FORESTRY
FUTURES
TRUST
ONTARIO

ForestNow Tree Species Composition Identification Results for the Romeo Malette and Ned Lake Forests

Project: 11B-2018 - Species Composition Determined from Satellite Images and Machine Learning

A project funded through the Knowledge Transfer and Tool Development program administered by the Forestry Futures Trust Committee

Global Surface Intelligence Ltd.
125 Princes Street, Edinburgh
Scotland, UK, EH2 4AD

Registration No. SC43901

https://www.surfaceintelligence.com

Gavin Tweedie
Global Surface Intelligence
125 Princes Street,
EDINBURGH
EH2 4AD
Scotland

August 14th, 2020

Ms. Shelley Vescio
Programs Coordinator
Forestry Futures Trust Committee
Suite 2003 - 1294 Balmoral Street
Thunder Bay, ON
P7B 5Z5

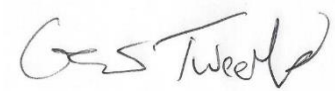
Dear Ms. Vescio,

We are pleased to submit the final report for the project 11B-2018 *"Species Composition Determined from Satellite Images and Machine Learning"*. Please find the enclosed document and final progress report.

In the near future, we will schedule some online workshops to present the results and we will send you an invite. As mentioned before, we are also interested in participating in any formal knowledge transfer sessions the Forestry Futures Trust may organize. Please keep us informed.

Please feel to reach out if you have any questions or comments.

Sincerely,



Gavin Tweedie

Executive Summary

This project has used the GSI Platform, ForestNow, to systematically identify the species composition in forests in Ontario and Michigan. ForestNow uses the power of high-performance computing and machine learning combined with actual ground observations/measurements and satellite data to provide an objective determination of tree species identification.

GSI is pleased to confirm that this project was highly successful with very strong results observed on multiple sites across a wide range of species. The results of this project prove that detection of tree species using machine learning is a viable method for enhanced forest inventory methods.

The tree attribute used in determining the species composition was proportional basal area. Several approaches were taken in our methodology where some were clearly better than others. The following report details the various approaches and results.

We have created an online geographic portal for ease of viewing without the need for specialty software knowledge (e.g. QGIS, ArcMap, etc.). Our portal offers users an easy and intuitive environment to view various prediction results and reference layers much like using Google Earth©.

GSI would welcome the opportunity from the OMNRF to apply the methodology developed here to large scale areas of Ontario's forest. GSI now has the capability of processing areas in excess of 15 million hectares in size in a matter of a couple of months resulting in extremely low cost per hectare.

GSI has also recently developed a very robust methodology for delineating stands of similar species composition in conjunction with other tree attributes such as tree size (DBH/height), total basal area, canopy closure, etc. This function can be tailored to individual clients' needs depending merging/splitting criteria.

Overall objectives

Improve the predictive capability of tree species composition at a forest operations' scale in an objective/systematic and automated process to enable both cost efficiencies and improve reliability of forest stand delineation through the FRI development process. This output can be used as stand-alone (in stand polygon format) or could be incorporated within the current process by aiding photo interpreters (in a raster format, i.e. heatmap) or work in conjunction with/or supplement new LiDAR initiatives for a potentially more robust forest inventory mapping.

Deliverables

All deliverables are in raster format which are pixel-based. Raster layers are an excellent format for displaying results and are flexible as it could allow forest inventory analysts to:

1. Calculate zonal statistics for existing forest inventory polygons by summing all pixels that fall within each polygon to produce a species composition typing label (e.g. 60% black spruce / 30% balsam fir / 10% white birch) similar to the outcomes from the traditional photo-interpretation method.
2. Auto-delineated polygons by aggregating pixels of similar species composition to create polygons. Since the process is automated, it allows to adjust delineation patterns based on the needs of the application. For example, in an operational application, there may not be the need to separate white and black spruce; however, it may be especially important wildlife habitat purposes. Raster layers allow for ultimate flexibility based on the end-use desired.

GSI has produced a raster layer for each of the species predicted where the sum of all species for each pixel sums to 100% (Figure 1 for example).

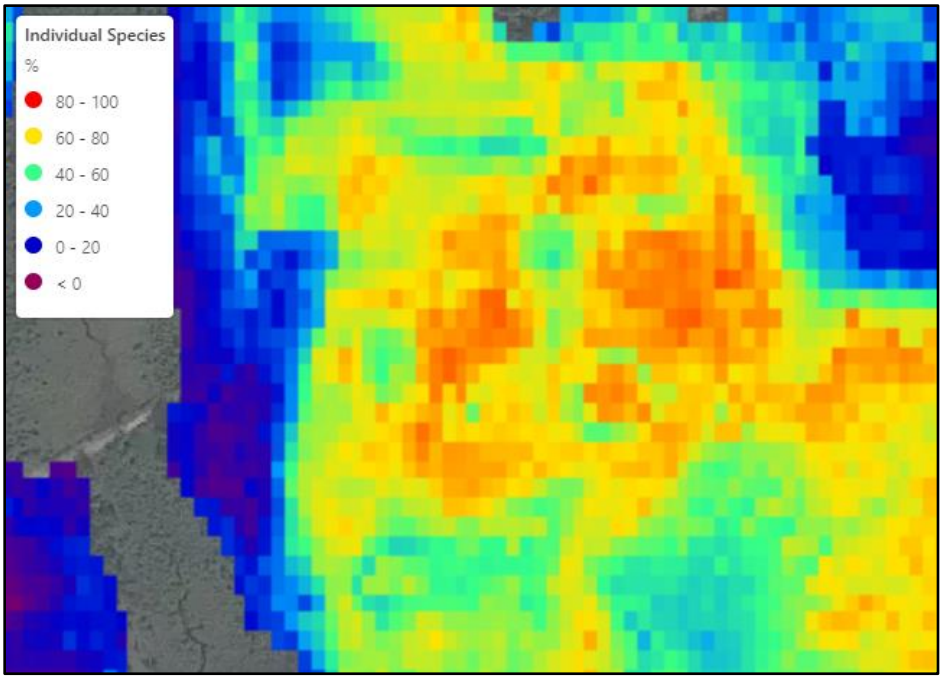


Figure 1. Sample image of a raster showing the percent composition of Jack pine.

For ease of viewing for general species distribution across each AOI, we also produced a “Leading Species” raster layer. This is a single raster layer in which each pixel is assigned a species based on the species with the highest percent (Figure 2).

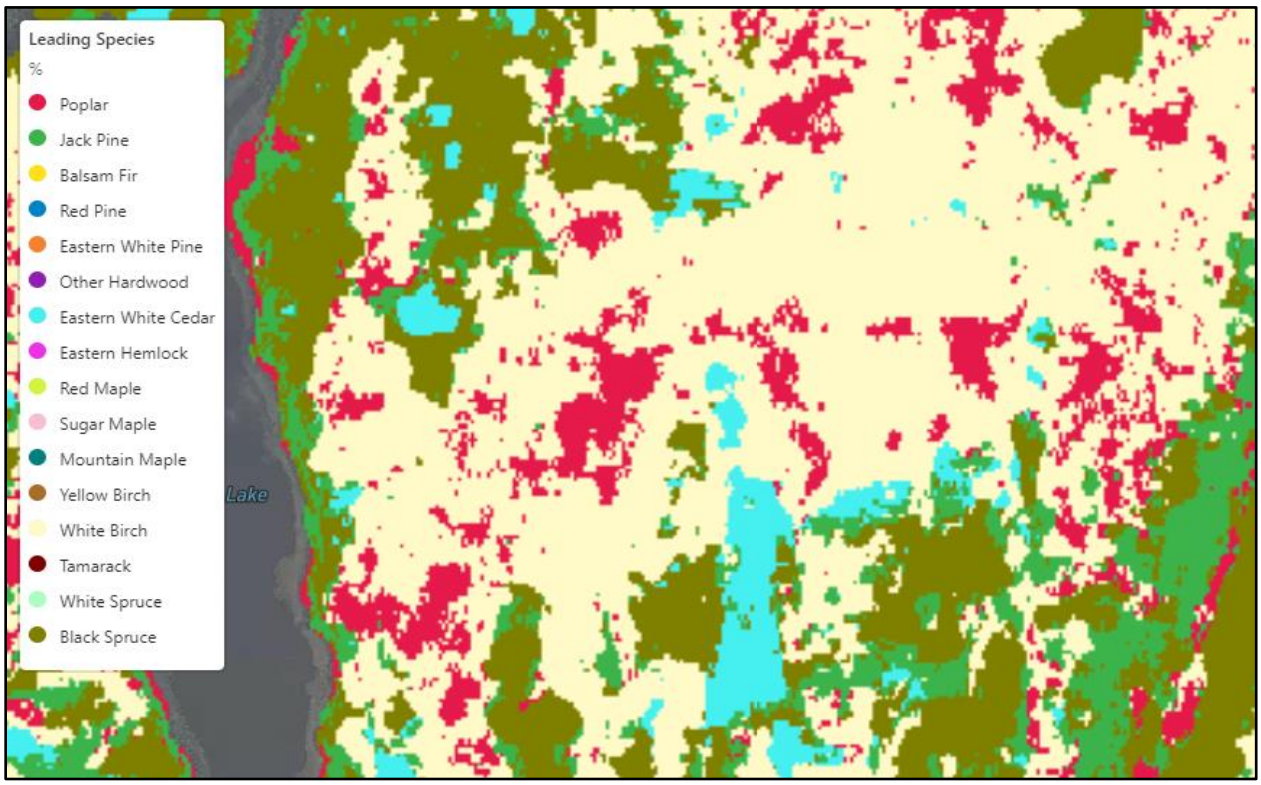


Figure 2. Sample of the Leading Species layer

Areas of Interest

Two Areas of interest (AOI) were used in this project for the purpose of covering multiple forest types; Boreal and Tolerant Hardwood (Figure 3).

Figure 3. Map showing the two areas of interest (AOIs), the Romeo Malette and Ned Lake Forests.

Romeo Malette Forest, Ontario

The Romeo Malette Forest (RMF) was chosen to represent the Boreal forest type as suggested by the OMNRF since there were several concurrent trials occurring on that forest with the acquisition of new single-photon LiDAR. Though the LiDAR was not used for this project, it was used on two other projects GSI led; therefore, it made sense to conduct all 3 projects on the same footprint.

Ned Lake, Michigan

One of GSI's clients had a project area in the northern Michigan area that had the desired species composition for a second AOI (tolerant hardwoods) and the OMNRF approved the use of this site as part of the project. The Ned Lake Forest (NLF) contains 237 high quality plots that were collected in 2018.

Target Audience and Benefits

All forest stakeholders can benefit from having a more accurate and flexible forest inventory. Species composition is no doubt part of this characterization.

- Ontario Ministry of Natural Resources and Forestry (OMNRF)
- Forest Industry Companies
- National Forest Inventory

Species composition of stands is especially important to managing and monitoring the forest resource from multiple interests:

- Timber management for effective sustainable supply for economic gain
- Wildlife and forest community management to maintain healthy populations
- Monitoring the effects of climate change on species presence and migration

In addition, accurate species composition can make for more effective and efficient timber extraction by better informing the operational planning process by:

- Better sequence the planning of harvest blocks based on the demand of specific species to processing facilities
- Better predict the best matched harvest system (i.e. prescription and machine type) to forest stands based on species composition
- Improve trucking efficiency by best matching the species mixture of a planned harvest block with mill destination timber specifications.

Model Data

Satellite Images

GSI used both reflectance (multi-spectral bands) and synthetic aperture radar (SAR) from various publicly available satellite images (e.g. Sentinel 1 and 2, Modis, Landsat, etc.). Reflectance data provides a much broader range of useful data; however, since it cannot penetrate through clouds, its frequency of clear usable scenes can be limited. GSI has some internal processes to reduce the impact caused by cloud cover to maximize the partial clear scenes. SAR on the other hand, can penetrate through cloud (therefore more frequently reliable); although, the data bands it provides do not allow for an as in-depth analysis compared to reflectance bands.

GSI continuously ingest images throughout the calendar year and as a result, it can detect the unique "phenology signature" by distinct species produced by changing seasons. This phenology is not exclusively programmed as part of the model; rather, it is the machine learning process that combines satellite observations with the training data and makes the correct association. These phenology observations could include:

- The presence of leaves or not on deciduous.
- The timing of new shoots in evergreens and/or bud break on deciduous.
- The color and timing of leaf fall in autumn.
- The color differences in twig bark color amongst deciduous species visible during leaf-off timing.

Ground Plots

GSI used ground plots provided by the OMNRF and a current client with last measurements from 2014 to 2018. Though some of these plots are fairly old, the fact that species composition is relatively stable, older plots can still be used; however, they are required to be screen for significant change since the last measurement. All plots were visually assessed against 2018 imagery where plots with clear signs of harvest or disturbance were screened out.

Natural disturbances are more difficult to assess visually; however, using mapped aerial survey disturbance data from 2015 to 2018, plots falling within these areas where not screened out but tagged

for identification. Plots identified as “disturbed” were accounted for in testing scenarios which is further explained in the Methodology section.

Methodology

Approaches

GSI has quantified the species composition at the plot level by assigning a percentage per species based on proportion of basal area for every single 10m pixel. GSI has developed a unique machine learning model that is based on multivariate output regression and which can force the sum of all predictions to equal 100%. This means that the machine learning process can model percentages without over or under prediction. The model also learns species grouping assemblages, where species that auto-correlate with others are proportionally predicted based on trained plot-level covariate patterns and quantities. This allows GSI to predict single species maps, where each separate map has an underlying group-level relationship. We then used this model with satellite data to develop a pixel-level individual species estimate.

This multivariate regression approach is more effective and flexible compared to the classification, another prediction method. Classification predicts based on a dominance approach so will assign each pixel a qualitative label representing the dominant species only or possibly a specific mixed condition, such as a spruce-fir combination. This method is effective in forest conditions where there is a low species diversity; however, is problematic in conditions of high species mixtures such as the two forests tested in this project where it could significantly under-represent minor species components. The other downfall classification analysis is that this method is only qualitative and not quantitative; therefore, some mathematical assumptions must be defined to compensate and develop a usable outcome for estimating species composition at a stand level.

Regression provides a more precise quantitative approach where it predicts continuous data, such as the proportion of all species across the entire composition mixture. By quantitative, it also means that any two pixels can have separate continuous values along mathematically comparable scale. Alternatively, classification analysis uses categorical data that represent forest qualities, but any two pixels do not exist along a mathematically comparable scale. This means GSI’s results can be validated on a plot-by-plot basis based using standard “goodness of fit” statistic tests such as R-squared. Classification analysis is limited to confusion matrices and binary diagnostic tests, such as sensitivity, specificity, positive predictive value, and negative predictive value.

Two main approaches were tested in this project where both approaches use ground plots: however, in a slightly different manner.

- Direct: This method uses the plot data to directly train on the presence of species based on basal area.
- Indirect: This method uses a synthetic set of data containing 20,000 data points based on the proportions found from the ground plots but independent from the plot locations.

Scenarios

Four elements were considered to design various scenarios where training on:

1. All trees within each plot vs canopy trees only

2. All plots vs removing ones below 20 m2/ha basal area
3. All plots vs removing ones older than 2007
4. All plots vs removing ones affected by a natural disturbance.

By testing all combinations, it resulted in 16 unique scenarios (2 x 2 x 2 x 2). This plot design was developed based on the plot availability for the RMF which was the first forest to be tested. Based on the result of this design and the difference in design for the NLF plots, not all scenarios where tested on the NLF. Plots on the NLF were all well stocked and very recent (2018); therefore, there was no reason for testing points #2 and #3. The NFL also did not have any disturbance data nor was there any known pest infestations which means point #4 was also not tested. Therefore, only 2 scenarios were tested on that site which were based on point #1.

Quantitative Validation

Validation of results is a crucial step to proving the accuracy of the results and the most important factor is the independence of the validation. With machine learning, it is important not to train and validate on the same ground plots. When the model trains with a ground plot, it tends to "over-fit" at that particular location; therefore, when validating on those over-fitted pixels, the results will typically show a very "high" accuracy. Those results though, are not a proper representation due to the lack of independence between the training and validation steps.

GSI believes in the true representation of the accuracy of the results; therefore, we employ methods best suited for the purpose of machine learning processing. The validation method differs depending on the two approaches described in the previous subsection.

Plot-Based Direct Approach

Since we have a relatively small number of plots to train with, reserving a significant subset of plots for validation only, would reduce the model's ability to accurately determine species composition. As a result, we use a K-Fold test which is common with machine learning processes where the test allows for using all the plots in the training while still validating against 100% of the same plots in an independent fashion. The test subsets the total dataset into different subsets where it trains on most plots and validates on the smaller reserved set. Then we run a k-fold analysis with plot data using GSI's developed automated procedures where we cycle through a different reserved validation subset until all plots have been validating against. For example, a 10-fold analysis would subset 100% of the plots into 10 separate 90:10% splits, where each 10% validation subset is unique. We then train on each 90% splits, cross-validate on each 10% split, and finally average accuracy across all 10 cross-validation sample. The result is a method that allows for using all plots in training and validation while preventing artificially over inflated accuracy from over-fitting. Over-fitting is analogous to writing a test with the answer sheet besides you where you would get excellent result on that particular test version but poor results when given an alternate test version with the previous answer sheet.

Plot-Based Indirect Approach

The indirect approach as explained earlier creates an independent training set that is spatially disconnected from the plot locations; therefore, the validation process is much simpler with this approach. As a result, this method cannot overfit the species composition prediction at the pixels associated with the plots directly; therefore, validation can then be performed on 100% of the plots without the need of the K-Fold approach for independence.

In both approaches, the R-squared values are derived comparing the individual species proportions at the plot level; therefore, the "goodness of fit" is determined for each species summarized across all plots.

One thing to note, some species will show a decent significance in one run and poor or non-significant on other runs. This type of contrast generally occurs for species where few plots contains that species and because of the nature of the K-fold test. When few plots contain a specific species, some of the K-fold runs will not have any plots containing that species in the reserved 10% of plots for the validation set which will create sporadic results. Therefore, species with a low presence are difficult to validate appropriately even though it is possible that the species could in fact be predicted reasonably well.

Visual Validation

Visual validation is another effective method for validation to confirm that the quantitative validation is effective and not giving results that are over-fitting at the plot location showing that results are consistent across the area. Two methods were examined with compelling support to the results.

Comparison to Plots

A simple overlay of plot locations overtop species prediction layer is an easy to compare. Figure 4 shows the comparison of the leading species per plot overtop the equivalent species prediction layer where the species color representation is the same in both for a quick visual comparison. It is important to note that this comparison is on the leading species only so it does ignore any minor species component that may be present; however, still effective at showing overall agreement.

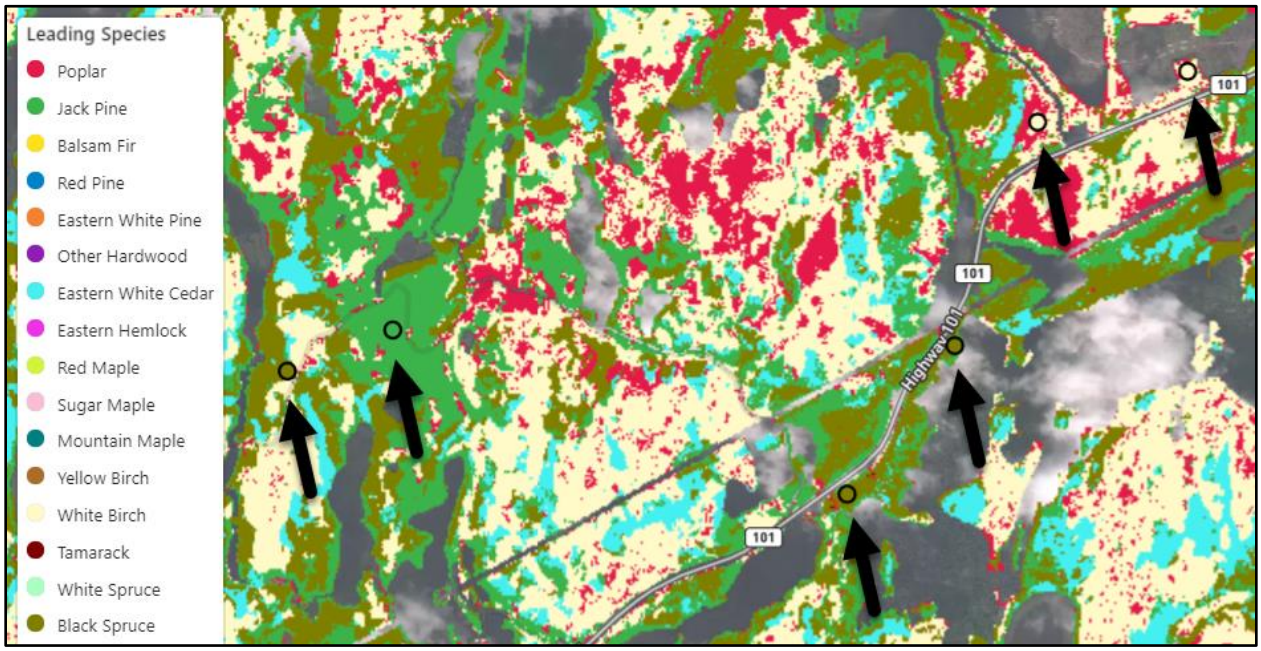


Figure 4. Overlaying of plots on top of predicted results for leading species

Comparison to Stand Delineation

Another effective visual comparison is to compare to the stand delineation from an up-to-date forest inventory for a fully independent visual. Figure 5 shows the leading species layer overlaid on top of the forest inventory stand boundaries provided by the OMNRF where there is a clear alignment between the two layers. It is important to note that while species is an important reason why stands are delineated, the size (e.g. diameter at breast height, height) and stocking (e.g. stem density in trees per hectare and/or crown closure) are other possible attributes for grouping or separating stands.

Both visual methods can be verified for the forests in the project specific portal described in the next section.

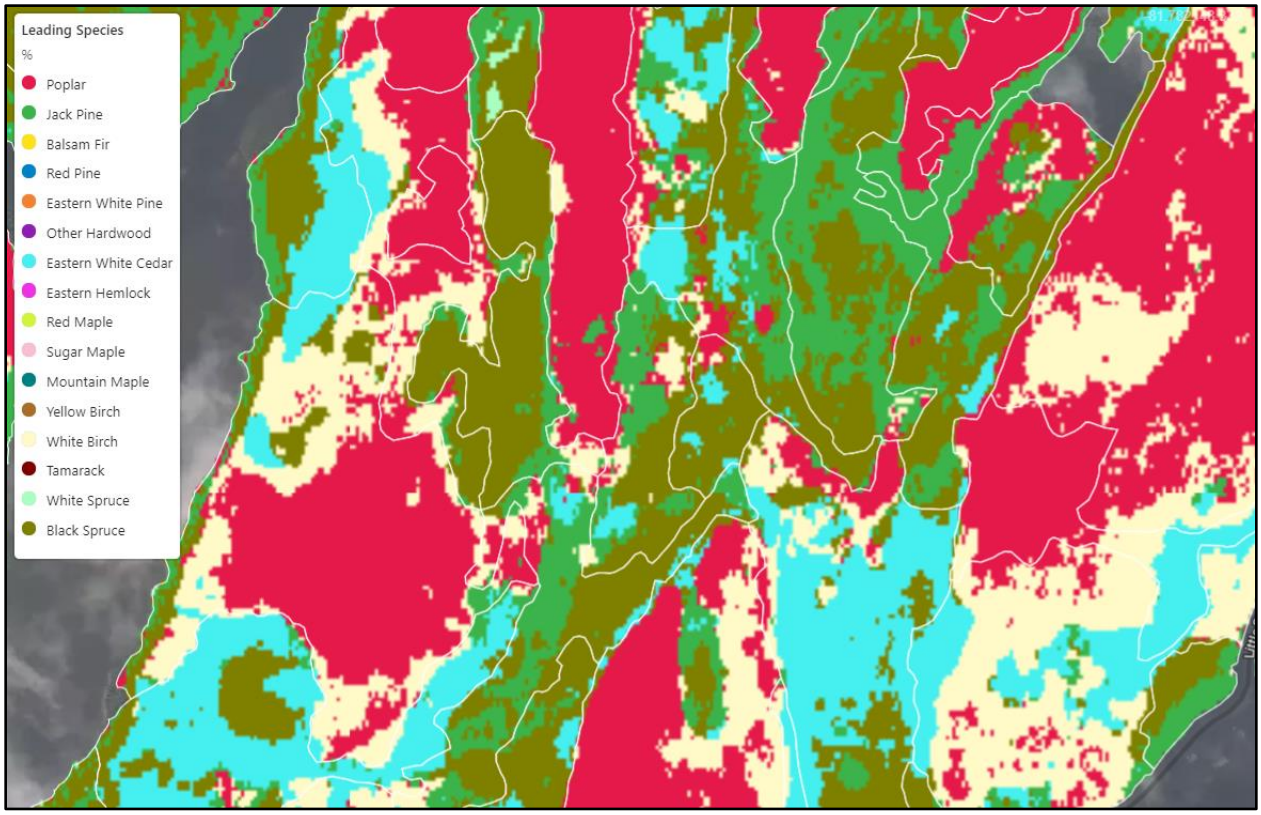


Figure 5. Overlaying of previous forest inventory stands on top of predicted results for leading species

Results

Portal

An online geographic has been setup for viewing some of the results. The portal format allows for easy viewing by anyone with no need for specialty software knowledge (e.g. QGIS, ArcMap, etc.). Our portal offers users an easy and intuitive environment to view various result and reference layers much like using Google Earth (refer to Figure 6 for example setup).

The layers presented are scenarios 5-8 for the RMF and both scenarios from the NFL. Each scenario contains a layer for each species composition and a single composite layer showing the leading species for each pixel as described in the Deliverables section.

We have also included the plot locations which are colored the same as the leading species layer for ease of comparing the resulting layers at the plot level.

When the *"Previous Forest Inventory Stands - RMF"* is turned on, it is possible to click on the individual stands to produce a composition label to compare against any of the species layers.

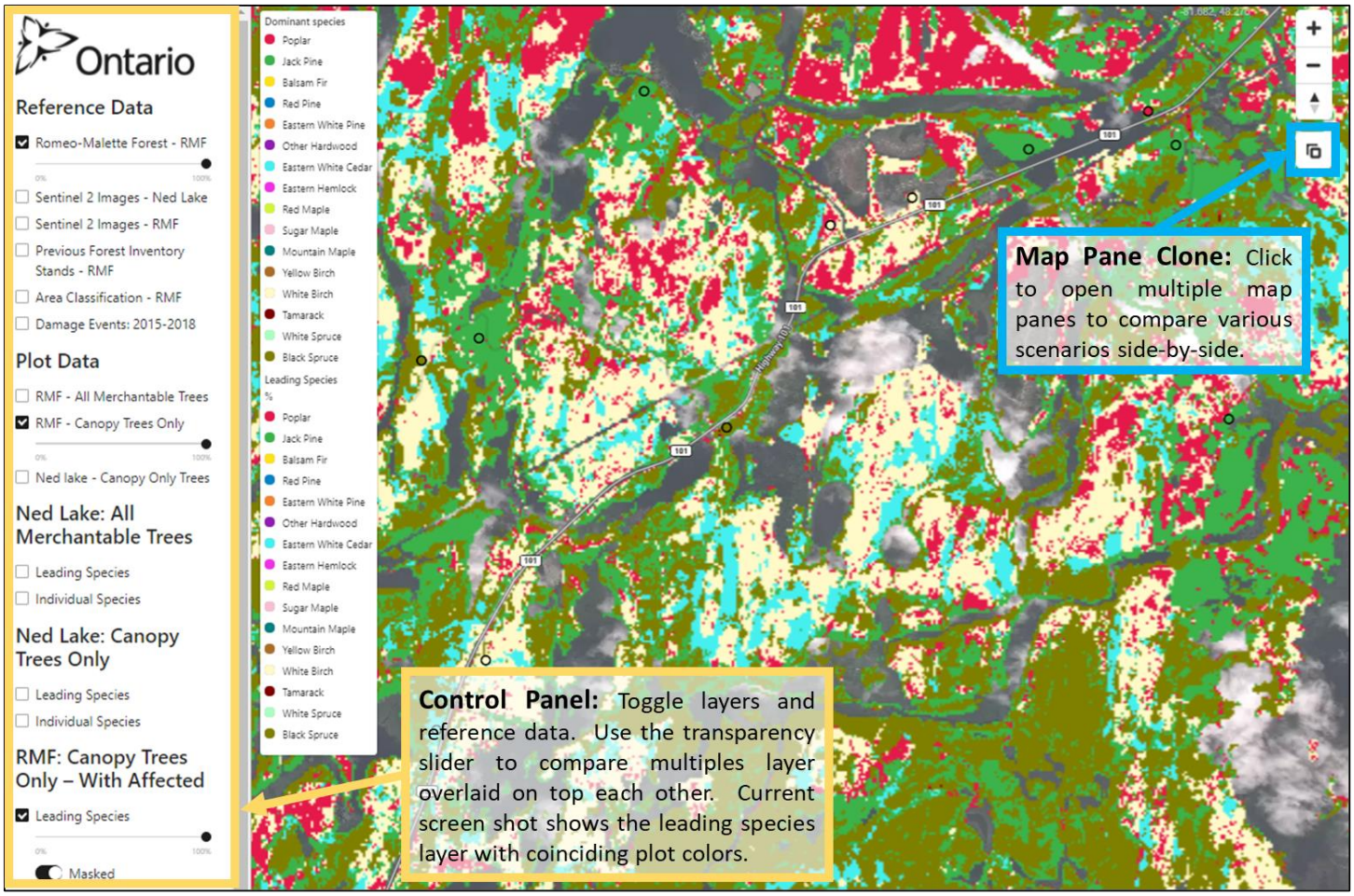


Figure 6. Screen capture of the GSI portal available for viewing.

The site can be accessed by contacting the company at the following link: https://www.surfaceintelligence.com/contact

Approaches

When comparing the direct and indirect approaches, there was a clear difference in significance where the direct approach was best by far. The indirect approach was only tested on the RMF; however, it was clear that when dealing with multiple species, the results were much poorer. As a result, only the direct approach was tested on the NLF.

Accuracy by Species

Results vary by scenario and by forest; however, the model results clearly show some impressive significance, especially on the RMF with many species showing R-squared values of +85% (Tables 1 and 2). The most significant results are seen in the RMF with nine species showing reasonable significance overall; though, the NLF also has some significant results on some select species.

Romeo Malette Forest

When comparing scenario runs with and without affected plots, there is an interesting trend that occurs when looking at the results of balsam fir where generally the runs without the affected plots seem to give better result. Since the majority of the affected plots where in the spruce budworm infested area and that insect's preferred host is balsam fir, it supports the possibility that the balsam fir in the infected area looks different or has been altered significantly since the last plot measurement. This notion was not explored further in this project since it is out of scope; however, Project #2 focusses on trying to predict areas of infection specifically.

In the comparison of scenario runs between all merchantable versus dominant trees only, generally the runs with dominant trees only, score best. This pattern makes sense since smaller understory trees are not visible in the satellite images; therefore, the accuracy of those trees would purely be by association with the canopy only. Another observation is that some comparisons show much lower significance on the dominant scenario in contrast to all merchantable trees. This erratic result is likely due to the nature of the K-fold test for species with few plots with that species presence as explained in the Validation section above. Modelling canopy trees only results in fewer trees per plot; therefore, it is possible for plots to lose the presence of certain species.

When assessing the comparison of scenarios filtering plots for minimum basal area and maximum plot age, the results are slightly better with a minimum basal area filter of 20 m2/ha filter, but no significant difference occurs when filtering based on age of the plots. The minimum basal area filter impact is likely the result of low basal area plots showing ground vegetation through the main canopy and the model might be getting confused with this understory vegetation.

Table 1. The R-squared results by species for the Romeo Malette Forest.

Series Name	A1 1st Series				A2 2nd Series				B1 1st Series				B2 2nd Series			
Plot Criteria	No plot age or min BA cut-off				No plot age cut-off and min 20 m2/ha BA				Plot age >= 2007 and no min BA cut-off				Plot age >= 2007 and min 20 m2/ha BA			
Tree Criteria	All Merchantable Trees		Dominant Trees Only		All Merchantable Trees		Dominant Trees Only		All Merchantable Trees		Dominant Trees Only		All Merchantable Trees		Dominant Trees Only	
Scenario Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Condition / Species	With Affected	Without Affected	With Affected	Without Affected	With Affected	Without Affected	With Affected	Without Affected	With Affected	Without Affected	With Affected	Without Affected	With Affected	Without Affected	With Affected	Without Affected
Balsam Fir	50%	65%	26%	56%	52%	52%	38%	59%	51%	44%	39%	43%	66%	44%	42%	56%
White Birch	72%	62%	68%	71%	77%	79%	74%	81%	83%	82%	82%	72%	64%	81%	80%	63%
White Cedar	80%	64%	68%	31%	70%	28%	72%	39%	53%	61%	80%	58%	63%	68%	73%	85%
Tamarack	90%	50%	90%	74%	97%	98%	80%	82%	98%	56%	98%	91%	97%	84%	91%	92%
Jack Pine	87%	86%	89%	84%	85%	90%	88%	87%	83%	90%	93%	82%	86%	84%	79%	82%
Trembling Aspen	41%	81%	61%	65%	66%	57%	57%	83%	34%	31%	32%	55%	61%	65%	65%	74%
Black Spruce	79%	79%	68%	76%	77%	80%	74%	79%	82%	80%	85%	72%	84%	70%	75%	72%
White Spruce	85%	68%	80%	86%	48%	52%	85%	97%	82%	76%	84%	82%	32%	87%	76%	95%
Yellow Birch	11%	41%	26%	N/S	69%	67%	N/S	N/S	56%	13%	33%	47%	N/S	N/S	N/S	10%
Black Ash	N/S	N/S	N/A	N/A	N/S	N/S	N/A	N/A	N/S	N/S	N/A	N/A	N/S	58%	N/A	N/A
Red Maple	N/S	N/S	N/S	N/S	38%	35%	N/A	N/S	N/S	N/S	N/S	N/S	17%	N/S	N/A	59%
Balsam Poplar	N/S	17%	41%	12%	N/S	N/S	14%	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S
Red Pine	N/S	16%	N/A	25%	N/A	N/S	N/A	N/S	13%	N/A	N/A	N/A	N/A	N/S	18%	N/S
White Pine	N/S	N/S	22%	21%	N/S	N/S	N/S	12%	N/S	16%	27%	N/S	N/S	N/S	N/S	N/S
Run Average	66%	66%	64%	68%	71%	67%	71%	76%	69%	59%	70%	67%	69%	73%	73%	70%

Please note: N/A means species was not present in the plots for a given scenario, N/S means species presence was not significant, and species highlighted in orange were species of lower significance overall. The runs highlighted in green (5-8) are displayed on the GSI portal for viewing.

Table 2. The R-squared results by species for the Ned Lake Forest.

Series Name	A1 1st Series	
Plot Criteria	No plot age or min BA cut-off	
Tree Criteria	All Merchantable Trees	Dominant Trees Only
Condition / Species	N/A	N/A
Sugar Maple	70%	65%
White Cedar	38%	45%
Balsam Fir	34%	16%
Black Spruce	29%	16%
Red Maple	13%	11%
Black Ash	10%	N/S
White Pine	N/S	12%
Trembling Aspen	25%	N/S
Eastern Hophornbean	1%	9%
White Spruce	5%	6%
White Birch	N/S	N/S
Tamarack	N/S	N/S
Yellow Birch	N/S	N/S
Balsam Poplar	N/S	N/S
Run Average	17%	12%

Please note: N/A means species was not present in the plots for a given scenario, N/S means species presence was not significant, and species highlighted in orange were species of lower significance overall

Ned Lake Forest

The results of NFL were more mixed compared to the RMF. Generally, results were of lower accuracy and individual species varied between the two scenarios on which performed better. The most common species (sugar maple) did perform relatively well. However, apart from a handful of species, most species tested occurred in relatively low proportions resulting in few plots with those species present. These low counts caused accuracy issues and lower than expected accuracy with the K-fold test as explained in the Validation section.

One highlight of the NFL results was the fact that the model can differentiate between sugar and red maple.

Despite the lower accuracy numbers, the client was still pleased with the results and chose GSI's species prediction as opposed to the landowner's most current species typing.

Conclusion

The results of this project prove that detection of tree species using machine learning is a viable method. Though the NFL results were more disappointing, GSI has since applied the same methodology in northern Maine where a similar species presence exist with much better results. The method was also used in southern and pacific US with very good results there as well.

GSI would welcome the opportunity from the OMNRF to apply the methodology developed here to large scale areas of Ontario's forest. GSI now has the capability of processing areas in excess of 15 million hectares in size in a matter of a couple of months resulting in extremely low cost per hectare.

GSI has also recently developed a very robust methodology for delineating stands of similar species composition in conjunction with other tree attributes such as tree size (DBH/height), total basal area, canopy closure, etc. This function can be tailored to individual clients' needs depending merging/splitting criteria.