Northern Minnesota Ecological Conservation

Exploring Classification of Forest Cover Types in Northern Minnesota Using Earth Observations

Summer 2025 | Idaho – Pocatello
August 8th, 2025

Authors: Betty Brown (Analytical Mechanics Associates), Nettie Hitt (Analytical Mechanics Associates), Isabelle LeConche (Analytical Mechanics Associates), Hailey Phillips (Analytical Mechanics Associates)

Abstract:
The Superior National Forest encompasses the northeastern part of Minnesota and contains critical ecological and economic resources. Responsible for assessing the forest's resources, the Minnesota Department of Natural Resources (MNDNR) requires data about the spatial distribution of tree species in order to inform their management decisions. Remote sensing can provide spectral and topographic data across large spatial extents, and machine learning models can leverage these data to generate spatial distribution maps of forest cover types needed to inform these decisions. The MNDNR partnered with NASA DEVELOP to explore a machine-learning method to classify the Superior National Forest into a cover type map. The team developed a supervised machine learning model that combined the MNDNR's plot-based forest inventory data with multispectral Harmonized Landsat Sentinel-2 imagery and topographic data to classify the study area into 12 different land cover classes. With an overall accuracy of 61.1%, the best performing model demonstrated some feasibility of our methods but was insufficiently reliable to produce a highly accurate land cover map containing classes with individual tree species. However, this project provided important insights on mapping methods. We found that topographic data had variable impact on accuracy, and the gradient boosting technique improved accuracy compared to the random forest technique. We also found a positive correlation between sample size and model accuracy, highlighting the need for additional and more proportionate distribution of training data across the 12 land cover classes. These findings will be used to support the MNDNR's efforts to generate land cover classifications needed for assessing detecting changes in forest composition.

Key Terms: Harmonized Landsat Sentinel-2 (HLS), lidar, forest ecology, natural resource management, machine learning, forest classification, mixed conifer and hardwood forest, Minnesota

Advisor: Keith Weber (Idaho State University, GIS Training and Research Center)

Lead: Isaac Goldings (Idaho – Pocatello)

1. Introduction

Maintaining biodiversity has become an important strategy in the sustainable use of forest resources (Lindenmayer et al., 2000). By mapping the composition of tree species in forests, researchers and managers can track changes in ecosystem health and biodiversity over time (Immitzer et al., 2012; Noss & Cooperrider, 1994; Persson et al., 2018). Spatial distribution maps of tree species offer insight into fire vulnerability, pest management, invasive species monitoring, and wildlife habitat mapping in addition to inventorying resources (Fassnacht et al., 2016). While field-based inventory methods are widely used to generate species distribution maps, complete in situ classification of public forests proves labor-intensive and unfeasible to conduct across large study areas (Immitzer et al., 2012). Remote sensing facilitates the scaling-up of species classification from plot to regional level. After training a classifier with plot-based in situ species data, machine-learning models can efficiently categorize Earth observation data into maps of different tree species and/or species associations.

Previous studies have demonstrated both the powerful capability and potential shortcomings of using machine-learning models to classify remotely sensed imagery (Grossmann et al., 2010). Classification accuracy varies across different decision tree-based models, such as random forest models and gradient-boosted models. Gradient-boosting is a machine-learning technique in which each decision tree learns from the previous iteration, potentially improving accuracy (Friedman 2001). Preceding literature has compared the efficacy and accuracy of gradient boosted decision tree models to non-gradient boosted decision tree models. Łos et al. (2021) compared the accuracy of a gradient-boosted model using the XGBoost library and random forest model for species classification and achieved an F-score 3% higher with the gradient boosted model (F-score = 95%).

In addition to model type, different combinations of spectral bands and temporal and spatial resolutions produce variable model accuracy and separability between tree species. In their comparison of three random forest models using Sentinel-2 and Landsat 8 derived imagery, Wang et al. (2022) identified short wave infrared bands (SWIR) and tasseled cap (TC) coefficients as important features in random forest model classification. Persson et al. (2018) also identified Sentinel-2 SWIR as bands of importance in addition to red edge and near infrared bands (NIR). Using Worldview-2 imagery, Immitzer et al. (2019) determined green, NIR, and blue bands were the most important in their model. When comparing models that used different combinations of spectral bands, both Immitzer et al. (2012) and Persson et al. (2018) achieved the highest overall accuracy using all bands (OA=95.9% and OA=88.2% respectively). Persson et al. (2018) also examined the impact of various phenological stages and temporal resolutions on the classification model. For a mixed conifer/deciduous forest, they determined that exploiting phenological variation between species is important for the success of the model. They achieved highest accuracy by including imagery from three seasons, which included the months April, May, July, and October. Existing literature also suggests the efficacy of combining spectral data with topographic and texture data to enhance classification model accuracy (Dalponte et al., 2012; Naidoo et al., 2012). In testing various model configurations of multispectral, topographic, and texture data, Ma et al. (2021) achieved highest classification accuracy using all three data types (OA=86.49%). These research insights provide avenues to attain high model accuracy, a critical standard for the useful application of machine learning models in forest management. While accuracy thresholds depend on the specific context of model application, 85% overall accuracy is consistently cited as the minimum accuracy required for high-confidence land classification mapping (Anderson et al., 1976).

Public forest management in the United States commonly falls under the responsibility of state authorities, such as the Minnesota Department of Natural Resources (MNDNR). The MNDNR conducts plot-based inventories (PBI) to maintain a publicly available, in situ record of forest traits and species composition. Maintaining a highly accurate and up-to-date record of cover type enables the MNDNR to monitor forest dynamics, including shifts in species diversity, and to manage and sustain forest resources. Specifically, this data informs management decisions surrounding biodiversity conservation, timber harvesting, and wildfire prevention and recovery. To better identify tree species across broad forest areas, the MNDNR has tested a

random forest machine learning approach in Google Earth Engine, but their model does not yet have a reliable accuracy (OA=63.9%). By refining a model to more accurately classify tree species, we aim to reduce the investment of time and cost needed for forest inventories in the field by the MNDNR, while providing an adaptable and efficient method of monitoring changes in forest composition over time.

In this study, we combined spectral imagery from Harmonized Landsat Sentinel-2 (HLS S30) with United States Geological Survey (USGS) topographic data and developed a Gradient Boosted decision tree model to classify 12 tree species groupings across Superior National Forest (NF). Located in Northeastern Minnesota, Superior NF spans an area of approximately3 million acres (USDA Forest Service, 2004). consisting of mixed coniferous and hardwood species. To obtain the most up-to-date cover type data for Superior NF, we trained our model with PBI data from 2020 to 2022 and applied these results to classify satellite imagery from 2024. Using the gradient boosted decision tree, we aimed to (1) create a map of the current species composition and distribution for 12 common tree species classes in the Superior NF, (2) determine how the inclusion of different spectral bands, topographic data, and seasonality affects the accuracy and feasibility of the gradient boosted model, and (3) assess how forest cover type has changed in Superior National Forest over time. Answers to these questions will enable agencies, like the MNDNR, to harness Earth observation data to monitor and better manage forests.

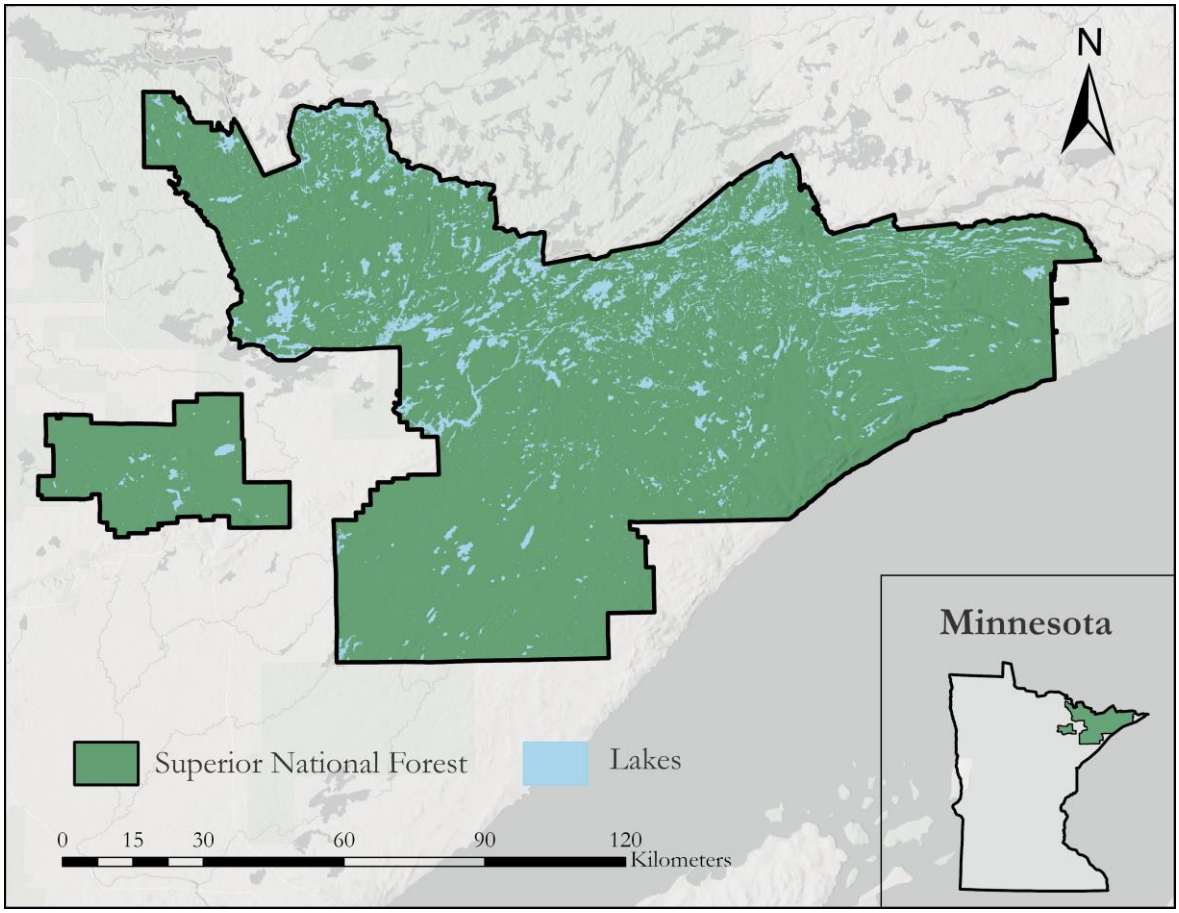

Figure 1. Spatial extent of the study area, Superior National Forest (Esri, CGIAR, USGS, Sources: Esri, TomTom, Garmin, GAO, NOAA, USGS, OpenStreetMap contributors, the GIS UserCommunity, U.S. Department of Commerce, Census Bureau; U.S. Department of Commerce (DOC), National Ocean Service (NOS), National Geodetic Survey (NGS)).

2. Methodology

2.1 Data Acquisition

We used Harmonized Landsat Sentinel (HLS S30) Multi-Spectral Instrument Surface Reflectance Daily Global 30m v2.0 Level 3 products from Earthdata (Table A1). For HLS data, the Landsat data is from Landsat 8 Operational Land Imager (OLI), Landsat 9 OLI-2, and the Sentinel data is from the Sentinel-2 Multispectral Instrument (MSI). To achieve full spatial coverage, we extracted scenes for nine days across spring, summer, and fall. Ultimately, we excluded summer imagery from our models due to the cloudiness of the scenes. Additionally, a previous study on a similar mixed broadleaf/coniferous forest revealed that adding summer imagery to a joint spring and fall model was redundant (Persson et al., 2018). We downloaded spectral imagery from 2021 to train our model because this fell within the middle of the PBI data collection range from 2020 to 2022. We then acquired imagery from 2024 for up-to-date comparison. After setting the search parameters to retrieve scenes with less than 20% cloud cover, we downloaded eight granules for each selected day to cover the Superior NF and maximize the associated PBI sample size. We acquired 2024 imagery for both spring and fall using the same parameters for the generation of an up-to-date cover type map. We used water boundaries from the USGS National Hydrography Dataset to mask out water bodies across the 2021 and 2024 spectral data (Table A1). For topographic data, we downloaded Digital Elevation Model (DEM) rasters from USGS 3D Elevation Program at 1 meter resolution, spanning 2018 to 2022 to cover the extent of our study area (Table A1). Our training data consisted of plot-based inventory (PBI) data from our partner, the MNDNR (Table A1). The MNDNR used two methodologies to collect PBI data: in situ data collection and remote sensing. To collect the ground truth PBI data, they assigned one of 12 tree species classes to a plot based on the majority tree species within its 11.33 meter radius. Secondarily, they also used true-color imagery from the fall to create representative plots for each class. The MNDNR developed these cover type classes informed by their management needs and commonly occurring stand compositions. The 12 classes were as follows: Ash, Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar, Northern/Central Hardwood, Oak, White Pine, Norway Pine, Jack Pine, White Spruce/Balsam Fir, Upland Larch (Tamarack), White/Red Cedar, Black Spruce Upland/Lowland, and Other (Upland/lowland brush, upland/lowland grass, marsh, and muskeg).

2.2 Data Processing

We processed all HLS granules in the ArcGIS Pro 3.5.2 ArcPy interface. To avoid erroneous spectral values, we used the Quality Assessment (QA) to convert every cloudy pixel to NODATA. Cloud-free granules were mosaicked across dates by season to cover the extent of the study area. Many pixels in the mosaics returned values outside the spectral reflectance range of HLS (0 to 10,000). Spectral values from 0 to 10,000 were preserved, and the rest were converted to NODATA. We reprojected the mosaics to the coordinate system of our Lidar data, NAD 83 UTM Zone 15N.

We derived several vegetation indices from HLS spectral bands including: Normalized Difference Vegetation Index (NDVI; Kriegler et al., 1969), Normalized Difference Moisture Index (NDMI; Jin & Sader, 2005), and Normalized Difference Water Index (NDWI; McFeeters, 1996). We created two NDVI rasters, one using the Broad NIR band (bNIR) and the other using the Narrow NIR (nNIR) (Equation 1-4).

$$Broad\ NDVI = \frac{(bNIR - RED)}{(bNIR + RED)} \tag{1}$$

$$Narrow\ NDVI = \frac{(nNIR - RED)}{(nNIR + RED)} \tag{2}$$

$$NDMI = \frac{(NIR - SWIR)}{(NIR + SWIR)} \tag{3}$$

$$NDWI = \frac{(GREEN - NIR)}{(GREEN + NIR)} \quad (4)$$

Using all spectral bands per season, we conducted a principal component analysis (PCA) in TerrSet liberaGIS version 20. PCA produced one composite multi-band raster file per season, each containing 99% of our unique data for fall and for spring. We used the results as an input layer for our models.

We created mosaics of both bare Earth elevation and maximum surface height rasters from Lidar-derived data for our study area. The bare earth mosaic provided elevation data that was then used to derive slope and aspect. We also calculated canopy height by subtracting the bare earth mosaic from the maximum surface height mosaic. This layer was developed to help differentiate cover types by height differences that may exist between land cover classes. Since the PBI polygons were smaller than an HLS pixel (30 meters), we converted the PBI data to points to ensure that each plot is attributed to a single pixel within the HLS imagery. For each model iteration, we clipped the PBI points twice. First, the PBI data was clipped to the extent of our Lidar data, and second, these data were clipped to the cloud mask for the associated season or season combination. Additionally, two points that had excessive (>11 meters) positional (ie., locational) departures were excluded (this threshold was based on the distribution of overall PBI point positional accuracy). Using the clipped PBI data, we randomly split this dataset (n=~830, varying by each temporal cloud mask) ensuring each of the 12 cover type classes contained 60% of our PBI data (used for training) and another dataset containing 40% of our PBI data (used for validation).

2.3 Data Analysis

To train a model that can accurately classify the 12 species classes, we used the gradient boosted model in the Forest-based and Boosted Classification and Regression tool in ArcGIS Pro. The 60% subset of PBI points for the proper season combination were input as the training data. We ran four variations using different data layers and compared their accuracies (Table 1). These model variations focused on Spring and Fall/Spring combinations following methodologies reported in Persson et al. (2018). We imported all raw spectral spring bands and spring training data into TerrSet and used MakeSig and SigComp modules to compare the spectral separability across species groupings.

We trained each model using 50 trees (iterations). We made this decision based on two exploratory Fall/Spring Spectral + Topographic models, one with 50 trees and one with 100 trees. The model that used 100 trees did not outperform the model that used 50 trees, justifying the use of 50 trees for the rest of the model variations. Each model trained and then created a prediction raster, a map visualizing where it predicted the spatial distribution of each class.

Table 1.
Different model variations for 2021 (see Table A1 for spectral imagery acquisition dates).

Model	**Included variables**
Fall/Spring Spectral + Topographic	Spring and Fall bands 1, 2, 3, 4, 5, 6, 7, 8, 8A, 11, 12 Spring and Fall PCA layers Spring and Fall broad NDVI, narrow NDVI, NDWI, and NDMI Canopy height, slope, aspect, elevation
Fall/Spring Spectral	Spring and Fall bands 1, 2, 3, 4, 5, 6, 7, 8, 8A, 11, 12 Spring and Fall PCA layers Spring and Fall broad NDVI, narrow NDVI, NDWI, and NDMI
Spring Spectral + Topographic	Spring bands 1, 2, 3, 4, 5, 6, 7, 8, 8A, 11, 12 Spring PCA layers Spring broad NDVI, narrow NDVI, NDWI, and NDMI

	Canopy height, slope, aspect, elevation
Spring Spectral	Spring bands 1, 2, 3, 4, 5, 6, 7, 8, 8A, 11, 12 Spring PCA layers Spring broad NDVI, narrow NDVI, NDWI, and NDMI

Overall Accuracy (OA) and Kappa Index of Agreement (KIA) statistics for each model were calculated in TerrSet v.20. This was done by creating error matrices using the respective 40% subsets of PBI data and the forest cover type predictions for independent validation. We used the output prediction raster from our highest accuracy model to create a 2021 cover type map. Additionally, we used the Top Variable Importance table produced in ArcGIS Pro to compare the Importance Gain (%) of variables within the model.

To create a forest cover classification map for 2024, we trained another model that used the same variables as the highest-accuracy model and predicted forest cover type for the study area. The model used Spring and Fall bands 1, 2, 3, 4, 5, 6, 7, 8, 8A, 11, 12, Spring and Fall PCA layers, and Spring and Fall broad NDVI, narrow NDVI, NDWI, and NDMI from HLS. This 2024 version used the same topographic data as the 2021 models and was trained on the same PBI dataset, but cloud masked to 2024 spectral imagery. We calculated the total area predicted for each species grouping for 2021 and 2024 to assess how the forest cover type may have changed across the landscape.

To evaluate how sample size affects model accuracy, we created a linear regression using the validation data from the highest accuracy model for each species class and their associated user accuracies. Furthermore, to assess the impact of independent validation, we trained the model with the highest accuracy again and performed a dependent validation without manually splitting and withholding any data. Using the entire PBI dataset, we trained the model with 90% of the data points and used the remaining 10% for the ArcGIS tool's in-built, dependent validation mechanism. To produce an error matrix, we validated the model using 100% of the PBI data.

Using the Forest-based and Boosted Classification and Regression tool in ArcGIS Pro, we created a random-forest based classification model to compare to the highest accuracy gradient boosted classification model. The random forest version was trained on the same subset of our PBI data and consisted of all the same inputs as our highest accuracy gradient boosted model.

3. Results

3.1 Analysis of Results

Among the four model variations for 2021 spectral imagery, the Fall/Spring Spectral + Topographic model achieved the highest OA and KIA (OA=61.1%, KIA= 0.51), followed by Spring Spectral (OA=59.9%, KIA= 0.50), Fall/Spring Spectral (OA=57.4%, KIA= 0.46), and Spring Spectral + Topographic (OA=55.4%, KIA= 0.44) (Figures 2 and 3). User Accuracy (UA) for the Fall/Spring Spectral + Topographic model ranged from 0.45-86.2% and Producer Accuracy (PA) ranged from NA-79.1%, where NA represents the oak class that was not validated because of insufficient sample sizes within the independent validation split (Table 2). Amongst the tree species classes, the (1) Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar and (2) Black Spruce Upland/Lowland groups were classified with the highest user accuracies at 86.2% and 77.4%, respectively (Table 2).

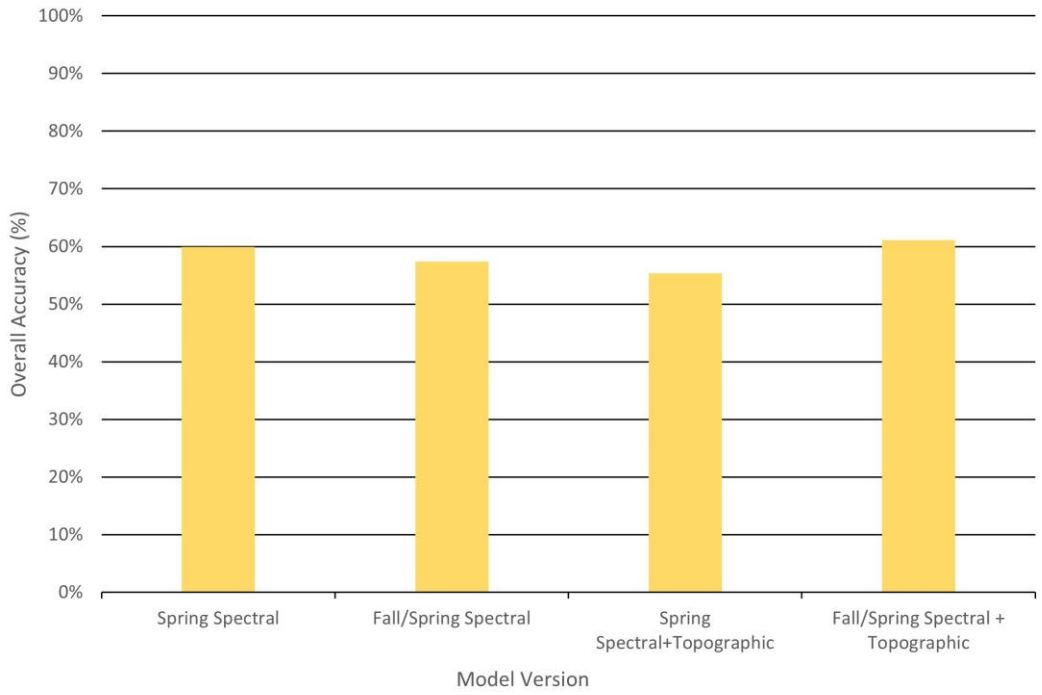

Figure 2. Overall Accuracy (OA) of 12 cover classes from the final model variations; (left to right) Spring Spectral data only, Fall/Spring spectral data only, Spring Spectral data and Topographic layers, and lastly, Fall/Spring Spectral data and Topographic layers.

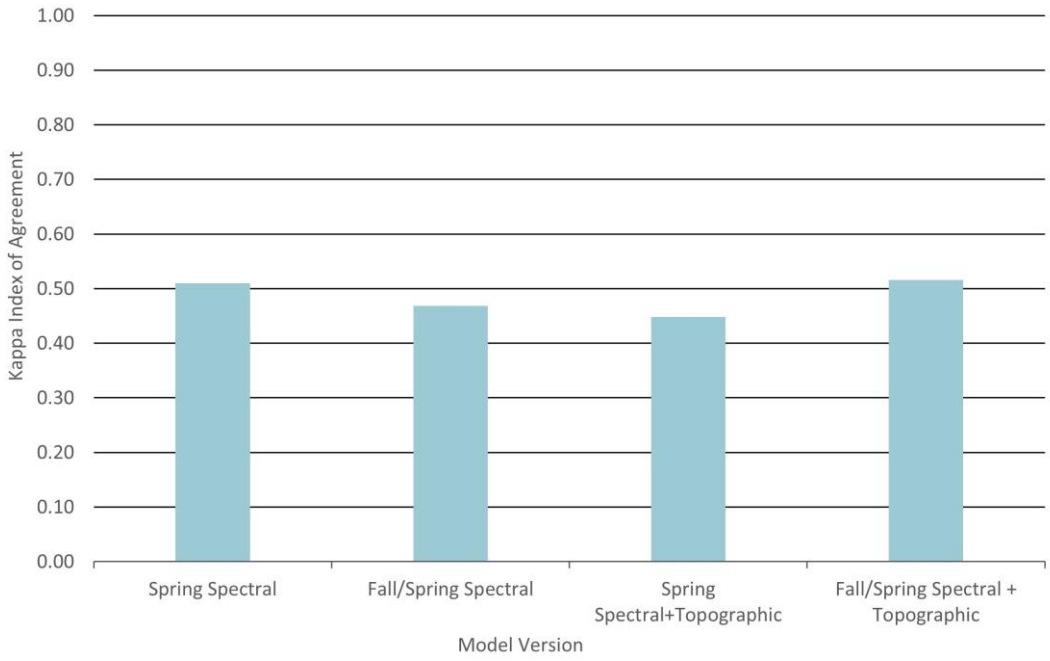

Figure 3. Kappa Index of Agreement (KIA) of 12 cover classes from the final model variations; (left to right) Spring Spectral data only, Fall/Spring Spectral data only, Spring Spectral data and Topographic layers, and lastly, Fall/Spring Spectral data and Topographic layers.

Table 2.
Confusion Matrix for the Fall/Spring Spectral + Topographic model. Each predicted class appears across the x-axis where 1 = Ash, 2 = Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar, 3 = Black Spruce Upland/Lowland, 4 = Jack Pine, 5 = Northern/Central Hardwood, 6 = Norway Pine, 7 = Oak, 8 = Other, 9 = Upland Larch (Tamarack), 10 = White Pine, 11 = White Spruce/Balsam Fir, and 12 = White/Red Cedar. Each class present in the validation data appears along the y-axis. The total number of correctly classified plots per class is represented by the green cells along the diagonal. PA stands for producer accuracy, UA for user accuracy, and OA for overall accuracy. The total along the x-axis represents the number of validation plots per class while the total along the y-axis represents the predicted number of plots per class.

	1	2	3	4	5	6		8	9	10	11	12	Total	PA (%)
1	12	4	0	0	0	0		1	2	0	1	1	21	57.1
2	17	156	9	2	11	4		13	11	1	20	6	250	62.4
3	5	3	96	3	0	1		2	11	0	9	5	135	71.1
4	0	1	0	4	0	1		0	0	0	0	0	6	66.7
5	0	3	0	0	19	1		0	0	0	1	0	24	79.1
6	0	1	3	1	0	11		0	0	2	1	2	21	52.3
7	0	0	0	0	0	0		0	0	0	0	0	0	NA
8	3	1	4	0	1	0		7	1	0	2	0	19	36.8
9	0	1	5	1	0	0		5	24	0	0	2	38	63.2
10	0	1	0	0	0	3		0	0	10	0	0	14	71.4
11	0	2	0	0	0	1		0	0	8	1	0	12	66.7
12	0	3	3	3	2	0		1	1	2	4	21	40	52.5
Total	38	181	124	18	34	20		32	50	13	44	40	594	
UA (%)	31.6	86.2	77.4	22.2	55.8	55		21.9	48	61.5	11.4	52.5		**OA = 61.1%**

Canopy height ranked as the most important variable in both models that used topographic data, including the model that achieved the highest accuracy (Importance Gain=8-9%). Across all four models, Short-wave Near Infrared consistently ranked amongst the top five most important variables (Importance Gain = 6-8%). All other variables displayed inconsistent levels of importance within model variations.

We predicted a 2021 forest cover type map of Superior NF using the Fall/Spring Spectral + Topographic model (Figure 4). Because the Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar and Black Spruce Upland/Lowland groups were classified with the highest accuracy, we computed a map of only these two classes, thereby mapping 59.4% of the useable imagery of Superior NF with a mean user accuracy of 81.8% (Figure 5).

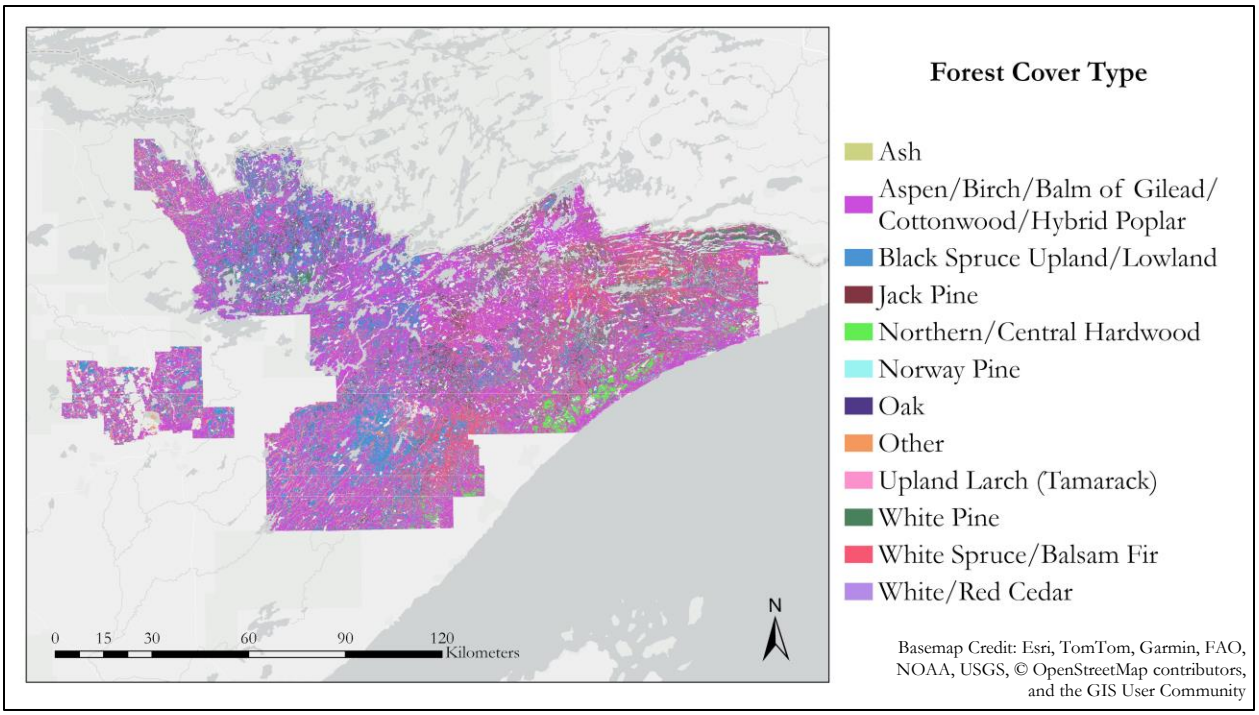

Figure 4. Forest cover classification prediction of Superior National Forest for 2021 by 12 cover type classes. This map was created using the highest accuracy model, Fall/Spring Spectral + Topographic (OA=61.1%).

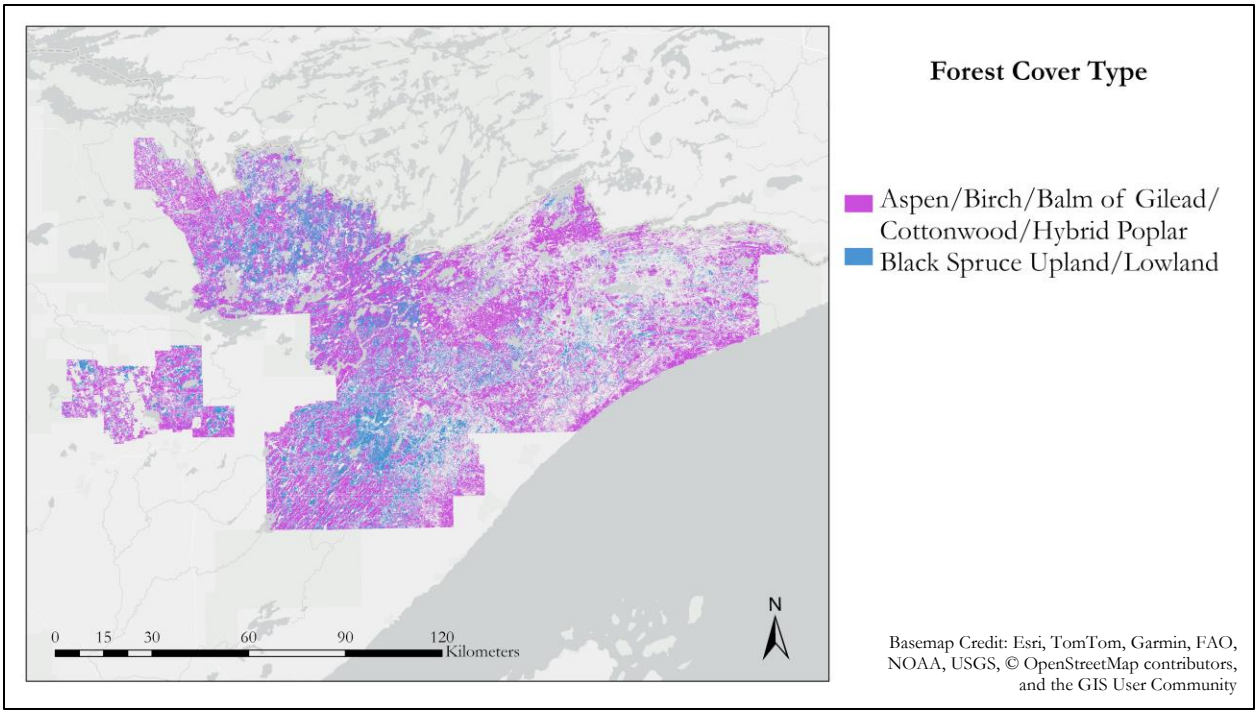

Figure 5. Forest cover classification prediction of Superior National Forest for 2021 for two cover types with the highest user accuracies, Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar (UA= 86.2%) and Black Spruce Upland/Lowland (UA=77.4%).

Forest cover type maps of Superior NF were also produced for 2024 to compare with 2021 (Figure 6). The model trained with 2024 spectral data achieved a lower OA and KIA (OA=54.6%, KIA=0.43). In both 2021 and 2024, the Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar class was the dominant class in the study area, more than double the area of the second most abundant cover type class, Black Spruce Upland/Lowland (Figure 7). White Spruce/Balsam Fir and Others followed Black Spruce Upland/Lowland in abundance in both years. All other classes varied in area between 2021 and 2024 but remained under 500 km² area each. Oak was the least abundant class with extremely low predicted area in both 2021 and 2024 (<2 km²).

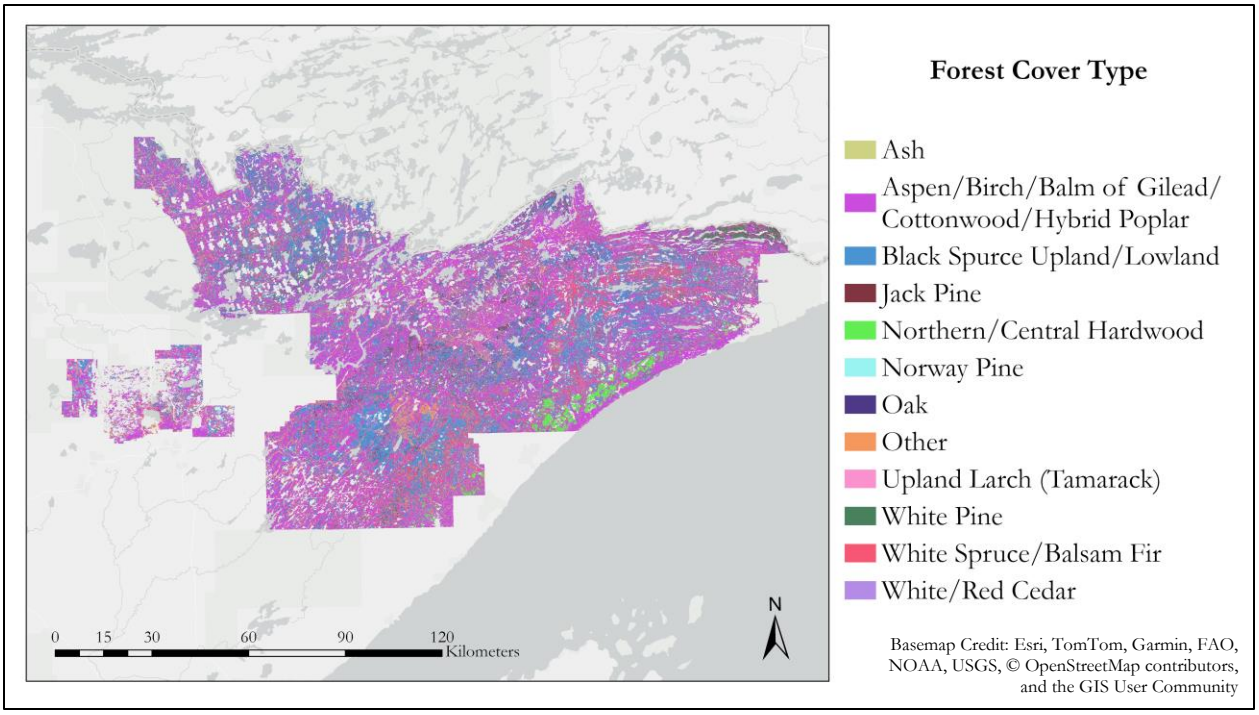

Figure 6. Forest cover classification prediction of Superior National Forest for 2024. This map was created using the 2024 model, which included fall and spectral data for 2024 and topographic features (OA=54.6%).

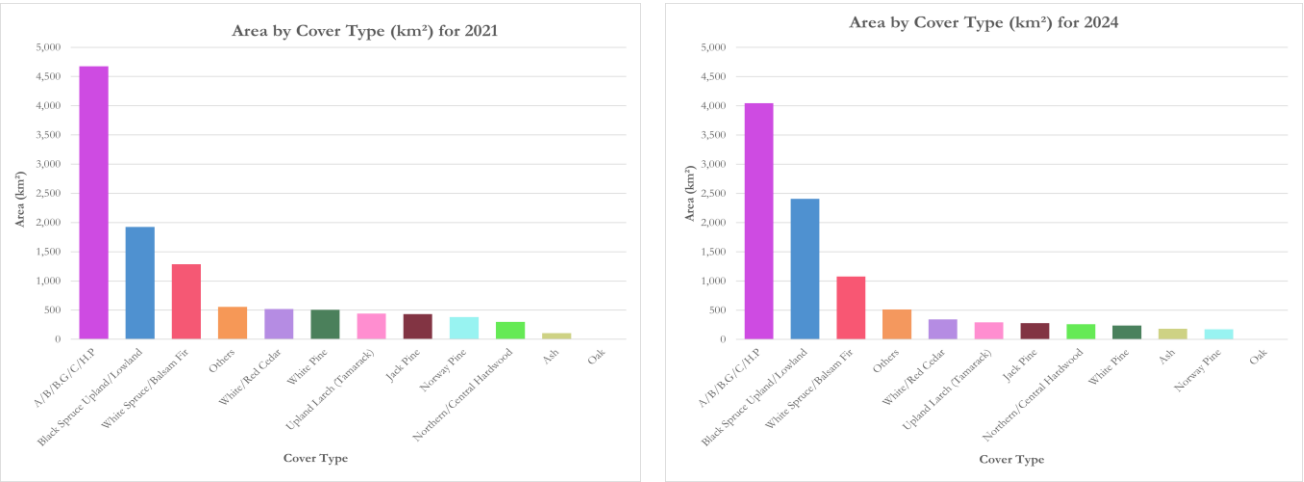

Figure 7. Area (km²) by forest cover type predictions of Superior National Forest for 2021(left) and 2024 (right). A/B/B.G/C/H.P represents Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar. Area of Oak is 0.03 km² (left) and 1.15 km² (right).

In the highest accuracy model developed for this study, OA was positively correlated with the validation sample size across cover type classes (Figure 1B, Appendix B). When assessing the impact of independent validation, we obtained an OA of 74.9% and a KIA of 0.69 for the model, when using dependent validation. The OA was 13.9% higher than observed in the independently validated model, suggesting overfitting. With an OA of 50% and a KIA of 0.36, the random forest-based version of the model had an OA 11% lower than the gradient boosted version and a KIA 0.15 lower.

3.2 Errors & Uncertainties

3.2.1 Input Variables

The Lidar-derived data, specifically canopy height, introduced uncertainty in multiple ways. To calculate canopy height for the entire study area, we subtracted the maximum digital surface height model from the bare earth model. This methodology, however, does not account for non-canopy returns: the maximum return could come from confounding sources, such as a building or a flying bird. While most of our Lidar-derived data was recorded in 2022, some parts of our study area only had Lidar data available for 2018 or for 2021. Due to tree growth and phenology, the temporal mismatch between these data and the rest of the data could result in potentially inaccurate differences in canopy height across the study area.

3.2.2 Training Data

Uneven and generally low sample sizes of the training data introduced the largest uncertainty. The total number of training plots per class ranged from two plots for Oak to 293 plots for the Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid poplar group. Within this vast range, Jack Pine, White Pine, and Oak had less than 30 total training plots, beneath the threshold for statistically significant training data. On average, there were only 0.08 PBI points per square kilometer of our study area. Additionally, after applying the fall and spring cloud mask, the total number of training plots decreased by approximately 20% (n = 830). The correlation graph characterizes this relationship (Figure 8).

Between the two different PBI creation methods, we received positional accuracy estimates for only the field PBI data. Although we excluded plots with positional accuracies over 11 meters, this value still signifies a large potential for inaccurate plot locations. The model treats these plots as definitively located, and as such, could associate surface reflectance values of another species with the training plot pixel. Without positional accuracy for the PBI derived with remote sensing, we are not certain these plots are accurately associated with corresponding spectral and topographic data.

Since the plots are smaller than the pixel size of the spectral imagery, many of our sample plots were initially excluded from the model. To avoid training data exclusion, we converted the plots to their centroid point. This transformation introduced some uncertainty, specifically the relationship between the training data and the high spatial resolution lidar data (1-meter pixels). We assumed the center point of the plot fell on the same type of tree that the plot was classified as. However, if the point happened to fall on a different tree, the model associated this unrepresentative value with the entire plot. Extrapolating one point to the whole plot area might not be representative of the true height composition within the plot.

The species groupings provided by MNDNR could also contribute to model confusion. Developed primarily for functional management applications, these groupings combine species with potentially very different spectral signatures. Visual inspection of the spectral separability graph in TerrSet revealed that the cover classes each had a wide range of reflectance values. The mean spectral values for each class are very similar across bands, preventing clear, quantitative differentiation (Figure C1, Appendix C). Low spectral separability between classes can confuse the model and lower model accuracy. Increasing the sample size in these groups would decrease variation and increase spectral separability, thereby enabling the model to create more accurate splits in decision trees. Comparing project maps to other land cover maps from other sources (e.g., from the USFS and USGS) may also be useful for understanding map errors and uncertainties.

4. Conclusions

4.1 Interpretation of Results

The final results elucidate the potential of using Earth observations to classify forests by overstory tree species or species groups, but also the limitations to achieving reliable overall map accuracy at the full classification scheme specificity. With an overall accuracy of 61.1%, our best performing map falls short of both the initial map computed by the MNDNR, and the 85% OA standard for reliable classification models, although this threshold is generalized and varies by application (Foody, 2008). A comparison of different model variations revealed inconsistent contributions of topographic features to model accuracy. However, our highest accuracy model, though only slightly higher, included topographic data and canopy height consistently ranked as the most important variable, suggesting topographic data should be included in future model development. We also generated other critical findings that will enable our partner to adapt their model design. Comparing independent versus dependent validation models supports independent validation methodology to minimize overfitting and curb potentially misleading model accuracy. The gradient boosted version of our highest accuracy model had a greater overall accuracy and kappa index score than the random forest version. This comparison indicates the potential importance in using the gradient boosting technique to improve classification accuracy.

In addition to revealing the factors contributing to model accuracy, our project illuminated factors that can result in relative inaccuracy. Our auxiliary analysis of the spring spectral separability between classes revealed substantial overlap in spectral reflectance values across classes. While further analysis is required, this finding suggests that a lack of separability between classes contributed to low model accuracy. Moving forward, MNDNR can consider refining these classes to express greater separability. We also discovered a positive correlation between sample size per class and user accuracy, and from this analysis, we determined that low sample sizes likely resulted in low model accuracy. This insight reveals the limitations of using machine-learning models to classify large forest areas. Our findings suggest that producing an accurate classifier requires robust plot-based inventory data to adequately differentiate species. Following this trend, we anticipate the methodology described in this paper could produce a higher accuracy model once a sufficient number of field data have been collected.

4.2 Feasibility & Partner Implementation

While it was feasible to produce a model to classify Superior NF, the final model did not achieve reliable accuracy to be used in the generation of a high-confidence land classification map. Despite unreliable final model accuracy, we suggest several critical, methodological insights to the MNDNR. Now aware of the potential of topographic data to improve model accuracy, our partner can include topographic data in further model development. The MNDNR can confidently switch from using a random forest model to using a gradient-boosted model and expect higher overall accuracy. Finally, by characterizing the limitations of unbalanced training data, we highlighted a key driver of model inaccuracy. This study identified a path toward achieving a reliable classification model for the Superior National Forest.

5. Acknowledgements

Project Partners

- Kangsan Lee & Lucas Spaete (Minnesota Department of Natural Resources (MNDNR))

Science Advisor

- Keith Weber (Idaho State University, GIS Training and Research Center)

DEVELOP Personnel

- Isaac Goldings (Node Lead, Idaho – Pocatello)

This material contains modified Copernicus Sentinel data (2021 & 2024), processed by ESA.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration.

This material is based upon work supported by NASA through contract 80LARC23FA024.

6. Glossary

Band – A wavelength range in the spectrum of reflected electromagnetic energy. Sensors measure reflectance values within a given band.
Confusion Matrix – A table that quantifies exactly how many of the validation data samples the model correctly classified and how many it incorrectly classified per class.
DEM – Digital Elevation Model. a 3D terrain model derived from aerial Lidar imagery.
Dependent validation – An accuracy calculation method in which all training data is inputted into the model tool and the tool itself divides the data and performs in-built validation.
Earth observations – Satellites and sensors that collect information about the Earth's physical, chemical, and biological systems over space and time
Gradient Boosted model – A machine-learning classification model that creates decision trees sequentially, using the previous tree to refine the next.
HLS – Harmonized Landsat Sentinel-2. A hybrid Landsat and Sentinel-2 product that combines imagery from both satellites to produce spectral imagery every 2-3 days at 30-meter pixel resolution.
Independent validation – An accuracy calculation method in which a subset of the training data is completely withheld from model development and used only to validate the model after its creation.
KIA – Kappa Index of Agreement. A sample-size adjusted index ranging from –1 to 1, where –1 indicates the model performed worse than a randomized model and +1 indicates perfect agreement between the prediction and truth.
Lidar – Light Detection and Ranging. A remote sensing technology in which an active sensor emits laser pulses, measures the time elapsed before their return, and uses this information to derive topographic data.
MSI – Multispectral Imager. A Sentinel 2 satellite multispectral satellite sensor.
NDMI – Normalized Difference Moisture Index. A band ratio ranging from –1 to 1, where –1 represents how moisture and +1 represents high moisture content.
NDVI – Normalized Difference Vegetation Index. A band ratio ranging from –1 to 1, where –1 represents no vegetation cover and +1 represents high vegetation cover.
NDWI – Normalized Difference Water Index. A band ratio ranging from –1 to 1, where –1 represents no vegetation cover and +1 represents high vegetation cover.
OA – Overall accuracy. A general measure of model accuracy calculated by dividing the total number of correctly predicted validation samples by the total number of predicted samples.
OLI – Operational Land Imager. A Landsat satellite sensor that measures visible, near infrared, and short wave infrared wavelengths of light.

PA – Producer Accuracy, A measure of accuracy calculated by dividing the total number of correctly predicted plots per class by the total number of predicted plots per class.
PBI – Plot Based Inventory. A method of data collection in which species composition within a defined boundary is recorded.
PCA – Principal Component Analysis. A technique that extracts the unique data from spectral bands and compresses these data into a single raster.
S30- HLS imagery that aligns with the same band system as Sentinel 2 as opposed to L30, which uses the same bands as Landsat.
Spectral separability – The measure of how distinct each species spectral signatures are.
Spectral signatures – A combination of reflectance values across visible and infrared wavelengths of light that is unique to each species and can be used for its identification.
UA – User Accuracy. A measure of accuracy calculated by dividing the total number of correctly predicted samples per class by the total number of input validation samples per class.

7. References

Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data* (Vol. 964). US Government Printing Office. https://doi.org/10.3133/pp964

Dalponte, M., Bruzzone, L., & Gianelle, D. (2012). Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data. *Remote Sensing of Environment*, *123*, 258–270. https://doi.org/10.1016/j.rse.2012.03.013

Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., Straub, C., & Ghosh, A. (2016). Review of studies on tree species classification from remotely sensed data. In *Remote Sensing of Environment* (Vol. 186, pp. 64–87). Elsevier Inc. https://doi.org/10.1016/j.rse.2016.08.013

Foody, G. M. (2008). Harshness in image classification accuracy assessment. *International Journal of Remote Sensing*, *29*(11), 3137-3158. https://doi.org/10.1080/01431160701442120

Friedman, J. H. (2001). 999 Reitz Lecture Greedy Function Approximation: A Gradient Boosting Machine 1. In *The Annals of Statistics* (Vol. 29, Issue 5).

Grossmann, E., Ohmann, J., Kagan, J., May, H., & Gregory, M. (2010). *Mapping Ecological Systems with a Random Forest Model: Tradeoffs between Errors and Bias*.

Immitzer, M., Atzberger, C., & Koukal, T. (2012). Tree species classification with Random forest using very high spatial resolution 8-band WorldView-2 satellite data. *Remote Sensing*, *4*(9), 2661–2693. https://doi.org/10.3390/rs4092661

Immitzer, M., Neuwirth, M., Böck, S., Brenner, H., Vuolo, F., & Atzberger, C. (2019). Optimal input features for tree species classification in Central Europe based on multi-temporal Sentinel-2 data. *Remote Sensing*, *11*(22). https://doi.org/10.3390/rs11222599

Jin, S., & Sader, S. A. (2005). Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances. *Remote Sensing of Environment, 94*(3), 364–372. https://doi.org/10.1016/j.rse.2004.10.012

Kriegler, F., Malila, W., Nalepka, R., & Richardson, W. (1969). Preprocessing transformations and their effect on multispectral recognition. *Proceedings of the 6th International Symposium on Remote Sensing of Environment*. Ann Arbor, MI: University of Michigan, 97-131.

Lindenmayer, D. B., Margules, C. R., & Botkin, D. B. (2000). Indicators of biodiversity for ecologically sustainable forest management. In *Conservation Biology* (Vol. 14, Issue 4, pp. 941–950). https://doi.org/10.1046/j.1523-1739.2000.98533.x

Łos, H., Mendes, G. S., Cordeiro, D., Grosso, N., Costa, H., Benevides, P., & Caetano, M. (2021). Evaluation of XGBOOST and LGBM Performance in Tree Species Classification with Sentinel-2 DATA. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *2021-July*, 5803–5806. https://doi.org/10.1109/IGARSS47720.2021.9553031

Ma, M., Liu, J., Liu, M., Zeng, J., & Li, Y. (2021). Tree species classification based on Sentinel-2 imagery and random forest classifier in the eastern regions of the Qilian mountains. *Forests*, *12*(12). https://doi.org/10.3390/f12121736

McFeeters, S.K. (1995). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Remote Sensing Letters, 17*(7), 1425-1423. https://doi.org/10.1080/01431169608948714

Naidoo, L., Cho, M. A., Mathieu, R., & Asner, G. (2012). Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, *69*, 167–179. https://doi.org/10.1016/j.isprsjprs.2012.03.005

Persson, M., Lindberg, E., & Reese, H. (2018). Tree species classification with multi-temporal Sentinel-2 data. *Remote Sensing*, *10*(11). https://doi.org/10.3390/rs10111794

USDA Forest Service (2004). Land and Resource Management Plan - Superior National Forest. Eastern Region. Milwaukee, WI: USDA Forest Service.

Wang, M., Zheng, Y., Huang, C., Meng, R., Pang, Y., Jia, W., Zhou, J., Huang, Z., Fang, L., & Zhao, F. (2022). Assessing Landsat-8 and Sentinel-2 spectral-temporal features for mapping tree species of northern plantation forests in Heilongjiang Province, China. *Forest Ecosystems*, *9*. https://doi.org/10.1016/j.fecs.2022.100032

8. Appendices

Appendix A: *Specifications for Data used in the Model*

Table A1.
Specifications for data used in the model

Data Product	Data Source	Spatial Resolution	Temporal Resolution	Acquisition Date	Acquisition Method
3DEP Lidar data	National Elevation Data 3DEP Program	1 meter	project dependent	2018-2022	3DEP LidarExplorer
Harmonized Landsat and Sentinel-2 (HLS S30)	USGS and NASA Earth Observing Systems	30 meters	2-3 days	Spring 2021: Apr 29, May 1 Fall 2021: Sep 28, Oct 18 & 21 Spring 2024: Apr 25, May 8 & 13 Fall 2024: Sep 25, 27 & 30 Oct 25	Earthdata
Water Bodies	USGS National Hydrography Dataset	1:24,000 scale	-	2017-2020	TNM Download (v2.0)
Plot-Based Inventory data	Minnesota Department of Natural Resources	11.33 meter radius plots	-	2020-2022	Field Inventory, Fall Color Imagery

Appendix B: *Sample Size and User Accuracy*

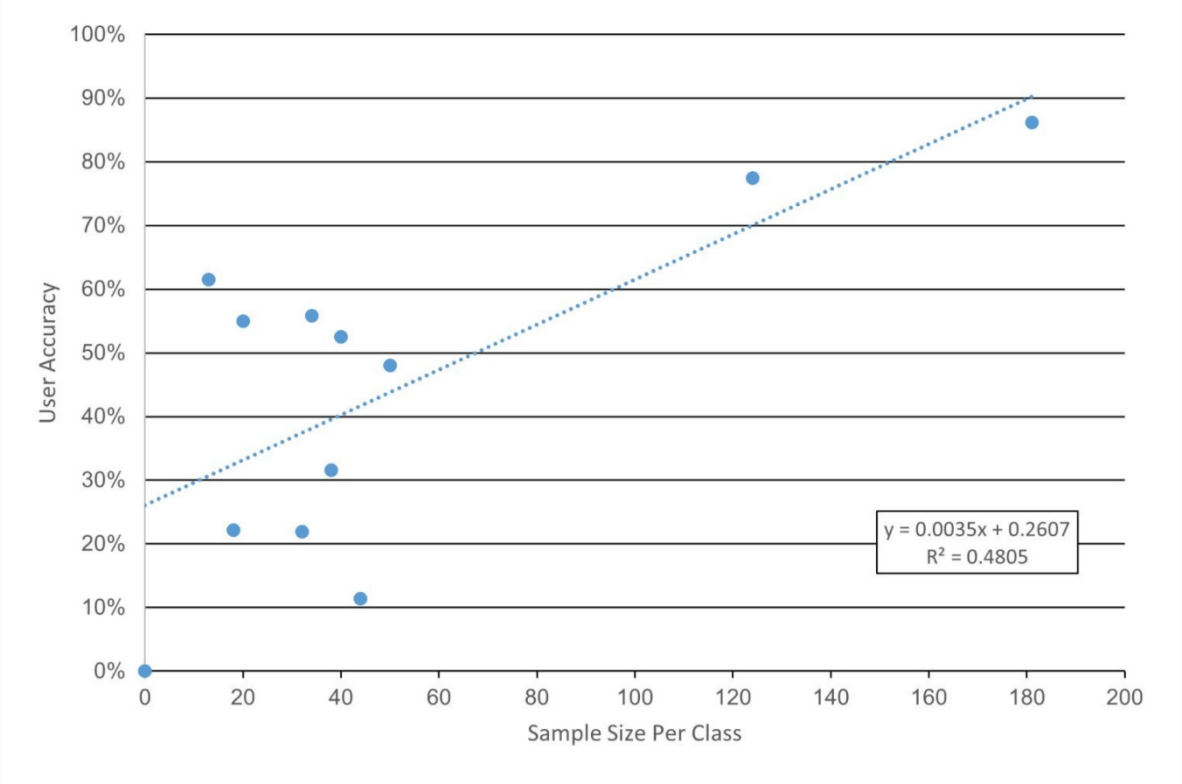

Figure B1. Relationship between sample size of cover type class in the validation data and user accuracy per class in the Fall/Spring Spectral + Topographic model.

Appendix C: *Spectral Separability*

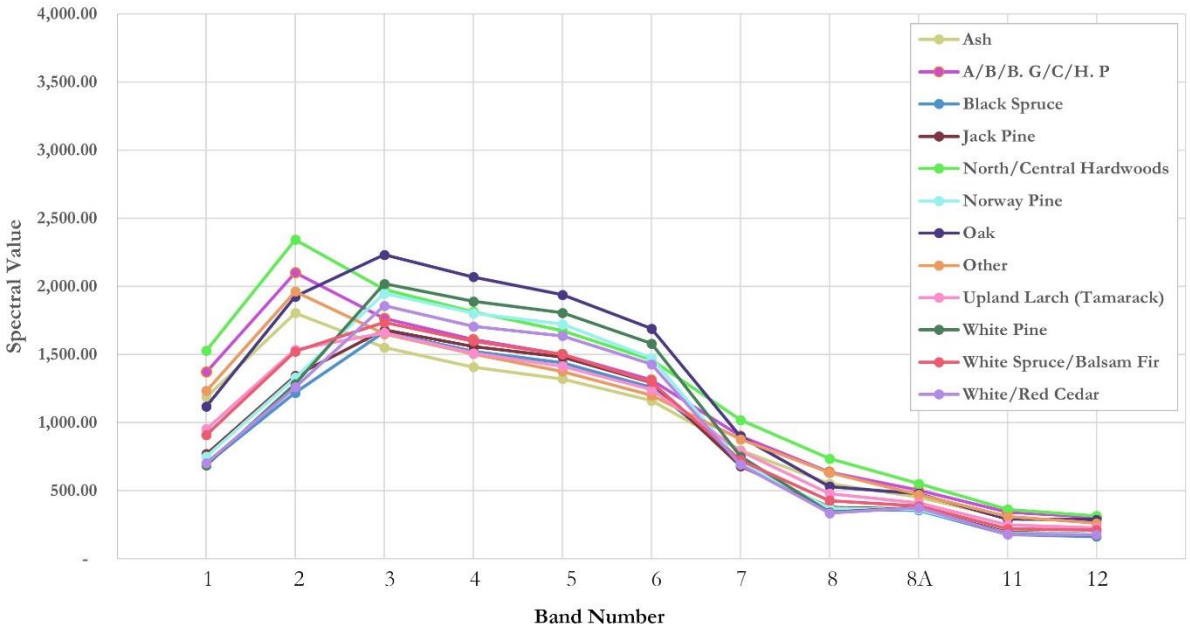

Figure C1. Mean spectral reflectance value for each cover type class across bands 1, 2, 3, 4, 5, 6, 7, 8, 8A, 11, and 12. A/B.G/C/H.P represents the Aspen/Birch/Balm of Gilead/Cottonwood/Hybrid Poplar group. The spectral values were rescaled by 10,000 according to the HLS S2 product convention.