Predicting forest carbon content and tree canopy cover on Forest Inventory and Analysis plots along a forest-prairie gradient

A Thesis
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA BY

Jennifer Jane Nolan

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

Matthew Russell

December 2021

ACKNOWLEDGMENTS

Throughout my Master's program and through the writing of this thesis, I have received a great deal of support. I would like to thank my advisor, Matthew Russell, whose advice and encouragement was invaluable. I would also like to thank my committee members, Grant Domke and Eric Shook, for their guidance through revisions and discussions. This work was supported by the US Department of Agriculture—Forest Service, Northern Research Station. Last, I would like to thank my family and friends for their continual support through the research and writing of this thesis.

ABSTRACT

Trees outside forests (TOF) are increasingly recognized as an important resource for carbon storage. TOF include trees used in agricultural windbreaks, urban ornamentals, and trees in non-forest ecosystems like prairie. This study used tree data from the US Forest Service's Forest Inventory and Analysis (FIA) database to detect variation in forest carbon stocks and tree canopy cover between trees in forest ecosystems and trees in prairie ecosystems.

A longitudinal transect was established, extending from Bismarck, ND to Duluth, MN to capture a gradient of mixed temperate forest in the east, prairie in the west, and a transitional zone in the middle. A number of potential transect sizes were evaluated using a series of power analyses (alpha = 0.05, power = 0.88, 0.83) and these determined that a transect radius of 50 km (total transect height 100 km) and transect length of 660 km was sufficient to capture the gradient and provide statistically significant results if differences existed.

All FIA plots within the transect which had tree measurements taken from 2012 – 2018 (n = 4,155) were then used in a series of random forest analyses. The response variables of interest were carbon content (in megagrams per hectare) and percent live canopy cover, both at the plot level. Twenty-seven predictor variables were assessed: a few plot condition indicators from the FIA data, but mostly climate variables (30-year climate normals). Six random forest analyses were run: three examining canopy cover as the response variable using all plots, forest plots, and non-forest plots, and three examining carbon content within the same groupings.

The power analysis lent confidence to the establishment of an effective study area and transect size; however, the random forest analyses were ultimately unable to consistently predict tree canopy cover or carbon content at the plot level. Although the random forest analyses did not provide statistically significant evidence for variation at the plot level, the patterns they revealed between which climate predictors performed best under forested and non-forested conditions are intriguing and may invite further investigation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: A power analysis of tree canopy cover and aboveground carbon content along a forest-to-prairie gradient

Introduction

Trees outside forests (TOF) are an interesting and understudied class. TOF are often overlooked in silviculture even though collectively, they compose a very significant segment of the vegetation that we rely on for ecosystem services. Trees outside forests include agricultural windbreaks, urban ornamentals, and trees in non-forest ecosystems, among others (Rossi et al., 2016). An important benefit of measuring TOF is the ability to better manage TOF (MacFarland, 2020).

Two categories of TOF that are being paid attention to more closely now are riparian buffers and agricultural windbreaks. Kellerman et al. created in 2019 an Esri StoryMap discussing TOF as a phenomenon and showing these windbreaks and riparian buffers at work in plains states like Kansas, North and South Dakota. Because TOF are understudied, Kellerman et al. (2019) needed to perform their own image classification to locate these phenomena, as they were not included in pre-existing inventories. Kellerman et al. points out how the National Land Cover Database (NLCD), which is developed from 30-meter Landsat data, while useful for regional or large-scale assessments, fails to pick up many TOF because of their small, thin shapes. For this reason, they propose the use of special shape-based indexes within the object-based image analysis (or OBIA) classification process to better capture riparian buffers, windbreaks, and to tell the difference between the two.

The USDA Forest Service has implemented a program called the Trees Outside Forests Image-based Inventory (TOFii) dedicated to identifying and measuring trees outside forests in the panhandle of Texas, Nebraska, Kansas, and North and South Dakota. These states were chosen because of their plains geography and a historic effort related to planting windbreaks in the central US to reduce the effects of soil erosion (USDA Forest Service, 2021). This inventory not only measures the oft-neglected TOF, it

works to identify the most effective methods of measuring TOF. Specifically, TOFii uses a custom rapid assessment line-intersect sampling tool "developed for use in estimating total length of windbreaks" (USDA Forest Service 2021).

In the absence of a national TOF inventory, can we use FIA database to draw conclusions about TOF? The power analysis detailed in this chapter is an attempt at doing just that. Using FIA data provides a consistent set of measurements taken across state boundaries and across time that can be used to draw conclusions about both TOF and trees within forests.

Cadenasso et al. (2003) describes three core features of ecological boundaries - they are three dimensional, feature a steeper gradient than the ecosystems they connect, and the width of the boundaries are relative to its steepness. A study of the prairie-forest boundary (PFB) in Minnesota suggests that many factors play a role in the location of the boundary, and the interplay of these factors depends largely on spatial scale (Danz et al, 2011).

The goal of this study is not to find the reasons for the location of the PFB - that will be left in the capable hands of climatologists. The goal here is rather to examine how tree characteristics vary within and to either side of the PFB using pre-existing forest inventory data. Chapter 1 will begin this process by using a statistical power analysis to assess the sample size needed to draw statistically significant conclusions from specific FIA variables along a transect that crosses the PFB.

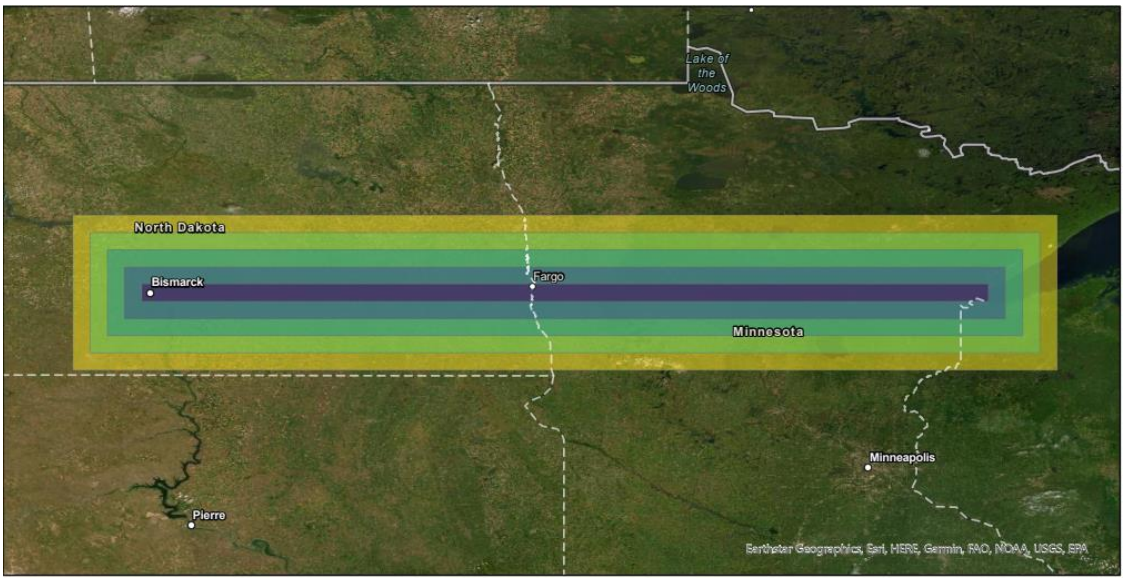
Methods

Determination of Study Area

A longitudinal transect encompassing forest, prairie, and forest-to-prairie transition environments was established, with one endpoint in Bismarck, ND, and the other in Duluth, MN. Plot density is higher in Minnesota than in North Dakota, as individual states can choose to pay for increased sampling intensity (USDA Forest Service, 2005).

Figure 1: Map of Potential Transect Sizes

Geodesic transects at 10, 30, 50, 70, and 90 kilometers from a transect line between Bismarck, ND and Duluth, MN



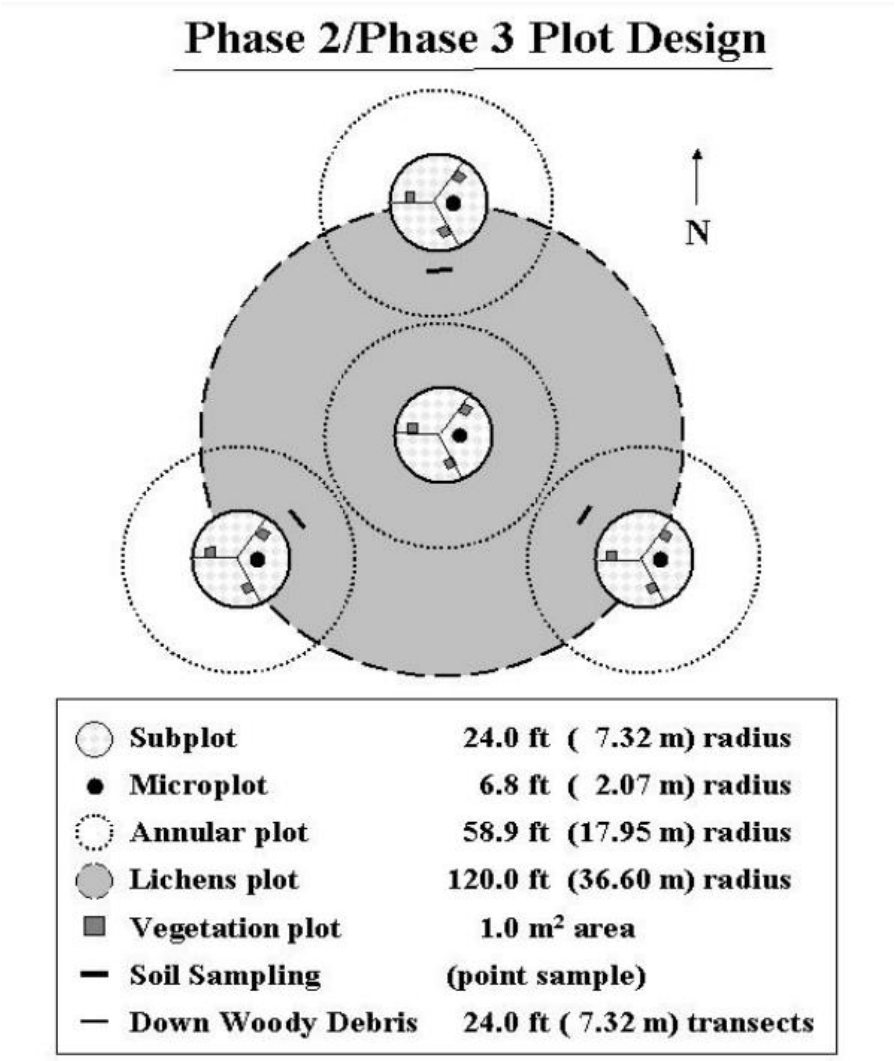
Field Sample Protocols

The data for this project comes from the **Forest Inventory and Analysis** program. The FIA's mission is to "make and keep current a comprehensive inventory and analysis of the present and prospective conditions of and requirements for the renewable resources of the forest and rangelands of the US" (USDA Forest Service, 2021).

Figure 2 displays the shape and arrangement of an FIA plot, described as "cluster of four circular subplots spaced out in a fixed pattern" (USDA Forest Service, 2005).

Figure 2: FIA Plot Diagram

Plot sampling design from the FIA Sampling and Plot Design Fact Sheet.



The two response variables being analyzed are the percentage of live canopy cover per plot and the aboveground carbon content per tree which is then scaled to represent carbon on a plot level.

- **live_canopy_cvr_pct** Eight methods of estimating live canopy cover percent are indicated in FIA documentation. A code (CANOPY_CVR_SAMPLE_METHOD_CD) is given for each observation indicating which method was used to estimate that observation's live canopy

cover percent. Half are field-measured and half are generated using imagery (USDA Forest Service, 2018).

- **carbon_ag** Aboveground carbon on a per tree basis. This estimate must be multiplied by the TPA_UNADJ expansion factor and then summed by plot, then converted from pounds-per-acre to megagrams-per-hectare.“The carbon (pounds) in the aboveground portion, excluding foliage, of live and standing dead trees greater than 1.0 inch d.b.h/d.r.c. Carbon is assumed to be one-half the value of biomass and is derived by summing the aboveground biomass estimates and multiplying by 0.5 as follows: CARBON_AG = 0.5 * (DRYBIO_BOLE + DRYBIO_STUMP + DRYBIO_TOP + DRYBIO_SAPLING + DRYBIO_WDLD_SPP)” (USDA Forest Service, 2018)

Data

FIA data was brought into an R code environment using SilviaTerra’s “tidyFIA” package (Rodman et al., 2020). Data was utilized from North Dakota, Minnesota, and Wisconsin from 2012 - 2018. This time range was selected because the FIA began measuring the live canopy cover percent of nonforest plots in 2012 [Reimann 2016], and the most recent inventory at the time of writing was completed in 2018. Data comes primarily from the FIADB “plot”, “condition”, and “tree” tables. Plot measurements with null geometry or null values in the live_canopy_cvr_pct or carbon_ag fields, respectively, were excluded from analysis. 13.8% of plot measurements had either null geometry or null values, which reduced the number of total observations from 227,164 to 196,010. Only the most recent measurement for each plot or tree was used. Only plots with the condition codes for “forest” or “nonforest” land were used (this excludes water, non-census, and unmeasured plots.)

Transects were created using a function which converts kilometer distances to decimal-degree distances that can then be added to or subtracted from the decimal degree coordinates of each endpoint of the transect line. Because the endpoints are intentionally similar in latitude, bounding boxes for each transect were established using the averaged latitude of the endpoints.

Power Analysis

A power analysis is a statistical test to determine the sample size needed to achieve statistical significance given a set of specific parameters about the data set. The parameters used in this analysis are n (sample size), delta (change threshold), sd (standard deviation), type (one or two sample test), and alternative (one- or two-sided curve).

Summary statistics were obtained from canopy cover and aboveground carbon content data summed by plot, then by transect. Power analyses were conducted at a range of potential sample sizes from 1 to 4,000, and at different alpha and delta levels to determine the best option for minimizing alpha, delta, and transect size to the extent possible while still achieving a power greater than 0.8. Figures 3 and 4 demonstrate an array of these power analyses at delta values equal to 5, 10, and 20 percent of the sd for each variable and at alpha values of 0.05, 1.0, and 2.0 to show the relationship between delta, alpha, and power in accordance with sample size.

Results

Table 1: Summary Statistics

Summary table of percent live canopy cover and carbon content statistics at transect widths of 10, 30, 50, 70, and 90 km.

	Percent Live Canopy Cover			Carbon Content (Mg/ha)		
Transect Width (km)	**Mean**	**n**	**Standard Deviation**	**Mean**	**n**	**Standard Deviation**
10	50.66	387	35.64	13.82	308	9.00
30	52.45	1151	35.23	13.63	972	9.03
50	53.56	1991	34.88	14.11	1695	9.31
70	54.28	2866	34.32	13.76	2500	9.20
90	54.82	3762	34.15	13.81	3332	9.41

Figure 3: Power curves for percent live canopy cover

Blue vertical lines indicate the sample size at transect widths of 10, 30, 50, 70, and 90 kilometers. Dashed horizontal line shows the power threshold of 0.8, commonly accepted as the minimum threshold for detection of statistically significant phenomena.

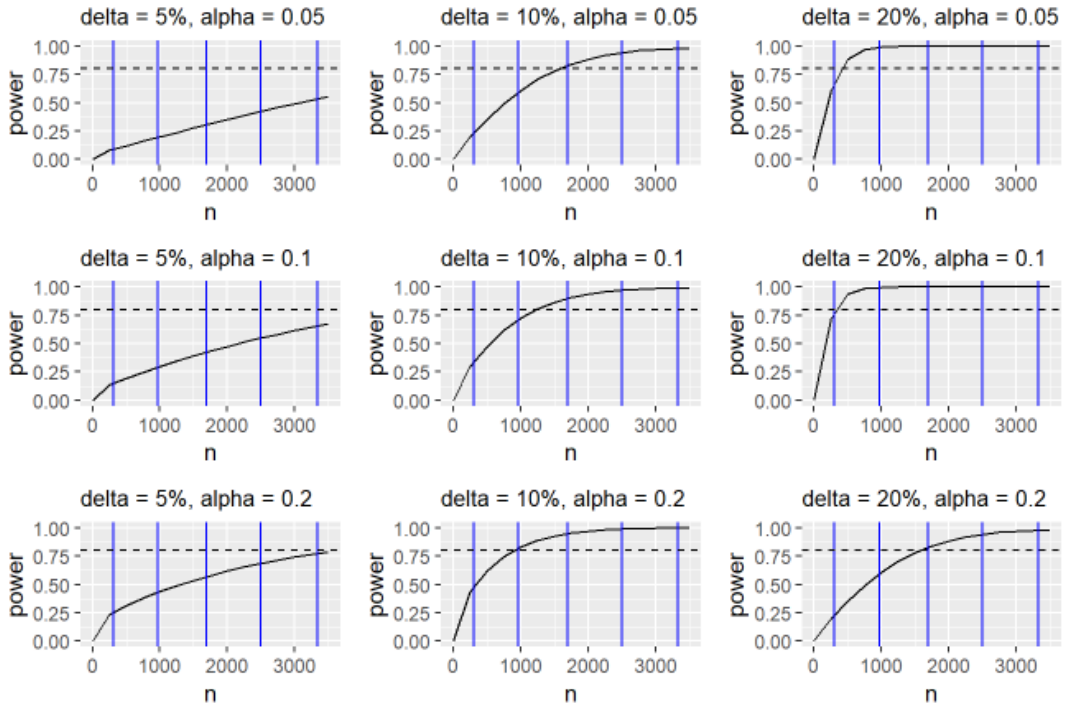
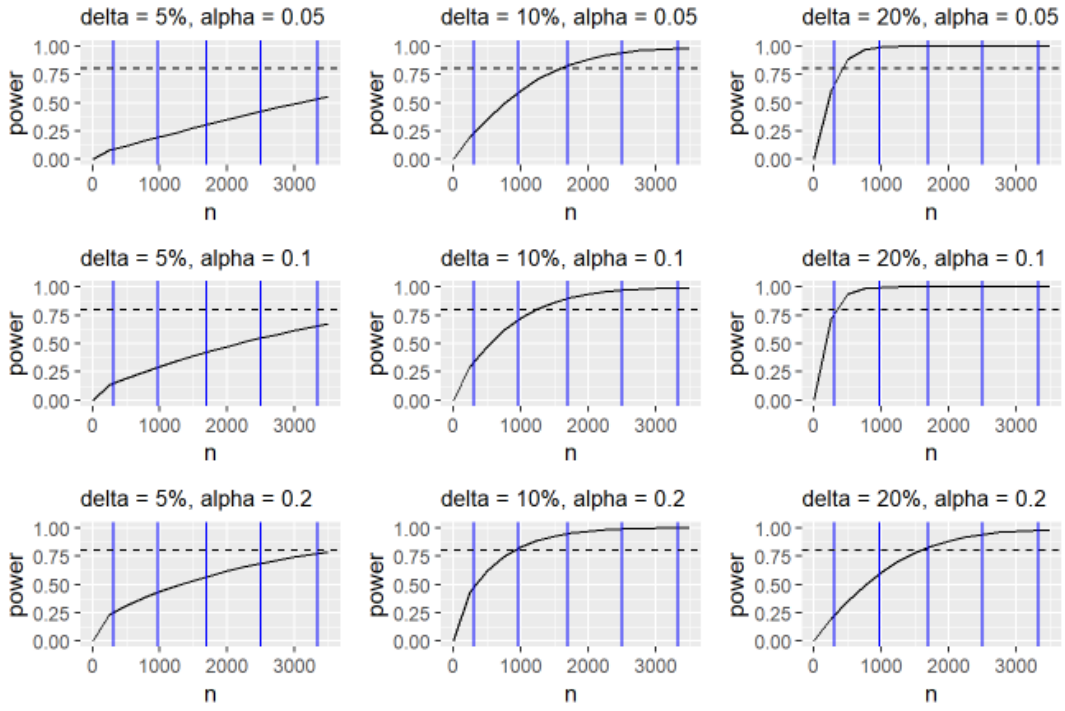


Figure 4: Power curves aboveground carbon content (Mg/ha)

Blue vertical lines indicate the sample size at transect widths of 10, 30, 50, 70, and 90 kilometers. Dashed horizontal line shows the power threshold of 0.8, commonly accepted as the minimum threshold for detection of statistically significant phenomena.



The power curve plots (Figures 3 & 4) indicate the combination of alpha = 0.05 and delta = 10% of the sd to be the ideal combination for minimizing alpha, delta, and transect size both for canopy cover and carbon content.

Attempting to minimize delta, alpha, and transect size to the greatest extent possible results in an ideal transect size of 50 kilometers. At this size, a delta value equal to 10% of the sd (3.49% for the canopy cover variable and 0.93 Mg/ha for carbon content) and a significance level (alpha) of 0.05 is sufficient to reach the established power threshold of 0.8. A power analysis with these parameters indicates a predicted power of 0.88 for canopy cover and 0.83 for carbon content. Higher power can be achieved by increasing the size of the transect, alpha, or delta - but doing so increases

uncertainty in other areas. A 50 kilometer transect is thus statistically sufficient and shall be used for subsequent analysis.

The 50 kilometer transect contains a total of 4,155 plots: 2,657 plots in Minnesota, 1,357 plots in North Dakota, and 141 plots in Wisconsin. From these, 1,400 plots are classified as forest and 2,775 are classified as non-forest land. This transect captures 16 distinct ecoregion subsections in 5 ecological provinces (Cleland et al., 2007).

Table 2: Bailey's Eco-Subregions in the 50km transect

Code, count, name, and ecological province for plot membership in ecoregion subsections found in the 50 kilometer transect.

Eco-subregion code	Count	Subsection name	Province
212Nc	753	Pine Moraines and Outwash Plains	Laurentian Mixed Forest
251Aa	633	Lake Agassiz Plain	Prairie Parkland (Temperate)
331Ea	532	Missouri Couteau	Great Plains – Palouse Dry Steppe
332Aa	452	Glaciated Drift Plains	Great Plains Steppe
212Nd	375	Tamarack Lowlands	Laurentian Mixed Forest
212Nb	312	St. Louis Moraines	Laurentian Mixed Forest
212Kb	225	Mille Lacs Uplands	Laurentian Mixed Forest
222Ma	223	Alexandria Moraine-Hardwood Hills	Midwest Broadleaf Forest
212Lb	218	North Shore Highlands	Laurentian Mixed Forest
331Mc	97	Missouri Plateau	Great Plains – Palouse Dry Steppe
212Ya	95	Superior-Ashland Clay Plain	Laurentian Mixed Forest
212Na	69	Chippewa Plains	Laurentian Mixed Forest
251Ab	68	Souris-Agassiz Stratified Fan Deposits	Prairie Parkland (Temperate)
331Mf	50	River Breaks	Great Plains – Palouse Dry Steppe
212Ld	42	Toimi Uplands	Laurentian Mixed Forest
212Ka	11	Bayfield Sand Plains	Laurentian Mixed Forest

Discussion

A longitudinal transect buffered at a width of fifty kilometers sufficiently captures a mix of forest, prairie, and forest-to-prairie transition zones and allows the drawing of

statistically significant conclusions, should they exist, regarding variation in percent of live canopy cover and aboveground carbon content. In Chapter 2, the plot measurements contained in the fifty-kilometer transect will be used to analyze effects on these two variables as a product of environmental factors. This will provide insight into whether TOF display characteristics distinct from trees within forests. If such a distinction is shown, it may indicate that methods developed to measure, analyze, or care for trees within forests may have to be adapted to be as accurate and effective for TOF.

The results of this power analysis indicate that FIA data can be used to examine specific phenomena in TOF even though it was not designed for that purpose. While a national TOF inventory would certainly improve the quality and scalability of a study such as this, in its absence the FIA data is an adequate surrogate. Should the conclusions of this study reveal something important about TOF, it will stand as evidence that TOF ought to be measured as a phenomenon in themselves, at as large a scale as is practical.

TOF provide important ecosystem services, especially as riparian buffers, agricultural windbreaks, and urban ornamentals. Knowing more about these trees as a category allows for more informed decision-making on the part of land managers and foresters.

Chapter 2: Predicting Tree Canopy Cover and Carbon Content on a forest-prairie gradient using Random Forest Analysis and 30-year climate normals.

Introduction

Trees outside forests (TOF) are an interesting and understudied class. TOF are often overlooked in forest resource assessments even though collectively, they compose a very significant segment of the vegetation that we rely on for ecosystem services. Trees outside forests include agricultural windbreaks, urban ornamentals, and trees in non-forest ecosystems, among others (Rossi et al., 2016). An important benefit of measuring TOF is the ability to better manage TOF (MacFarland, 2020). This analysis attempts to detect variation in percent canopy cover and carbon content across groups of trees contained in the Forest Inventory and Analysis (FIA) database spanning a longitudinal transect which spans from Bismarck, ND, to Duluth, MN.

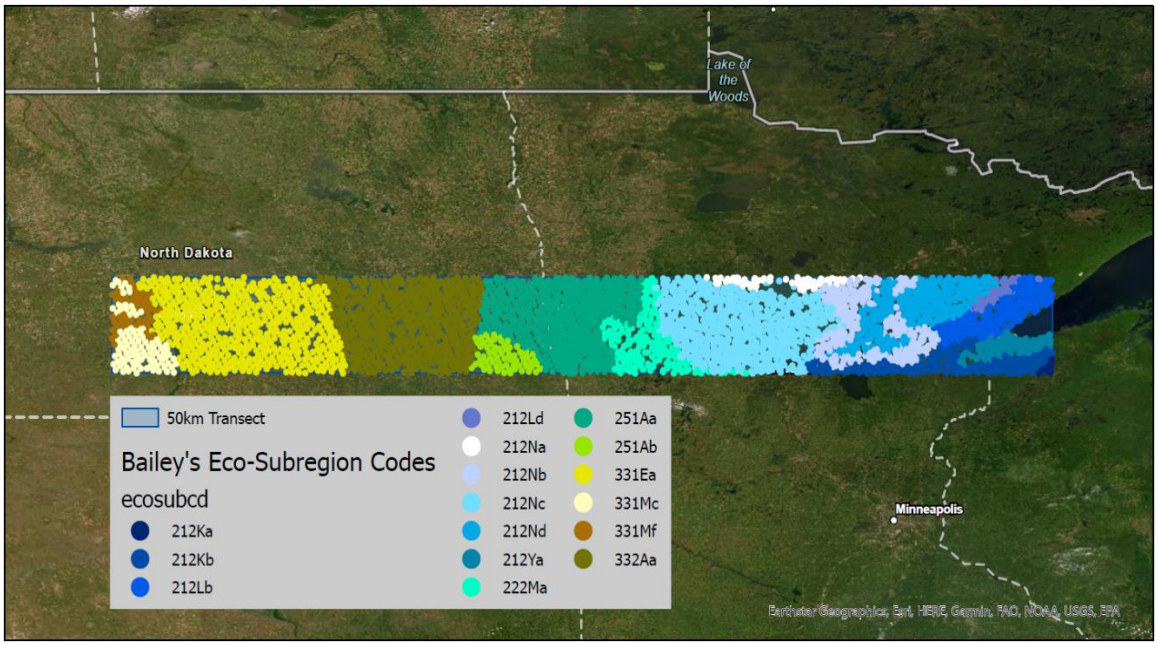
Chapter 1 established that a transect radius height of 50 kilometers (resulting in a total transect height of 100 kilometers) was sufficient to capture variation in tree canopy cover and carbon content across the prairie-forest border (PFB) if such variation exists. Chapter 2 uses random forest analysis to attempt to detect variation in these two variables at the plot level using local 30-year climate normals and plot condition information as predictors.

Local climate has an intuitive link to vegetation patterns: annual precipitation and mean temperature are established predictors of what species of trees may be found in a given area (Iverson 1998). This analysis tests whether climate variables can predict more complex characteristics than tree species: whether we can use climate information to predict the amount of tree carbon or percent live canopy cover is present on a given FIA plot. Additionally, this analysis seeks to know how these variables change across the PFB. It is established that trees of different species store different amounts of carbon (Kirby 2007), but it is unclear whether these variations also occur in line with hyper-local climate differences.

Figure 5 displays all FIA plots present in the 50 kilometer transect, colored by eco-subregion. Because not all plots had measurements taken during the study's temporal window (2012-2018), not all plots pictured were able to be used in the random forest: Figure 5 is meant to show the diversity of ecosystem types found along the transect.

Figure 5: 50-kilometer Transect Map of Eco-Subregions

50 kilometer transect map of plots categorized by Bailey's eco-subregion code



Methods

Determination of Study Area

A longitudinal transect encompassing forest, prairie, and forest-to-prairie transition environments was established, with one endpoint in Bismarck, ND, and the other in Duluth, MN. Plot density is higher in Minnesota than in North Dakota, as individual states can choose to pay for increased sampling intensity (USDA Forest Service, 2005). In Chapter 1, a power analysis was conducted on various transect sizes and determined that a transect width of 50 kilometers was statistically appropriate.

Field Sample Protocols

The data for this project comes from the **Forest Inventory and Analysis** program. The FIA's mission is to "make and keep current a comprehensive inventory and analysis of the present and prospective conditions of and requirements for the renewable resources of the forest and rangelands of the US" (USDA Forest Service, 2021).

The two variables being analyzed are the percentage of live canopy cover per plot and the aboveground carbon content per tree.

- **live_canopy_cvr_pct** Eight methods of estimating live canopy cover percent are indicated in FIA documentation. A code (CANOPY_CVR_SAMPLE_METHOD_CD) is given for each observation indicating which method was used to estimate that observation's live canopy cover percent. Half are field-measured and half are generated using imagery (USDA Forest Service, 2018).
- **carbon_ag** Aboveground carbon on a per tree basis. This value must be multiplied by the TPA_UNADJ expansion factor and then summed by plot, then converted from pounds-per-acre to megagrams-per-hectare."The carbon (pounds) in the aboveground portion, excluding foliage, of live and standing dead trees greater than 1.0 inch d.b.h/d.r.c. Carbon is assumed to be one-half the value of biomass and is derived by summing the aboveground biomass estimates and multiplying by 0.5 as follows: CARBON_AG = 0.5 * (DRYBIO_BOLE + DRYBIO_STUMP + DRYBIO_TOP + DRYBIO_SAPLING + DRYBIO_WDLD_SPP)"(USDA Forest Service, 2018).

Data

FIA data were brought into an R environment using SilviaTerra's "tidyFIA" package. (Rodman et al., 2020) Data were utilized from North Dakota, Minnesota, and Wisconsin from 2012 - 2018. This time range was selected because the FIA began measuring the live canopy cover percent of nonforest plots in 2012 (Reimann 2016), and the most recent inventory at the time of writing was completed in 2018. Data comes primarily from the FIADB "plot", "condition", and "tree" tables. Plots with null geometry or null values in the live_canopy_cvr_pct or carbon_ag fields, respectively, were excluded from analysis. Only the most recent measurement for each plot or tree was

used. Only plots with the condition codes for “forest” or “nonforest” land were used (this excludes water, non-census, and unmeasured plots).

The fifty kilometer transect was created using a function which converts kilometer distances to decimal-degree distances, that can then be added to or subtracted from the decimal degree coordinates of each endpoint of the transect line. Because the endpoints are intentionally similar in latitude, bounding boxes for the transect were established using the averaged latitude of the endpoints.

Predictor Variable Selection

Three topographical variables were selected from the FIA data: **elev**, elevation of the plot in meters, **long**, longitude at the center of the plot, and **lat**, latitude at the center of the plot.

Three additional variables were obtained directly from the FIA data. The **spcd_per_plot** variable identifies the number of unique tree species present on an individual FIA plot as a representation of tree species diversity. The **dist_bin** and **dist_count** variables represent the occurrence of any disturbances (like fire, disease, insect damage, or logging) that occurred on the plot during the study time frame: a binary indicator for whether any disturbance occurred and a count of total disturbances (range: 0-2), respectively.

An additional seventeen climate variables were used as predictors. These variables were accessed from a USDA and Virginia Tech partnered web tool (USDA-Virginia Tech, 2021) that provides point estimates of climate measures. The web tool allows submission of a list of locations with a high degree of specificity (~1 km grid cell size) and returns climate information for those locations based on selected desired output. The output used here were the 30-year climate normals and derived variables of those normals with a separate entry for each plot in the transect.

Table 3 lists all numerical predictors along with their mean, min, and max values found within the subset of data used in the following analyses (n = 1,633). Three categorical variables were not included in the table: ecological sub-section code

(**ecosubcd**, see Table 2), plot condition status code **(cond_status_cd**: 1 = forest, 2 = non-forest), and a binary indicator variable for plot disturbance (**dist_bin**: 0 = no disturbance, 1 = one or more disturbances).

Table 3: Summary Statistics of Predictor Variables

Summary statistics of quantitative predictor variables and description of variables.

variable_name	abbreviation	mean	min	max
Carbon Content (megagrams per hectare)	carb_mgha	14.07	0.05	67.32
Percent Live Canopy Cover	live_canopy_cvr_pc	62.5	0	95
Disturbance Count	dist_count	0.11	0	2
Tree Species Diversity	spcd_per_plot	3.97	1	10
Trees per Hectare	Tpha	162.33	2.44	980.54
Elevation	elev	403.33	187	612
Longitude	long	-93.72	-101.14	-91.66
Latitude	lat	46.83	46.36	47.26
Mean Annual Temperature	mat	4.03	2.8	6
Mean Annual Precipitation	map	699.3	398	810
Growing Season Precipitation	gsp	504.77	314	558
Mean Temperature During Coldest	mtcm	-14.58	-16.2	-11.9
Mean Minimum Temperature	mmin	-20.82	-22.5	-18.1
Mean Temperature During Warmest	mtwm	19.78	17.6	22.1
Mean Maximum Temperature	mmax	26.69	24.6	29.9
First Freezing Date	fday	263.41	256	271
Frost-Free Period	ffp	120.59	95	142
Degree Days > 5 C	dd5	1823.32	1468	2233
Degree Days > 5 C During Growing	gsdd5	1463.57	1046	1946
Julian Date When 100 dd5 Reached	d100	126.92	122	135
Degree Days < 0 C	dd0	1443.01	1159	1640
Summer Precipitation Balance	smrpb	1.15	0.79	1.33
Summer/Spring Precipitation	smrsprpb	1.44	0.96	1.58
Winter Precipitation	winp	96.72	41	147
Annual Dryness Index	adi	2.64	1.88	5.45

Random Forest Analysis

Data was randomly subset into a training set (70% of the data) and a test set (30%). The analysis was performed in R using the randomForest package (Liaw & Wiener). A total of 27 predictor variables were used in each analysis: the 26 detailed in the previous section, as well as either the percent canopy cover or carbon content in megagrams per hectare depending on which of those response variables was being analyzed: when analyzing canopy cover, carbon content was included as a predictor and vice versa.

Parameters for the random forest analysis were set to an ntree (number of trees in the random forest) of 500 and an mtry (number of variables tested at each split) of 3. Plotting error rates for each variable showed stabilization after approximately 100 trees, so an ntree of 500 was used to ensure the analysis was performed on stabilized trees. The default mtry is equal to the square root of the number of observations, which in this case with 26 predictors would be around 5. Given the chance of multicollinearity with climate variables, a more conservative mtry was used to lessen the likelihood of overfitting.

A random forest analysis was performed on all plots for both response variables, then an additional random forest analysis was performed where the training data was split into forest plots and non-forest plots and tested for both response variables. Each of these six analyses were performed twenty times: the results shown for each analysis include the mean predictions and standard errors across those twenty iterations.

Results

The random forest analyses showed some interesting patterns, but ultimately the relationship between the predictor and response variables for most models was insufficient to explain variation in tree canopy cover or carbon content at the plot level. The highest percentage of variation explained in any of the analyses was 62.1%, followed by 14.3%, and the remaining four analyses each resulted in a negative percentage of variation explained.

Predictor Variable Importance

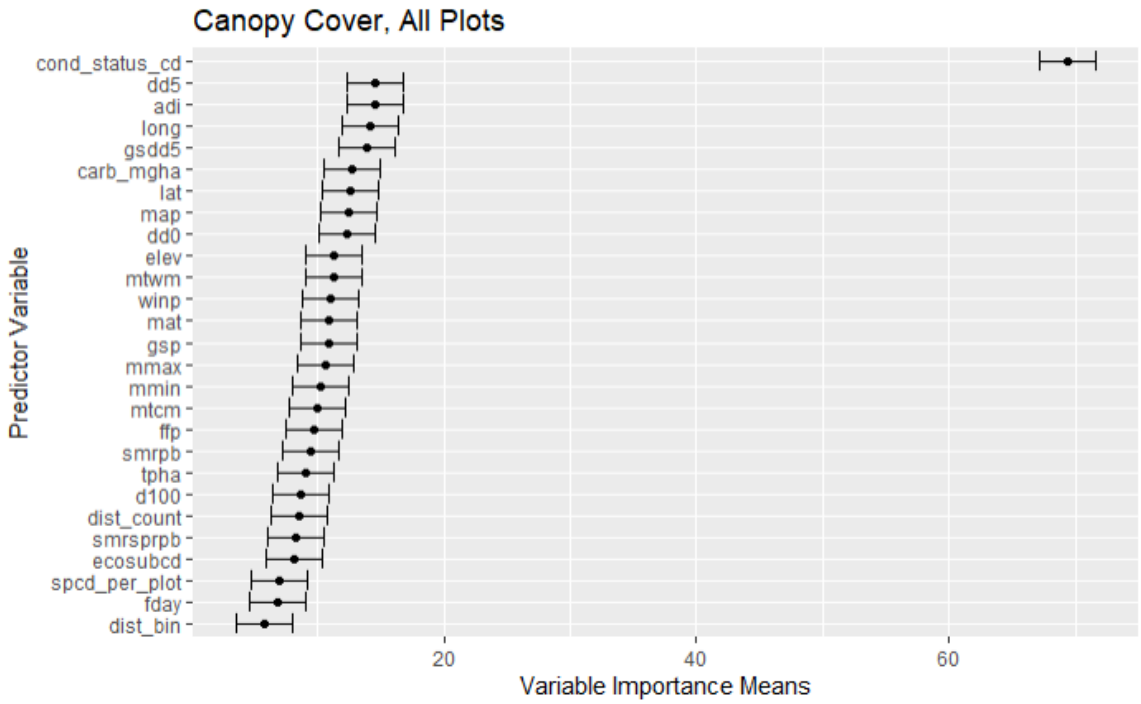
Canopy Cover

The random forest analysis for canopy cover using all plots was by far the highest performing model with a mean percentage of variation explained of 62.12% and a mean squared residuals value (MSR) of 366.28. One predictor variable, **cond_status_cd**, which is a binary indicator of whether a plot is classified as forest or nonforest, accounts for most of the explanatory power of this analysis. This result prompted the splitting of the training data further into forest and non-forest to test if other predictors would perform better under one or the other conditions.

Figure 6 shows the results of running this random forest analysis twenty times and documenting the percent variation explained for each variable at each iteration. No variable besides **cond_status_cd** showed clear explanatory power for canopy cover on all plots: the remaining 25 variables shuffled in order across the twenty iterations with only slightly higher variation explained by the **dd5** (degree days above 5°C) and **adi** (annual dryness index.)

Figure 6: Predictor Variable Importance for Percent Live Canopy Cover

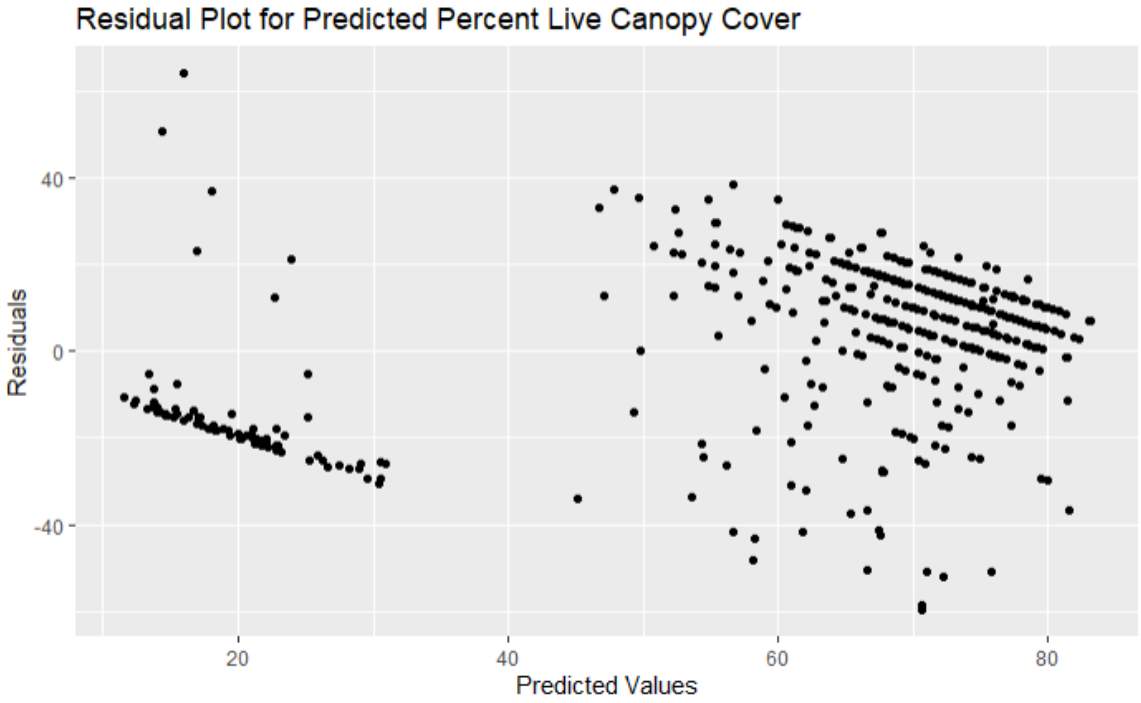
Variable importance plot means of twenty random forest analyses of canopy cover. Error bars derived from the mean standard error for each variable.



As this was the highest performing model, a residual plot was produced for just this analysis. The residuals varied widely, from around +60 to -60 percent. Given that the response variable is a percentage out of 100, residual variation at that scale showed the predictions to be notably unreliable. The striping pattern in the residual plot may be an artifact of how canopy cover in the FIA database is recorded in increments of five percent.

Figure 7: Residual Plot for Predicted Percent Live Canopy Cover

Plot of residual values between random forest prediction of the test data subset and the actual test data subset points.



The forested subset of the canopy cover test data resulted in a mean percentage of variation explained of -7.96 % and a MSR of 61.06. Three predictors did consistently appear as the top three variables: out of 20 iterations, longitude was in the top three 13 times, **adi** 12 times, and **dd5** 12 times. The next highest number of appearances for a single variable was 6, so these three seem to perform consistently better than the other predictors.

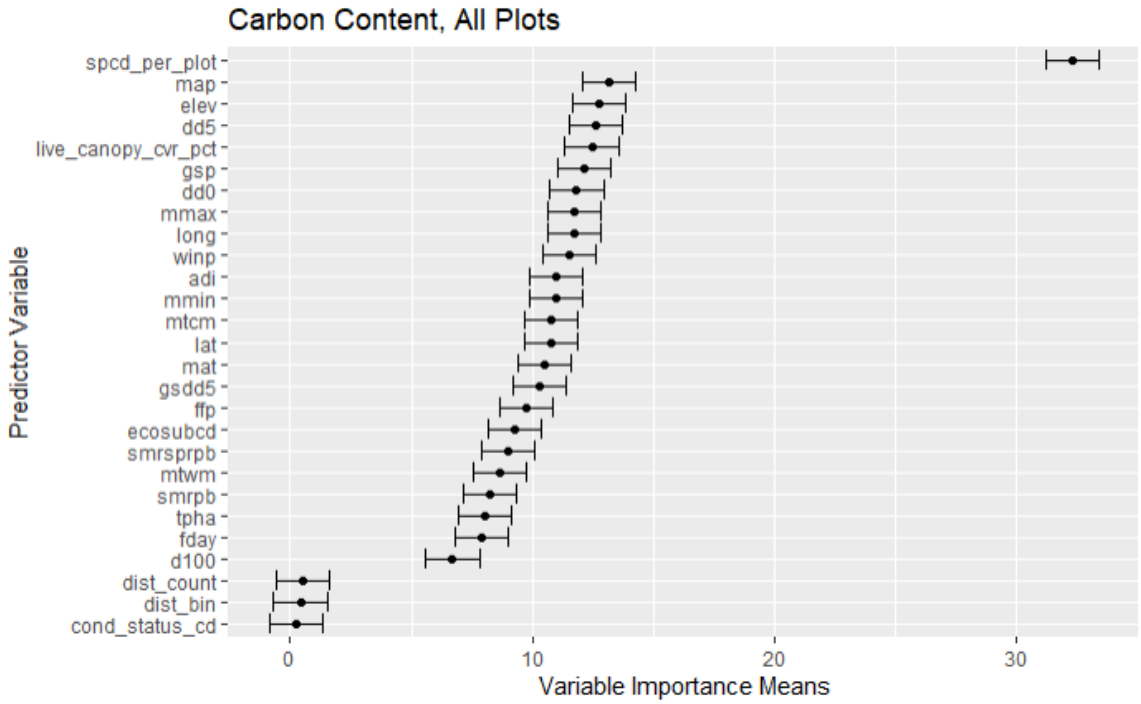
The random forest analysis of the nonforest subset of canopy cover test data yielded a percentage of variation explained of -7.6% and an MSR of 60.85. Here, two predictors consistently appeared in the top three: longitude appeared 14 times and **adi** appeared 10 times. The variable **dd5** only appeared in the top three just 5 times out of 20, which was an interesting difference considering its comparatively better performance in the forested subset.

Carbon Content

The random forest analysis examining carbon content across all plots was the second-best performing model, with a variation explained of 14.34% and an MSR of 76.34. Similar to the canopy cover/all plots analysis, there was one variable that was the top predictor by a notable margin in every iteration. For carbon content, that variable is **spcd_per_plot**, an indicator of tree species diversity at the plot level. Figure 8 shows **spcd_per_plot** performing significantly better than the rest of the predictors, a few predictors performing consistently worse than the others, and most of the predictors lumped in the middle, where they shuffled positions over the 20 iterations.

Figure 8: Predictor Variable Importance for Carbon Content (mg/ha)

Variable importance plot means of twenty random forest analyses of carbon content. Error bars derived from the mean standard error for each variable.



The forested subset of carbon content test data produced a variation explained of -6.69% and an MSR of 94.78. As in the all plots carbon content analysis, spcd_per_plot held the number one predictor spot in all twenty iterations. The **mmin** variable appeared in the top three 13 times, and no other variables regularly appeared in the top 3.

The carbon content nonforest subset resulted in a variation explained of -6.39% and an MSR of 94.71. Similar to the forested carbon content subset, **spcd_per_plot** was the top predictor across every iteration and the only other predictor that was consistently in the top three was **mmin** (this time, it appeared in 12 of the twenty iterations.)

Discussion

Random forest analysis of carbon content and canopy cover at the plot level using local climate data struggled to reliably predict these attributes. While the results indicate an overall inability to accurately predict canopy cover and carbon content, some predictors consistently appeared as the most explanatory and, therefore, may offer some insight into what factors are most important to consider to evaluate trends across the prairie-forest gradient.

The clear best predictors for canopy cover and carbon content were **cond_status_cd** (forest or non-forest indicator) and **spcd_per_plot** (tree species diversity) respectively. Tree species diversity was not normalized by the number of trees present on a plot, which could be a confounding factor, but the related variable of trees per hectare did not perform well in the random forest analyses, which is a hint that the species diversity is important regardless.

The relationship between the percentage of sky covered by canopy and whether a plot is categorized as forest or non-forest seems intuitive: forests have more canopy cover than non-forest. It was interesting, though, that the same intuitive notion does not hold for carbon content: **cond_status_cd** was in fact reliably one of the worst predictors for carbon content at the plot level. Because carbon content was derived from the per-tree measurement of aboveground carbon estimates in the FIA dataset, it likely is not the case

that this poor performance of **cond_status_cd** can be explained by other vegetation making up the difference.

The consistency with which **mmin** (mean minimum annual temperature) appeared as a secondary predictor for carbon content and longitude/annual dryness index appeared as secondary predictors for canopy cover also may bear further investigation. Mean minimum temperature is a natural predictor for carbon content in that it is a factor in the well-documented latitudinal gradient of biodiversity (Willig 2003). It may be more predictive than the similar measures of mean maximum temperature or number of degree days above zero in that some tree species do not thrive when winters dip below certain minimum temperatures (Schenker 2014.)

Longitude is of particular interest given the longitudinal transect shape – while latitude as proximity to the equator is generally more predictive of ecosystem characteristics, the transect here with its small range of latitudes and its clear longitudinal ecosystem changes as it crosses the PFB may explain why longitude appeared as a comparatively good predictor and latitude did not.

The prairie-forest ecotone is characterized by a shift in precipitation-to-evaporation ratio (Frelich 2010), where more evaporation and a subsequently drier climate is likely to give rise to prairie and less evaporation more likely to produce forest. The presence of the annual dryness index as a comparatively strong predictor for canopy cover then lends confidence that despite the overall low predictive power of the analyses, they may still be picking up on genuine ecological trends. The annual dryness index, for this reason, and mean minimum annual temperature, given the relationship between vegetation growth and freezing temperatures, may be regional effects, while longitude, forest status, and tree species diversity may still perform well as predictors if a similarly shaped transect were created in another region.

Some alterations that could be made to the random forest analysis to build better models in the future might be a larger sample size through increasing the temporal and/or geographic range of the data, making predictions at the tree level instead of plot level, or

using a more comprehensive vegetation survey that examines the impact of non-tree vegetation.

One concern in using FIA data is the practice of plot perturbance: in order to protect the privacy of individual plot locations, many FIA plot locations are perturbed either by having the location fuzzed or have the data swapped with another FIA plot in the same county (Coulston et al 2006). With point predictions of climate data, the perturbance of plot locations was a concern at first, but Sabor et al. (2007) determined that plot perturbance should generally only be a concern if data is both high resolution and exhibits low spatial autocorrelation. The climate data here is high resolution, but has high spatial autocorrelation, so plots being swapped within a small geographic area is not a major concern. Further studies should ensure that data does not have low spatial autocorrelation if using perturbed plots.

The high spatial autocorrelation that was observed is a natural characteristic of climate data: locations that are very close to each other tend to have a similar climate. This was one reason that despite the minimal results from the random forest, the transect size was not increased to include more data. Increasing the transect north or south risks the interference of latitudinal climate gradients, while increasing east or west risks muddling the prairie-forest border focus as you include more and more ecoregions and the variation they bring. Additionally, regional characteristics must be considered: this transect in the Northern US features quite cold winter temperatures. If a new transect was taken in another location, perhaps the southern edge of the PFB, canopy cover and carbon content may exhibit different patterns without the harsh winters.

Last, another potential avenue to further this analysis would be to include more variables: particularly, if the analysis were done on a per-tree basis, information like basal area, tree height, and species characteristics could be incorporated. Auxiliary data from remote sensing could provide land and tree characteristics not captured in the FIA dataset, and in particular could provide data on non-forest areas, which were somewhat lacking overall in the FIA dataset. More complete data could bolster (or upend) the

patterns observed in this analysis. Regardless, additional research on TOF will help better our understanding of this understudied resource.

Bibliography

Cadanesso M. L., Pickett S. T. A., Weathers K. C., Bell S. S., Benning T. L., Carreiro M. M., and Dawson T. E. 2003. "An interdisciplinary and synthetic approach to ecological boundaries. *BioScience* 53: 717-722

Cleland, D.T.; Freeouf, J.A.; Keys, J.E.; Nowacki, G.J.; Carpenter, C.A.; and McNab, W.H. 2007. Ecological Subregions: Sections and Subsections for the conterminous United States. Gen. Tech. Report WO-76D [Map on CD-ROM] (A.M. Sloan, cartographer). Washington, DC:: U.S. Department of Agriculture, Forest Service, presentation scale 1:3,500,000; colored.

Coulston, J.W.; Riitters, K.H.; McRoberts, R.E.; Reams, G.A.; and Smith, W.D. "True versus perturbed forest inventory plots for modeling: a simulation study." Canadian Journal of Forest Research 36:801-807. [doi:10.1139/X05-265]

Danz, N.P., Frelich L. E., Reich, P. B., and Niemi, G. J. 2013. "Abrupt prairie-forest transition across a smooth climate gradient in presettlement Minnesota, USA." *Journal of Vegetation Science* 24:1129-1140. [doi: 10.1111/jvs.12028]

Danz, N.P., Reich, P. B., Frelich, L. E., and Niemi, G. J. 2011. "Vegetation controls vary across space and spatial scale in a historic grassland-forest biome boundary." *Ecography* 34:402-414. [doi: 10.1111/j.1600-0587.2010.06561.x]

Freeman E. A., Moisen G. G., Coulston J. W., and Wilson B. T. 2014. "Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance." Canadian Journal of Forest Research 46: 323-339. [doi: 10.1139/cjfr-2014-0562]

Frelich, L. E. and Reich, P. B. (2010). "Will environmental changes reinforce the impact of global warming on the prairie–forest border of central North America?" Frontiers in Ecology and the Environment, 8(7), 371–378. [doi:10.1890/080191]

Iverson L. R. and Prasad A. M. 1998. "Predicting abundance of 80 tree species following climate change in the eastern united states." Ecological Monographs, 68(4), 465–485. [doi:10.1890/0012-9615(1998)068[0465:paotsf]2.0.co;2]

Kellerman, T., Benegas J., Meneguzzo D., and Liknes G. 2019. "Making Trees Outside Forests Count." U.S. Dept of Agriculture Forest Service National Agroforestry Center Story Map.

Kirby K. R. and Potvin C. 2007. "Variation in carbon storage among tree species: Implications for the manangement of a small-scale carbon sink project." Forest Ecology and Management 246: 208-221. [doi: 10.1016/j.foreco.2007.03.072]

Lin D., Anderson-Teixeira K. J., Lai J., Mi X., Ren H., Ma K. 2016. "Traits of dominant tree species predict local scale variation in forest aboveground and topsoil carbon stocks." Plant Soil 409: 435-446. [doi: 10.1007/s11104-016-2967-0]

MacFarland, K. 2020. “Why the trees outside forests count.” US Department of Agriculture Forestry Blog. https://www.usda.gov/media/blog/2020/03/10/why-trees-outside-forests-count

Meneguzzo, D. M., Liknes, G. C., and Nelson, M. D. 2013. “Mapping trees outside forests using high-resolution aerial imagery: a comparison of pixel- and object based classification approaches.” *Environmental Monitoring and Assessment*. 185: 6261-6275.

Rodman H., Clough B., Rutenbeck N., and Pond N. 2020. tidyFIA: Assemble Forest Inventory and Analysis (FIA) Data for Analysis. R package version 0.12. https://github.com/SilviaTerra/tidyFIA

Rossi J.P., Garcia J., Roques A. and Rousselet J. 2016. “Trees outside forests in agricultural landscapes: spatial distribution and impact on habitat connectivity for forest organisms”. *Landscape Ecology* 31: 243-254.

Sabor A. A., Radeloff V.C., McRoberts R.E., Clayton M., and Stewart S.I. 2007. “Adding uncertainty to forest inventory plot locations: effects on analyses using geospatial data.” Canadian Journal of Forest Research, 37: 2313-2325. [doi:10.1139/X07-067]

Schenker G., Lenz A., Korner C. and Hoch G. 2014. “Physiological minimum temperatures for root growth in seven common European broad-leaved tree species.” Tree Physiology, 34(3), 302–313. [doi:10.1093/treephys/tpu003]

USDA Forest Service, 2021. “Forest Inventory and Analysis: About Us.” https://www.fia.fs.fed.us/about/about_us/

USDA Forest Service, 2021. “Trees Outside Forests Image-based Inventory (TOFii).” https://www.nrs.fs.fed.us/inventory_monitoring/monitoring_assessment/trees-outside-of-forests/

USDA Forest Service, 2018. “FIADB Description and User Guide”, v8.0, Section 2.4.97

USDA Forest Service, 2018. “FIADB Description and User Guide”, v8.0, Section 3.1.121

USDA Forest Service, 2005. “Forest Inventory and Analysis: Sampling and Plot Design”. https://www.fia.fs.fed.us/library/fact-sheets/data-collections/Sampling%20and%20Plot%20Design.pdf

USDA-Virginia Tech, 2021. “Custom Climate Data Requests.” *Climate Estimates and Plant-Climate Relationships.* http://charcoal.cnre.vt.edu/climate/customData/index.php accessed 1 September 2021.

Van Deusen, P.C. “Forest inventory estimation with mapped plots.” 2004. Canadian Journal of Forest Research 34: 493-497. [doi:10.1139/X03-209]

Westfall, J.A.; Woodall, C.W.; and Hatfield, M.A. “A statistical power analysis of woody carbon flux from forest inventory data.” 2013. Climatic Change 118:919-931. [doi:10.1007/s10584-012-0686-z]

Willig M.R., Kaufman D.M.; Stevens R.D. (2003). “Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis.” Annual Review of Ecology, Evolution, and Systematics 34(1), 273–309. [doi:10.1146/annurev.ecolsys.34.012103.144032]

Wyckoff P. H. and Bowers R. 2010. “Response of the prairie–forest border to climate change: impacts of increasing drought may be mitigated by increasing CO2.” Journal of Ecology 98(1), 197–208. [doi:10.1111/j.1365-2745.2009.01602.x]