# Native Plant Community Random Forest Model

Bart Richardson| MNIT@DNR / Ecological Waters Resources

MINNESOTA IT SERVICES

# The Beautiful

- Smooth
- Soft
- Flowing
- Peaceful
- Safe

# The Sublime

- Rough
- Hard
- Angular
- Scary
- Dangerous

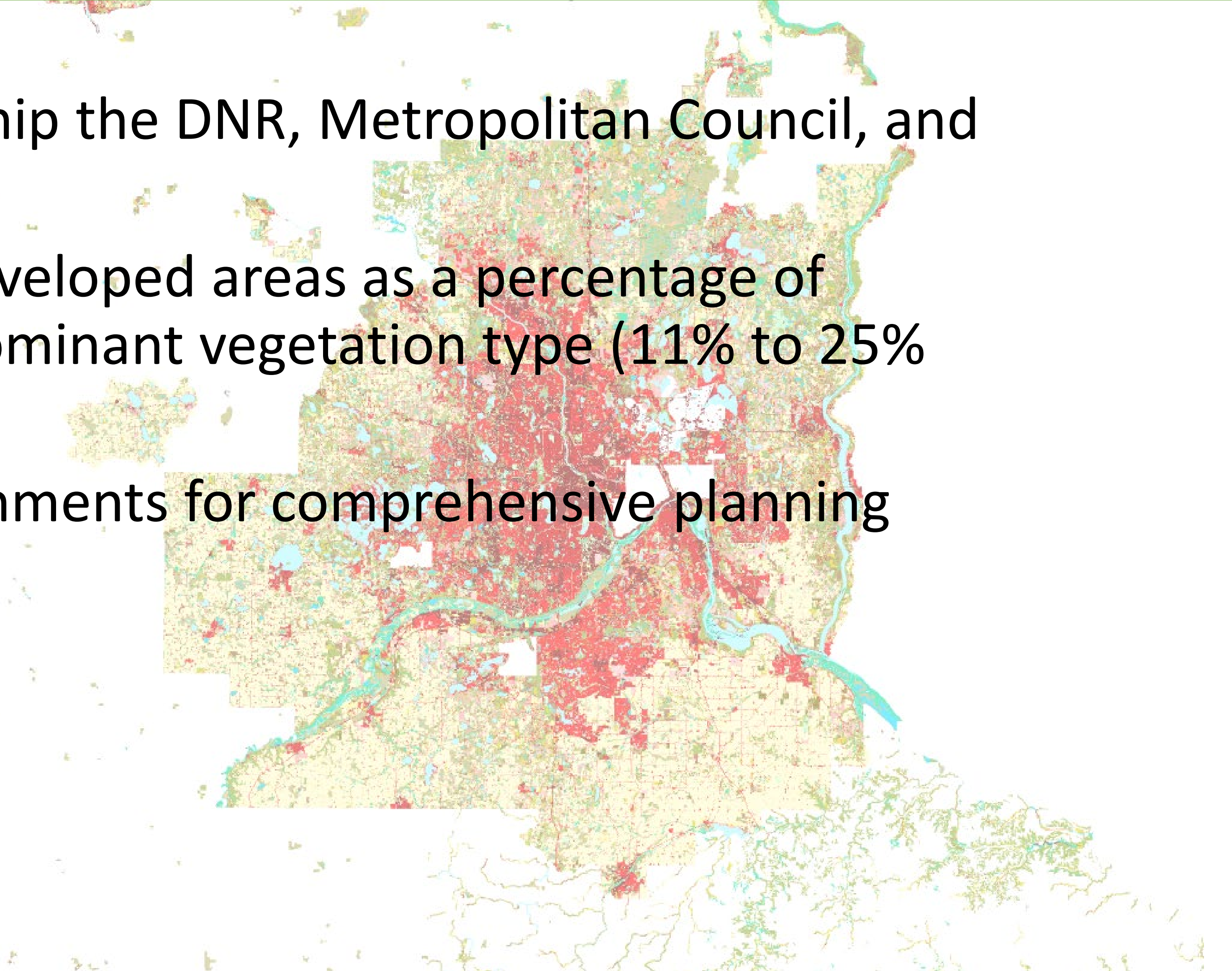
# The Picturesque

# Modeling Native Plant Communities (NPC) with Random Forest

- **Background**
  - The Minnesota Land Cover Classification (MLCCS)
  - The DNR's Native Plant Communities (NPC)
  - Combining the MLCCS and NPC – MLCCS version 2.0
- **GIS Models for the MLCCS & NPC**
  - Previous NPC models – Anoka Sandplains & Tamarac National Wildlife Refuge
  - Statewide MLCCS Landsat data (2013) – University of Minnesota
  - Statewide NPC model using Random Forest (2017) – University of North Dakota
  - MetroGIS NPC model using Random Forest
- **Combine the Random Forest model with Aerial Photo Interpretation**
  - Going from grids to polygons…

# Minnesota Land Cover Classification System (MLCCS)

- Developed in 2000 in partnership the DNR, Metropolitan Council, and the National Park Service
- Unique approach to classify developed areas as a percentage of impervious surface with the dominant vegetation type (11% to 25% impervious cover with trees)
- Used by county and city governments for comprehensive planning



# DNR Mn Biological Survey Native Plant Communities (NPC)

- Developed in 2003 based on ecological and environmental conditions, expanding on the DNR's Ecological Classification System (ECS)
- Community classification is dependent on environmental factors, such as soil types, slopes, aspect, and the geographical location

# Combining the MLCCS with the NPC

- In 2022 the Data and Technology team of the Metro Conservation Network (MCNet) merged the NPC system with the MLCCS, replacing all of the original MLCCS natural community codes with NPC codes
- New codes were created for altered/non-native native plant communities, such as ‘FFn98b - Altered/Non-native Floodplain Forest’
- The MLCCS natural community codes went from a 5 tiered system to a 7 tiered system
  - Original code - 32141; 3.de.UP.nAB.nAN; Aspen-birch forest northern hardwoods subtype
  - New code – 3111522; 3.1.FDn43b2; Aspen - Birch Woodland Hardwood Subtype
- **There isn’t a 1:1 relationship between the old codes and the new codes.**
- **Hence the need for a NPC model**

# GIS Models for the MLCCS & NPC

- **GIS Models for the MLCCS & NPC**
  - Anoka Sandplains – analysis of site conditions, Jason Husveth
  - Tamarac National Wildlife Refuge – analysis of site conditions, Scott Zager
  - Statewide MLCCS Landsat data (2013) – satellite image interpretation, University of Minnesota
  - Statewide NPC model using Random Forest (2017) – model training with environmental variables, University of North Dakota
  - MetroGIS NPC model using Random Forest – this project….

# The sublime random forest

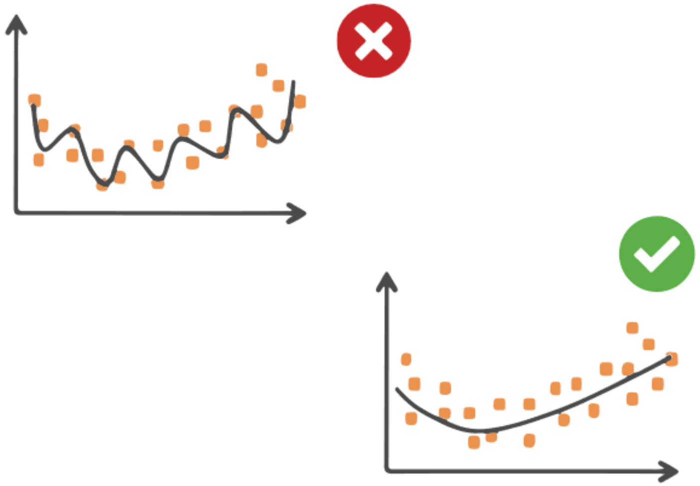
# Random Forest Model

- Machine learning algorithm for classification and regression analysis
- An ensemble learning method made up of classifiers, or **decision trees**, and their predictions
- Bagging or grouping the random decision trees together creates a **‘random forest’**
- Environmental variables (such as grid layers) are used to train the model
- *The average of randomly generated decision trees helps eliminate data noise and creates a more accurate prediction*

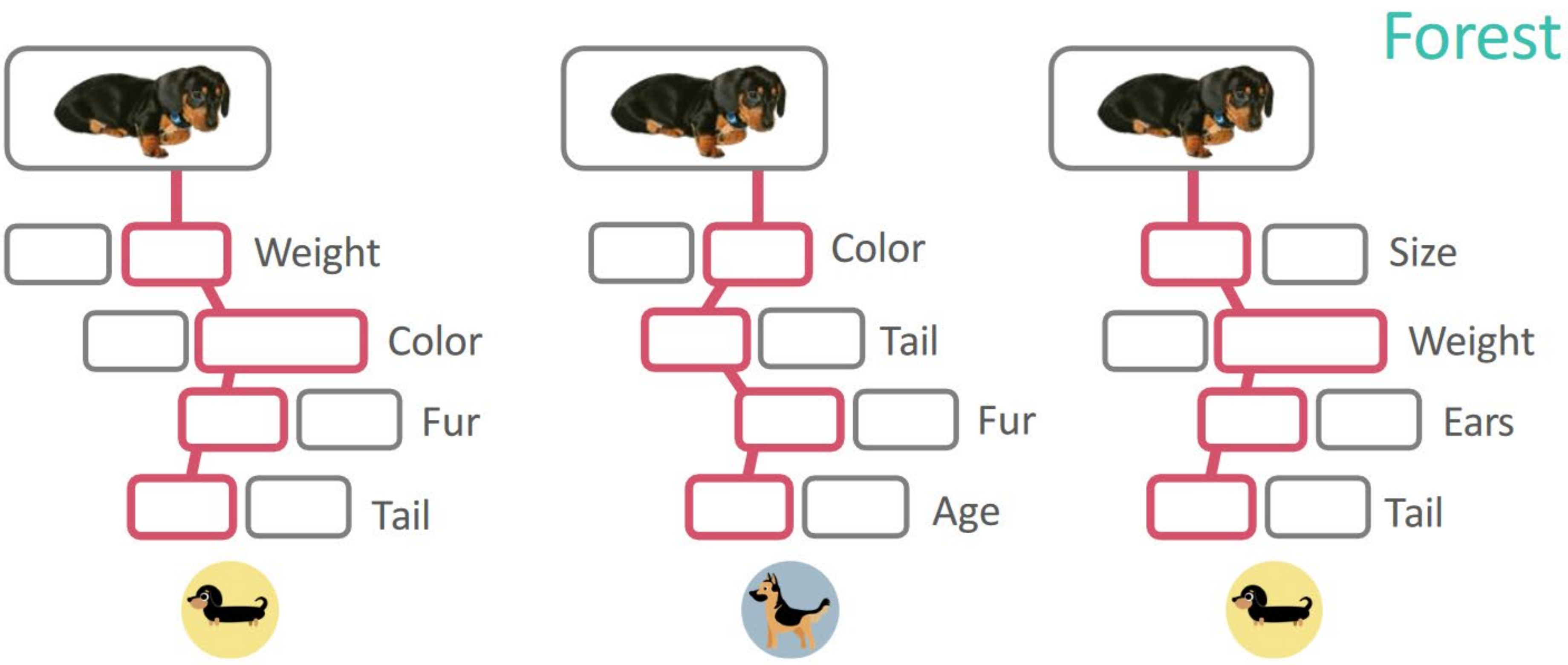
# Random Forest Model

## When we can't trust a model

Mimics training dataset and models noise instead of generalizing a trend

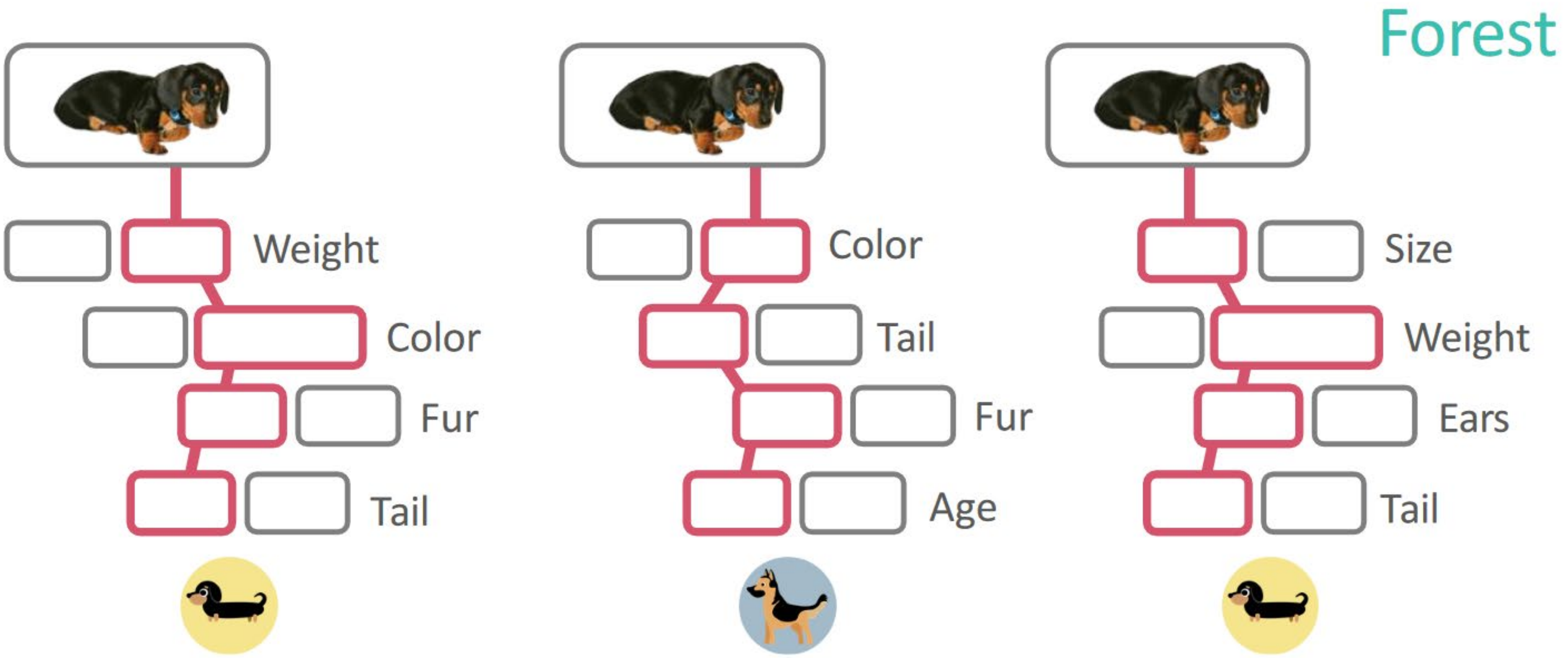


# Random Forest Model




Random subset of data and variables used in each tree

# Random Forest Model




Majority vote wins

# Random Forest Model

## Modeling workflow

Step 0. Prepare your data

Step 1. Train a model

Step 2. Evaluate model performance

Step 3. Train again with different parameters

Step 4. Compare models

Step 5. Repeat…

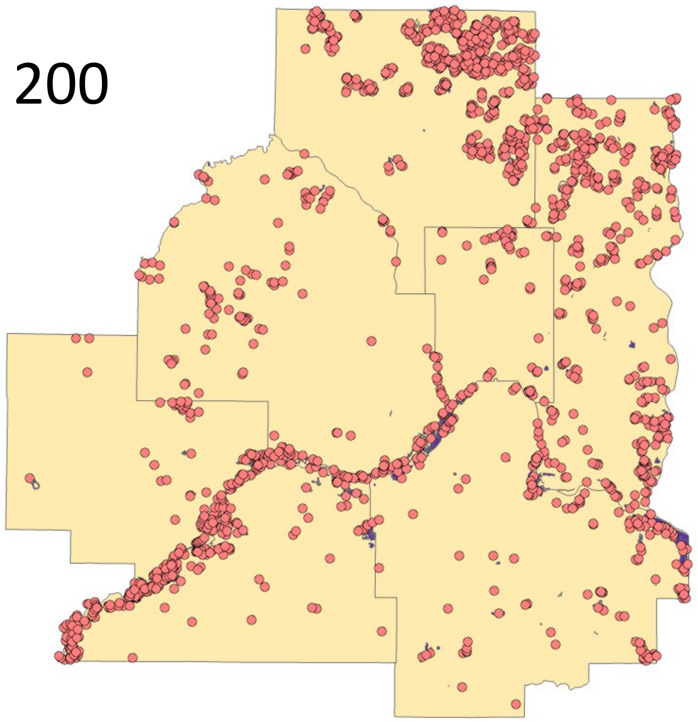
Step 6. Use best model to predict unknown values

# Random Forest Model

## Prepare MBS Native Plant Community Data

- Using National Land Cover Data and UM MLCCS data, delete NPC polygons that no longer exist
- Create a training dataset with a balanced representation of NPC classes
    - Remove community types that have less than 100 sites
    - If a community has more than 100 sites select the largest 200
- Create a training point layer from polygon centroids

| NPC class code | NPC Class | Count |
|---|---|---|
| FDs37 | Southern Dry-Mesic Oak (Maple) Woodland | 217 |
| FFs68 | Southern Floodplain Forest | 151 |
| FPs63 | Southern Rich Conifer Swamp | 129 |
| MHs38 | Southern Mesic Oak-Basswood Forest | 170 |
| MHs39 | Southern Mesic Maple-Basswood Forest | 149 |
| MRn83 | Northern Mixed Cattail Marsh | 136 |
| MRn93 | Northern Bulrush-Spikerush Marsh | 151 |
| OPn92 | Northern Rich Fen (Basin) | 97 |
| UPs13 | Southern Dry Prairie | 246 |
| WFn55 | Northern Wet Ash Swamp | 108 |
| WMn82 | Northern Wet Meadow/Carr | 227 |



# Random Forest Model

## Prepare Environmental Variable Training Grids

- Referencing past NPC models, select data layers associated with native plant communities
- Convert polygons to grids, snapping everything to the 30 meter NLCD grid
- Eliminate NODATA values

Soils Data:
- Soil texture
- Soil order
- Organic soils
- Hydric soils
- Soil slope group
- Soil vertical relief

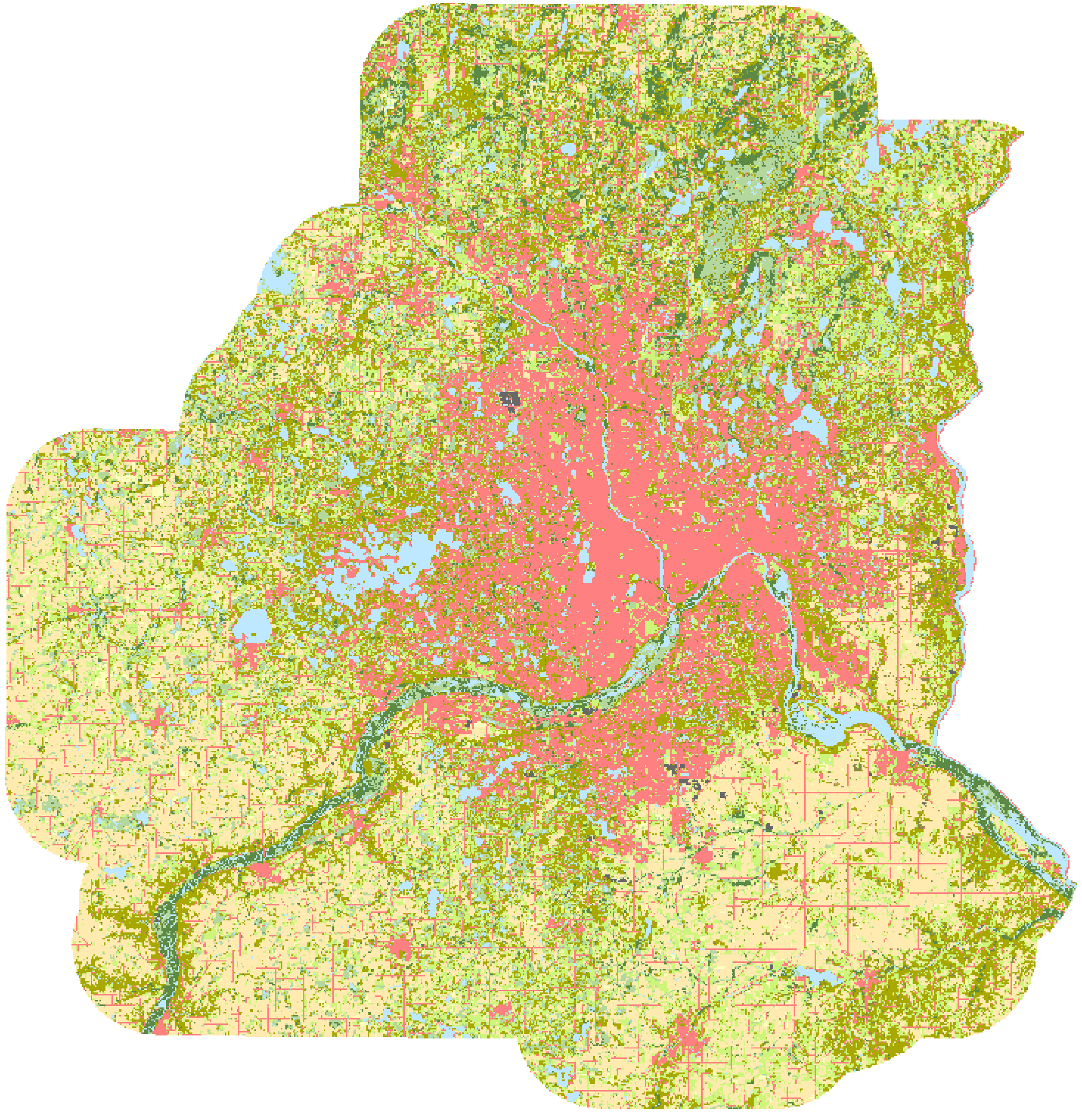
Topographical Data:
- Aspect
- Curvature
- Slope
- Landform
- Topographic position
- Distance to river
- Riparian areas
- Geomorphic
- Sedimentary
- Glacial phase

Vegetation and land cover:
- Canopy height
- NLCD
- NWI
- MLCCS
- Normalized diff. veg. indexes

Temperature:
- Surface temp.
- Solar area
- Continuous heat index

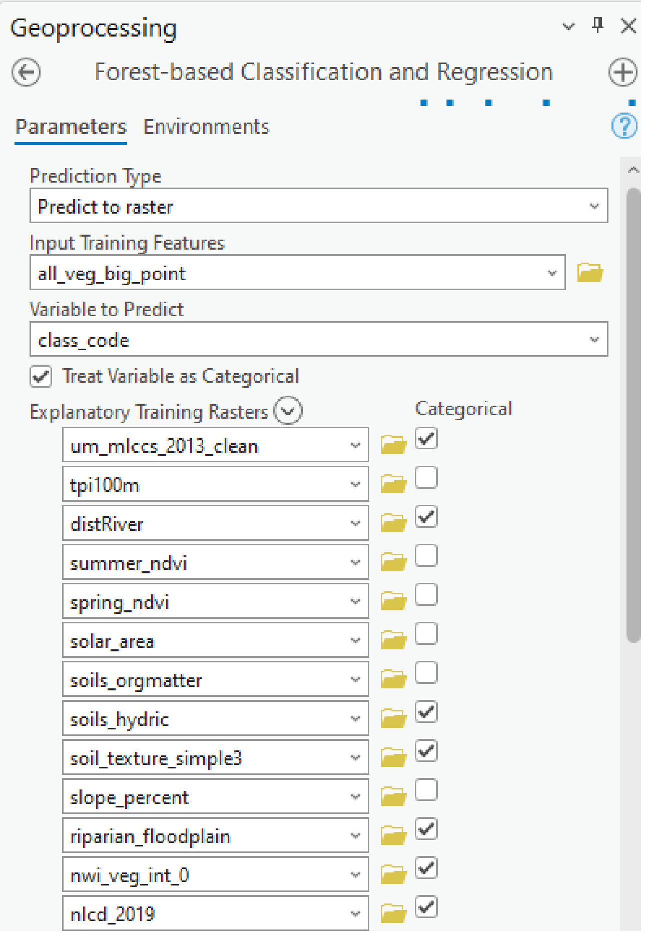
# Random Forest Model – 29 Training Grids

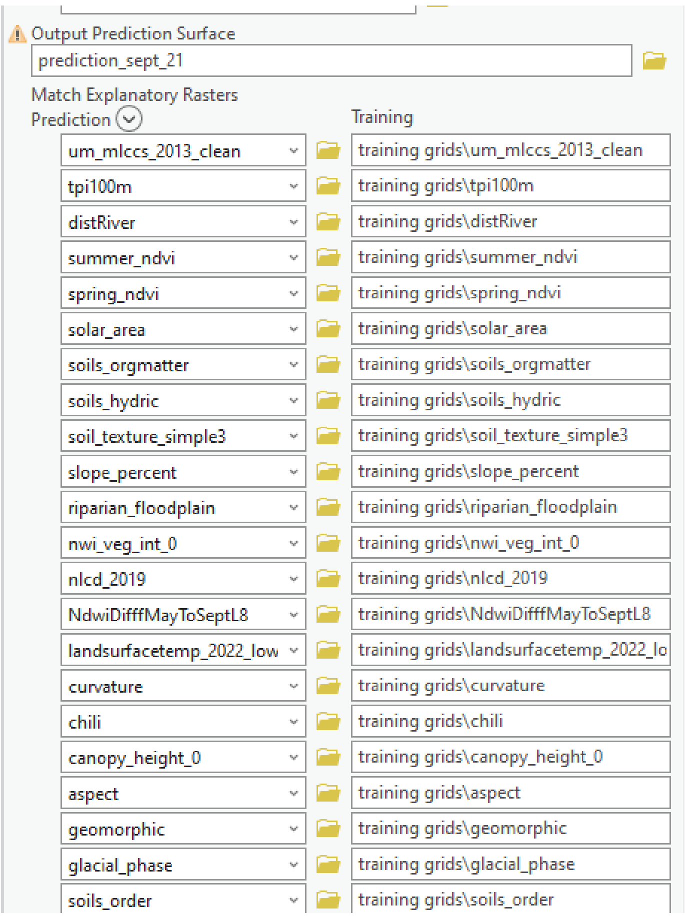
| variable | Full name |
|---|---|
| aspect_group | Aspect in four groups - north (315 to 45), east (45 to 135), south (135 to 225), west (225 to 315) |
| aspect_real | Aspect in real numbers |
| canopy_height | Canopy height, LiDAR derived |
| | |
| chili | Continuous heat-insolation load index |
| curvature | curvature of raster surface |
| distRiver | Distance to nearest flowing water body |
| geomorphic | Geomorphology of Minnesota - Geomorphic Association |
| glacial_phase | Geomorphology of Minnesota - Glacial Phase |
| landsurfacetemp_2022_low | Land Surface Temperature 2022, Twin Cities - |
| NdwiDifffMayToSeptL8 | Seasonal wetness: Normalized diffrence water index of May-Sept |
| nlcd_2019 | National Land Cover Dataset (NLCD) from 2019 |
| nwi_veg_int_0 | National Wetlands Inventory (NWI), Simplified Plant Community Classification |
| riparian_floodplain | FEMA Floodplain and Floodway combined with DNR streams intesecting flat areas |
| sedimentary | Geomorphology of Minnesota - Sedimentary Association |
| slope_percent | landform slope, percent |
| soil_texture_simple | SSURGO soils - Surface Texture, simplified to 13 classes |
| soils_grigal_relief | Cummins Grigal Soils of Minnesota - DESC attribute |
| soils_grigal_texture | Cummins Grigal Soils of Minnesota - TEXT_1 attribute |
| soils_grigal | Cummins Grigal Soils of Minnesota - GEISSOIL_K attribute |
| soils_hydric | SSURGO soils - HydrcRatng attribute |
| soils_order | USDA STATSGO - Dominant Soil Orders |
| soils_orgmatter | SSURGO soils - OrgMatter attribute |
| soils_slope_group | SSURGO soils - Slope, reclass to 5 groups |
| solar_area | Area Solar Radiation - watt hours per square meter |
| spring_ndvi | Normalized difference vegetation index for spring |
| summer_ndvi | Normalized difference vegetation index for summer |
| topo_landform | Geomorphology of Minnesota - Topographic Expression |
| | |
| tpi100m | Topographic position index at 100 m |
| um_mlccs_2013 | MLCCS and Impervious Surface by Landsat and Lidar: 2013 |

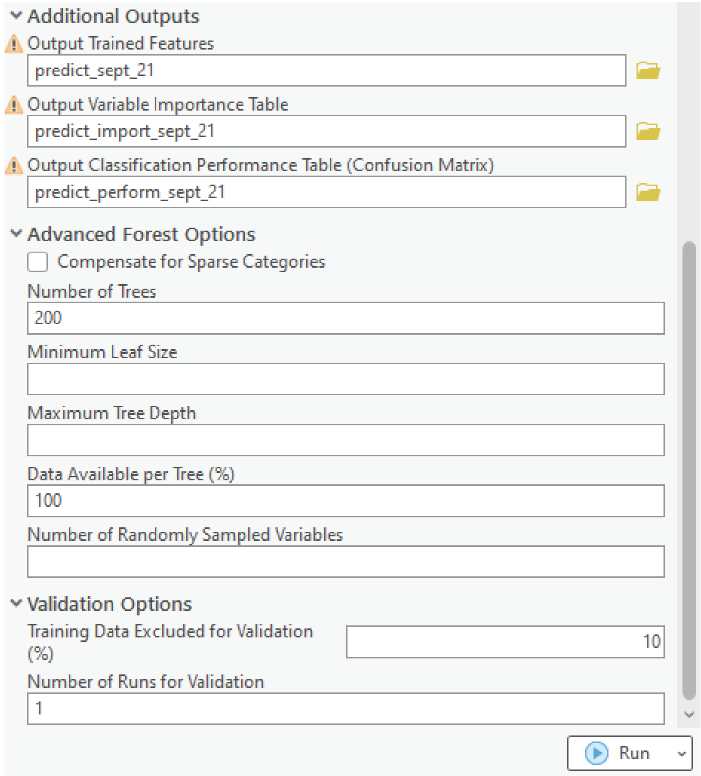


# Random Forest Model

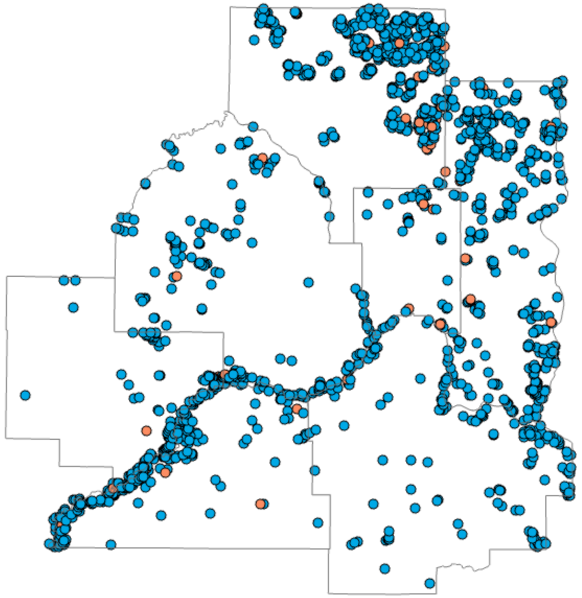
## ESRI Forest-based Classification and Regression (Spatial Statistics)

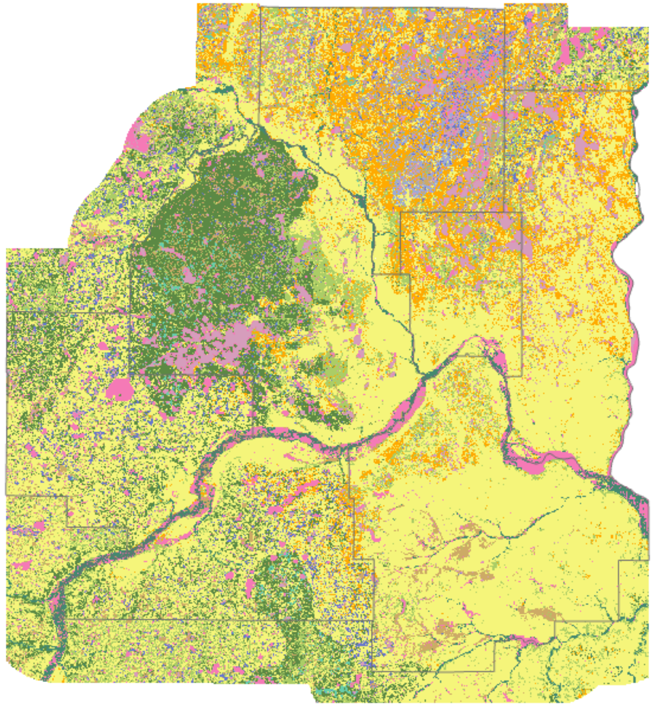


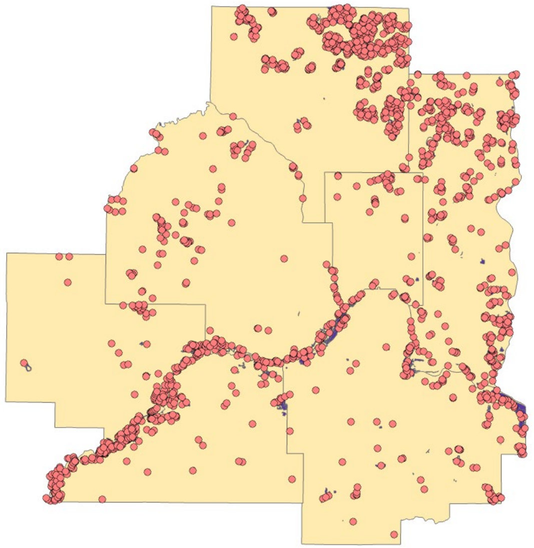



# The Picturesque Results

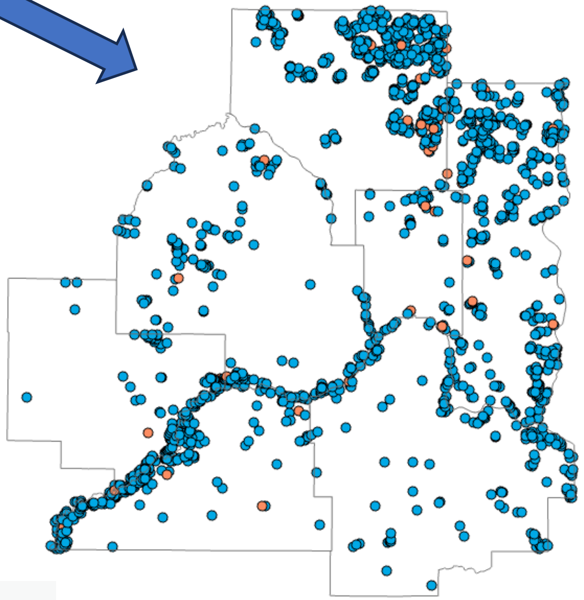




# Random Forest Model - Results

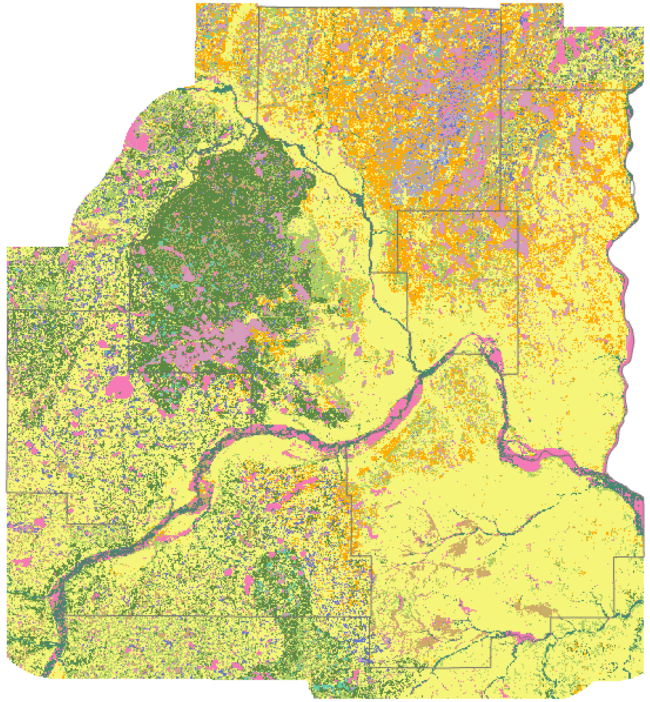


NPC points



CLASS_CODE(SAME_CLASS)
- Misclassified
- Correctly Classified

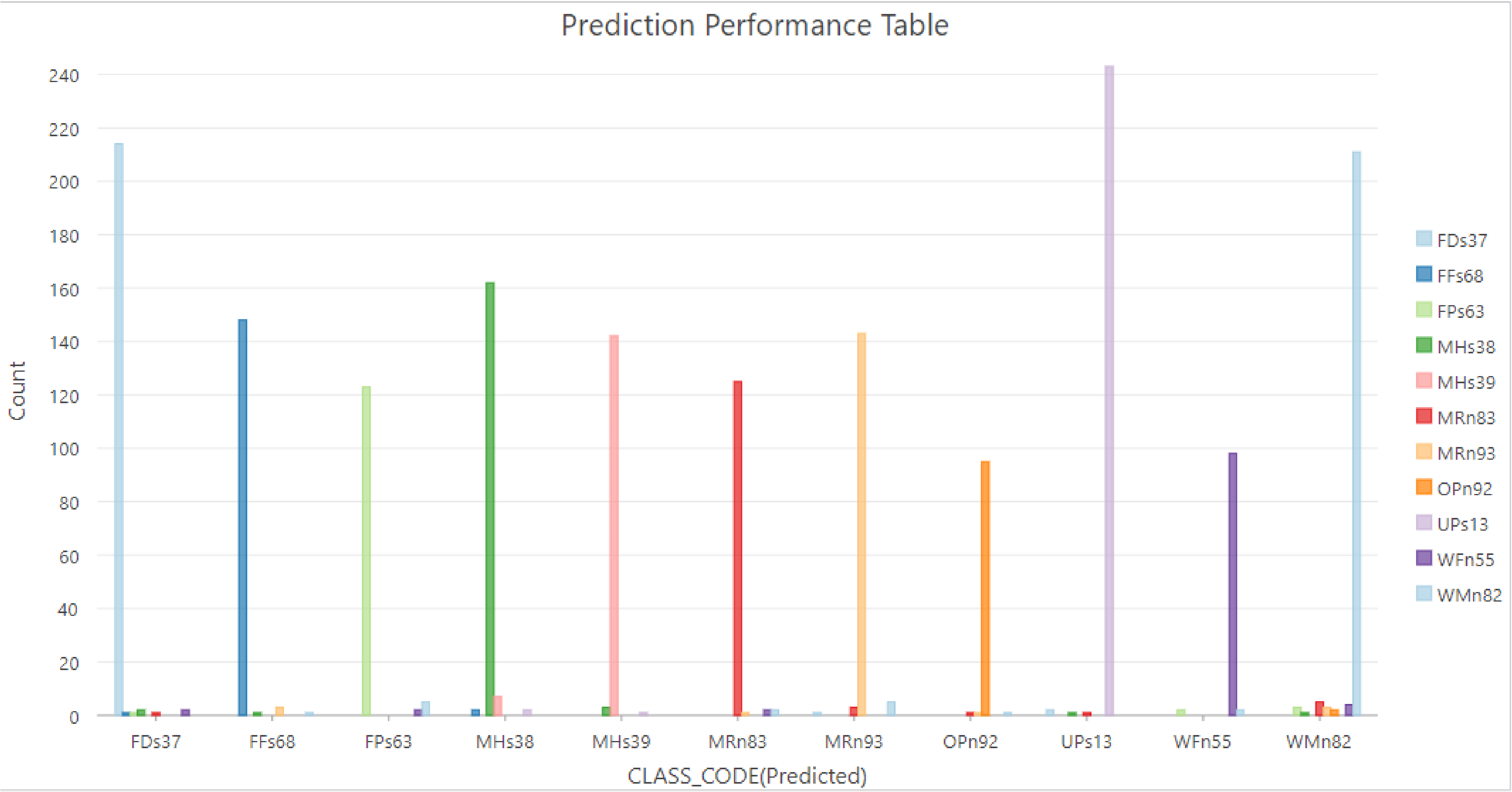
Prediction points



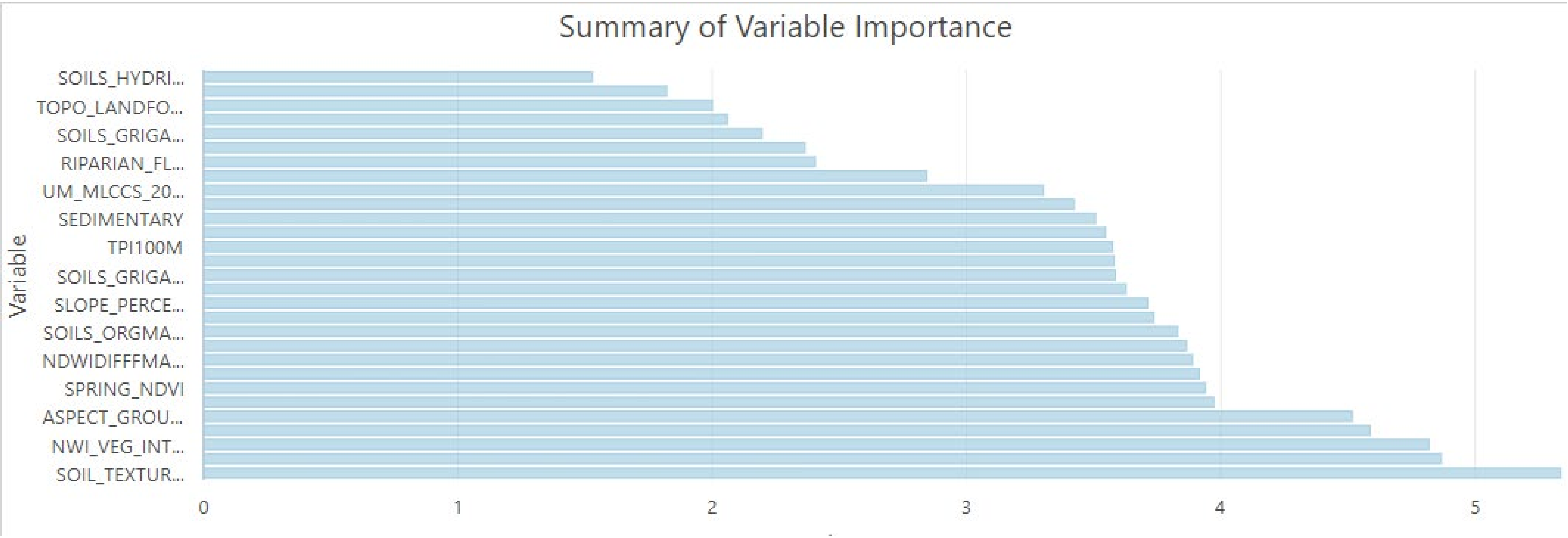
CLASS_CODE(Predicted)
- FDs37 - Southern Dry-Mesic Oak (Maple) Woodland
- FFs68 - Southern Floodplain Forest
- FPs63 - Southern Rich Conifer Swamp
- MHs38 - Southern Mesic Oak-Basswood Forest
- MHs39 - Southern Mesic Maple-Basswood Forest
- MRn83 - Northern Mixed Cattail Marsh
- MRn93 - Northern Bulrush-Spikerush Marsh
- OPn92 - Northern Rich Fen (Basin)
- UPs13 - Southern Dry Prairie
- WFn55 - Northern Wet Ash Swamp
- WMn82 - Northern Wet Meadow/Carr

Prediction layer

# Random Forest Model - Results



# Random Forest Model - Results



# Random Forest Model - Results

| Model Characteristics | |
|---|---|
| Number of Trees | 200 |
| Leaf Size | 1 |
| Tree Depth Range | 17-30 |
| Mean Tree Depth | 22 |
| % of Training Available per Tree | 100 |
| Number of Randomly Sampled Variables | 5 |
| % of Training Data Excluded for Validation | 10 |

| Model Out of Bag Errors | | |
|---|---|---|
| Number of Trees | 100 | 200 |
| MSE | 42.156 | 40.661 |
| | | |
| FDs37 | 33.012 | 31.039 |
| FFs68 | 27.439 | 25.460 |
| FPs63 | 33.930 | 32.689 |
| MHs38 | 64.487 | 63.656 |
| MHs39 | 31.791 | 30.719 |
| MRn83 | 62.103 | 60.239 |
| MRn93 | 38.238 | 36.662 |
| OPn92 | 57.755 | 57.399 |
| UPs13 | 13.251 | 11.168 |
| WFn55 | 67.031 | 67.561 |
| WMn82 | 60.281 | 58.312 |

# Random Forest Model - Results

## Top Variable Importance

| Variable | Importance | % |
|---|---|---|
| SOIL_TEXTURE_SIMPLE | 7.04 | 5 |
| CANOPY_HEIGHT_0 | 6.31 | 5 |
| ASPECT_GROUP4 | 6.21 | 5 |
| NWI_VEG_INT_0 | 6.19 | 5 |
| SOILS_GRIGAL_RELIEF5 | 6.18 | 5 |
| SUMMER_NDVI | 5.53 | 4 |
| DISTRIVER | 5.40 | 4 |
| NLCD_2019 | 5.36 | 4 |
| SPRING_NDVI | 5.34 | 4 |
| NDWIDIFFFMAYTOSEPTL8 | 5.18 | 4 |
| SLOPE_PERCENT_NEW | 5.16 | 4 |
| SOILS_ORGMATTER | 5.14 | 4 |
| SOLAR_AREA | 5.14 | 4 |
| LANDSURFACETEMP_2022_LOW2 | 5.13 | 4 |
| TPI100M | 4.96 | 4 |
| CHILI | 4.93 | 4 |
| ASPECT_REAL | 4.89 | 4 |
| CURVATURE | 4.86 | 4 |
| SEDIMENTARY | 4.80 | 4 |
| SOILS_GRIGAL_TEXTURE3 | 4.77 | 4 |

# Random Forest Model - Results

## Training Data: Classification Diagnostics

*Predictions for the data used to train the model compared to the observed categories for those features

| Category | F1-Score | MCC | Sensitivity | Accuracy |
|---|---|---|---|---|
| FDs37 | 1.00 | 1.00 | 1.00 | 1.00 |
| FFs68 | 1.00 | 1.00 | 1.00 | 1.00 |
| FPs63 | 1.00 | 1.00 | 1.00 | 1.00 |
| MHs38 | 1.00 | 1.00 | 1.00 | 1.00 |
| MHs39 | 1.00 | 1.00 | 1.00 | 1.00 |
| MRn83 | 1.00 | 1.00 | 0.99 | 1.00 |
| MRn93 | 0.99 | 0.99 | 1.00 | 1.00 |
| OPn92 | 1.00 | 1.00 | 1.00 | 1.00 |
| UPs13 | 1.00 | 1.00 | 1.00 | 1.00 |
| WFn55 | 1.00 | 1.00 | 1.00 | 1.00 |
| WMn82 | 1.00 | 0.99 | 0.99 | 1.00 |

# Random Forest Model - Results

## Validation Data: Classification Diagnostics

*Predictions for the test data (excluded from model training) compared to the observed values for those test features

| Category | F1-Score | MCC | Sensitivity | Accuracy |
|---|---|---|---|---|
| FDs37 | 0.62 | 0.59 | 0.73 | 0.94 |
| FFs68 | 0.78 | 0.75 | 0.82 | 0.96 |
| FPs63 | 0.58 | 0.54 | 0.60 | 0.93 |
| MHs38 | 0.46 | 0.40 | 0.50 | 0.89 |
| MHs39 | 0.52 | 0.49 | 0.46 | 0.94 |
| MRn83 | 0.52 | 0.48 | 0.44 | 0.92 |
| MRn93 | 0.65 | 0.61 | 0.62 | 0.92 |
| OPn92 | 0.55 | 0.53 | 0.60 | 0.97 |
| UPs13 | 0.85 | 0.83 | 0.87 | 0.96 |
| WFn55 | 0.42 | 0.39 | 0.33 | 0.92 |
| WMn82 | 0.38 | 0.28 | 0.42 | 0.82 |

# Random Forest Model - Results

**Explanatory Variable Range Diagnostics**

(a) % of overlap between the ranges of the training data and the input explanatory variable
(b) % of overlap between the ranges of the validation data and the training data
(c) % of overlap between the ranges of the training data and the prediction data
* Data ranges do not coincide. Training or validation is occurring with incomplete data.
+ Ranges of the training data and prediction data do not coincide and the tool is attempting to extrapolate.
*Succeeded at Thursday, September 21, 2023 5:10:24 PM (Elapsed Time: 14 minutes 11 seconds)*

| Variable | Training | | Validation | | Prediction | | Share | | |
|---|---|---|---|---|---|---|---|---|---|
| | Minimum | Maximum | Minimum | Maximum | Minimum | Maximum | Training[a] | Validation[b] | Prediction[c] |
| TPI100M | -11.34 | 7.52 | -4.80 | 3.85 | -17.86 | 15.37 | 1.00 | 0.46* | 1.76+ |
| DISTRIVER | 0.00 | 0.08 | 0.00 | 0.06 | -0.00 | 0.10 | 1.00 | 0.74* | 1.20+ |
| SUMMER_NDVI | 0.01 | 0.51 | 0.03 | 0.50 | -0.11 | 0.55 | 1.00 | 0.94* | 1.32+ |
| SPRING_NDVI | -0.03 | 0.39 | -0.02 | 0.25 | -0.12 | 0.49 | 1.00 | 0.66* | 1.47+ |
| SOLAR_AREA | 4717.21 | 5976.16 | 4722.52 | 5946.24 | 4158.86 | 6035.09 | 1.00 | 0.97* | 1.49+ |
| SOILS_ORGMATTER | 0.00 | 87.00 | 0.00 | 87.00 | 0.00 | 87.00 | 1.00* | 1.00 | 1.00 |
| SLOPE_PERCENT_NEW | 0.00 | 29.39 | 0.00 | 30.10 | 0.00 | 40.26 | 0.98* | 1.02 | 1.37+ |
| NDWIDIFFMAYTOSEPTL8 | -0.25 | 0.12 | -0.20 | 0.07 | -0.32 | 0.25 | 1.00 | 0.75* | 1.58+ |
| LANDSURFACETEMP_2022_LOW2 | 81.00 | 99.18 | 81.00 | 95.32 | 81.00 | 109.50 | 1.00 | 0.79* | 1.57+ |
| CURVATURE | -3.51 | 2.73 | -1.57 | 1.40 | -8.00 | 7.51 | 1.00 | 0.48* | 2.49+ |
| CHILI | 47.25 | 239.08 | 96.61 | 225.20 | 4.02 | 263.65 | 1.00 | 0.67* | 1.35+ |
| CANOPY_HEIGHT_0 | 0.00 | 34.38 | 0.00 | 29.39 | 0.00 | 118.81 | 1.00 | 0.85* | 3.46+ |
| ASPECT_REAL | -1.00 | 358.38 | -1.00 | 345.52 | -1.00 | 359.99 | 1.00 | 0.96* | 1.00+ |

# Next Steps…

# Next Steps…

## Going from the predictive grid to polygons

- Where MLCCS data exists, intersect the natural polygons with the NPC predictive grid
  - Create a protocol that maintains the hydrological regime between the MLCCS and NPC
  - An upland MLCCS polygon should align with an NPC upland community

- Use LandSat, LiDAR, and aerial photos to generate polygons based on vegetation patterns
  - Create a process that aids aerial photo interpretation and heads up digitizing

# Next Steps…

**Run the model statewide(?)**

- Greatly increase the number of native plant communities
- Find tune the training layers for different geographic regions – north, central, and south
- Try modeling to the NPC Type and Subtype level
- Use Trimble eCognition to help create NPC polygons

# Thank you!

A special thank you to:

- Amy Kendig
- Dustin Graham
- Bob Dunlap
- Anne Espeset
- Jennifer Corcoran
- Ram Deo
- Molly Shoberg
- Ben Gosack
- Jason Husveth
- Scott Zager
- Seth Fore