0.379    to the left,  agree=0.979, adj=0.915, (0 split)
##       SiteCode splits as  LL-LLLLRRLLLLRLLLRRLR-LLLRLLLRLLRLLLL--LRLRRRRLLLLR-LL
LLLLLLRL--LRLLLLRLLL-LRLLLLLLRLLLLLRLLLLLLLLLLLRLLLLLLLLLL-LLLLRLLLLRLLLLLRLLLLLRLL
L-RLLRLRLLLLLLL-L-LR-LLRLLLLLLRLL-L, agree=0.816, adj=0.233, (0 split)
##       OP       < 0.27197  to the left,  agree=0.813, adj=0.221, (0 split)
##       SE       < 0.000395 to the left,  agree=0.807, adj=0.193, (0 split)
##       ZN       < 0.004075 to the left,  agree=0.800, adj=0.164, (0 split)
##
## Node number 3: 257 observations,    complexity param=0.1222049
##   mean=21.91918, MSE=190.4247
##   left son=6 (214 obs) right son=7 (43 obs)
##   Primary splits:
##       OP       < 2.8095   to the left,  improve=0.5468624, (0 missing)
##       OC       < 15.0474  to the left,  improve=0.5137808, (0 missing)
##       K        < 0.18656  to the left,  improve=0.3267747, (0 missing)
##       SiteCode splits as  L--LR---LLL-LLLLLLL-R--LLLLLLL---RLLR--LL--L--LLRR--R-
RLR-------LLLL-RLLRL-RLRLLRL-----RL-LLR---LL-LL-L-LLR-LLL-----L-L--LL-L-R----L-LLL
---LLRL-LL-LLLL--L-R--LLL--L--LLLL-L, improve=0.3126466, (0 missing)
##       CL       < 0.014375 to the left,  improve=0.2974700, (0 missing)
##   Surrogate splits:
##       OC       < 12.1005  to the left,  agree=0.988, adj=0.930, (0 split)
##       SiteCode splits as  L--LL---LLL-LLLLLLL-L--LLLLLLL---LLRL--LL--L--LLRL--L-
LLR-------LLLL-RLLRL-LLRLLLL-----LL-LLR---LL-LL-L-LLL-LLL-----L-L--LL-L-R----L-LLL
---LLRL-LL-LLRL--L-R--LLL--L--LLLL-L, agree=0.879, adj=0.279, (0 split)
##       EC       < 2.921035 to the left,  agree=0.844, adj=0.070, (0 split)
##       Date     splits as  L-LL--L-LLL-L-L---L-LL--LLLLLL-LLLLL--LLLLRL-LLLLLL--L
--LLLLL--L--------------LLL-L-LL-L-LL-LL-LL-LLLLLLLLLLLLRLLL-L-LLLL, agree=0.840, a
dj=0.047, (0 split)
##       NI       < -6.5e-05 to the right, agree=0.837, adj=0.023, (0 split)
##
## Node number 4: 6378 observations,    complexity param=0.04649899
##   mean=2.890826, MSE=4.189719
##   left son=8 (4237 obs) right son=9 (2141 obs)
##   Primary splits:
##       K  < 0.02811  to the left,  improve=0.3810835, (0 missing)
##       OC < 0.93938  to the left,  improve=0.3553606, (0 missing)
##       OP < 0.16584  to the left,  improve=0.3182269, (0 missing)
##       BR < 0.001165 to the left,  improve=0.2981336, (0 missing)
##       EC < 0.11325  to the left,  improve=0.2856111, (0 missing)
##   Surrogate splits:
##       MN < 0.000965 to the left,  agree=0.814, adj=0.446, (0 split)
##       FE < 0.037735 to the left,  agree=0.813, adj=0.444, (0 split)
##       SI < 0.14337  to the left,  agree=0.811, adj=0.437, (0 split)
##       TI < 0.003515 to the left,  agree=0.811, adj=0.437, (0 split)
##       AL < 0.058885 to the left,  agree=0.809, adj=0.430, (0 split)
##
## Node number 5: 2012 observations,    complexity param=0.03463939
##   mean=7.589211, MSE=14.74337
```

```
##   left son=10 (1728 obs) right son=11 (284 obs)
##   Primary splits:
##       SO4 < 2.38635  to the left,  improve=0.2557358, (0 missing)
##       CR  < 0.000665 to the left,  improve=0.2396845, (0 missing)
##       S   < 0.67805  to the left,  improve=0.2374953, (0 missing)
##       V   < 0.0029   to the left,  improve=0.2366923, (0 missing)
##       K   < 0.148315 to the left,  improve=0.2313183, (0 missing)
##   Surrogate splits:
##       S        < 0.874975 to the left,  agree=0.970, adj=0.785, (0 split)
##       SiteCode splits as  LL-LLLLLLLLL-LL--LLLR-LLLL--LLL-LLLL---LLLLLLLL-LLL-L-
L-LLL-LL--LLLLL-LLL--LL-LLL-LL-LLLLL-L-L--L-L--LLLLLLLLLLL-LLLLLLLLLL-LLLLLLL-LLL-L
L-LLLL-LL-LLLLL-L-LL-LLL-LLL-LL-L-L, agree=0.869, adj=0.074, (0 split)
##       NI       < 0.0017   to the left,  agree=0.869, adj=0.074, (0 split)
##       V        < 0.004775 to the left,  agree=0.869, adj=0.074, (0 split)
##       SE       < 0.002    to the left,  agree=0.867, adj=0.060, (0 split)
##
## Node number 6: 214 observations,    complexity param=0.02313744
##   mean=17.34485, MSE=60.01467
##   left son=12 (201 obs) right son=13 (13 obs)
##   Primary splits:
##       NO3      < 5.11497  to the left,  improve=0.3945392, (0 missing)
##       SiteCode splits as  L--LL---LLL-LLLLLLL-R--LLLLLLL---LLLL--LL--L--LLLR--L-
LL--------LLLL--LLRL-LL-LLRL-----RL-LLL---LL-LL-R-LLL-LLL-----L-L--LL-L------L-LLL
---LLLL-LL-LLLL--L----LLL--L--LLLL-L, improve=0.3915086, (0 missing)
##       Date     splits as  R-LL--L-LRL-L-L---L-LL--LLLLLL-RLRLL--LLRL-L-RLRLRL--L
--LLLLL--L--------------LLL-L-LL-L-LL-LL-LL-LLLLLLLLLLLL-LLL-L-LLLL, improve=0.3623
386, (0 missing)
##       K        < 0.25385  to the left,  improve=0.3110933, (0 missing)
##       SO4      < 4.978265 to the left,  improve=0.2807596, (0 missing)
##   Surrogate splits:
##       Date     splits as  L-LL--L-LRL-L-L---L-LL--LLLLLL-LLRLL--LLLL-L-LRRLLL--L
--LLLLL--L--------------LLL-L-LL-L-LL-LL-LL-LLLLLLLLLLLL-LLL-L-LLLL, agree=0.958, a
dj=0.308, (0 split)
##       SE       < 0.00251  to the left,  agree=0.958, adj=0.308, (0 split)
##       S        < 2.37981  to the left,  agree=0.958, adj=0.308, (0 split)
##       SO4      < 6.5974   to the left,  agree=0.953, adj=0.231, (0 split)
##       SiteCode splits as  L--LL---LLL-LLLLLLL-L--LLLLLLL---LLLL--LL--L--LLLR--L-
LL--------LLLL--LLLL-LL-LLLL-----LL-LLL---LL-LL-L-LLL-LLL-----L-L--LL-L------L-LLL
---LLLL-LL-LLLL--L----LLL--L--LLLL-L, agree=0.949, adj=0.154, (0 split)
##
## Node number 7: 43 observations,    complexity param=0.02850279
##   mean=44.68448, MSE=217.0472
##   left son=14 (33 obs) right son=15 (10 obs)
##   Primary splits:
##       OC       < 25.39485 to the left,  improve=0.6688224, (0 missing)
##       OP       < 7.920375 to the left,  improve=0.6323192, (0 missing)
##       EC       < 1.432395 to the left,  improve=0.5185635, (0 missing)
##       SiteCode splits as  ----R----L----------R--L---------L-LR---------L-L---R-
R-L-------R----L--R--L-L--L------L----L-------------R------------------R----L-L
-------L-------L-----L-------------L--, improve=0.4939978, (0 missing)
##       CL       < 0.02494  to the left,  improve=0.4923968, (0 missing)
```

```
##   Surrogate splits:
##       OP       < 6.893135 to the left,  agree=0.953, adj=0.8, (0 split)
##       SiteCode splits as  ----R----L----------L--L---------L-LR---------L-L---R-
L-L-------L----L--L--L-L--L------L----L-------------R------------------R----L-L
-------L-------L-----L-------------L--, agree=0.884, adj=0.5, (0 split)
##       EC       < 1.903125 to the left,  agree=0.884, adj=0.5, (0 split)
##       Date     splits as  ----------------------------------------L----------L
------------------------------R--L--L--L----LL---LLRLR--LL------LR-, agree=0.860, a
dj=0.4, (0 split)
##       CL       < 0.031235 to the left,  agree=0.837, adj=0.3, (0 split)
##
## Node number 8: 4237 observations,    complexity param=0.01140478
##   mean=1.992608, MSE=1.44871
##   left son=16 (2207 obs) right son=17 (2030 obs)
##   Primary splits:
##       OP  < 0.09625  to the left,  improve=0.4069047, (0 missing)
##       K   < 0.012955 to the left,  improve=0.4025839, (0 missing)
##       OC  < 0.532155 to the left,  improve=0.3954182, (0 missing)
##       S   < 0.098965 to the left,  improve=0.3856653, (0 missing)
##       SO4 < 0.26405  to the left,  improve=0.3640022, (0 missing)
##   Surrogate splits:
##       OC  < 0.423575 to the left,  agree=0.875, adj=0.739, (0 split)
##       EC  < 0.048955 to the left,  agree=0.793, adj=0.568, (0 split)
##       K   < 0.012705 to the left,  agree=0.745, adj=0.468, (0 split)
##       S   < 0.11291  to the left,  agree=0.734, adj=0.445, (0 split)
##       SO4 < 0.29305  to the left,  agree=0.726, adj=0.429, (0 split)
##
## Node number 9: 2141 observations,    complexity param=0.01306348
##   mean=4.668382, MSE=4.857784
##   left son=18 (1819 obs) right son=19 (322 obs)
##   Primary splits:
##       OC < 1.81129  to the left,  improve=0.2750742, (0 missing)
##       OP < 0.28626  to the left,  improve=0.1937325, (0 missing)
##       FE < 0.20788  to the left,  improve=0.1722669, (0 missing)
##       K  < 0.06402  to the left,  improve=0.1667441, (0 missing)
##       EC < 0.22105  to the left,  improve=0.1650228, (0 missing)
##   Surrogate splits:
##       OP   < 0.38288  to the left,  agree=0.921, adj=0.475, (0 split)
##       EC   < 0.31343  to the left,  agree=0.880, adj=0.202, (0 split)
##       Date splits as  LLLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLLL--LLLLLLLLLLLLLLLLLLL
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLRLLRRLLLLLLLLLLLLLL, agree=0.853, adj=
0.022, (0 split)
##       CR   < 0.000805 to the left,  agree=0.852, adj=0.016, (0 split)
##       SE   < 0.00147  to the left,  agree=0.851, adj=0.012, (0 split)
##
## Node number 10: 1728 observations,    complexity param=0.01685673
##   mean=6.802017, MSE=8.031354
##   left son=20 (964 obs) right son=21 (764 obs)
##   Primary splits:
##       OC < 1.16163  to the left,  improve=0.2660032, (0 missing)
##       K  < 0.0549   to the left,  improve=0.2628756, (0 missing)
```

```
##       FE < 0.404225 to the left,  improve=0.2057942, (0 missing)
##       SI < 1.209615 to the left,  improve=0.2053201, (0 missing)
##       AL < 0.636465 to the left,  improve=0.2052434, (0 missing)
##   Surrogate splits:
##       OP       < 0.32231  to the left,  agree=0.850, adj=0.661, (0 split)
##       EC       < 0.206845 to the left,  agree=0.789, adj=0.522, (0 split)
##       SiteCode splits as  RL-LLLLLRLLL-LL--RLLR-RLLL--LLL-RRLL---RLLLRLLL-LRL-L-
L-LLL-RL--LLLLL-RLL--LL-RRL-LR-LLLLR-R-L--L-L--RLLLLLLRLLL-RLLRLRLLLL-LLLLLLL-LRL-R
L-RLLR-LL-RLLLL-L-LR-LLR-LLL-LL-R-L, agree=0.683, adj=0.284, (0 split)
##       CU       < 0.000695 to the left,  agree=0.676, adj=0.267, (0 split)
##       ZN       < 0.00406  to the left,  agree=0.674, adj=0.263, (0 split)
##
## Node number 11: 284 observations,    complexity param=0.0173121
##   mean=12.37889, MSE=28.8712
##   left son=22 (270 obs) right son=23 (14 obs)
##   Primary splits:
##       SO4 < 6.76665  to the left,  improve=0.4623944, (0 missing)
##       PB  < 0.016615 to the left,  improve=0.4265695, (0 missing)
##       MN  < 0.008395 to the left,  improve=0.4185620, (0 missing)
##       ZN  < 0.04141  to the left,  improve=0.4171126, (0 missing)
##       S   < 1.93679  to the left,  improve=0.4099518, (0 missing)
##   Surrogate splits:
##       S    < 2.573415 to the left,  agree=0.993, adj=0.857, (0 split)
##       NI   < 0.00304  to the left,  agree=0.975, adj=0.500, (0 split)
##       V    < 0.008205 to the left,  agree=0.975, adj=0.500, (0 split)
##       ZN   < 0.053945 to the left,  agree=0.975, adj=0.500, (0 split)
##       Date splits as  LL-LLLLL--L-R-----L-L---------L-------LLL-L--LL-LLLL-LLLL-
LLL-LL-LLLLLLLL-LLLRLLLLRLLLLRLLLLLLLLLLLLLLLLLLRLLLLLLRL-LLLLLL, agree=0.972, adj=
0.429, (0 split)
##
## Node number 12: 201 observations
##   mean=16.10734, MSE=28.57004
##
## Node number 13: 13 observations
##   mean=36.4786, MSE=156.4184
##
## Node number 14: 33 observations
##   mean=38.052, MSE=73.78158
##
## Node number 15: 10 observations
##   mean=66.57164, MSE=65.60985
##
## Node number 16: 2207 observations
##   mean=1.256258, MSE=0.627724
##
## Node number 17: 2030 observations
##   mean=2.793161, MSE=1.110907
##
## Node number 18: 1819 observations
##   mean=4.182025, MSE=3.344767
##
```

```
## Node number 19: 322 observations
##   mean=7.415849, MSE=4.520096
##
## Node number 20: 964 observations
##   mean=5.500812, MSE=4.404893
##
## Node number 21: 764 observations
##   mean=8.443853, MSE=7.775161
##
## Node number 22: 270 observations
##   mean=11.5469, MSE=12.47624
##
## Node number 23: 14 observations
##   mean=28.42452, MSE=74.24786
```

```
pred <- predict(fit1, US_DATA_LRG_test)

ModelMetrics::rmse(pred,US_DATA_LRG_test$PM2.5_UNC)
```

```
## [1] 6.17434
```

```
ModelMetrics::gini(pred,US_DATA_LRG_test$PM2.5_UNC)
```

```
## [1] 0.9282768
```

```
#0.03234565
fit1$cptable
```

```
##            CP nsplit rel error    xerror       xstd
## 1  0.36489610      0 1.0000000 1.0003518 0.07605944
## 2  0.15417066      1 0.6351039 0.6501399 0.04173009
## 3  0.12220485      2 0.4809332 0.4985611 0.03998255
## 4  0.04649899      3 0.3587284 0.3941103 0.02711683
## 5  0.03463939      4 0.3122294 0.3421852 0.02628566
## 6  0.02850279      5 0.2775900 0.3213551 0.02532643
## 7  0.02313744      6 0.2490872 0.2992325 0.02255095
## 8  0.01731210      7 0.2259498 0.2773869 0.02106606
## 9  0.01685673      8 0.2086377 0.2520433 0.01790144
## 10 0.01306348      9 0.1917809 0.2305255 0.01659266
## 11 0.01140478     10 0.1787175 0.2197864 0.01640755
## 12 0.01000000     11 0.1673127 0.2171949 0.01646286
```

```
 fit2 <- rpart(
    formula = PM2.5 ~ .-PM2.5_UNC,
    data    = US_DATA_LRG,
    method  = "anova",
    control = list(minsplit = 10, maxdepth = 12, xval = 10)
)
fit2
```

```
## n= 8647
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 8647 219000.9000  4.549602
##    2) OC< 3.693035 8390  90149.2100  4.017542
##      4) SO4< 1.0019 6378  26722.0200  2.890826
##        8) K< 0.02811 4237   6138.1860  1.992608
##         16) OP< 0.09625 2207   1385.3870  1.256258 *
##         17) OP>=0.09625 2030   2255.1420  2.793161 *
##        9) K>=0.02811 2141  10400.5200  4.668382
##         18) OC< 1.81129 1819   6084.1310  4.182025 *
##         19) OC>=1.81129 322   1455.4710  7.415849 *
##      5) SO4>=1.0019 2012  29663.6600  7.589211
##       10) SO4< 2.38635 1728  13878.1800  6.802017
##         20) OC< 1.16163 964   4246.3170  5.500812 *
##         21) OC>=1.16163 764   5940.2230  8.443853 *
##       11) SO4>=2.38635 284   8199.4220 12.378890
##         22) SO4< 6.76665 270   3368.5850 11.546900 *
##         23) SO4>=6.76665 14   1039.4700 28.424520 *
##    3) OC>=3.693035 257  48939.1500 21.919180
##      6) OP< 2.8095 214  12843.1400 17.344850
##       12) NO3< 5.11497 201   5742.5780 16.107340 *
##       13) NO3>=5.11497 13   2033.4390 36.478600 *
##      7) OP>=2.8095 43   9333.0300 44.684480
##       14) OC< 25.39485 33   2434.7920 38.052000 *
##       15) OC>=25.39485 10    656.0985 66.571640 *
```

```
fit2$cptable
```

```
##            CP nsplit rel error    xerror       xstd
## 1  0.36489610      0 1.0000000 1.0001045 0.07604518
## 2  0.15417066      1 0.6351039 0.6436667 0.04050769
## 3  0.12220485      2 0.4809332 0.4934282 0.03891217
## 4  0.04649899      3 0.3587284 0.3756975 0.02425669
## 5  0.03463939      4 0.3122294 0.3228140 0.02394118
## 6  0.02850279      5 0.2775900 0.2981885 0.02293460
## 7  0.02313744      6 0.2490872 0.2716021 0.02061702
## 8  0.01731210      7 0.2259498 0.2524880 0.01702999
## 9  0.01685673      8 0.2086377 0.2389885 0.01607148
## 10 0.01306348      9 0.1917809 0.2217493 0.01557486
## 11 0.01140478     10 0.1787175 0.2090751 0.01550154
## 12 0.01000000     11 0.1673127 0.2050727 0.01531889
```

2. use a grid search method to find the optimal hyper-parameters for a single tree model

```
hyper_grid <- expand.grid(
  minsplit = seq(5, 20, 1),
  maxdepth = seq(8, 15, 1)
)

head(hyper_grid)
```

```
##   minsplit maxdepth
## 1        5        8
## 2        6        8
## 3        7        8
## 4        8        8
## 5        9        8
## 6       10        8
```

```
# total number of combinations
nrow(hyper_grid)
```

```
## [1] 128
```

```
models <- list() #best method i've found for doing this--but computationally expens
ive...

for (i in 1:nrow(hyper_grid)) {

  # get minsplit, maxdepth values at row i
  minsplit <- hyper_grid$minsplit[i]
  maxdepth <- hyper_grid$maxdepth[i]

  # train a model and store in the list
  models[[i]] <- rpart(
   formula = PM2.5 ~ .-PM2.5_UNC,
    data    = US_DATA_LRG,
    method  = "anova",
    control = list(minsplit = minsplit, maxdepth = maxdepth)
    )
}
```

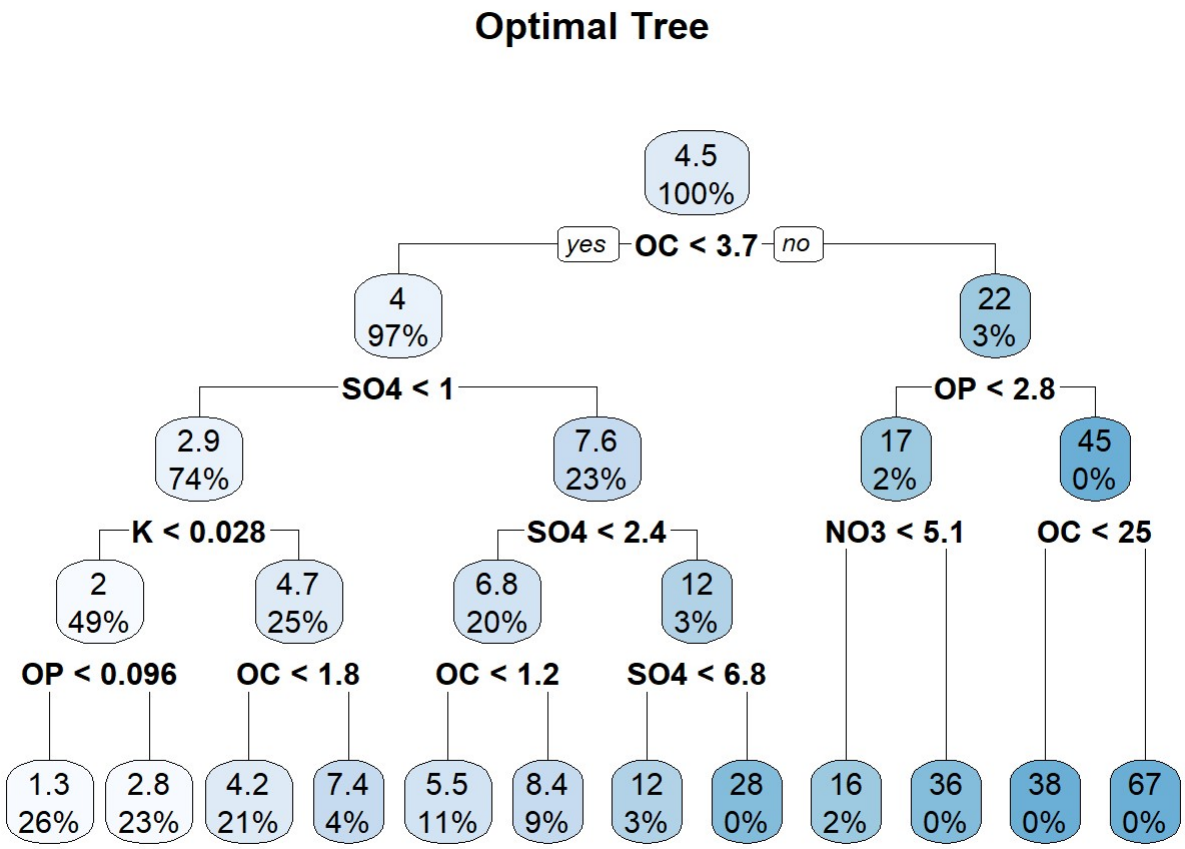
```
# function to get optimal cp
get_cp <- function(x) {
  min    <- which.min(x$cptable[, "xerror"])
  cp <- x$cptable[min, "CP"]
}

# function to get minimum error
get_min_error <- function(x) {
  min    <- which.min(x$cptable[, "xerror"])
  xerror <- x$cptable[min, "xerror"]
}

hyper_grid %>%
  mutate(
    cp    = purrr::map_dbl(models, get_cp),
    error = purrr::map_dbl(models, get_min_error)
    ) %>%
  arrange(error) %>%
  top_n(-5, wt = error)
```

```
##   minsplit maxdepth   cp     error
## 1       10        8 0.01 0.1973481
## 2       14       15 0.01 0.2001115
## 3       14       14 0.01 0.2009031
## 4        7       12 0.01 0.2020608
## 5       20       13 0.01 0.2027092
```

```
optimal_tree <- rpart(
    formula = PM2.5 ~ .-PM2.5_UNC,
    data    = US_DATA_LRG,
    method  = "anova",
    control = list(minsplit = 10, maxdepth = 8, cp = 0.01, xval=10)
    )

pred <- predict(optimal_tree, newdata = US_DATA_LRG_test)


rmse_op=RMSE(pred = pred, obs = US_DATA_LRG_test$PM2.5)
ModelMetrics::gini(pred,US_DATA_LRG_test$PM2.5)
```

```
## [1] 0.9329352
```

```
mae_op=MAE(pred = pred, obs = US_DATA_LRG_test$PM2.5)

rpart.plot(optimal_tree, main='Optimal Tree') #optimal tree determined througt grid search
```




```
summary(optimal_tree)
```

```
## Call:
## rpart(formula = PM2.5 ~ . - PM2.5_UNC, data = US_DATA_LRG, method = "anova",
##     control = list(minsplit = 10, maxdepth = 8, cp = 0.01, xval = 10))
##   n= 8647
##
##            CP nsplit rel error    xerror       xstd
## 1  0.36489610      0 1.0000000 1.0003115 0.07605303
## 2  0.15417066      1 0.6351039 0.6633240 0.04142286
## 3  0.12220485      2 0.4809332 0.5096679 0.03960257
## 4  0.04649899      3 0.3587284 0.3915864 0.02526373
## 5  0.03463939      4 0.3122294 0.3327126 0.02471715
## 6  0.02850279      5 0.2775900 0.3103268 0.02371196
## 7  0.02313744      6 0.2490872 0.2873444 0.02177956
## 8  0.01731210      7 0.2259498 0.2581629 0.01895770
## 9  0.01685673      8 0.2086377 0.2392860 0.01490210
## 10 0.01306348      9 0.1917809 0.2257883 0.01409075
## 11 0.01140478     10 0.1787175 0.2145467 0.01502649
## 12 0.01000000     11 0.1673127 0.2063296 0.01453057
##
## Variable importance
##       OC       OP      SO4        S       EC SiteCode        K       SE
##       28       21       11       10        7        5        3        2
##       ZN      NO3     Date       MN       FE       SI       TI       AL
##        2        2        2        1        1        1        1        1
##       NI        V       PB
##        1        1        1
##
## Node number 1: 8647 observations,    complexity param=0.3648961
##   mean=4.549602, MSE=25.32681
##   left son=2 (8390 obs) right son=3 (257 obs)
##   Primary splits:
##       OC < 3.693035 to the left,  improve=0.3648961, (0 missing)
##       OP < 0.838775 to the left,  improve=0.3426816, (0 missing)
##       EC < 0.477345 to the left,  improve=0.3353746, (0 missing)
##       K  < 0.078975 to the left,  improve=0.3108213, (0 missing)
##       ZN < 0.003445 to the left,  improve=0.2197968, (0 missing)
##   Surrogate splits:
##       OP  < 0.815255 to the left,  agree=0.986, adj=0.545, (0 split)
##       EC  < 0.8572   to the left,  agree=0.978, adj=0.257, (0 split)
##       PB  < 0.03269  to the left,  agree=0.971, adj=0.027, (0 split)
##       K   < 0.70079  to the left,  agree=0.971, adj=0.027, (0 split)
##       NO3 < 10.94156 to the left,  agree=0.971, adj=0.027, (0 split)
##
## Node number 2: 8390 observations,    complexity param=0.1541707
##   mean=4.017542, MSE=10.74484
##   left son=4 (6378 obs) right son=5 (2012 obs)
##   Primary splits:
##       SO4 < 1.0019   to the left,  improve=0.3745293, (0 missing)
##       S   < 0.35026  to the left,  improve=0.3738907, (0 missing)
##       K   < 0.032105 to the left,  improve=0.3698196, (0 missing)
```

```
##       OC  < 1.06415  to the left,  improve=0.3622052, (0 missing)
##       OP  < 0.237865 to the left,  improve=0.3482584, (0 missing)
##   Surrogate splits:
##       S        < 0.379    to the left,  agree=0.979, adj=0.915, (0 split)
##       SiteCode splits as  LL-LLLLRRLLLLRLLLRRLR-LLLRLLLRLLRLLLL--LRLRRRRLLLLR-LL
LLLLLLRL--LRLLLLRLLL-LRLLLLLLRLLLLLRLLLLLLLLLLLRLLLLLLLLLL-LLLLRLLLLRLLLLLRLLLLLRLL
L-RLLRLRLLLLLLL-L-LR-LLRLLLLLLRLL-L, agree=0.816, adj=0.233, (0 split)
##       OP       < 0.27197  to the left,  agree=0.813, adj=0.221, (0 split)
##       SE       < 0.000395 to the left,  agree=0.807, adj=0.193, (0 split)
##       ZN       < 0.004075 to the left,  agree=0.800, adj=0.164, (0 split)
##
## Node number 3: 257 observations,    complexity param=0.1222049
##   mean=21.91918, MSE=190.4247
##   left son=6 (214 obs) right son=7 (43 obs)
##   Primary splits:
##       OP       < 2.8095   to the left,  improve=0.5468624, (0 missing)
##       OC       < 15.0474  to the left,  improve=0.5137808, (0 missing)
##       K        < 0.18656  to the left,  improve=0.3267747, (0 missing)
##       SiteCode splits as  L--LR---LLL-LLLLLLL-R--LLLLLLL---RLLR--LL--L--LLRR--R-
RLR-------LLLL-RLLRL-RLRLLRL-----RL-LLR---LL-LL-L-LLR-LLL-----L-L--LL-L-R----L-LLL
---LLRL-LL-LLLL--L-R--LLL--L--LLLL-L, improve=0.3126466, (0 missing)
##       CL       < 0.014375 to the left,  improve=0.2974700, (0 missing)
##   Surrogate splits:
##       OC       < 12.1005  to the left,  agree=0.988, adj=0.930, (0 split)
##       SiteCode splits as  L--LL---LLL-LLLLLLL-L--LLLLLLL---LLRL--LL--L--LLRL--L-
LLR-------LLLL-RLLRL-LLRLLLL-----LL-LLR---LL-LL-L-LLL-LLL-----L-L--LL-L-R----L-LLL
---LLRL-LL-LLRL--L-R--LLL--L--LLLL-L, agree=0.879, adj=0.279, (0 split)
##       EC       < 2.921035 to the left,  agree=0.844, adj=0.070, (0 split)
##       Date     splits as  L-LL--L-LLL-L-L---L-LL--LLLLLL-LLLLL--LLLLRL-LLLLLL--L
--LLLLL--L--------------LLL-L-LL-L-LL-LL-LL-LLLLLLLLLLLLRLLL-L-LLLL, agree=0.840, a
dj=0.047, (0 split)
##       NI       < -6.5e-05 to the right, agree=0.837, adj=0.023, (0 split)
##
## Node number 4: 6378 observations,    complexity param=0.04649899
##   mean=2.890826, MSE=4.189719
##   left son=8 (4237 obs) right son=9 (2141 obs)
##   Primary splits:
##       K  < 0.02811  to the left,  improve=0.3810835, (0 missing)
##       OC < 0.93938  to the left,  improve=0.3553606, (0 missing)
##       OP < 0.16584  to the left,  improve=0.3182269, (0 missing)
##       BR < 0.001165 to the left,  improve=0.2981336, (0 missing)
##       EC < 0.11325  to the left,  improve=0.2856111, (0 missing)
##   Surrogate splits:
##       MN < 0.000965 to the left,  agree=0.814, adj=0.446, (0 split)
##       FE < 0.037735 to the left,  agree=0.813, adj=0.444, (0 split)
##       SI < 0.14337  to the left,  agree=0.811, adj=0.437, (0 split)
##       TI < 0.003515 to the left,  agree=0.811, adj=0.437, (0 split)
##       AL < 0.058885 to the left,  agree=0.809, adj=0.430, (0 split)
##
## Node number 5: 2012 observations,    complexity param=0.03463939
##   mean=7.589211, MSE=14.74337
```

```
##   left son=10 (1728 obs) right son=11 (284 obs)
##   Primary splits:
##       SO4 < 2.38635  to the left,  improve=0.2557358, (0 missing)
##       CR  < 0.000665 to the left,  improve=0.2396845, (0 missing)
##       S   < 0.67805  to the left,  improve=0.2374953, (0 missing)
##       V   < 0.0029   to the left,  improve=0.2366923, (0 missing)
##       K   < 0.148315 to the left,  improve=0.2313183, (0 missing)
##   Surrogate splits:
##       S        < 0.874975 to the left,  agree=0.970, adj=0.785, (0 split)
##       SiteCode splits as  LL-LLLLLLLLL-LL--LLLR-LLLL--LLL-LLLL---LLLLLLLL-LLL-L-
L-LLL-LL--LLLLL-LLL--LL-LLL-LL-LLLLL-L-L--L-L--LLLLLLLLLLL-LLLLLLLLLL-LLLLLLL-LLL-L
L-LLLL-LL-LLLLL-L-LL-LLL-LLL-LL-L-L, agree=0.869, adj=0.074, (0 split)
##       NI       < 0.0017   to the left,  agree=0.869, adj=0.074, (0 split)
##       V        < 0.004775 to the left,  agree=0.869, adj=0.074, (0 split)
##       SE       < 0.002    to the left,  agree=0.867, adj=0.060, (0 split)
##
## Node number 6: 214 observations,    complexity param=0.02313744
##   mean=17.34485, MSE=60.01467
##   left son=12 (201 obs) right son=13 (13 obs)
##   Primary splits:
##       NO3      < 5.11497  to the left,  improve=0.3945392, (0 missing)
##       SiteCode splits as  L--LL---LLL-LLLLLLL-R--LLLLLLL---LLLL--LL--L--LLLR--L-
LL--------LLLL--LLRL-LL-LLRL-----RL-LLL---LL-LL-R-LLL-LLL-----L-L--LL-L------L-LLL
---LLLL-LL-LLLL--L----LLL--L--LLLL-L, improve=0.3915086, (0 missing)
##       Date     splits as  R-LL--L-LRL-L-L---L-LL--LLLLLL-RLRLL--LLRL-L-RLRLRL--L
--LLLLL--L--------------LLL-L-LL-L-LL-LL-LL-LLLLLLLLLLLL-LLL-L-LLLL, improve=0.3623
386, (0 missing)
##       K        < 0.25385  to the left,  improve=0.3110933, (0 missing)
##       SO4      < 4.978265 to the left,  improve=0.2807596, (0 missing)
##   Surrogate splits:
##       Date     splits as  L-LL--L-LRL-L-L---L-LL--LLLLLL-LLRLL--LLLL-L-LRRLLL--L
--LLLLL--L--------------LLL-L-LL-L-LL-LL-LL-LLLLLLLLLLLL-LLL-L-LLLL, agree=0.958, a
dj=0.308, (0 split)
##       SE       < 0.00251  to the left,  agree=0.958, adj=0.308, (0 split)
##       S        < 2.37981  to the left,  agree=0.958, adj=0.308, (0 split)
##       SO4      < 6.5974   to the left,  agree=0.953, adj=0.231, (0 split)
##       SiteCode splits as  L--LL---LLL-LLLLLLL-L--LLLLLLL---LLLL--LL--L--LLLR--L-
LL--------LLLL--LLLL-LL-LLLL-----LL-LLL---LL-LL-L-LLL-LLL-----L-L--LL-L------L-LLL
---LLLL-LL-LLLL--L----LLL--L--LLLL-L, agree=0.949, adj=0.154, (0 split)
##
## Node number 7: 43 observations,    complexity param=0.02850279
##   mean=44.68448, MSE=217.0472
##   left son=14 (33 obs) right son=15 (10 obs)
##   Primary splits:
##       OC       < 25.39485 to the left,  improve=0.6688224, (0 missing)
##       OP       < 7.920375 to the left,  improve=0.6323192, (0 missing)
##       EC       < 1.432395 to the left,  improve=0.5185635, (0 missing)
##       SiteCode splits as  ----R----L----------R--L---------L-LR---------L-L---R-
R-L-------R----L--R--L-L--L------L----L-------------R------------------R----L-L
-------L-------L-----L-------------L--, improve=0.4939978, (0 missing)
##       CL       < 0.02494  to the left,  improve=0.4923968, (0 missing)
```

```
##   Surrogate splits:
##       OP       < 6.893135 to the left,  agree=0.953, adj=0.8, (0 split)
##       SiteCode splits as  ----R----L----------L--L---------L-LR---------L-L---R-
L-L-------L----L--L--L-L--L------L----L-------------R------------------R----L-L
-------L-------L-----L-------------L--, agree=0.884, adj=0.5, (0 split)
##       EC       < 1.903125 to the left,  agree=0.884, adj=0.5, (0 split)
##       Date     splits as  ----------------------------------------L----------L
------------------------------R--L--L--L----LL---LLRLR--LL------LR-, agree=0.860, a
dj=0.4, (0 split)
##       CL       < 0.031235 to the left,  agree=0.837, adj=0.3, (0 split)
##
## Node number 8: 4237 observations,    complexity param=0.01140478
##   mean=1.992608, MSE=1.44871
##   left son=16 (2207 obs) right son=17 (2030 obs)
##   Primary splits:
##       OP  < 0.09625  to the left,  improve=0.4069047, (0 missing)
##       K   < 0.012955 to the left,  improve=0.4025839, (0 missing)
##       OC  < 0.532155 to the left,  improve=0.3954182, (0 missing)
##       S   < 0.098965 to the left,  improve=0.3856653, (0 missing)
##       SO4 < 0.26405  to the left,  improve=0.3640022, (0 missing)
##   Surrogate splits:
##       OC  < 0.423575 to the left,  agree=0.875, adj=0.739, (0 split)
##       EC  < 0.048955 to the left,  agree=0.793, adj=0.568, (0 split)
##       K   < 0.012705 to the left,  agree=0.745, adj=0.468, (0 split)
##       S   < 0.11291  to the left,  agree=0.734, adj=0.445, (0 split)
##       SO4 < 0.29305  to the left,  agree=0.726, adj=0.429, (0 split)
##
## Node number 9: 2141 observations,    complexity param=0.01306348
##   mean=4.668382, MSE=4.857784
##   left son=18 (1819 obs) right son=19 (322 obs)
##   Primary splits:
##       OC < 1.81129  to the left,  improve=0.2750742, (0 missing)
##       OP < 0.28626  to the left,  improve=0.1937325, (0 missing)
##       FE < 0.20788  to the left,  improve=0.1722669, (0 missing)
##       K  < 0.06402  to the left,  improve=0.1667441, (0 missing)
##       EC < 0.22105  to the left,  improve=0.1650228, (0 missing)
##   Surrogate splits:
##       OP   < 0.38288  to the left,  agree=0.921, adj=0.475, (0 split)
##       EC   < 0.31343  to the left,  agree=0.880, adj=0.202, (0 split)
##       Date splits as  LLLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLLL--LLLLLLLLLLLLLLLLLLL
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLRLLRRLLLLLLLLLLLLLL, agree=0.853, adj=
0.022, (0 split)
##       CR   < 0.000805 to the left,  agree=0.852, adj=0.016, (0 split)
##       SE   < 0.00147  to the left,  agree=0.851, adj=0.012, (0 split)
##
## Node number 10: 1728 observations,    complexity param=0.01685673
##   mean=6.802017, MSE=8.031354
##   left son=20 (964 obs) right son=21 (764 obs)
##   Primary splits:
##       OC < 1.16163  to the left,  improve=0.2660032, (0 missing)
##       K  < 0.0549   to the left,  improve=0.2628756, (0 missing)
```

```
##       FE < 0.404225 to the left,  improve=0.2057942, (0 missing)
##       SI < 1.209615 to the left,  improve=0.2053201, (0 missing)
##       AL < 0.636465 to the left,  improve=0.2052434, (0 missing)
##   Surrogate splits:
##       OP       < 0.32231  to the left,  agree=0.850, adj=0.661, (0 split)
##       EC       < 0.206845 to the left,  agree=0.789, adj=0.522, (0 split)
##       SiteCode splits as  RL-LLLLLRLLL-LL--RLLR-RLLL--LLL-RRLL---RLLLRLLL-LRL-L-
L-LLL-RL--LLLLL-RLL--LL-RRL-LR-LLLLR-R-L--L-L--RLLLLLLRLLL-RLLRLRLLLL-LLLLLLL-LRL-R
L-RLLR-LL-RLLLL-L-LR-LLR-LLL-LL-R-L, agree=0.683, adj=0.284, (0 split)
##       CU       < 0.000695 to the left,  agree=0.676, adj=0.267, (0 split)
##       ZN       < 0.00406  to the left,  agree=0.674, adj=0.263, (0 split)
##
## Node number 11: 284 observations,    complexity param=0.0173121
##   mean=12.37889, MSE=28.8712
##   left son=22 (270 obs) right son=23 (14 obs)
##   Primary splits:
##       SO4 < 6.76665  to the left,  improve=0.4623944, (0 missing)
##       PB  < 0.016615 to the left,  improve=0.4265695, (0 missing)
##       MN  < 0.008395 to the left,  improve=0.4185620, (0 missing)
##       ZN  < 0.04141  to the left,  improve=0.4171126, (0 missing)
##       S   < 1.93679  to the left,  improve=0.4099518, (0 missing)
##   Surrogate splits:
##       S    < 2.573415 to the left,  agree=0.993, adj=0.857, (0 split)
##       NI   < 0.00304  to the left,  agree=0.975, adj=0.500, (0 split)
##       V    < 0.008205 to the left,  agree=0.975, adj=0.500, (0 split)
##       ZN   < 0.053945 to the left,  agree=0.975, adj=0.500, (0 split)
##       Date splits as  LL-LLLLL--L-R-----L-L---------L-------LLL-L--LL-LLLL-LLLL-
LLL-LL-LLLLLLLL-LLLRLLLLRLLLLRLLLLLLLLLLLLLLLLLLRLLLLLLRL-LLLLLL, agree=0.972, adj=
0.429, (0 split)
##
## Node number 12: 201 observations
##   mean=16.10734, MSE=28.57004
##
## Node number 13: 13 observations
##   mean=36.4786, MSE=156.4184
##
## Node number 14: 33 observations
##   mean=38.052, MSE=73.78158
##
## Node number 15: 10 observations
##   mean=66.57164, MSE=65.60985
##
## Node number 16: 2207 observations
##   mean=1.256258, MSE=0.627724
##
## Node number 17: 2030 observations
##   mean=2.793161, MSE=1.110907
##
## Node number 18: 1819 observations
##   mean=4.182025, MSE=3.344767
##
```

```
## Node number 19: 322 observations
##   mean=7.415849, MSE=4.520096
##
## Node number 20: 964 observations
##   mean=5.500812, MSE=4.404893
##
## Node number 21: 764 observations
##   mean=8.443853, MSE=7.775161
##
## Node number 22: 270 observations
##   mean=11.5469, MSE=12.47624
##
## Node number 23: 14 observations
##   mean=28.42452, MSE=74.24786
```

```
optimal_tree
```

```
## n= 8647
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 8647 219000.9000  4.549602
##    2) OC< 3.693035 8390  90149.2100  4.017542
##      4) SO4< 1.0019 6378  26722.0200  2.890826
##        8) K< 0.02811 4237   6138.1860  1.992608
##         16) OP< 0.09625 2207   1385.3870  1.256258 *
##         17) OP>=0.09625 2030   2255.1420  2.793161 *
##        9) K>=0.02811 2141  10400.5200  4.668382
##         18) OC< 1.81129 1819   6084.1310  4.182025 *
##         19) OC>=1.81129 322   1455.4710  7.415849 *
##      5) SO4>=1.0019 2012  29663.6600  7.589211
##       10) SO4< 2.38635 1728  13878.1800  6.802017
##         20) OC< 1.16163 964   4246.3170  5.500812 *
##         21) OC>=1.16163 764   5940.2230  8.443853 *
##       11) SO4>=2.38635 284   8199.4220 12.378890
##         22) SO4< 6.76665 270   3368.5850 11.546900 *
##         23) SO4>=6.76665 14   1039.4700 28.424520 *
##    3) OC>=3.693035 257  48939.1500 21.919180
##      6) OP< 2.8095 214  12843.1400 17.344850
##       12) NO3< 5.11497 201   5742.5780 16.107340 *
##       13) NO3>=5.11497 13   2033.4390 36.478600 *
##      7) OP>=2.8095 43   9333.0300 44.684480
##       14) OC< 25.39485 33   2434.7920 38.052000 *
##       15) OC>=25.39485 10    656.0985 66.571640 *
```

```
tmp <- printcp(optimal_tree)
```

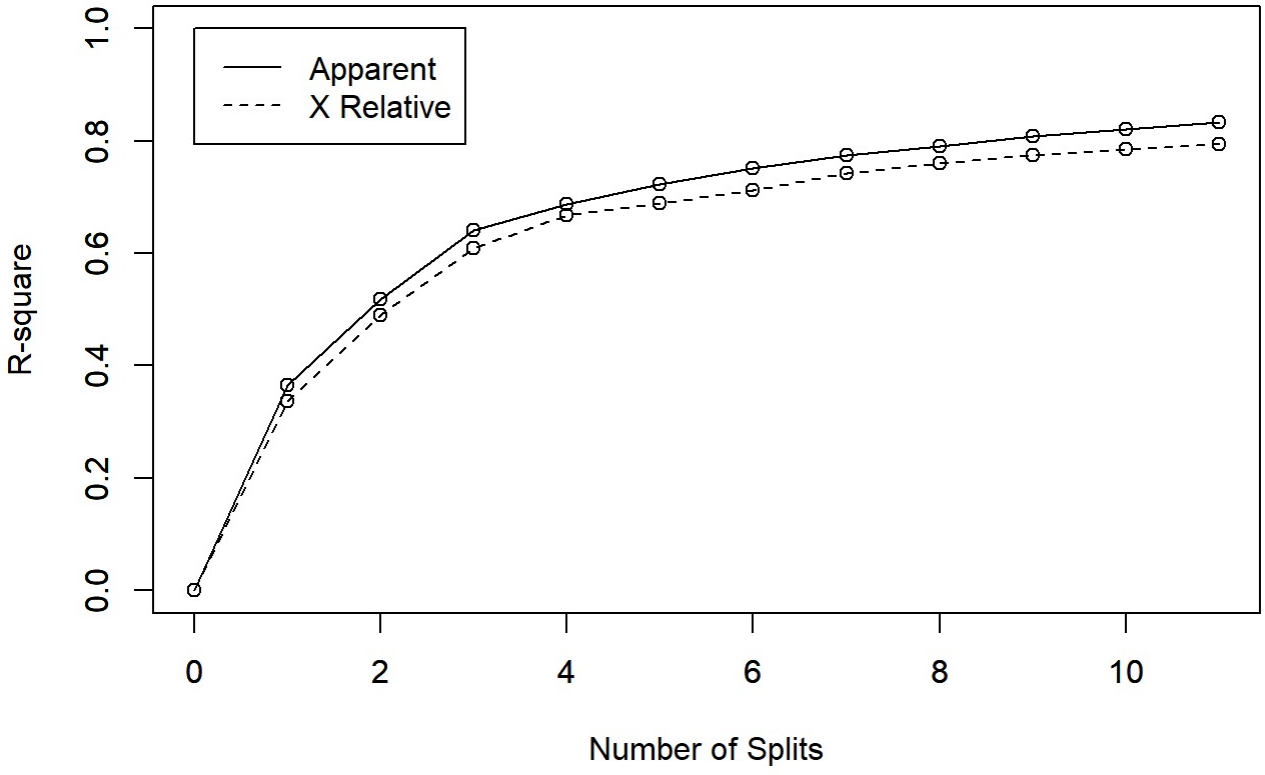
```
##
## Regression tree:
## rpart(formula = PM2.5 ~ . - PM2.5_UNC, data = US_DATA_LRG, method = "anova",
##     control = list(minsplit = 10, maxdepth = 8, cp = 0.01, xval = 10))
##
## Variables actually used in tree construction:
## [1] K   NO3 OC  OP  SO4
##
## Root node error: 219001/8647 = 25.327
##
## n= 8647
##
##          CP nsplit rel error  xerror     xstd
## 1  0.364896      0   1.00000 1.00031 0.076053
## 2  0.154171      1   0.63510 0.66332 0.041423
## 3  0.122205      2   0.48093 0.50967 0.039603
## 4  0.046499      3   0.35873 0.39159 0.025264
## 5  0.034639      4   0.31223 0.33271 0.024717
## 6  0.028503      5   0.27759 0.31033 0.023712
## 7  0.023137      6   0.24909 0.28734 0.021780
## 8  0.017312      7   0.22595 0.25816 0.018958
## 9  0.016857      8   0.20864 0.23929 0.014902
## 10 0.013063      9   0.19178 0.22579 0.014091
## 11 0.011405     10   0.17872 0.21455 0.015026
## 12 0.010000     11   0.16731 0.20633 0.014531
```

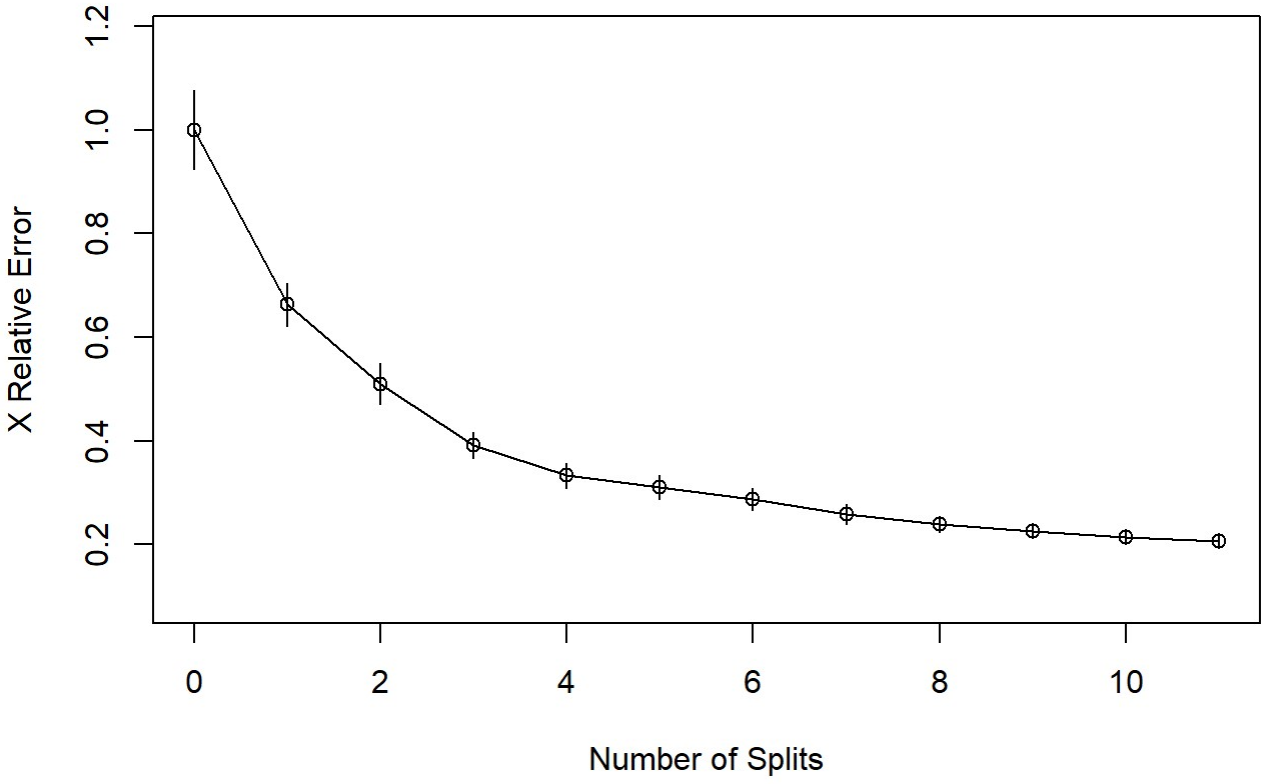
```
rsq.val <- 1-tmp[,c(3,4)]
rsq.val #rquared and xerror for each split
```

```
##    rel error        xerror
## 1  0.0000000 -0.0003114605
## 2  0.3648961  0.3366760066
## 3  0.5190668  0.4903321091
## 4  0.6412716  0.6084136421
## 5  0.6877706  0.6672873817
## 6  0.7224100  0.6896732000
## 7  0.7509128  0.7126555606
## 8  0.7740502  0.7418370628
## 9  0.7913623  0.7607139969
## 10 0.8082191  0.7742116659
## 11 0.8212825  0.7854533238
## 12 0.8326873  0.7936703885
```

```
rsq_op = rsq.val[nrow(rsq.val),] #final rquared and xerror
rsq.rpart(optimal_tree)#xerror and rsqu vs splits plot
```

```
##
## Regression tree:
## rpart(formula = PM2.5 ~ . - PM2.5_UNC, data = US_DATA_LRG, method = "anova",
##     control = list(minsplit = 10, maxdepth = 8, cp = 0.01, xval = 10))
##
## Variables actually used in tree construction:
## [1] K   NO3 OC  OP  SO4
##
## Root node error: 219001/8647 = 25.327
##
## n= 8647
##
##          CP nsplit rel error  xerror     xstd
## 1  0.364896      0   1.00000 1.00031 0.076053
## 2  0.154171      1   0.63510 0.66332 0.041423
## 3  0.122205      2   0.48093 0.50967 0.039603
## 4  0.046499      3   0.35873 0.39159 0.025264
## 5  0.034639      4   0.31223 0.33271 0.024717
## 6  0.028503      5   0.27759 0.31033 0.023712
## 7  0.023137      6   0.24909 0.28734 0.021780
## 8  0.017312      7   0.22595 0.25816 0.018958
## 9  0.016857      8   0.20864 0.23929 0.014902
## 10 0.013063      9   0.19178 0.22579 0.014091
## 11 0.011405     10   0.17872 0.21455 0.015026
## 12 0.010000     11   0.16731 0.20633 0.014531
```


Apparent
X Relative
R-square
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Number of Splits


X Relative Error
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10
Number of Splits

```
metrics_op = c(rmse_op,rsq_op[1],mae_op)
metrics_op
```

```
##           rel error
## 2.1568766 0.8326873 1.2141927
```

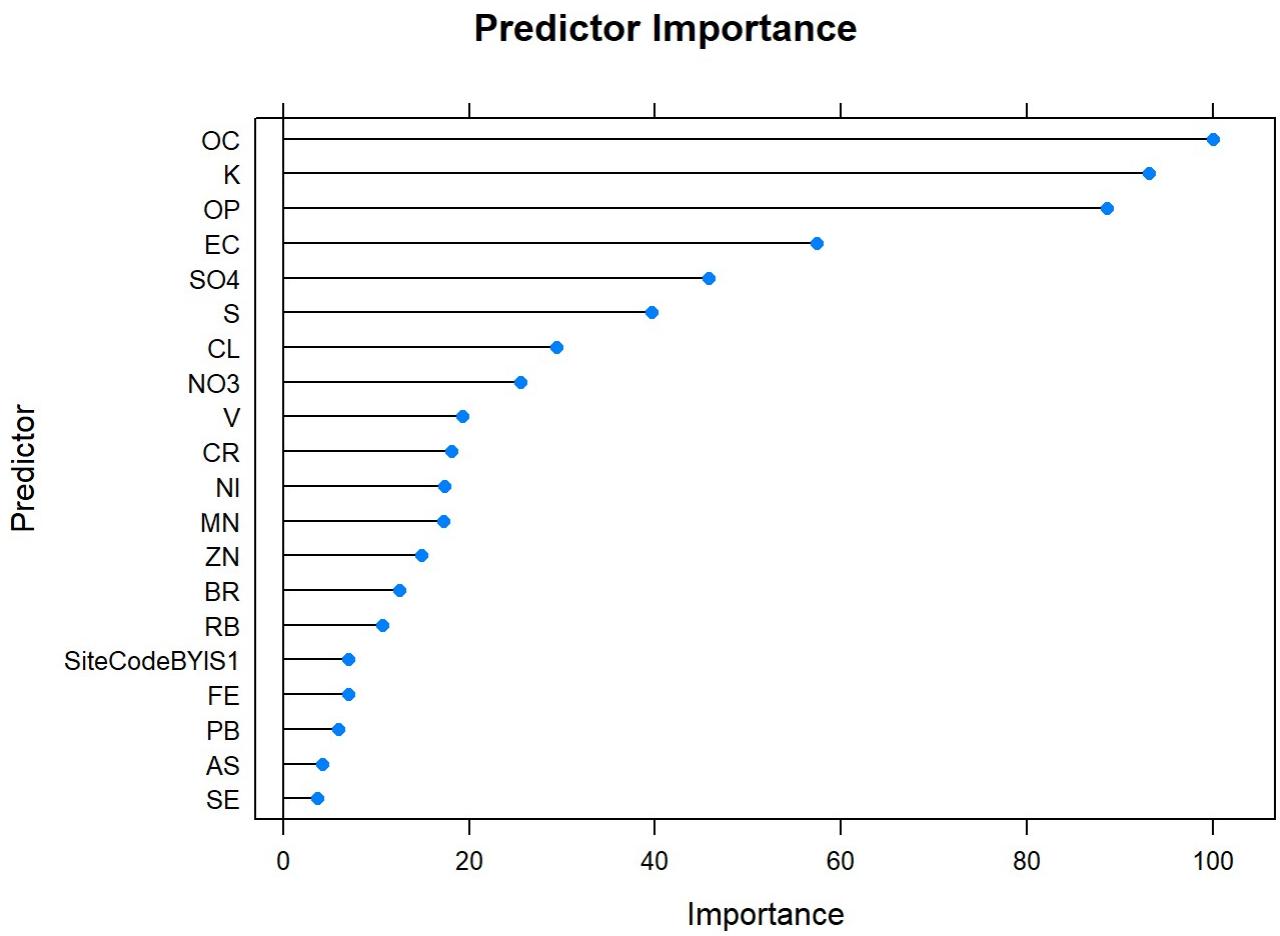
```
colnames(metrics_op)
```

```
## NULL
```

```
# Specify 10-fold cross validation
ctrl <- trainControl(method = "cv",  number = 10)

# CV bagged model
bagged_cv <- train(
  PM2.5 ~ .-PM2.5_UNC,
  data    = US_DATA_LRG,
  method = "treebag",
  trControl = ctrl,
  importance = TRUE
  )

# assess results
bagged_cv  #this is an object with many useful items
```

```
## Bagged CART
##
## 8647 samples
##   33 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 7781, 7783, 7781, 7783, 7783, 7783, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.793048  0.8762557  0.9628197
```

```
# plot most important variables
plot(varImp(bagged_cv),20, main="Predictor Importance", ylab="Predictor")
```




```
varImp(bagged_cv)
```

```
## treebag variable importance
##
##   only 20 most important variables shown (out of 321)
##
##               Overall
## OC            100.000
## K              93.120
## OP             88.612
## EC             57.398
## SO4            45.777
## S              39.605
## CL             29.379
## NO3            25.574
## V              19.323
## CR             18.043
## NI             17.345
## MN             17.256
## ZN             14.872
## BR             12.469
## RB             10.654
## SiteCodeBYIS1   7.050
## FE              6.964
## PB              5.935
## AS              4.208
## SE              3.699
```

```
metric_bag= bagged_cv$results[1,][2:4]
bagged_cv
```

```
## Bagged CART
##
## 8647 samples
##   33 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 7781, 7783, 7781, 7783, 7783, 7783, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.793048  0.8762557  0.9628197
```

```
metric_bag
```

```
##       RMSE  Rsquared       MAE
## 1 1.793048 0.8762557 0.9628197
```

```
metrics_fin = rbind(metric_bag,metrics_op)

rownames(metrics_fin)=c('Bagged_Tree_10cv','Optimal_GrdSrh_Tree_10cv' )
metrics_fin
```

```
##                               RMSE  Rsquared       MAE
## Bagged_Tree_10cv          1.793048 0.8762557 0.9628197
## Optimal_GrdSrh_Tree_10cv 2.156877 0.8326873 1.2141927
```