Lecture 17:
Random Forests using EPA Stream Data Set
STAT 452, Spring 2021

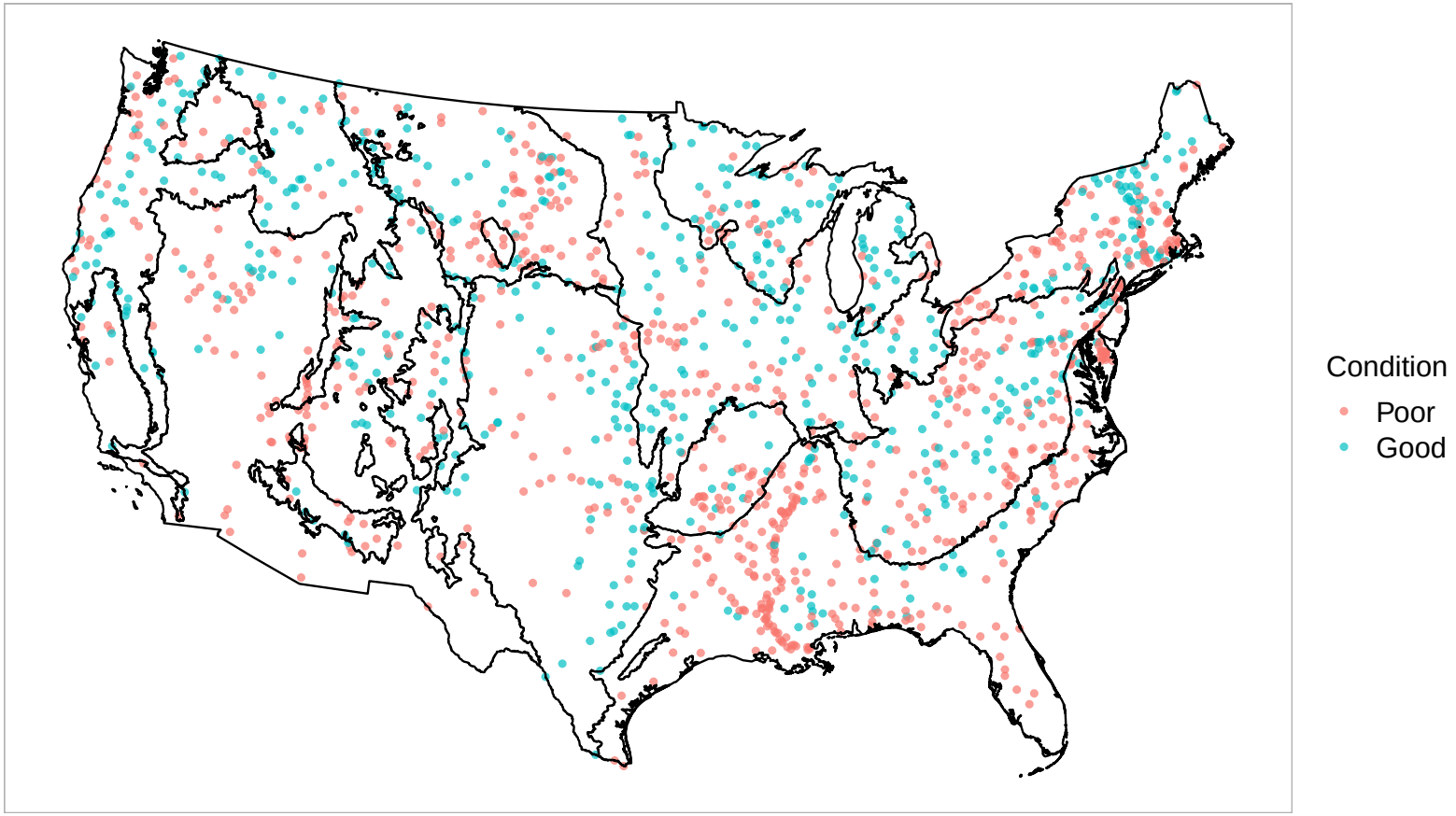
EPA Stream Condition Data

- The Environmental Protection Agency (EPA) randomly sampled stream sites across the conterminous US during the summer months of 2008/09. The effort was part of a larger environmental monitoring program called the National Rivers and Stream Assessment (NRSA).[1]

- Streams were classified as being in `Good` or `Poor` condition according to an aquatic health index. Sampled macroinvertebrates were used as the primary indicator of aquatic health.

[1] `https://www.epa.gov/national-aquatic-resource-surveys/nrsa`

Map of sampled stream sites from 2008/09 National Rivers and Stream Assessment.



Random Forest Model

Response: Binary `Good`/`Poor` condition of $n = 1433$ sampled stream sites.

Predictors: $p = 210$ landscape features for each stream's catchment (i.e., local drainage area around each stream segment). The predictors are from the StreamCat data set.[2]

- Examples: % urbanization, % agriculture, road densities, dams, temperature, precipitation, forest change, etc.

[2] `https://www.epa.gov/national-aquatic-resource-surveys/streamcat`

```
# load EPA stream data set
> streams <- readRDS(url("https://ericwfox.github.io/data/streams.rds"))

> dim(streams)
[1] 1433  211

> table(streams$Condition)
Poor Good
 862  571

# fit random forest model using all predictors
> library(randomForest)
> set.seed(999) # set seed for reproducibility
> rf1 <- randomForest(Condition ~ ., data=streams)

# make predictions for stream condition (Good/Poor)
# predictions are made using the OOB data
rf_preds <- predict(rf1, type = "response")
```

```
# make confusion matrix
> tb <- table(predicted = rf_preds, actual = streams$Condition)
> addmargins(tb)
         actual
predicted Poor Good  Sum
     Poor  752  211  963
     Good  110  360  470
     Sum   862  571 1433

# Accuracy (percent correctly classified)
> (752 + 360) / 1433
[1] 0.7759944

# Sensitivity (percent of good streams correctly classified)
> 360 / 571
[1] 0.6304729

# Specificity (percent of poor streams correctly classified)
> 752 / 862
[1] 0.8723898
```

Balanced Random Forests

- Machine learning algorithms such as random forests can be affected by class imbalances in the response variable.
- The EPA stream data is moderately imbalanced (about 60% of streams were in poor condition, and 40% in good condition)
- The random forest model using the defaults performed much better at predicting the streams in poor condition than the streams in good condition.

Balanced Random Forests

- One way to deal with class imbalances is to use a down sampling approach.
- Each tree in the ensemble is built by drawing a bootstrap sample with the same number of cases from each class. In practice, the number of cases drawn from each class is set to the size of the minority class.
- Since the 571 good streams are the minority class, we would build each random forest tree using a balanced bootstrap sample with 571 poor and 571 good streams.
- We can fit a balanced random forest model by specifying the `sampsize` argument of the `randomForest()` function.

```
> table(streams$Condition)
Poor Good
 862  571

> nmin <- min(table(streams$Condition))
> nmin
[1] 571

# fit balanced RF model
> set.seed(999) # set seed for reproducibility
> rf2 <- randomForest(Condition ~ ., data=streams,
                        sampsize = c(nmin,nmin))

# make prediction for stream condition (Good/Poor)
# predictions are made using the OOB data
> rf_preds2 <- predict(rf2, type = "response")
```

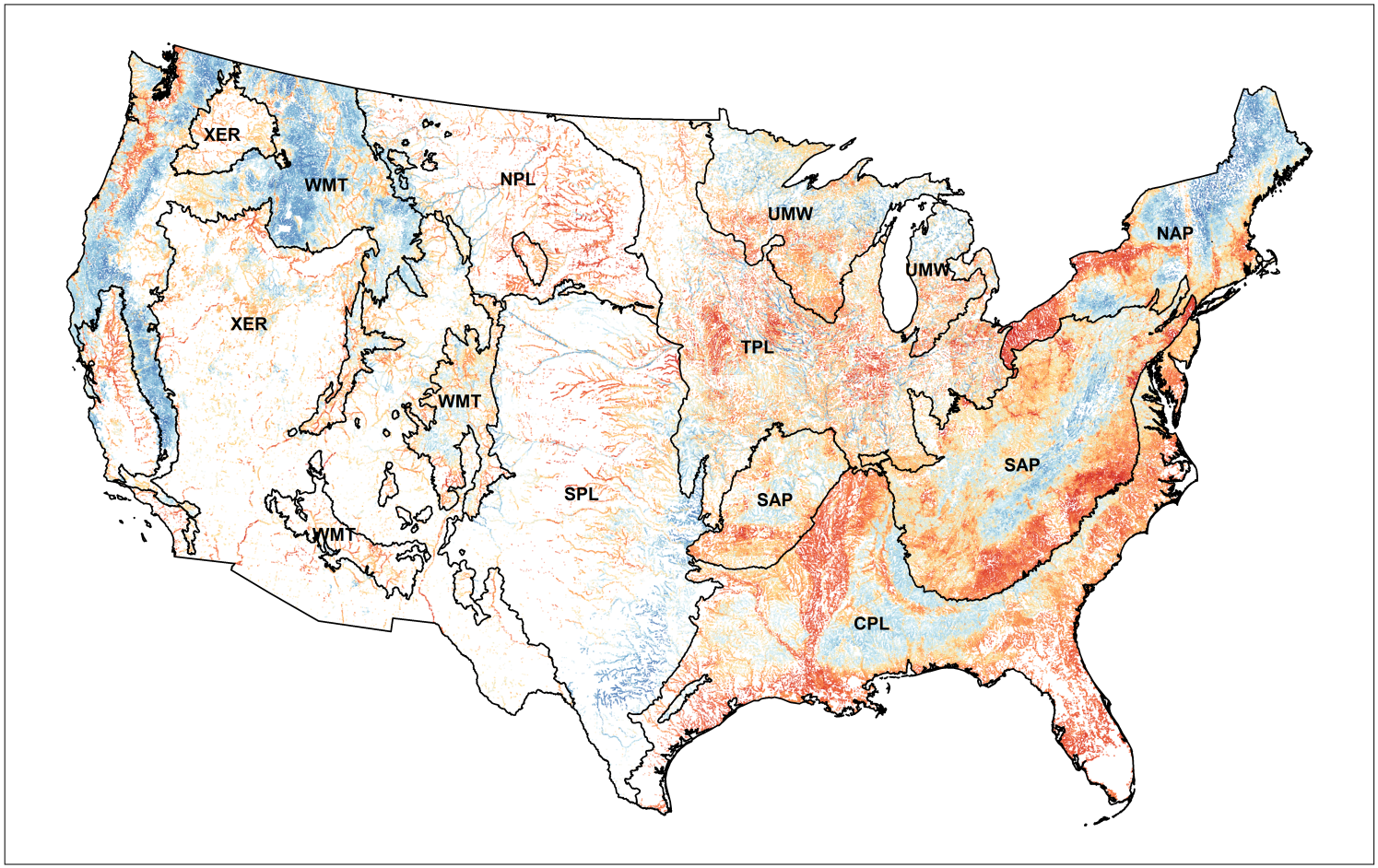
```
> # make confusion matrix
> tb <- table(predicted = rf_preds2, actual = streams$Condition)
> addmargins(tb)
         actual
predicted Poor Good  Sum
     Poor  687  159  846
     Good  175  412  587
     Sum   862  571 1433

> # Accuracy (percent correctly classified)
> (687 + 412) / 1433
[1] 0.7669225

> # Sensitivity (percent of good streams correctly classified)
> 412 / 571
[1] 0.7215412

> # Specificity (percent of poor streams correctly classified)
> 687 / 862
[1] 0.7969838
```

Mapping Predictions

- One application of the random forest model is to map the predicted probability of good stream condition for all 1.1 million perennial stream reaches in the US.
- The $p = 210$ predictors (i.e., landscape features) from the StreamCat data set are available for all 1.1 million perennial stream reaches across the US, and thus make mapping of the predictions possible.
- Let $\boldsymbol{x}$ be the predictor values at a new, unsampled location and $T_b(\boldsymbol{x}) \in \{0, 1\}$ the predicted `Poor`/`Good` condition of the b^{th} tree in the random forest model. Then we define the predicted probability of good stream condition as

$$\hat{p}(\boldsymbol{x}) = \frac{1}{B} \sum_{i=1}^{B} T_b(\boldsymbol{x})$$

That is, the predicted probability is the proportion of the random forest trees that voted `Good`.

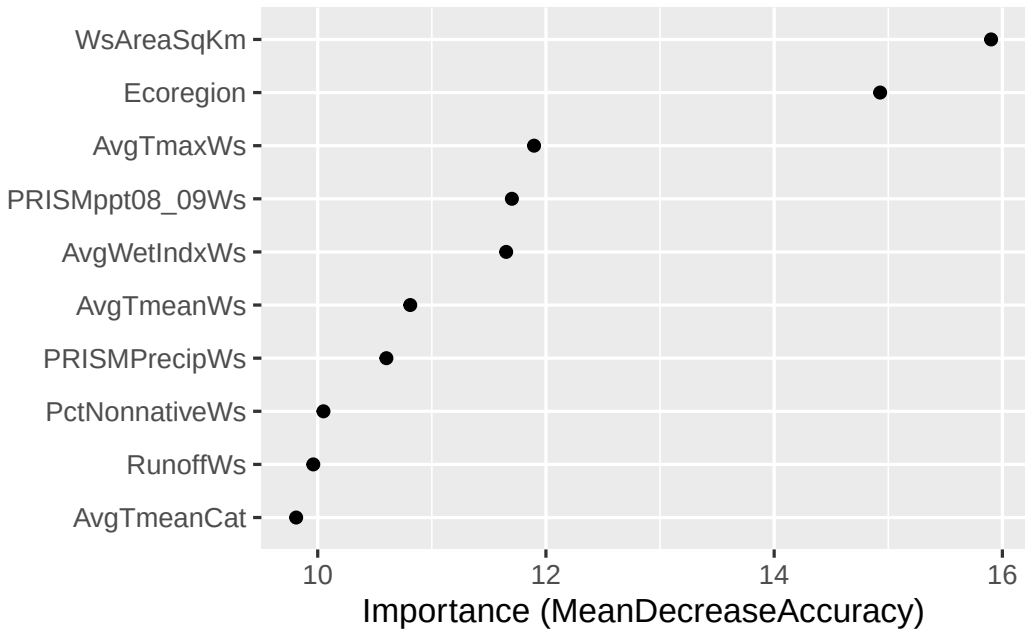


Ecoregions: Coastal Plains (CPL), Northern Appalachians (NAP), Northern Plains (NPL), Southern Appalachians (SAP), Southern Plains (SPL), Temperate Plains (TPL), Upper Midwest (UMW), Western Mountains (WMT), and Xeric (XER)

Variable Importance

For modeling stream condition, the two most important predictors are the watershed area and ecoregion. Essentially, this means that the size and location of the stream are important.

```
> library(vip)
> set.seed(999)
> rf1 <- randomForest(Sale_Price ~ ., importance = TRUE, data = streams)
> vip(rf1, num_features = 10, geom = "point", include_type = TRUE)
```



References

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

Hill, R.A., Fox, E.W., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., and Weber, M.H. (2017). Predictive mapping of the biotic condition of conterminous-USA rivers and streams. *Ecological Applications*, 27(8), 2397-2415.

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R., and Thornbrugh, D.J. (2016). The Stream-Catchment (StreamCat) Dataset: A database of watershed metrics for the conterminous Unites States. *Journal of the American Water Resources Association*, 52(1):120–218.