QUANTIFYING ERROR IN VEGETATION MAPPING

A Thesis presented to the Faculty of the Graduate School at the University of Missouri-Columbia

---

In Partial Fulfillment of the Requirements for the Degree Master of Sciences

---

by
ERICA SERNA
Dr. Hong He, Dr. Daniel Dey and Dr. John Fresen, Thesis Supervisors
May 2011

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

QUANTIFYING ERROR IN VEGETATION MAPPING

Presented by Erica Serna,

a candidate for the degree of master of Forestry,

and hereby certify that, in their opinion it is worthy of acceptance.

Professor Hong He

Professor Daniel Dey

Professor John Fresen

Thank you Jesus Christ, Holy Spirit, and Father

ACKNOWLEDGEMENTS

First, I would like to thank my thesis committee for their input and efforts over the last years. Dr. Hong He worked many hours assisting with computer programs, thesis writing and process development. Dr. Daniel Dey was an asset as a skilled forester and also for the detailed comments made in both the project proposal and thesis. Dr. John Fresen provided encouragement and guidance far beyond expected. His involvement was crucial to developing me as a researcher and also providing statistical wisdom.

I would also like to thank the many people that made this project possible, like those in the USFS and NRCS who collected the data in previous years and make research at landscape scales possible. Also, thank you to Dr. Brice Hanberry who worked countless hours alongside me as a mentor.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

QUANTIFYING ERROR IN VEGETATION MAPPING

Erica Serna
Drs. Hong He, Daniel Dey and John Fresen, Thesis Advisers

Abstract

Understanding the current distribution and structure of forest vegetation is important for designing forest management plans and prioritizing restoration at landscape scales. This project provides information on Random Forest, a relatively new statistical package in the field of forestry, and patterns in mapping errors, a less explored field of study particularly in the forests of the Midwest United States. Vegetation maps can be made from classification and regression trees, such as Random Forest, by integrating environmental variables with vegetation information. An understanding of the accuracy of the maps is important because management plans and restoration efforts are more effective with accurate data. This study was done in forested regions in Minnesota with the purpose of 1) analyzing physiographic factors controlling tree species distribution; 2) mapping potential species distributions; 3) quantifying error in vegetation mapping; and 4) understanding map accuracy by evaluating minimum amounts of sample data necessary for reliable mapping. The results from Random Forest were found to be realistic ecologically and biologically. Also, tree species required records of 1-2 trees per 10,000 ha to produce accurate maps. Knowing the minimum amount of data points necessary for acceptable accuracy assists scientists mapping vegetation. This study demonstrates the effectiveness of Random Forest in vegetation mapping, which can be useful for future vegetation mapping.

Chapter 1

MAPPING TREE SPECIES DISTRIBUTIONS USING FOREST INVENTORY DATA AND ENVIRONMENTAL VARIABLES

Introduction

Understanding the current distribution and structure of forest vegetation is important for designing forest management plans and prioritizing restoration at landscape scales (Mladenoff, White, Pastor, & Crow, 1993). Humans have severely altered North American forests, including the boreal and northern hardwood forests in Minnesota, through decades of harvest, fire suppression, and other anthropogenic land uses (Friedman & Reich, 2005). Restoring these altered forests requires a landscape scale mapping assessment of current forest composition and structure.

Forest inventories provide the vegetation information to describe current conditions. A major source of forest inventory data nationally is the Forest Inventory and Analysis database administered by the USDA Forest Service (Data and Tools, 2010). The FIA database contains inventory plots distributed across the nation that have various parameters ranging from individual tree level information to the entire 0.4 hectare field plot (Forest Inventory and Analysis Program, 2008).

Researchers often use the FIA data in assessing and monitoring the nation's forest ecosystem health, though the point data has two fundamental limitations. First, FIA data are at a density of 1-2 plots/10 km^2. The distance between points is generally too far for a typical ecological restoration task that is often conducted at individual sites less than a square kilometer in size. Second, point data is limiting in depicting spatially continuous vegetation (He, et al., 2007). GIS and spatial inference are often needed to convert point data into spatially continuous data forms. With additional environmental variables, spatial inference can also be effective in interpolating data from coarse resolutions to fine resolutions by determining vegetation in locations that data were not recorded.

Besides forest field inventories, researchers use remote sensing as another common vegetation mapping method at the regional and landscape scale (Jensen, 2000). Remote sensing is effective in capturing vegetation distribution but has limitations such as the inability to detect understory species (He, Mladenoff, Radeloff, & Crow, 1998; Stenback & Congalton, 1990). This prevents an accurate description of the forest structure and species composition.

Scientists have developed statistical models that integrate environmental variables with forest inventory plots to derive vegetation maps at fine resolutions and over large areas (Cutler, et al., 2007). Linear models, such as logistic regression, are often used in vegetation mapping by determining linear combinations of predictor variables to classify the data (Cutler, et al., 2007). Linear models limit accuracy in

mapping by the removal of correlating variables, which hinders interactions between variables to be expressed (Miller, Turner, Smithwick, Dent, & Stanley, 2004).

Classification and regression trees, such as Random Forest (RF), are a nonparametric alternative to the often inadequate linear models in modeling ecological data (De'ath & Fabricius, 2000). Classification trees develop rules for partitioning observations into two or more classes using predictor variables (Cutler, et al., 2007). The partitioning creates branches resulting in a tree-like appearance, hence the name classification tree. Random Forest exceeds the use of other common classification and regression methods in ecology (Cutler, et al., 2007; Prasad, Iverson, & Liaw, 2006) because the classification accuracy is high, it does not overfit the data and is very stable to small perturbations in the data (Cutler, et al., 2007). Other advantages include the ability to input missing values and determine variable importance.

Objectives

The first objective of this project is to analyze physiographic factors controlling tree species distribution. The second objective is to predict and map potential distributions of individual tree species based on the identified relationships between tree species and physiographic factors.

Study Area

According to the United States Forest Service (USFS) National Hierarchical Framework of Ecological Units, the study area includes a section of the Laurentian Mixed Forest (LMF) province in Minnesota (MN) as shown in Appendix A (Avers, et al.,

1994). This landscape was selected because forests are the dominant vegetation type and because there are differences in common tree species that vary with the environment.

The LMF province covers 9.3 million hectares of the northeastern region of MN. Rolling to steep ridges, low bedrock knobs, peatlands and glacial lake plains dominate the landscape (Albert, 1995). The average elevation is between 200 and 400 meters above sea level. Annual precipitation ranges from 53 cm in the western portion of the region to 81 cm in the east. Vegetation is influenced by the low precipitation in the winter. The climate is characterized by short, mild summers with a mean temperature of 18°C and long, cold winters with a mean of -11°C. The soils are mainly alfisols, entisols and histisols (Albert, 1995). Conifer and hardwood-conifer forests dominate the province with species such as quaking aspen (*Populus tremuloides*), black spruce (*Picea mariana*) and balsam fir (*Abies balsamea*). Black spruce is most abundant according to FIA records (50%), followed by quaking aspen (13%), balsam fir (6%), and paper birch (*Betula papyrifera*) (5.7%).

Methods

Data Preparation

The Forest Service updates the FIA database containing tree data across the US every five to ten years. Forest Service employees collected the data for this project between 2000 and 2004. Field crews used fixed-radius plot layouts for sample tree selection in ground plots 0.4 ha in size. Across the landscape, the plot density is 1-2

plots/10 km^2. For more information on FIA sampling design, please refer to the FIA manual (Miles, et al., 2001). Due to privacy laws, the Forest Service fuzzed and swapped the exact plot locations (Lister, et al., 2005). All coordinates were fuzzed within 1.6 km of the exact plot location while most were within 0.8 km. Also, up to 20% of the private plot coordinates were swapped with similar private plots in the same county (Forest Inventory and Analysis Program, 2008). A Forest Service employee overlaid the accurate FIA data and prepared environmental data to produce accurate site conditions for tree species while maintaining privacy for private landowners.

The Natural Resources Conservation Service (NRCS) provides county level soil information available across the nation at mapping scales between 1:12,000 and 1:63,360 through the Soil Survey Geographic (SSURGO) Database. Limitations include error from digitizing the soil maps, which were hand drawn and unrepeatable. Also, the surveys are time consuming and done infrequently. The NRCS has not completed the SSRUGO database in the entire LMF province, so the study area is only a portion of the province, where the soil survey has been completed. Five environmental variables were selected from SSURGO to predict tree distributions: available water capacity, organic matter, pH, and percent sand and clay. These covariates were selected because (1) they influence the basic needs of vegetation (e.g. sunlight, water and nutrients), (2) they were available in the study area (NRCS), and (3) they are relatively stable over time, and it is better to include more variables than necessary because Random Forest hierarchically determines the importance of each variable. All SSURGO calculations were at the map unit level, which contain one to three soil components each. All

environmental variables, including the raster data calculated at a resolution between thirty and sixty meters, were generalized to the SSURGO map unit level for determining potential species distributions.

Digital elevation models (DEM) were used at thirty meter resolution to calculate the topographic position index (TPI), topographic convergence (wetness) index (TCI), terrain ruggedness, mean slope in percent, aspect (transformed) in degrees and elevation for each SSURGO polygon. A sixty meter DEM was used to calculate solar radiation index due to the computation limitation.

These twelve environmental variables from SSURGO and DEMs are chosen to reflect the three main types of influences on species: limiting factors, disturbances and resources (Guisan & Thuiller, 2005). Though no actual disturbance variable is used, like fire frequency, this influence is accounted for by other variables like terrain ruggedness since highly rugged terrains typically have less disturbance than non-rugged terrains (Guyette, Spetich, & Stambaugh, 2006). Also, locations with high water availability have little disturbance. Random Forest statistical package performs best with quantitative data so qualitative variables like anthropogenic and natural disturbances were not used.

The TPI and TCI tools were downloaded from the Environmental Systems Research Institute (ESRI) website in a toolbox called Topography Tools written by Thomas Dilts (ESRI SUPPORT CENTER). The TPI categorizes raster cells into topographic position (i.e. ridge top, valley bottom, mid-slope, etc.) by finding the difference between a cell elevation value and the average elevation of the neighborhood around that cell.

Positive values mean the cell is higher than its surroundings while negative values indicate the cell is lower. Significantly higher values than the surrounding neighborhood indicate the cell is likely at or near the top of a hill or ridge. Significantly lower values suggest the cell is at or near the bottom of a valley. Values near zero could mean either a flat area or a mid-slope area. The TCI quantifies topographic control on soil moisture by calculating the upslope contribution area in hectares in relation to slope in percent by the equation $TCI = \ln(a/\tan\beta)$ where a is the upslope contribution and β is the local slope angle. Locations with high topographic convergence accumulate water more than areas with low convergence.

The terrain ruggedness was calculated with a vector ruggedness measure (VRM) for use in GIS written by Mark Sappington (Sappington M. , 2008). The script is available through the ESRI website. The VRM utilizes the heterogeneity of slope and aspect in the dispersion of vectors in three dimensions (Sappington, Longshore, & Thompson, 2007). Values in the output raster range from 0 (no variation in the terrain) to 1 (complete variation). Natural terrains typically range between 0 and 0.4.

ESRI provided all other needed calculation tools within the GIS software ArcMap® version 9.3 (Environmental Systems Research Institute, 2008). Aspect was transformed using the equation $A = \sin(A_1 + 45) + 1$, where A is the transformed aspect and A_1 is the original aspect in azimuth degrees (Trimble & Weitzman, 1956). A value of zero is equivalent to southwest.

After calculating all the environmental variables from raster data, the results were joined to the SSURGO data and generalized to the map unit level. All twelve environmental variables were then at a resolution equal to the SSURGO map unit.

Similar species were combined into the following species groups: ash, elm, maple, populus, red oak, spruce and white oak (Appendix B). The ash group contains black ash (*Fraxinus nigra*) and green ash (*Fraxinus pennsylvanica*). The elm group contains American elm (*Ulmus americana*) and slippery elm (*Ulmus rubra*). The maple group comprises red maple (*Acer rubrum*), silver maple (*Acer saccharinum*) and sugar maple (*Acer saccharum*). The populus group comprises balsam poplar (*Populus balsam*), bigtooth aspen (*Populus grandidentata*) and quaking aspen (*Populus tremuloides*). The red oak group contains northern pin oak (*Quercus ellipsoidalis*) and northern red oak (*Quercus rubra*). The spruce group contains white spruce (*Picea glauca*) and black spruce (*Picea mariana*). The white oak group comprises white oak (*Quercus alba*) and bur oak (*Quercus macrocarpa*).

The species groups were partitioned into three spatial patterns, aggregately, sparsely and widely distributed species based on species habit and current MN distribution (Appendix B). The aggregated species include jack pine (*Pinus banksiana*), northern white cedar (*Thuja occidentalis*), tamarack (*Larix laricina*) and white pine (*Pinus strobus*). The sparsely distributed species include the elm group, the red oak group, red pine (*Pinus resinosa*) and the white oak group. The widely distributed species

include American basswood (*Tilia Americana*), the ash group, balsam fir, the maple group, paper birch, the populus group and the spruce group.

Spatial modeling

Random Forest in R® statistical software was used for determining presence probability of species distribution (R Development Core Team, 2009). Random Forest was selected for this study because of the very high classification accuracy, and its ability to determine variable importance while performing statistical analyses like regression and classification. Another important feature of Random Forest for this study is the algorithm for imputing missing values.

Random Forest classifies observations by recursive binary partitioning into regions that are increasingly homogeneous (Breiman, 2001). Each classification tree that is created has ending branches called nodes. The tree is fully grown when further subdivision no longer reduces the Gini index (Cutler, et al., 2007). A difference between Random Forest and other classification trees is that Random Forest trees are not pruned once fully grown (Breiman, 2001). Many trees are grown from the data set and the predictions from all trees are combined to produce more accurate classifications.

Random Forest grows multiple trees by drawing bootstrap samples with in-bag observations (typically 63% of the data points are selected to develop trees) and out-of-bag observations (the unused data points to calculate error and variable importance later). Bagging increases accuracy when randomizing methods are used (Breiman, 2001). Each bootstrap sample grows one tree and at each node only a small amount of

randomly selected variables are used to grow the individual branches. Often the number of variables is the square root of the total amount of variables used. The random and small amount of variables at each node ensures that correlations among the grown trees are small (Cutler, et al., 2007). The fully grown trees are used to predict the out-of-bag observations, performing a cross-validation of the results.

Fifteen maps, one for each species, were created based on the FIA tree data and prepared environmental variables using Random Forest. The FIA data provides tree species presence information and the environmental variables provide information on relevant site conditions which Random Forest uses to predict probability of tree presence. The Random Forest results were imported into ArcMap® to create the fifteen individual species group maps (Environmental Systems Research Institute, 2008).

The accuracy and error rates were computed for each observation using the out-of-bag predictions, and the results simplified by averaging all observations creating a kind of cross-validation of accuracy estimates without needing a set aside data set (Cutler, et al., 2007). Sixty three percent of the inputs were used for modeling and 37% for the validation. Cross-validation creates bias without a known extent but the out-of-bag estimates are unbiased (Breiman, 2001). In Random Forest, each tree is given a misclassification rate according to the out-of-bag observations. Then the values of the predictor variables are randomly altered for the out-of-bag data and pass through the tree to get new predictions. To determine the importance value of a variable, the

difference between misclassification rates (original and altered out-of-bag data) are divided by the standard error (Cutler, et al., 2007).

The results from Random Forest were joined to a Minnesota map in ArcMap to create presence probability maps. These maps display probability classes as follows: 0-0.25, 0.25-0.5, 0.5-0.75 and 0.75-1.

The map results were compared to the FIA known present locations using the receiver operating characteristic (ROC) package called ROCR in R statistical software to quantify the accuracy of each map. A curve is created that is a graphical plot of the sensitivity of the system when the value considered truly present from the Random Forest presence probability results is varied. This is done by plotting the true positive rate against the false positive rate. The true positive rate is the fraction of presences found in the FIA data and the Random Forest results out of the total presences from the FIA data. The false positive rate is the fraction of FIA locations without a species present that Random Forest predicted as present to the total absences from the FIA data. A true positive rate of 1 would imply that Random Forest perfectly classified each location where a species was present in the FIA data.

Results

Predicted potential species distribution

The widely distributed species that Random Forest predicted include American basswood, the ash group, balsam fir, the maple group, paper birch, the populus group and the spruce group (Figures 1-6). The most important variables for American basswood include topographic convergence index (TCI), terrain ruggedness, also known

as vector ruggedness measure (VRM), and clay. American basswood is most frequently found in a TCI around 5.16, VRM of $3x10^{-4}$ and percent clay around 12% (Figure 1). For the ash group, Random Forest found percent sand and aspect to be the most important variables. Ash is most frequently found in percent sand near 5% or 48% and a NW aspect (Figure 2). Ash is representative of a species with a wide distribution. This species group can be found throughout the study area, beyond the north-central area where there is a high density of high probability polygons. Solar radiation index, organic matter (OM) and elevation are most important for balsam fir. Balsam fir is most frequently found when SRI is near 5483 Watt hours per square meter (WH/m^2), OM is 7.2% and at elevations around 417 m (Figure 3). For the maple group, Random Forest found TCI and aspect to be the most important variables. Maples are typically found when TCI is around 5.0 and on NW aspects (Figure 4). The probability potential of paper birch occurrence is best predicted using TCI, solar radiation index (SRI), and elevation (Figure 5). Paper birch is most frequently found when TCI is near 5.1, SRI averages 5483 WH/m^2 and at elevations of 419 m. The most important variables for the populus group include TCI and SRI. Populus species are most frequently found when TCI is around 5.2 and SRI is 5487 WH/m^2 (Figure 6). The spruce group is best predicted according to Random Forest by OM and percent sand. The highest probabilities of occurrence for spruce occur when OM is near 10 percent and percent sand is 5 (Figure 7).

The aggregated species groups that Random Forest predicted include jack pine (Figure 8), northern white cedar (Figure 9), tamarack (Figure 10) and white pine (Figure 11). For jack pine, Random Forest found percent sand and clay to be the most important

variables. Jack pine is typically found in areas of 90% sand and 4% clay. Jack pine is a good representative of aggregated species because of the clumps of high probability polygons. Most polygons are aggregated as opposed to isolated. Percent OM and aspect are the most important predictor variables for northern white cedar. Northern white cedar is most frequently found when OM is near 9.5% and on NW aspects. The most important variables for tamarack include OM, slope and TCI. Tamarack is most frequently found in soils with 10% OM, on gentle (1.3%) slopes and where TCI is 6.5. White pine is best predicted according to Random Forest by TCI, sand and slope. White pine corresponds to a TCI value of 5.0, 71% to 93% sand and 3.7% slope.

The sparsely distributed species that Random Forest predicted include the elm group (Figure 12), the red oak group (Figure 13) and red pine (Figure 14). The pH and SRI are the most important predictor variables for the elm group. The elm group is most frequently found in a pH of 6.5 and an SRI near 5476 WH/m^2. For the red oak group, Random Forest found the VRM, TCI and slope to be the most important variables. The red oak group is typical of a VRM of 0.000401, TCI near 5.0 and a slope around 1.7%. The most important variables for red pine include sand and available water capacity (AWC). Red pine corresponds with soils with 87% sand and 0.032 cm/cm AWC and is typical of sparsely distributed species. The majority of the landscape has less than 50% probability of containing red pine and only a small portion with larger than 75% probability. Large, aggregated sections of the study area do not show red pine as potentially present.

Random Forest created a map of the white oak group distributed widely throughout the study area when it is a sparsely distributed species group (Figure 15). White oaks are generalists and are sprinkled throughout the range in varying densities, making it difficult for Random Forest to map. The most important variables for the white oak group include TCI and OM. The white oak group is most frequently found in a TCI near 5.18 and OM around 1.06%.

The northern white cedar had the highest true positive rate (TPR) at 0.96, which is near perfect classification (Table 2). The American basswood, jack pine, red pine and tamarack are also at or above 0.90. All of the aggregately distributed species, except for white pine, have the highest TPR values. Balsam fir and the spruce group have TPR values of 0.88. The ash, maple, red oak and white oak groups also have TPR values above 0.80. Paper birch, the populus group and white pine have TPR values in the 0.70 to 0.79 range and elm has the lowest TPR at 0.63. These TPR values are based on a 0.75 cutoff, where all Random Forest presence probabilities above 0.75 are considered present.

Table 1 Importance value results produced by RF for the ten most important variables for each species group. Random Forest assigned how important each variable was for predicting the presence probability for each species.

	Most Important									Least Important
basswood	TCI	VRM	Clay	Aspect	AWC	Sand	Slope	OM	TPI	Elevation
IV	1	0.821	0.806	0.789	0.738	0.734	0.726	0.702	0.673	0.665
ash	Sand	Aspect	Slope	TCI	Elevation	TPI	OM	Clay	AWC	SRI
IV	1	0.72	0.691	0.688	0.654	0.648	0.64	0.639	0.576	0.568
balsam fir	SRI	OM	Elevation	pH	TCI	VRM	Clay	Sand	Aspect	Slope
IV	1	0.818	0.815	0.802	0.785	0.704	0.699	0.684	0.664	0.636
maple	TCI	Aspect	Slope	VRM	Sand	Elevation	SRI	Clay	TPI	AWC
IV	1	0.711	0.688	0.63	0.617	0.613	0.522	0.482	0.465	0.463
paper birch	TCI	SRI	Elevation	Slope	VRM	Aspect	OM	Sand	pH	TPI
IV	1	0.693	0.68	0.647	0.633	0.535	0.517	0.478	0.449	0.437
populus	TCI	SRI	Elevation	Aspect	Clay	Sand	OM	AWC	TPI	VRM
IV	1	0.853	0.633	0.564	0.555	0.525	0.513	0.513	0.502	0.495
spruce	OM	Sand	Slope	Aspect	AWC	TCI	VRM	Clay	SRI	Elevation
IV	1	0.837	0.787	0.742	0.68	0.648	0.593	0.401	0.373	0.373
jack pine	Sand	Clay	AWC	SRI	Elevation	Aspect	pH	TPI	TCI	Slope
IV	1	0.605	0.478	0.431	0.412	0.396	0.301	0.3	0.247	0.186
cedar	OM	Aspect	Slope	TCI	VRM	Sand	AWC	TPI	Elevation	SRI
IV	1	0.885	0.615	0.564	0.517	0.482	0.461	0.427	0.426	0.411
tamarack	OM	Slope	TCI	Aspect	Sand	VRM	AWC	TPI	Clay	SRI
IV	1	0.795	0.743	0.696	0.487	0.461	0.42	0.385	0.36	0.282
white pine	TCI	Slope	Sand	Elevation	VRM	TPI	OM	AWC	SRI	Aspect
IV	1	0.63	0.625	0.578	0.577	0.57	0.531	0.51	0.474	0.454
elm	pH	SRI	Elevation	Clay	OM	AWC	Aspect	Slope	Sand	TCI
IV	1	0.943	0.863	0.861	0.836	0.785	0.784	0.784	0.754	0.722
red oak	TCI	VRM	Slope	TPI	Aspect	pH	Elevation	Sand	AWC	OM
IV	1	0.989	0.849	0.634	0.629	0.589	0.586	0.551	0.538	0.525
red pine	Sand	AWC	SRI	Clay	Elevation	TCI	OM	Aspect	TPI	pH
IV	1	0.823	0.781	0.734	0.69	0.672	0.573	0.538	0.519	0.513
white oak	TCI	OM	Elevation	SRI	Slope	pH	Aspect	AWC	TPI	Sand
IV	1	0.987	0.828	0.792	0.705	0.682	0.652	0.61	0.609	0.513

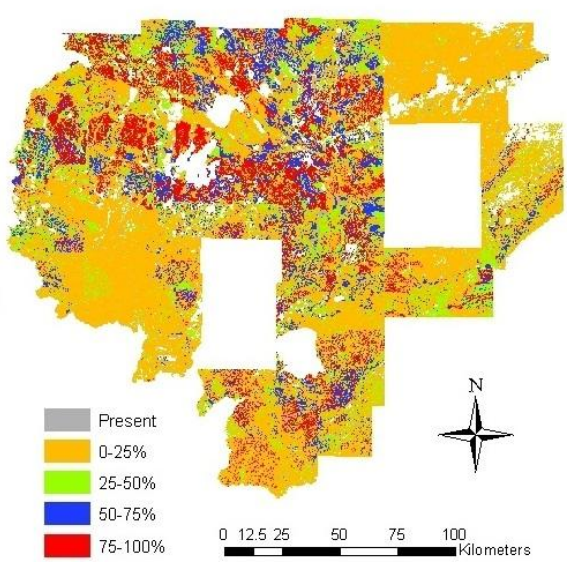


Figure 1 Probability classes for potential American basswood occurrence

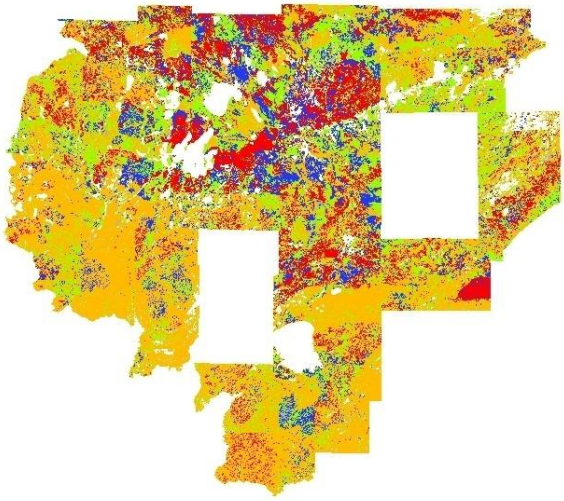


Figure 2 Probability classes for potential ash group occurrence

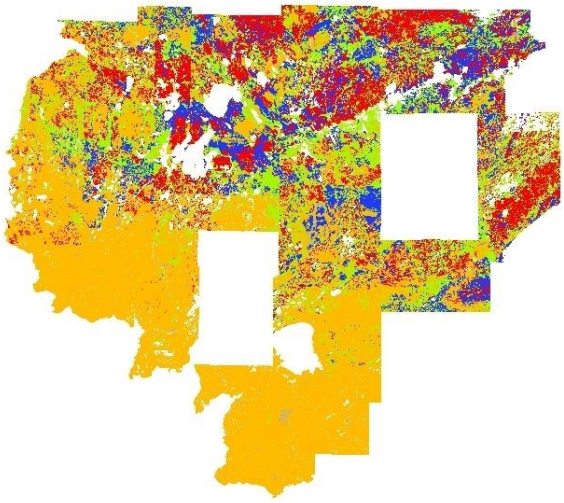


Figure 3 Probability classes for potential balsam fir occurrence

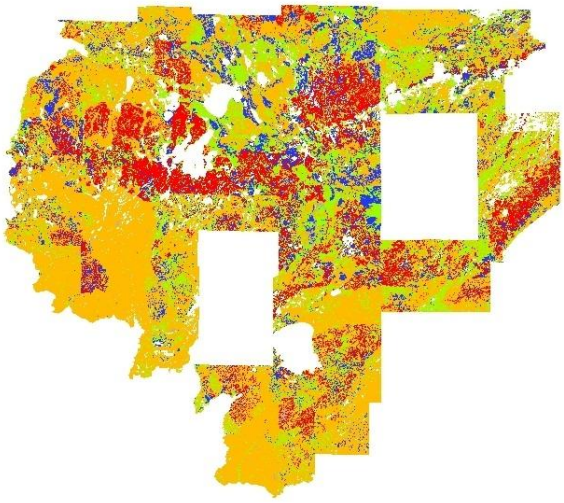


Figure 4 Probability classes for potential maple group occurrence

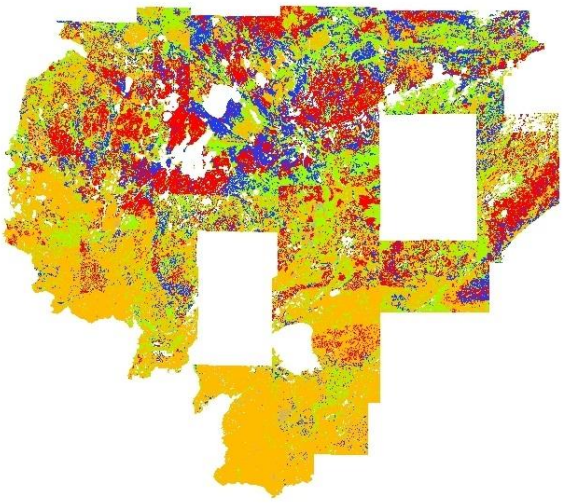


Figure 5 Probability classes for potential paper birch occurrence

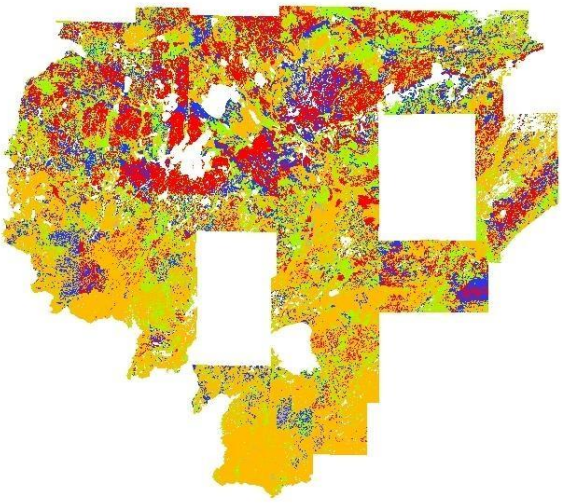


Figure 6 Probability classes for potential populus group occurrence

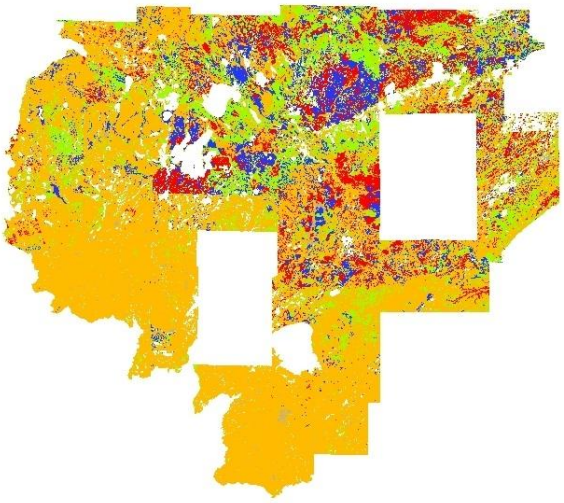


Figure 7 Probability classes for potential spruce group occurrence

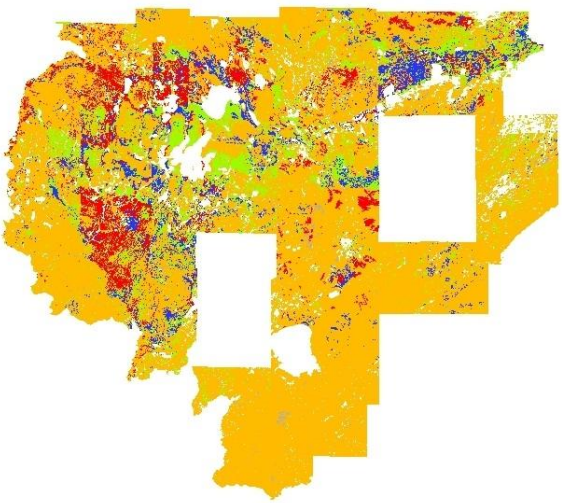


Figure 8 Probability classes for potential jack pine occurrence

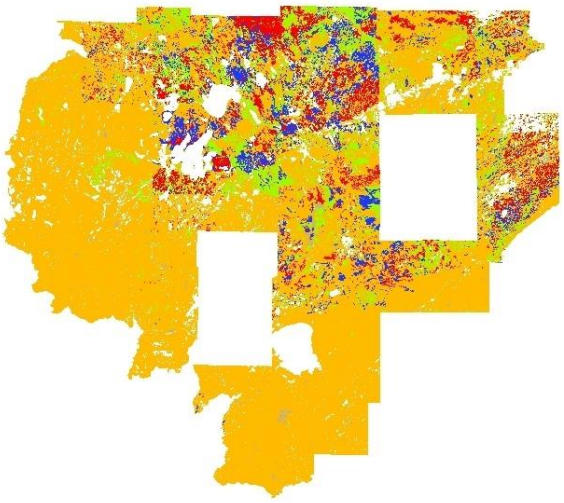


Figure 9 Probability classes for potential northern white cedar occurrence

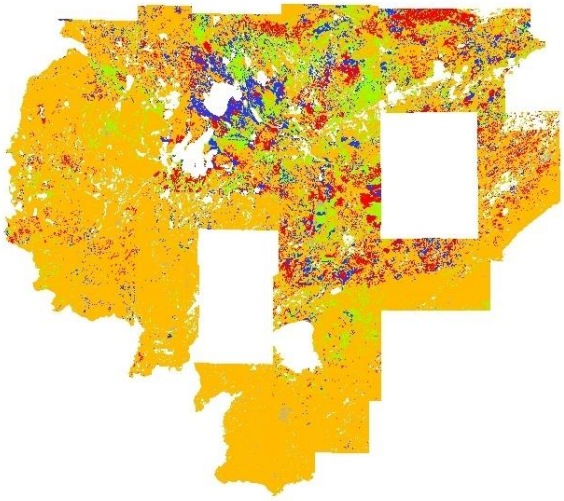


Figure 10 Probability classes for potential tamarack occurrence

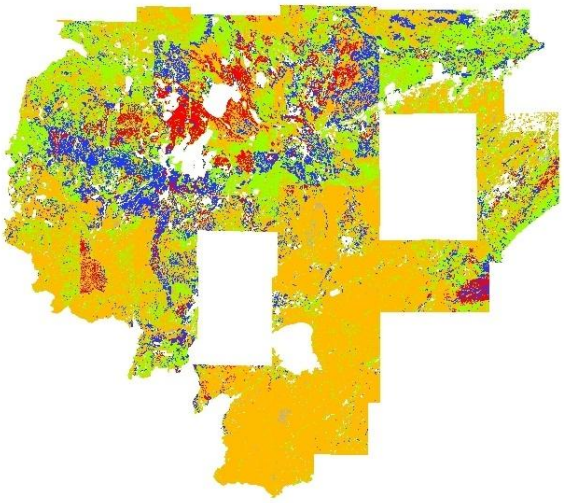


Figure 11 Probability classes for potential white pine occurrence

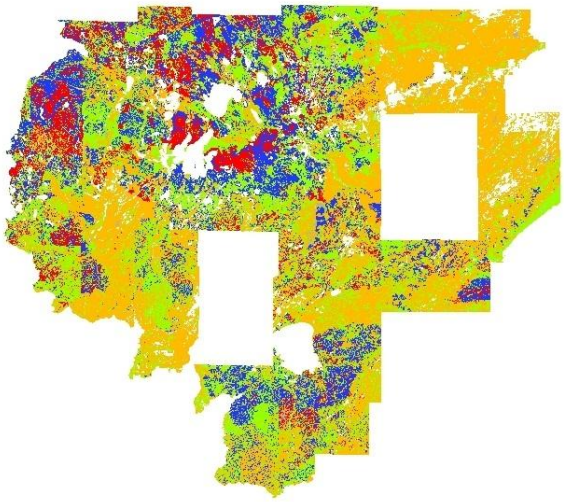


Figure 12 Probability classes for potential elm group occurrence

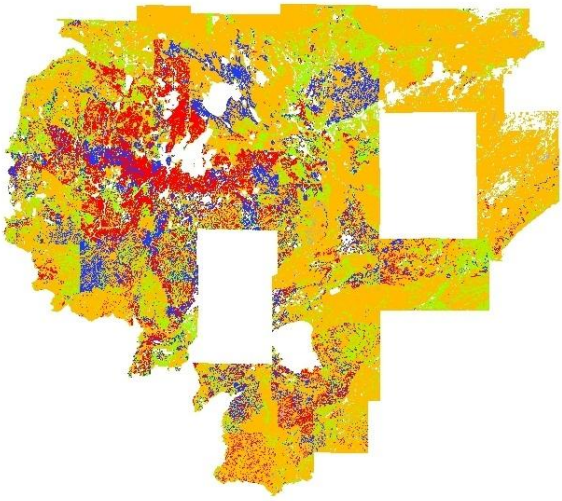


Figure 13 Probability classes for potential red oak group occurrence

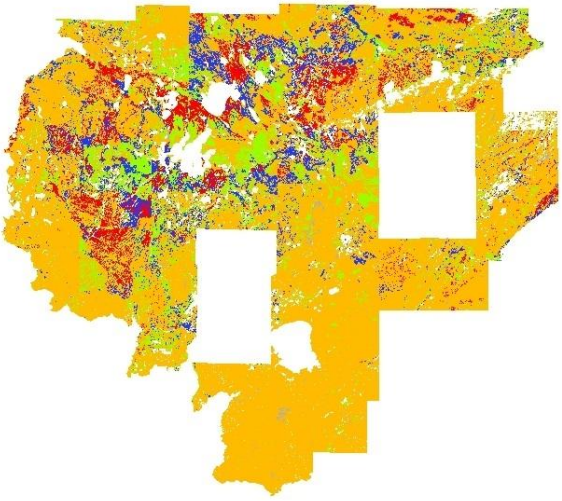


Figure 14 Probability classes for potential red pine occurrence

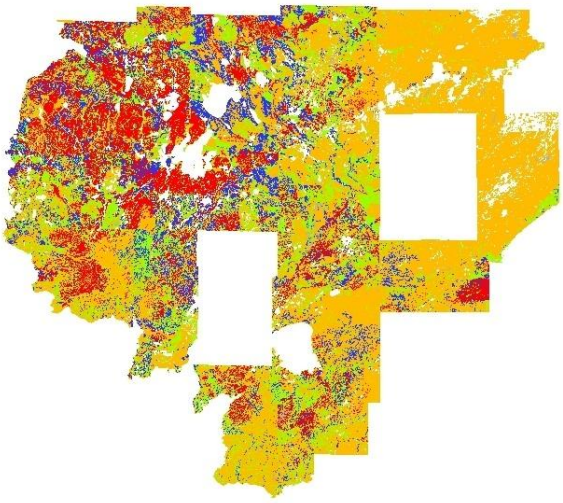


Figure 15 Probability classes for potential white oak group occurrence

Table 2 True positive rate for each species from the ROCR package in R statistical software.

Species	True Positive Rate
basswood	0.90
ash	0.82
balsam fir	0.88
maple	0.84
paper birch	0.78
populus	0.74
spruce	0.88
jack pine	0.93
cedar	0.96
tamarack	0.93
white pine	0.76
elm	0.63
red oak	0.83
red pine	0.93
white oak	0.83

Discussion

The results from Random Forest for widely distributed species (American basswood, ash, balsam fir, maple, paper birch, poplar, aspen and spruce) are supported by the species habitat characteristics. American basswoods are large, rapid growing trees found in deep moist soils (Fowells, 1965). They are shade tolerant. Random Forest predicted presence probability the highest in moist soils and low variations in terrain.

Black and green ashes grow in moist soil areas. They are shade intolerant, early successional species. Random Forest predicted the highest probability in sandy soils, that occur near streams (Burns & Honkala, 1990). There is also high probability of ash in sand at only 5%, typical of peats, bogs and lakes. Ashes are also predicted to grow on NW facing slopes, providing cool and moist soils.

Balsam firs are small to medium size trees that are very shade tolerant (Burns & Honkala, 1990; Bergeron, 2000). Soil moisture is an important environmental factor to the species, which explains Random Forest considering solar radiation the most important predictor variable. Higher incoming solar radiation dries out soils and lower amounts of solar radiation can provide an environment for cool and moist soils.

Maples grow on a wide variety of sites and are very shade tolerant (Allen, Molloy, Cooke, & Pendrel, 1999). Optimum growth is on soils that are not too wet and not too dry. They are not fire tolerant but grow well after disturbance. Random Forest predicted presence probability the highest in average to low TCI values compared to other species and slopes with a NW aspect, suggesting that shaded locations are more favorable for maples.

Paper birches are medium-sized, fast-growing trees found on almost any soil and topographic position, which are characteristics of species that have low mapping accuracy (Guisan, et al., 2007). They are shade intolerant and have thin bark (Bergeron, 2000; Burns & Honkala, 1990). Paper birch is predicted in similar locations to balsam fir but more wide spread.

Balsam poplar, bigtooth aspen and quaking aspen are early successional species found in floodplains. They are fast growing, shade intolerant and medium-sized trees (Huffman, Fajvan, & Wood, 1999). Random Forest predicted the highest probability in sunny and moist locations, like river bottoms.

Black and white spruces are found in a wide range of edaphic and climatic conditions. They are intermediate in tolerance to shade (Bergeron, 2000; Fowells, 1965). Spruces do not grow well in sandy soil which is supported by Random Forest by predicting spruces in about 5% sand locations.

The aggregated species (jack pine, northern white cedar, tamarack and white pine) are logically mapped by Random Forest according to previously known habitat characteristics of the tree species. The high true positive rate (TPR) values of the aggregately species, except for white pine, also demonstrate the accuracy of Random Forest predictions (Table 2). Jack pines are small to medium-sized, very shade intolerant trees that invade areas after major disturbances (Bergeron, 2000). They can grow on very dry, sandy soils that are inhospitable to other species but grow best on well drained loamy sands. Red pines often succeed jack pines on loamy sand soils, followed by white pine, then hardwoods such as sugar maple, basswood and northern red oak. Paper birch and quaking aspen frequently succeed jack pine instead of red and white pine, followed by the hardwoods or a spruce-fir association (Burns & Honkala, 1990). Random forest found percent sand and clay to be the most important variables and predicted highest probability of jack pines in 90% sand soils.

Northern white cedar is a shade tolerant and found in the northeastern part of MN (Bergeron, 2000). It grows most commonly on cool, moist, organic soils like near streams and drainage-ways (Burns & Honkala, 1990). Random Forest assigned OM and aspect as most important variables supporting previously known habitat characteristics

of northern white cedar. The preferable NW aspect provides the cool and moist environment on which the cedars thrive.

Tamaracks are small to medium-sized deciduous conifers that are very shade intolerant. They can tolerate a wide range of edaphic conditions but are most commonly found in moist, organic soils (Burns & Honkala, 1990). Tamaracks can grow in more extreme sites like peatlands. Random Forest correctly identified moisture and organic matter to be important variables with tamarack having higher probability of presence in organic, moist locations.

White pines are intermediate in shade tolerance and associated with somewhat excessively drained sandy deposits (Burns & Honkala, 1990). They can survive in soils with a pH value as low as 4.0 and are climax species in these extreme sites. Random Forest predicted highest probability of white pines on sandy, steep slopes. White pine has a low TPR value suggesting that the Random Forest model is not as strong for this species as it is for other species even though the previously known habitat characteristics of the tree species align with the Random Forest results.

The results from Random Forest for sparsely distributed species (elm, red oak and red pine) are supported by the species habitat characteristics. Elms are intermediate in shade tolerance and exist in a pH range of 5.5 to 8.0 (Fowells, 1965). In MN, elms are associated with plains, moraine hills, bottomlands and swamp margins. They can be found a variety of soils like well drained sands to poorly drained clays. Dutch elm disease has severely decreased the amount of American elms leaving a small

percentage of large diameter trees in mixed forest stands. Random Forest predicted elm at a mean pH of 6.5 and also found the average amount of solar radiation in MN suitable for elms. Even with the correlation between the species habitat characteristics and the Random Forest results, elm had the lowest TPR value (0.63) suggesting that the model used was not sufficient for mapping this species.

Northern red oak and northern pin oak grow on a wide range of soils and topography which proves more difficult to map than species that grow in narrow niches. Red oaks are intermediate in shade tolerance and often form pure stands. Random Forest considers lightly rough terrains on slopes with a TCI of 5.0 to be the most important environmental variables for mapping. Previous research determined aspect, slope position and shape to be important variables for red oaks (Burns & Honkala, 1990).

Red pines are medium-sized trees that are shade intolerant. They grow best on sandy soils adjacent to bodies of water. Due to changes in land management and timber harvesting, red pine populations declined drastically in previous decades (Leahy & Pregitzer, 2003). Random forest found percent sand to be the most important variable and predicted highest probability of red pines in 87% sand soils.

White oaks and bur oaks are very drought resistant and often dominate severe sites like thin soils, claypans and gravel ridges (Burns & Honkala, 1990). They are intermediate in shade tolerance and are often replaced by maple-basswood communities. Random Forest mapped the white oak group as widely distributed in the

study area when it is a sparsely distributed species group. One reason may be most of the land has a potential for white oaks according to the variables used to create the model. White oaks do not out-compete other species, like maples, well in colder climates like northern MN. A climate variable, like temperature, would be a good addition for predicting species distributions.

In summary, the predictions from Random Forest are realistic ecologically because the locations that were predicted to have high probability of presence align with the species biological and ecological attributes. Also, the important variables chosen by Random Forest align with the species known attributes.

A suggestion for future research is including weather variables to see if species, like white oak, are mapped more ecologically realistic. Another suggestion would be to use a different statistical approach, other than a CART, and adding disturbance variables like type of disturbance (fire, clearcut, shelterwood) and frequency of the disturbance.

Chapter 2

QUANTIFYING ERROR IN TREE SPECIES MAPPING

Introduction

Species maps are often used in management planning and restoration projects. Accurate estimates of species distributions are required to create effective management and restoration projects. The modeling of the species distribution is an important tool in the area of conservation (Guisan & Zimmermann, 2000).

Many statistical methods have been developed to compare created maps to data assumed to be true (Engler, Guisan, & Rechsteiner, 2004). The comparison can be done between forest inventory plots and a vegetation map, or between two vegetation maps. A comparison between point data and a vegetation map has limitations due to the lack of overlapping data. On the other hand, it is difficult to acquire an accurate vegetation map to which one can compare a second vegetation map.

To determine the accuracy of created maps, Kappa statistics have been used frequently in previous studies, though many have not differentiated between location and quantity errors (Pontius, 2000; Congalton, 1999). The Kappa statistic was found to be the most useful statistic for determining agreement between maps in previous research (Monserud & Leemans, 1992). Specific knowledge of the source of error,

location or quantity, can better assist the user in understanding map accuracy. Limitations to the Kappa statistic include an assumption that the data samples are independent, which is not always true in vegetation mapping since the created maps are often derived from the same data set (Leeuw, et al., 2006).

Other similar cross-tabulation methods based on locations throughout a map have been used (Foody, 2007). McNemar's test is based on a chi-square statistic. The null hypothesis states that the proportion of pixels will be correctly classified in both maps (Leeuw, et al., 2006). A limitation of McNemar's test is that the variable must be dichotomous. Explained deviance calculates the percentage of deviance but is limited to goodness of fit for generalized linear models (Finos, Brombin, & Salmaso, 2010). The receiver-operating characteristic plot method (ROC) creates a curve that is a graphical plot of the sensitivity of the system when the value considered truly present is varied. The ROC has the advantage of being threshold-independent but has limitations in models that do not have predictions across the entire study area (Peterson, Papes, & Soberon, 2008).

Mapping species with different spatial distribution patterns and life history characteristics can have diverse mapping accuracies (Guisan, et al., 2007). Species with high mapping accuracy include dominant, slow-growing species with narrow geographic distributions and that are good competitors (Guisan, et al., 2007). Also, species with contiguous distributions map more accurately (Reese, Wilson, Hoeting, & Flather, 2005). Conversely, early successional species that are rarely dominant and often scattered

among other species prove more difficult to map (Guisan, et al., 2007). Map accuracy can be affected by the spatial pattern of tree species distributions, such as aggregately, sparsely and widely distributed patterns. Understanding mapping accuracy is critical for the end user who will use the map for management plans.

The amount of data points used to create a vegetation map affects the accuracy of the map. Using larger sample sizes improves model performance (Reese, Wilson, Hoeting, & Flather, 2005). Guisan, et al. (2007) found that larger sample sizes produce more accurate results and that each species varies greatly in necessary data points, even within each tree genus.

Many studies have been done to determine the minimum amount of data required to create an accurate map, particularly for animals. In multiple studies, as low as 10 data points in a landscape scale have been found to be sufficient for mapping species (Hernandez, Graham, Master, & Albert, 2006; Wisz, et al., 2008). Hernandez, et al. (2006) studied insects, birds and mammals in California. Wisz, et al. (2008) studied a variety of animals and plant species around the globe. The results of the mobile and non-mobile species were comparable with as low as 10 data points necessary for accurate mapping. Kadmon, Farber, and Danin (2003) found 50 data points (a density of about 0.18 trees per 10,000 ha) to be sufficient for accurate mapping of woody vegetation in Israel.

This study aims to research the minimum data points necessary for mapping common tree species. Different maps were created in the study area for each tree

species by using varying amounts of tree data. These maps were then compared to determine the accuracy of each map. Comparing the similarity in location of categorical information and the quantity of locations with each category between two maps can be used to quantify map accuracy.

Objectives

The first objective of this project is to quantify error in vegetation mapping based on the common species distribution patterns, aggregately, widely and sparsely distributed. The second objective is to understand map accuracy by evaluating minimum amounts of sample data necessary for reliable mapping.

This project provides information on patterns in mapping errors, a less explored field of study particularly in the forests of the Midwest United States. Also, a fairly new technology in the forestry field was used, Random Forest statistical package, to gain more accurate results than in previous research. Random Forest is a nonparametric alternative to the often inadequate linear models in modeling ecological data (De'ath & Fabricius, 2000). Linear models, such as logistic regression, are often used in vegetation mapping by determining linear combinations of predictor variables to classify the data (Cutler, et al., 2007). Linear models limit accuracy in mapping by the removal of correlating variables, which hinders interactions between variables to be expressed (Miller, et al., 2004).

Random Forest exceeds the use of other common classification and regression methods in ecology (Cutler, et al., 2007; Prasad, Iverson, & Liaw, 2006) because the

classification accuracy is high, it does not overfit the data and is very stable to small perturbations in the data (Cutler, et al., 2007). More accurate maps and an understanding of the limitations of a map can aid in designing forest management plans and in prioritizing large scale restoration.

The tree species in the study area have diverse distributions and spatial patterns. The differences in mapping accuracy were determined between three spatial patterns in tree species distributions, aggregately, sparsely and widely distributed. Based on a set of unique site characteristics, presence probability maps in the Laurentian Mixed Forest (LMF) province in Minnesota were produced. Various amounts of data points were used for each species to generate a set of maps and compared individual map predictions among all maps for that species using kappa statistics. Knowing the minimum amount of data points necessary for acceptable accuracy will help scientists mapping vegetation. Separating the species into spatial patterns provides a more specific number of points tailored to individual species.

Study Area

According to the United States Forest Service (USFS) National Hierarchical Framework of Ecological Units, the study area includes a section of the Laurentian Mixed Forest (LMF) province in Minnesota (MN) as shown in Appendix A (Avers, et al., 1994). This landscape was selected because forests are the dominant vegetation type and because there are differences in common tree species in the environment.

The LMF province covers 9.3 million ha of the northeastern region of MN. Rolling to steep ridges, low bedrock knobs, peatlands and glacial lake plains dominate the landscape (Albert, 1995). The average elevation is between 200 and 400 meters above sea level. Annual precipitation ranges from 53 cm in the west to 81 cm in the east of the province. Vegetation is influenced by the low precipitation in the winter. The climate is characterized by short, mild summers with a mean temperature of 18°C and long, cold winters with a mean of -11°C. The soils are mainly alfisols, entisols and histisols (Albert, 1995). Conifer and hardwood-conifer forests dominate the province with species such as quaking aspen (*Populus tremuloides*), black spruce (*Picea mariana*) and balsam fir (*Abies balsamea*). Black spruce grows most abundantly according to FIA records (50%), followed by quaking aspen (13%), balsam fir (6%), and paper birch (*Betula papyrifera*) (5.7%).

Methods

Chapter 1 discusses the data preparation and creation of maps using the maximum amount of data points available in the study area. After creating a map for each species using the maximum amount of data points for the model in Random Forest, maps were produced with successively less data points. Seventy five percent of the total data points were used, then fifty percent, followed by twenty five percent. Also, two more sets of maps were created using different amounts of data points to find the minimum amount necessary to create accurate maps for each species by predicting the minimum based on results from the 75%, 50% and 25% trials. Amounts used can be found in Appendix C.

Random Forest produces a table including species presence probability for each SSURGO polygon. In ArcMap the table that Random Forest produced based on the maximum data points available was joined to the MN SSURGO shapefile. Then the 75% data point prediction table was also joined to the shapefile. Each polygon in the shapefile has two probability categories, one for the maximum data points and one for the 75%. These values were compared to determine the mapping accuracy. This process was repeated for the various trials of different data point amounts.

An error matrix was constructed to compare the maps of the same species with different sample amounts. From the matrix, Kno, Klocation and Khisto values were calculated to quantify error. Kappa is the amount of agreement between two maps after chance agreement has been removed ranging from -1, major difference, to 1, perfect similarity. Kno shows the proportion correctly classified in relation to the expected proportion classified correctly by a model with no ability to specify the quantity or location accurately (Pontius, 2000). Klocation and Khisto demonstrate the source of error, whether location or quantity. Klocation indicates the similarity in spatial distribution of categories but does not differentiate between categories that are close and categories that are distant and is independent of the total number of cells per category (Pontius, 2000). Values are similar to kappa except Klocation can go far below -1. Khisto is a measure of quantitative similarity between two maps (Hagen, 2002) and multiplying Khisto by Klocation yields kappa (Prasad, Iverson, & Liaw, 2006).

Kappa values between 0.81 and 1 represent almost perfect agreement, 0.61 to 0.8 represent substantial agreement, 0.41 to 0.6 represent moderate agreement, 0.21 to 0.4 represent fair agreement, 0 to 0.2 represent slight agreement and less than 0 represent poor agreement (Landis & Koch, 1977). Accurate maps are defined as having a K statistic above 0.61.

Another quantification error statistic, Kquantity (Pontius, 2000), is dependent on Klocation, defeating the purpose of separating quantity error from location error (Sousa, Caeiro, & Painho, 2002) and is not used in this study. Klocation and Khisto demonstrate the sources of error showing scientists where improvement is most needed (Pontius, 2000).

Each map is homogenous in species but produced from different data points and amounts. Species groups were partitioned into three spatial patterns, aggregately, sparsely and widely distributed species based on species habit and current MN distribution for further analysis (Appendix B). The aggregated species include jack pine, northern white cedar, tamarack and white pine. The sparsely distributed species include the elm group, the red oak group, red pine and the white oak group. The widely distributed species include American basswood, the ash group, balsam fir, the maple group, paper birch, the populus group and the spruce group.

Results

Comparison of prediction agreement

According to Kno, the aggregated species are mapped more accurately than the widely distributed species followed by the sparsely distributed species (Figure 16). The Kno displays four distinguished groups of accuracy, groups A_{kno}, B_{kno}, C_{kno} and D_{kno} (Figure 17). Group A represents the most accurate mapping and group D represents the least accurate. Group A_{kno} consists of only aggregated species, group B_{kno} and C_{kno} contain aggregately, sparsely and widely, and group D_{kno} contains only sparsely. The aggregated species produce maps with substantial agreement even with only 25% of the data points used in building the model. No other species attained this level of accuracy. Most species have a large drop in mapping accuracy after the data points used is below 50%, most likely because below 50% results in about 600 or less data points or 1.23 trees per 10,000 ha in the study area.

Klocation produced similar results. The aggregated species are mapped more accurately than the widely distributed species, followed by the sparsely distributed species (Figure 18). Group A_{kloc} comprises aggregately and widely distributed species, group B_{kloc} sparsely and widely, group C_{kloc} aggregately, sparsely and widely, and group D_{kloc} only sparsely distributed species (Table 3).

Khisto also shows that aggregated species are mapped more accurately than the widely distributed species, followed by the sparsely distributed species (Figure 19). Group A_{kloc} has only aggregated species, group B_{kloc} aggregated and widely, group C_{kloc} aggregated, sparsely and widely, and group D_{kloc} only sparsely distributed species (Table

4). Tamarack is found in group A for all three statistics while the elm group is found in group D for all three statistics (Figure 17, Table 3, and Table 4).

Sparsely distributed species require the fewest data points to produce accurate location maps. The widely distributed species require the most amount of data points, while aggregated species fall in the middle (Figure 20). Of the widely distributed species, the ash group and the populus group require 900 points or 1.84 trees per 10,000 ha, paper birch and the spruce group require 800 points or 1.63 trees per 10,000 ha, balsam fir and the maple group require 600 points or 1.23 trees per 10,000 ha and American basswood requires 400 points or 0.82 trees per 10,000 ha to map location accurately. Of the aggregated species, northern white cedar and tamarack require 700 points or 1.43 trees per 10,000 ha, jack pine requires 300 points or 0.61 trees per 10,000 ha and white pine requires 100 points or 0.20 trees per 10,000 ha. Of the sparsely distributed species, red pine requires 600 points or 1.23 trees per 10,000 ha and the elm group, the red oak group and the white oak group require 400 points or 0.82 trees per 10,000 ha to map accurately.

Random Forest mapped quantity better than location for all species and required fewer data points to accurately map quantity. As with the location maps, sparsely distributed species require the fewest points to produce accurate quantity maps. Aggregated species require the second fewest while widely distributed species require the most data points (Figure 21). The widely distributed species need between 400 and 700 points or 0.82 and 1.43 trees per 10,000 ha. The ash group, paper birch and the

populus group require 700 points or 1.43 trees per 10,000 ha, balsam fir and the maple group require 500 points or 1.02 trees per 10,000 ha and American basswood and the spruce group require 400 points or 0.82 trees per 10,000 ha. The aggregated species, northern white cedar and jack, require 300 points or 0.61 trees per 10,000 ha, tamarack requires 350 points or 0.71 trees per 10,000 ha and white pine a mere 40 points or 0.08 trees per 10,000 ha. The sparsely distributed species need as low as 40 points or 0.08 trees per 10,000 ha and as high as 400 points or 0.82 trees per 10,000 ha. Red pine requires 400 points or 0.82 trees per 10,000 ha, the elm group requires 200 points or 0.41 trees per 10,000 ha and red and white oak group require 150 points or 0.30 trees per 10,000 ha.

The widely distributed species had the largest amount of data records in the FIA database and did not map accurately when building the model with less than 700 records or 1.43 trees per 10,000 ha. This brings into question the accuracy of the map for white pine because there are less than 700 records present in the database. The map considered real is merely created from the maximum records available, up to 2,500 records due to computer limitations.

Exact amounts of data points necessary to accurately map aggregately, sparsely or widely distributed species could not be determined due to the vast differences between species within each spatial pattern group. With a conservative approach, the largest minimum amount of points required in each spatial pattern can be considered the minimum (Table 5). For example, the ash group, a widely distributed species group,

requires 1,100 points or 2.25 trees per 10,000 ha to produce an accurate map according to Kno so the minimum amount of data points for widely distributed species is 1,100. The widely distributed species require more data points than sparsely and aggregated to attain the same level of mapping accuracy. According to Kno, aggregated species require 700 points or 1.43 trees per 10,000 ha and sparsely require 500 points or 1.02 trees per 10,000 ha. To accurately map location, more data points are necessary than to accurately map quantities of each category. Widely distributed species require a minimum of 900 points or 1.84 trees per 10,000 ha to map location and 700 points or 1.43 trees per 10,000 ha to map quantity accurately. Aggregated species require 700 points or 1.43 trees per 10,000 ha for location and 350 points or 0.71 trees per 10,000 ha for quantity. Sparsely distributed species require 400 points or 0.82 trees per 10,000 ha for location and only 200 points or 0.41 trees per 10,000 ha to map quantity accurately (Table 5).

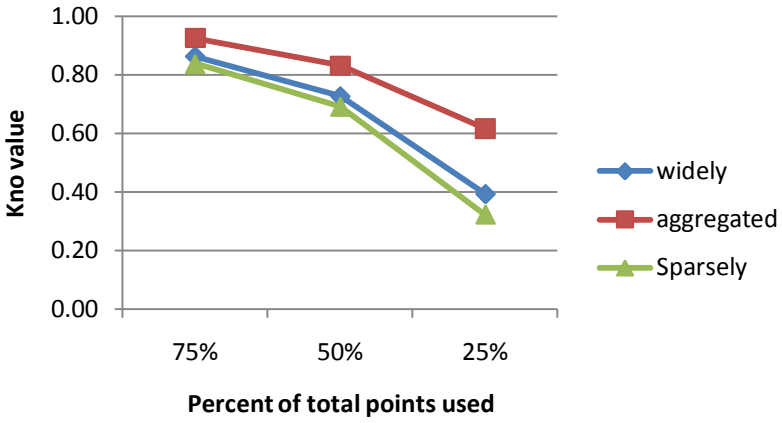


Figure 16 Average Kno at each comparison level, 75%, 50% and 25% of total FIA points by each spatial pattern group

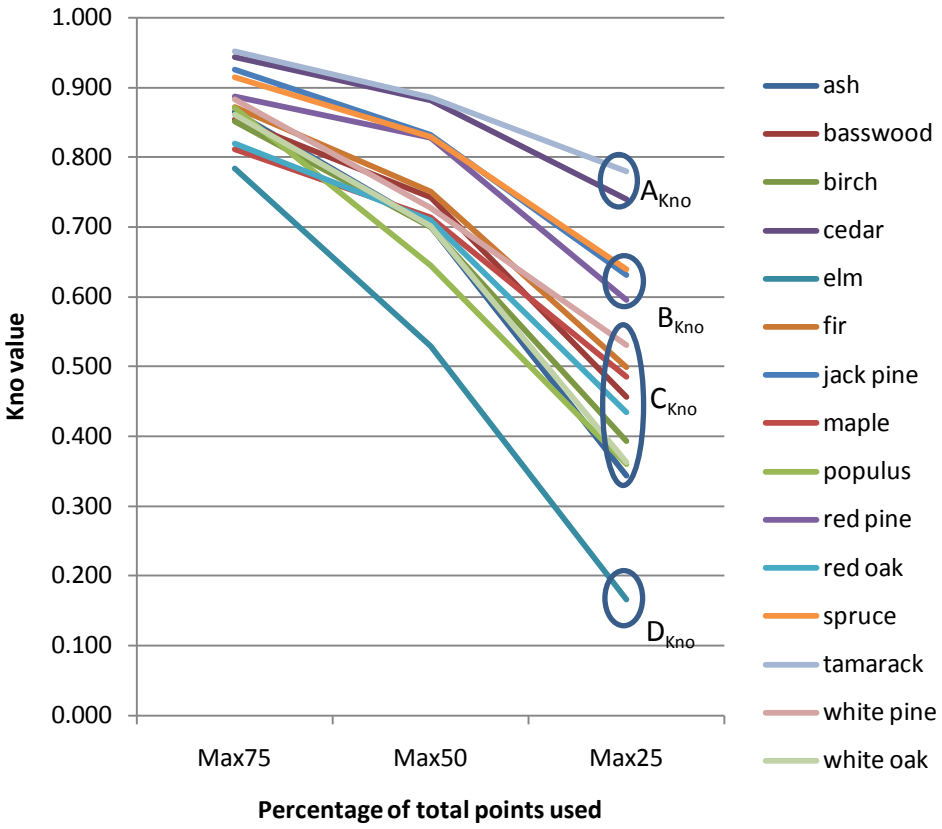


Figure 17 Kno values for all 15 tree species broken into four distinct accuracy groups

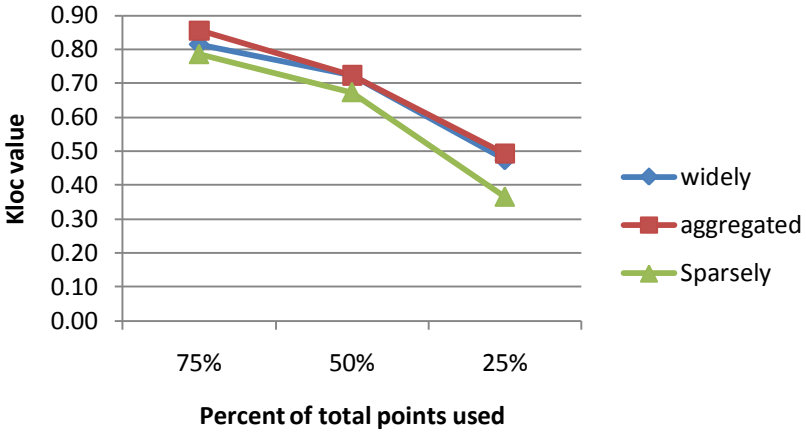


Figure 18 Average Klocation value by spatial pattern group

Table 3 Groupings based on accuracy according to Klocation

Group A	Group B	Group C	Group D
balsam fir	maple group	ash group	elm group
American basswood	red oak group	white oak group	
paper birch	red pine	white pine	
jack pine			
northern white cedar			
populus group			
spruce group			
tamarack			

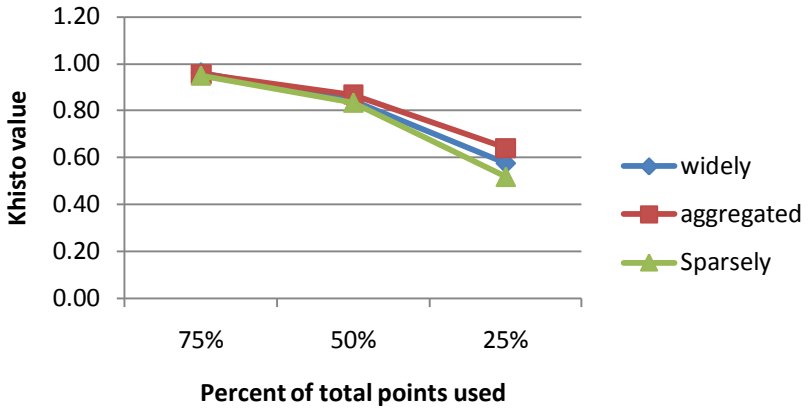


Figure 19 Average Khisto by spatial pattern group

Table 4 Groupings by accuracy based on Khisto statistic

Group A	Group B	Group C	Group D
tamarack	balsam fir	ash group	elm group
white pine	maple group	American basswood	
	northern white cedar	paper birch	
	spruce group	jack pine	
		populus group	
		red oak group	
		red pine	
		white oak group	

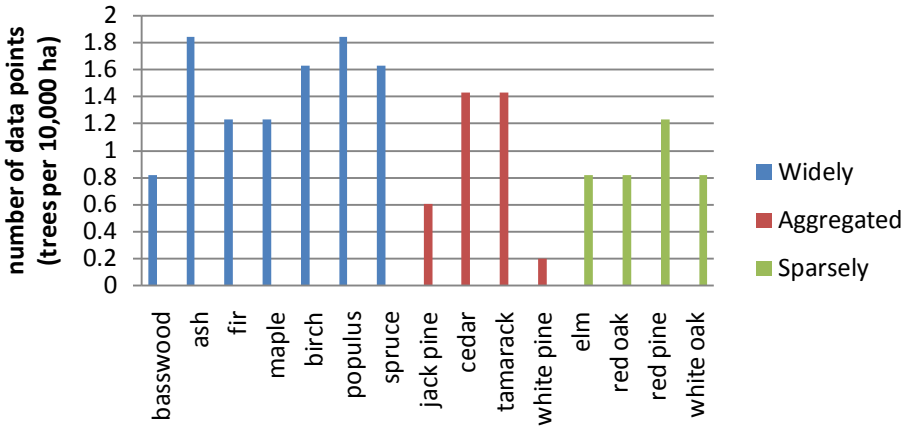


Figure 20 Amount of data points necessary to create accurate maps by species according to Klocation

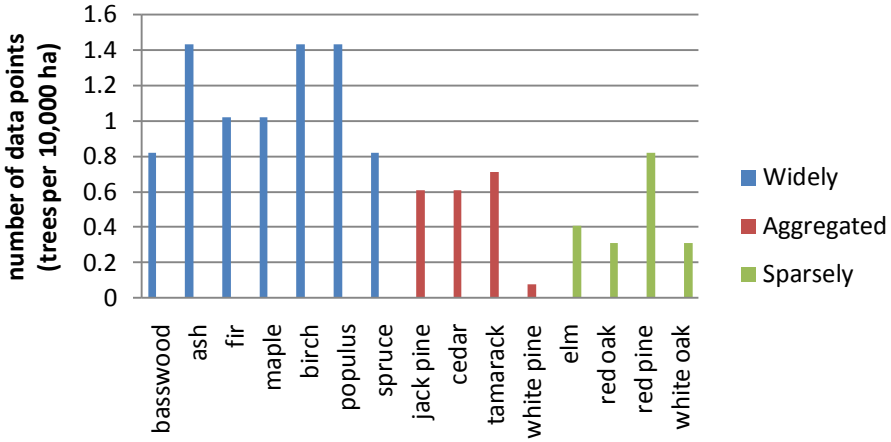


Figure 21 Amount of data points necessary to create accurate maps by species according to Khisto

Table 5 Largest minimum density of species that each spatial pattern requires in trees per 10,000 ha

	Kno	Location	Quantity
Widely	2.25	1.84	1.43
Aggregately	1.43	1.43	0.71
Sparsely	1.02	0.82	0.41

Discussion

Quantifying error

Aggregated species mapped most accurately in both location and quantity. This was expected because dominant species have previously been found to have higher mapping accuracy and aggregated species often out-compete and become dominant in stands (Guisan, et al., 2007). Widely distributed species fell second in accuracy and sparsely were the least accurate. Widely distributed species were probably difficult to map because they do not have a narrow geographic distribution. They can be found on a variety of landscapes making it difficult for Random Forest to accurately predict presence probability. Sparsely distributed species are difficult to map because they are the minority in stands.

According to all three statistics, Kno, Kloc and Khisto, elm had the lowest mapping accuracy and tamarack had the highest accuracy. Elms are moderately fast growing species, which have been previously found to have lower mapping accuracy (Guisan, et al., 2007). Tamaracks often form pure stands making them the dominant trees in a stand. In mixed stands, tamaracks are in the overstory. Both characteristics are associated with high mapping accuracy.

Sparsely distributed species require the fewest data points to accurately map location. This may be caused by the fewer data points available to create the true map since the species is not as common as widely distributed species. Widely distributed species require the largest amount of data points to map location accurately. A potential

reason for this is the small fraction of the total data points used to create the model in Random Forest and an overwhelming amount of points to test the accuracy of the model. Two thirds of the data should be used to build the model but due to computer limitations, a maximum of 2,500 data points could be used. This is a small fraction of the 24,440 points of trees in the populus group, resulting in a sample taken from a sample which can lead the data farther from being an accurate representation of reality.

The ash group and populus group, both widely distributed, require the most points to map location accurately. White pine, aggregated distribution, requires the least amount of data points. The accuracy of white pine's true map should be questioned because it was created with only 322 points at a density of 0.66 trees per 10,000 ha.

The quantity of polygons in each probability class mapped more accurately overall and with fewer data points than the location of each polygon when comparing the true map to the alternate maps with fewer data points used to build the model. As with location, sparsely distributed species required the fewest data points for quantity accuracy and widely distributed species required the most. Again the ash and populus groups require the most amount of data points while white pine requires the least. Paper birch is also among the species that require the largest amount of data points to map accurately.

Improving mapping accuracy

The creation of accurate species distribution maps needs a certain amount of data points. Each species is unique and has a different minimum amount of data points necessary. In this project, I attempted to determine the minimum amount necessary to create accurate maps using Random Forest based on the species distribution pattern. The results show that in general, widely distributed species require more data points, followed by aggregated then sparsely.

Widely distributed species had the most available data points to create the true map. It may be inferred that the widely distributed species true maps are the most accurate and it was found that the widely distributed species require a minimum of 1,100 points or 2.25 trees per 10,000 ha to maintain accuracy. Elm and white pine did not have 1,100 records in the study area.

This study produced a wide range of minimum data points within the three distribution groups making it difficult to determine a minimum amount necessary based on distribution pattern. According to the largest amount necessary of a species within each group, widely distributed species should have at least 1,100 points or 2.25 trees per 10,000 ha to produce an accurate map. Only 900 points or 1.84 trees per 10,000 ha are required to map location accurately and 700 points or 1.43 trees per 10,000 ha to map quantity accurately. Aggregated species require 700 points or 1.43 trees per 10,000 ha but if the user is only interested in quantity accuracy, only 350 points or 0.71 trees per 10,000 ha are required. Sparsely distributed species require 500 points or 1.02 trees

per 10,000 ha to map accurately according to Kno, 400 points or 0.82 trees per 10,000 ha for location accuracy and only 200 points or 0.41 trees per 10,000 ha for quantity accuracy.

Using larger sample sizes improves model performance (Reese, Wilson, Hoeting, & Flather, 2005) and with current technology, there is no need to decrease data with the goal of improving processing time. Kadmon, Farber, and Danin (2003) found 50 data points (a density of about 0.18 trees per 10,000 ha) to be sufficient for accurate mapping. In multiple studies, as low as 10 data points in a landscape scale have been found to be sufficient for mapping species (Hernandez, Graham, Master, & Albert, 2006; Wisz, et al., 2008). Guisan, et al. (2007) found that larger sample sizes produce more accurate results and that each species varies greatly in necessary data points, even within each tree genus. A limitation in this study may be in the grouping species together. Further studies should be done in Minnesota without grouping species to determine minimum data points necessary for accurate mapping.

Limitations in this study include relying solely on FIA data to demonstrate current tree species locations because of the spatial gaps in the FIA point data. Only species presence information was available. The absence of a species was not taken into account. Also, the environmental variables are at a coarse resolution and have heterogeneous environmental conditions within each map unit. The FIA plots are smaller than the map unit which may affect mapping accuracy. The model was developed using the observed distribution of tree species in relation to the calculated

environmental variables but is not extensive enough to describe all possible combinations of environmental variables. A model limitation can be seen in species like white oak, white pine and elm. The white oak species was expected to be sparsely distributed but Random Forest predicted it as a widely distributed species, white pine had very few data records in FIA and elm had the lowest Kappa values. A computer limitation occurred with large data sets, like the populus group. A sample of the FIA sample had to be taken, possibly leading the data set farther from an accurate representation of reality.

Suggestions for future research are including weather variables to see if species, like white oak, are mapped more ecologically realistic. Also, research should be done on why white oak was mapped as widely distributed and why elm had the lowest mapping accuracy. Future vegetation mapping projects could use a different statistical approach for the methodology of variable selection, regression and dealing with large data sets.

LITERATURE CITED

Albert, D. (1995). *Regional landscape ecosystems of Michigan, Minnesota, and Wisconsin: a working map and classification.* St. Paul, MN: USDA Forest Service North Central Forest Experiment Station.

Allen, D. C., Molloy, A. W., Cooke, R. R., & Pendrel, B. A. (1999). A ten-year regional assessment of sugar maple mortality. In S. B. Horsley, R. P. Long, & eds, *Sugar maple ecology and health: proceedings of an international symposium* (pp. 27-45). Warren, PA: Gen Tech Rep NE-261. Radnor, PA: U.S. Dept of Agriculture, Forest Service, Northeastern Research Station.

Avers, P., Cleland, D., McNab, W., Jensen, M., Bailey, R., King, T., et al. (1994). *The national hierarchical framework of ecological units.* Washington, D.C.: ECOMAP, USDA Forest Service.

Bergeron, Y. (2000). Species and stand dynamics in the mixed woods of Quebec's Southern Boreal Forest. *Ecology* , 1500-1516.

Breiman, L. (2001). Random Forests. *Machine Learning , 45*, 5-32.

Burns, R. M., & Honkala, B. H. (1990). *Silvics of North America.* Washington, DC: US Dept of Agriculture, Forest Service.

Congalton, R. G. (1999). *Assessing the accuracy of remotely sensed data: principles and practices.* Boca Raton: Lewis Publishers.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology , 88* (11), 2783-2792.

Data and Tools. (2010). (USDA Forest Service) Retrieved 2008, from FIA Program: http://fia.fs.fed.us/tools-data

De'ath, G., & Fabricius, K. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology , 81*, 3178-3192.

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* , 263-274.

Environmental Systems Research Institute. (2008). Arc ver. 9.3.1. Redlands, CA.

ESRI SUPPORT CENTER. (n.d.). *ArcScripts*. Retrieved from http://arcscripts.esri.com

Finos, L., Brombin, C., & Salmaso, L. (2010). Adjusting stepwise p-values in generalized linear models. *Communications in Statistics- Theory and Methods* , 1832-1846.

Foody, G. (2007). Map comparison in GIS. *Progress in Physical Geography* , 439-445.

Forest Inventory and Analysis Program. (2008). *The Forest Inventory and Analysis Database: Database Description and Users Manual Version 3.0 for Phase 2.* USDA Forest Service.

Fowells, H. A. (1965). *Silvics of forest trees of the United States.* Washington, DC: U.S. Department of Agriculture, Forest Service.

Friedman, S. K., & Reich, P. B. (2005). Regional legacies of logging: departure from presettlement forest conditions in northern Minnesota. *Ecological Applications* , 726-744.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters* , 993-1009.

Guisan, A., & Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling* , 147-186.

Guisan, A., Zimmermann, N., Elith, J., Graham, C., Phillips, S., & Peterson, A. (2007). What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? *Ecological Monographs , 77* (4), 615-630.

Guyette, R., Spetich, M., & Stambaugh, M. (2006). Historic fire regime dynamics and forcing factors in the Boston Mountains, Arkansas, USA. *Forest Ecology and Management* , 293-304.

Hagen, A. (2002). *Comparison of maps containing nominal data.* Maastricht: National Institute for Public Health and the Environment.

He, H., Dey, D., Fan, X., Hooten, M., Kabrick, J., Wikle, C., et al. (2007). Mapping pre-European settlement vegetation using a hierarchical Bayesian model and GIS. *Plant Ecology , 191*, 85-94.

He, H., Mladenoff, D., Radeloff, V., & Crow, T. (1998). Integration of GIS data and classified satellite imagery for regional forest assessment. *Ecological Applications , 8* (4), 1072-1083.

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* , 773-785.

Huffman, R. D., Fajvan, M. A., & Wood, P. B. (1999). Effects of residual overstory on aspen development in Minnesota. *Canadian Journal of Forest Research , 29*, 284-289.

Jensen, J. R. (2000). *Remote Sensing of the Environment, an earth resource perspective.* Upper Saddle River, NJ: Prentice-Hall, Inc. .

Kadmon, R., Farber, O., & Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications* , 853-867.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics* , 159-174.

Leahy, M. J., & Pregitzer, K. S. (2003). Comparison of presettlement and present-data in northeastern lower Michigan. *American Midland Naturalist* , 71-89.

Leeuw, J., Jia, H., Yang, L., Liu, X., Schmidt, K., & Skidmore, A. (2006). Comparing accuracy assessments to infer superiority of image classification methods. *International Journal of Remote Sensing* , 223-232.

Lister, A., Scott, C., King, S., Hoppus, M., Butler, B., & Griffith, D. (2005). Strategies for preserving owner privacy in the National Information Management System of the USDA Forest Service's Forest Inventory and Analysis Unit. *Proceedings of the fourth annual forest inventory and analysis symposium* (pp. 163-166). St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station.

Miles, P. D., Brand, G. J., Alerich, C. L., Bednar, L. F., Woudenberg, S. W., Glover, J. F., et al. (2001). *The Forest Inventory and Analysis Database: Database Description and Users Manual Version 1.0.* St. Paul: USDA Forest Service North Central Research Station.

Miller, J., Turner, M., Smithwick, E., Dent, C., & Stanley, E. (2004). Spatial extrapolation: The science of predicting ecological patterns and processes. *BioScience , 54*, 310-320.

Mladenoff, D. J., White, M. A., Pastor, J., & Crow, T. R. (1993). Comparing spatial pattern in unaltered old-grown and disturbed forest landscapes. *Ecological Applications , 3*, 294-306.

Monserud, R. A., & Leemans, R. (1992). Comparing global vegetation maps with the kappa statistic. *Ecological Modelling* , 275-293.

NRCS. *SSURGO Metadata - Table Column Descriptions Version 2.2.3.* USDA Natural Resource Conservation Service.

Peterson, A., Papes, M., & Soberon, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling* , 63-72.

Pontius, R. G. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing , 66*, 1011-1016.

Prasad, A., Iverson, L., & Liaw, A. (2006). New classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems , 9*, 181-199.

R Development Core Team. (2009, April 17). *http://www.R-project.org*. Retrieved from The R Project for Statistical Computing.

Reese, G., Wilson, K., Hoeting, J., & Flather, C. (2005). Factors affecting species distribution predictions: A simulation modeling experiment. *Ecological Applications , 15* (2), 554-564.

Sappington, J., Longshore, K., & Thompson, D. (2007). Quantifying landscape ruggedness for animal habitat analysis: a case study using bighorn sheep in the Mojave Desert. *Journal of Wildlife Management , 71* (5), 1419-1426.

Sappington, M. (2008). *Vector Ruggedness Measure*. Retrieved from http://arcscripts.esri.com/details.asp?dbid=15423

Sousa, S., Caeiro, S., & Painho, M. (2002). *Assessment of map similarity of categorical maps using kappa statistics.* ISEGI– Instituto Superior de Estatística e Gestão de Informação.

Stenback, J., & Congalton, R. (1990). Using thematic mapper imagery to examine forest understory. *Photogrammetric Engineering and Remote Sensing , 56*, 1285-1290.

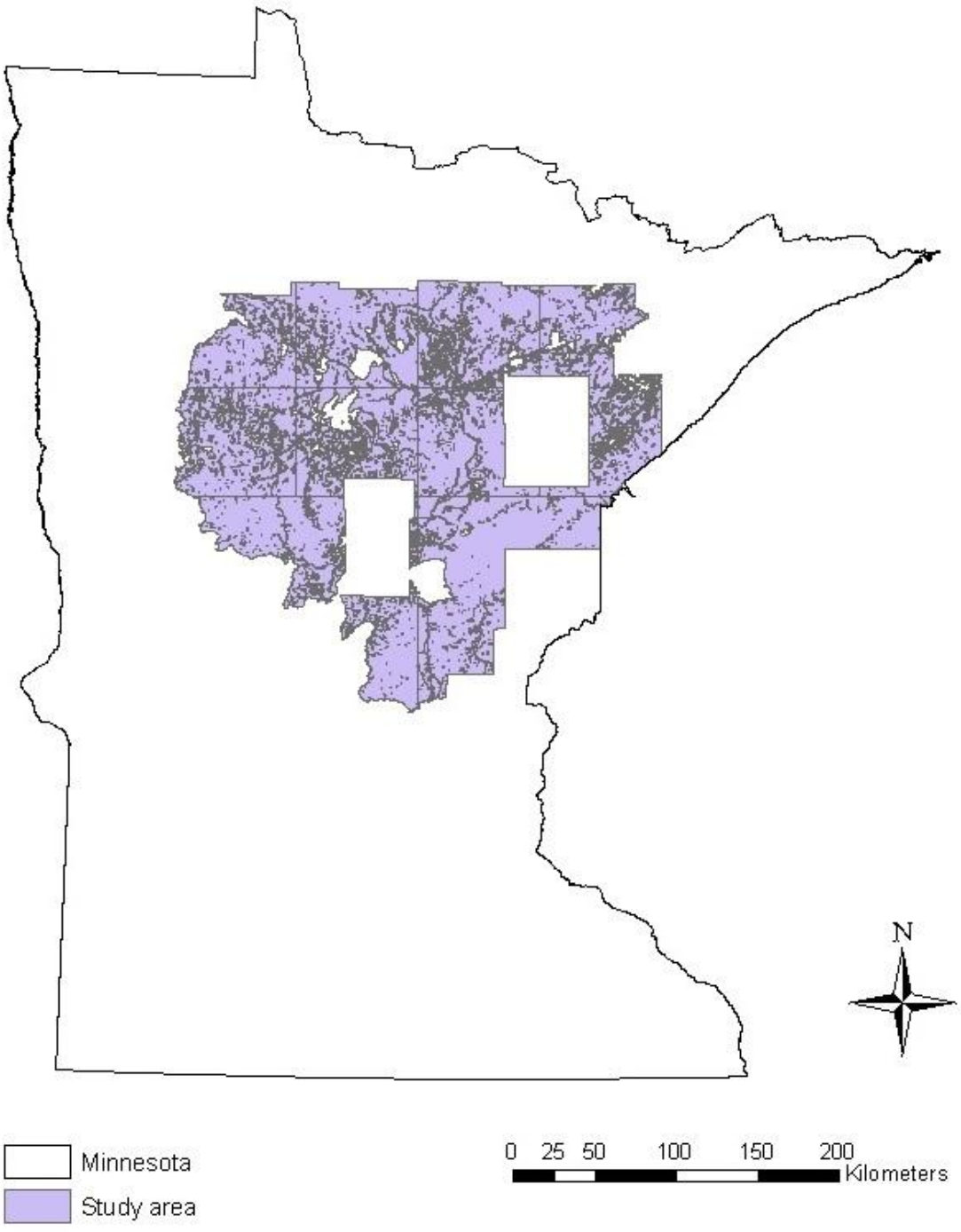
Trimble, G. R., & Weitzman, S. (1956). Site index studies of upland oaks in the northern Appalachians. *Forest Science* , 162-173.

Wisz, M., Hijmans, R., Li, J., Peterson, A., Graham, C., Guisan, A., et al. (2008). Effects of sample size on the performance of species distribution models. *Diversity of Distributions* , 763-773.

Appendix

Appendix A

Map of study area in the Laurentian Mixed Forest Province in Minnesota



Appendix B

Tree species in the study for MN.

Species		Species Group	Spatial Pattern
Larix laricina	tamarack	tamarack	aggregately
Pinus banksiana	jack pine	jack pine	aggregately
Pinus strobus	white pine	white pine	aggregately
Thuja occidentalis	northern white cedar	cedar	aggregately
Pinus resinosa	red pine	red pine	sparsely
Quercus alba	white oak	white oak	sparsely
Quercus ellipsoidalis	northern pin oak	red oak	sparsely
Quercus macrocarpa	bur oak	white oak	sparsely
Quercus rubra	northern red oak	red oak	sparsely
Ulmus americana	American elm	elm	sparsely
Ulmus rubra	slippery elm	elm	sparsely
Abies balsamea	balsam fir	fir	widely
Acer rubrum	red maple	maple	widely
Acer saccharinum	silver maple	maple	widely
Acer saccharum	sugar maple	maple	widely
Betula papyrifera	paper birch	birch	widely
Fraxinus nigra	black ash	ash	widely
Fraxinus pennsylvanica	green ash	ash	widely
Picea glauca	white spruce	spruce	widely
Picea mariana	black spruce	spruce	widely
Populus balsam	balsam poplar	populus	widely
Populus grandidentata	bigtooth aspen	populus	widely
Populus tremuloides	quaking aspen	populus	widely
Tilia americana	American basswood	basswood	widely

Appendix C

Amount of FIA data points used in each trial.

	Maximum data points available	75% of maximum data points	50% of maximum data points	25% of maximum data points	Test 1 of minimum data points necessary	Test 2 of minimum data points necessary
ash	2500	1875	1250	625	1000	800
basswood	1833	1375	916	458	400	300
birch	2500	1875	1250	625	1000	700
cedar	2500	1875	1250	625	400	300
elm	683	513	342	171	200	150
balsam fir	2500	1875	1250	625	1000	600
jack pine	1494	1121	747	374	100	300
maple	2500	1875	1250	625	1000	600
populus	2500	1875	1250	625	1000	800
red pine	2021	1516	1011	505	400	350
red oak	1179	885	590	295	200	150
spruce	2500	1875	1250	625	400	350
tamarack	2500	1875	1250	625	400	300
white pine	322	242	161	81	70	40
white oak	1733	1300	867	433	375	200