Not All Biomass is Created Equal:
An Assessment of Social and Biophysical Factors Constraining Wood Availability in Virginia

Pamela Hope Braff

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Forest Resources and Environmental Conservation

Stephen P. Prisley
Philip J. Radtke
John Coulston

April 24, 2013
Blacksburg, VA

Keywords: wood availability, forest inventory and analysis (FIA), classification and regression trees, random forest, logistic regression

Not All Biomass is Created Equal:
An Assessment of Social and Biophysical Factors Constraining Wood Availability in Virginia

Pamela Hope Braff

Most estimates of wood supply do not reflect the true availability of wood resources. The availability of wood resources ultimately depends on collective wood harvesting decisions across the landscape. Both social and biophysical constraints impact harvesting decisions and thus the availability of wood resources. While most constraints do not completely inhibit harvesting, they may significantly reduce the probability of harvest. Realistic assessments of woody availability and distribution are needed for effective forest management and planning. This study focuses on predicting the probability of harvest at forested FIA plot locations in Virginia. Classification and regression trees, conditional inferences trees, random forest, balanced random forest, conditional random forest, and logistic regression models were built to predict harvest as a function of social and biophysical availability constraints. All of the models were evaluated and compared to identify important variables constraining harvest, predict future harvests, and estimate the available wood supply. Variables related to population and resource quality seem to be the best predictors of future harvest. The balanced random forest and logistic regressions models are recommended for predicting future harvests. The balanced random forest model is the best predictor, while the logistic regression model can be most easily shared and replicated. Both models were applied to predict harvest at recently measured FIA plots. Based on the probability of harvest, we estimate that between 2012 and 2017, 10 – 21 percent of total wood volume on timberland will be available for harvesting.

DEDICATION

To my Zeyda for his unconditional love and support.

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Steve Prisley for his mentorship and guidance throughout this entire process. I especially appreciate his constant encouragement and enthusiasm for my research and other academic interests.

I would also like to thank my other committee members, Dr. Phil Radtke for his helpful advice early on and for the introduction to Random Forest in his Forest Analytics course as well as Dr. John Coulston for his help understanding and using FIA data.

Special thanks go to Sam Lambert at the Southern Research Station, FIA Spatial Data Services for his assistance with data collection.

Table of Contents

List of Figures

List of Tables

Chapter 1. INTRODUCTION

There are more than 15.9 million acres of forestland in Virginia (Virginia Department of Forestry, 2012) (Figure 1). Virginia's forests provide many valuable goods and services with economic, social and ecological significance. Forests provide critical ecosystem services such as carbon sequestration and water filtration as well as opportunities for recreation and enjoyment. When extracted, wood resources support a variety of forest product industries with cultural and economic importance. Although forest product industries support many communities in the rural South, including Virginia, some studies have raised concerns about their future under current market conditions (Wear et al. 2007). In order to appropriately manage Virginia's forestland to sustain important ecosystem services and support forest product industries, it is essential to understand the distribution and availability of wood resources.

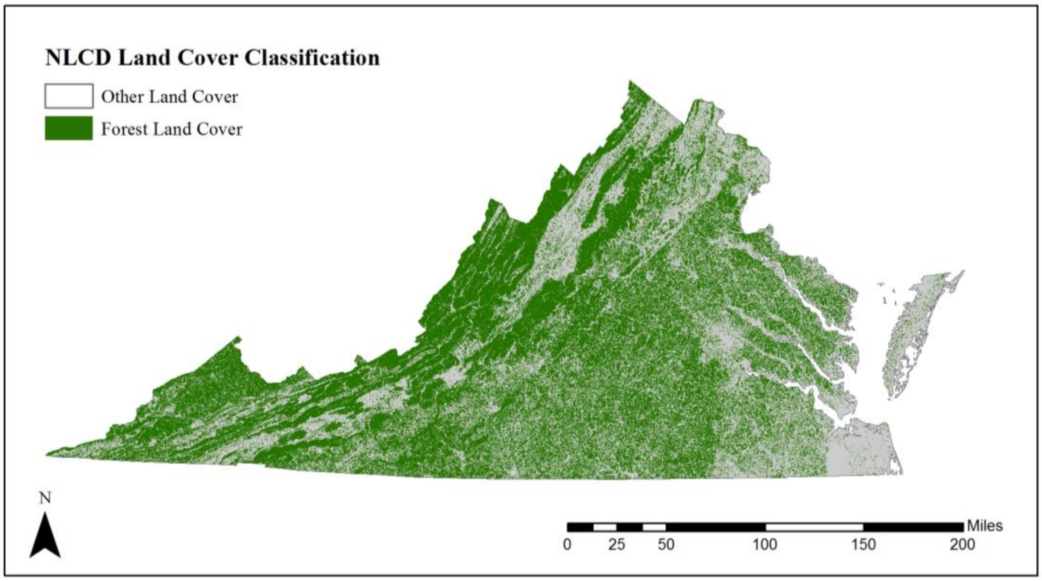

Figure 1. Forest cover in Virginia, data from National Land Cover Database (Fry J. et al. 2011).

Not all 15.9 million acres of Virginia's forestland are equally available for harvest. The available wood supply is constrained by both biophysical and social factors (Dennis 1990; Butler et al. 2010). Biophysical attributes such as timber stand quality/quantity impact the value of harvesting, while the natural setting may constrain resource access. Additionally, social constraints can impact the likelihood of a landowner choosing to make a (wood) resource available (i.e. choose to harvest). It is important to distinguish between the total and available wood supply, as they will rarely, if ever be equal (Butler et al. 2010).

Some recent estimates of biomass supply have accounted for availability constraints by first estimating the aggregate (total) supply and then applying some reduction rate to estimate the available supply. Following this general method, Perlack et al. (2005) estimated that in the United States, there are 8.4 billion dry tons of woody biomass suitable for bioenergy or biobased products. They then applied reduction rates, varying by ownership, to this initial estimate in order to account for accessibility and removal constraints. Perlack et al. (2005) estimated that out of 8.4 billion dry tons of suitable resources, no more than 60 million dry tons could be available for removals annually. This is a substantial reduction, three orders of magnitude smaller than their initial supply estimate.

In a similar fashion, Markowski-Lindsay et al. (2012) estimated biomass availability in Massachusetts by first determining total acreage of forestland. They then reduced this acreage based on the percentage of owners expected to harvest residual woody biomass (i.e. social availability). The resulting acreage was then further reduced to represent biophysical accessibility constraints. Markowski-Lindsay et al. (2012) estimated that 80,000 – 369,000 dry tons/year of woody biomass would be available from private and state owned forests. They emphasize that by incorporating social constraints their estimate is much smaller than a previous biomass availability estimate of 891,000 dry tons/year (Kelty et al. 2008).

Although Perlack et al. (2005) and Markowski-Lindsay et al. (2012) only provide rough estimates of biomass availability; both studies clearly indicate large differences between total and available supplies. There are a few notable issues with their estimation techniques that inhibit a more detailed assessment of availability. First, they treat social and biophysical constraints as independent events with no influence on each other. However, the combined impact of social and biophysical constraints could likely lead to increased or decreased availability estimates. For example, landowner intentions are irrelevant if the stand is not biophysically available, however in other situations high demand for wood resources may motivate new landowners to harvest. It is necessary to consider all constraints collectively because independent estimates do not account for such influence (Brinckman and Munsell 2012).

Secondly, both studies applied reduction rates to aggregate supply estimates. The factors, which constrain harvest, may vary greatly across the landscape and cannot be well represented with a single value or reduction rate. Furthermore, it is impossible to account for the combined influence of multiple constraints without knowing location-specific conditions. For a more realistic and detailed availability assessment, estimates must account for all location-specific harvesting constraints before estimating the aggregate supply.

Butler et al. (2010) considered location-specific constraints to estimate the percent of total wood available from family owned forests in the northern United States. Rather than reducing an aggregate supply as done by Perlack et al. (2005) and Markowski -Lindsay et al. (2012), they estimated woody availability of specific locations using forest inventory plots. They developed unique, plot dependent reduction rates based on the type and number of constraints at each plot. Multiple constraints progressively reduced the probability of harvest. To estimate the percent of available supply, wood weight was estimated at each plot and then reduced by the determined reduction rate. After accounting for social and biophysical constraints, Butler et al. (2010) estimated that only 38 percent of the total wood supply is available to be extracted. This estimate completely depends on the variables and thresholds selected to constrain harvest, which were somewhat arbitrarily determined, based on expert opinion. Depending on the selected threshold, availability estimates ranged from 18 – 60 percent. Although Butler et al. (2010) carefully selected variables to represent theoretical constraints described in the literature, they acknowledge that more research is needed to accurately select appropriate variables and define realistic constraint thresholds.

In an earlier study, Prestemon and Wear (2000) also investigated the impact of harvesting constraints at forest inventory plots for North Carolina coastal plain southern pine stands. However, rather than of estimating the percent of wood available based on theoretical availability constraints, Prestemon and Wear (2000) developed an empirical model to predict harvesting behavior at forest inventory plots. They then applied area expansion factors to the predicted harvest probability and representative plot volume to estimate the available supply and develop econometric supply curves. This approach directly reflects the aggregation of individual harvesting decisions across the landscape (Prestemon and Wear 2000).

Harvesting behavior ultimately determines wood resource availability, but it is unclear how well estimated reduction rates applied in recent availability estimates (e.g. Perlack et al. 2005; Butler et al. 2010; and Markowski-Lindsay et al. 2012) relate to observed harvests. Despite applying a variety of different approaches, all of these studies still indicate that depending on social and biophysical constraints, actual wood availability is much less than total wood supply estimates may indicate. The composition of wood resources is as important, if not more important than resource quantity (Prestemon and Wear 2000). Prestemon and Wear (2000) show that wood supply can be estimated by relating decision criteria to harvest occurrences. If harvest occurrences could be predicted on a larger scale, than those predictions could be used to estimate wood supply and inform conservation and industry management decisions.

The main objective of this study was to model the probability of harvest at permanent Forest Inventory and Analysis (FIA) plots in Virginia. Harvest probability models were built as a function of social and biophysical plot characteristics using both parametric and non-parametric modeling approaches. The second objective of this study was to use these models to identify important variables constraining harvest, predict future harvests and estimate the available wood supply in Virginia. This research can inform forest management by describing which factors impact harvest probability and the extent to which they constrain wood supply availability. Furthermore, the methods developed can provide a framework for future estimates of harvest probability and wood supply availability in other states/regions.

Chapter 2. LITERATURE REVIEW

Harvesting Behavior

Although few studies have evaluated the impact of harvesting behavior on wood availability, variables related to harvesting behavior have been extensively examined. Previous research has identified many variables correlated to harvest occurrences (Table 1). However, the variables examined and their reported impact on harvesting is generally inconsistent. Most studies of harvesting behavior have focused on small scales (i.e. town or county level). The variation among previous harvesting studies is likely due to differences in location and available data. Although the findings from these past studies cannot be widely applied, they provide a great foundation of knowledge regarding social and biophysical harvesting constraints. If harvesting behavior could be effectively predicted on a larger scale, those predictions could be used to compute regional availability estimates following the general methodology described by Prestemon and Wear (2000).

Table 1. Factors correlated to harvest as indicated by previous research.

Variable	Correlation	Study
Biophysical Constraints		
Tract/Plot Size	Positive	Dennis (1989); Cleaves and Bennet (1995); Cleaves et al. (2002); Conway et al. – Central VA (2000); Hyberg and Holthausen (1989); Jamnick and Beckett (1988); Kuuluvainen et al. (1996); Størdal et al. (2008); Wear et al. (2003)
	None	Dennis (1990); Conway et al. – Southwest VA (2000)
Timber Volume	Positive	Barlow et al. (1998); Dennis (1989); Jamnick and Beckett (1988)
Valuable Species	Positive	Dennis (1990)
Slope	Negative	Wear and Flamm (1993); Wear et al. (1999)
Social Constraints		
Population density	Negative	Barlow et al. (1998); Wear et al. (1999)
Distance to urban center	Positive	Barlow et al. (1998); Cleaves et al. (2002)
Road Access	Varied	Barlow et al. (1998); Wear and Flamm (1993)
Local Regulations	Negative	Paula et al. (2011)
Endangered Species	Positive	Zhang (2004)
Ownership Type	Management Differences	Alig and Wear (1992); Barlow et al. (1998); Prestemon and Wear (2000); Munn and Arano (2006); Wear and Flamm (1993)
Timber Price	None	Dennis (1989); Dennis (1990); Kuuluvainen et al. (1996); Prestemon and Wear (2000)
	Positive	Binkley (1981); Boyd (1984); Kuuluvainen and Salo (1991)
	Negative	Hyberg and Holthausen (1989)

Table 1 continued. Factors correlated to harvest as indicated by previous research.

Variable	Correlation	Study
Social Constraints - Nonindustrial Private Owners Only		
Income	Negative	Dennis (1989); Dennis (1990); Jamnick and Beckett (1988); Romm et al. (1987); Wear et al. (2003)
	None	Boyd (1984); Kuuluvainen et al. (1996)
Non Forest Related Wealth	Negative	Hyberg and Holthausen (1989); Størdal et al. (2008)
Debt	Positive	Størdal et al. (2008); Wear et al. (2003)
Age	Negative	Romm et al. (1987); Størdal et al. (2008)
Education	Positive	Dennis (1989); Størdal et al. (2008)
	None	Binkley (1981); Boyd (1984)
Length of Ownership	Positive	Vokoun et al. (2006)

Forest Inventory and Analysis

Previous wood availability studies have used Forest Inventory and Analysis (FIA) data to predict harvest occurrences (Prestemon and Wear 2000) and estimate percent availability (Butler et al. 2010). The FIA program provides the principal source of landscape level and real time information about the nation's forest resources(Reams et al. 1999; Smith 2002). FIA consists of permanently established inventory plots, distributed based on a systematic grid design. Approximately 20 percent of FIA plots from eastern states and 10 percent of plots from western states are surveyed annually to report on the status and trends of US forests. Field crews collect data from any partially forested ground plots (phase 2); forested areas must meet minimum stocking (10 percent) and size (1 acre, 120 feet) requirements (Smith 2002; Bechtold and Patterson 2005).

Each plot consists of one or more conditions defined by differences in land use and vegetation. Any differences in ownership, forest type, stand size, stand density, regeneration or reserved status creates a distinct plot condition (Figure 2). Many FIA variables are collected at the condition level and only reported for forested conditions (Forest Inventory and Analysis 2013b).

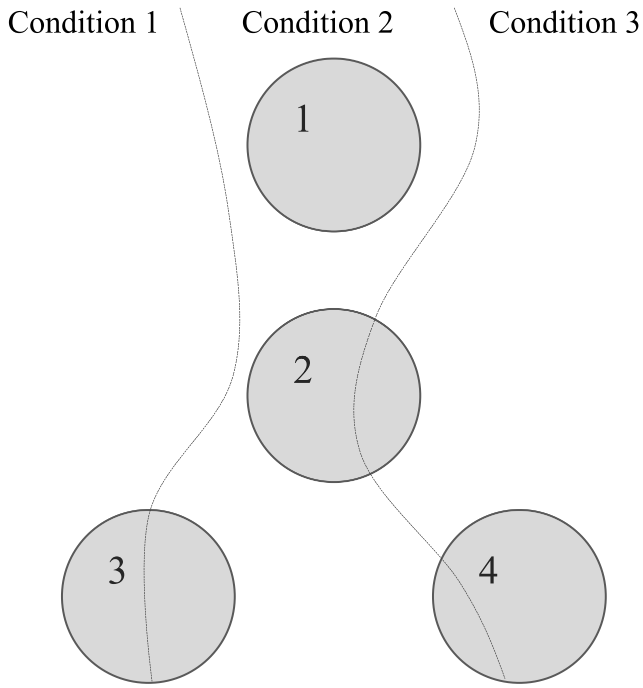

Figure 2. FIA plot conditions. Differences in land use or vegetation define conditions, which are mapped and recorded. Figure adapted from Forest Inventory and Analysis (2013b).

Publically available FIA data can be applied to various national and regional natural resource assessments(Reams et al. 1999). However, for a variety of reasons related to privacy and ecological/survey integrity concerns, the exact coordinates of plot locations and some ownership information is not publically available. Published FIA plot coordinates are perturbed up to one mile (generally within ½ mile) and up to 20 percent of plots are swapped (generally within the same county)(McRoberts et al. 2005; Woudenberg et al. 2010). This process purposely introduces uncertainty into the published FIA data (McRoberts et al. 2005). While this uncertainty should not impact estimates over large areas (FIA, 2013b), publicly available data may not be suitable for analyses requiring fine-scale spatial data (Prisley et al. 2009).

Model Descriptions

Classification and Regression Trees

Classification and Regression Trees (CART), developed by Breiman et al. (1984), is a nonparametric statistical model widely applied in variety of disciplines (Vayssières et al. 2000). CART models the response of a categorical (classification tree) or continuous (regression tree) response variable from one or more explanatory variables, which may also be categorical or continuous. Through recursive portioning, the data set is repeatedly split based on the explanatory variable which best divides the response into homogenous subgroups.

CART models consist of a collection of decision rules to group the response based on the explanatory variables (Breiman et al., 1984). Decision trees, interpreted much like dichotomous keys, are a graphical representation of the partitioning rules (Vayssières et al. 2000). Beginning with a single root node, each consecutive node represents a decision rule by which the data are split between branches. The terminal nodes (leaves) represent the final groups once all responses are classified (De'Ath and Fabricius 2000; Vayssières et al. 2000). After the largest possible tree is grown (e.g. each terminal node represents a single response class), the tree must be "pruned" to remove redundant and excessive branches (Breiman et al. 1984).

Unlike traditional parametric statistics, such as logistic regression, CART is well suited to model complex data with nonlinear relationships and high-order interactions (De'Ath and Fabricius 2000). Previous research has shown that classification trees better detect variable interactions and classify predictions as well as, if not better, than traditional logistic regression approaches (Vayssières et al. 2000). One disadvantage of CART is that small changes to the input data could result in major changes to the decision tree. With advances in machine learning, single-tree methods such as CART have been improved by using combinations (ensembles) of trees (Cutler et al. 2007).

Random Forest

Random forest (RF) is a basic ensemble tree method based on CART (Breiman 2001). The algorithm first randomly selects a bootstrap sample of observations with replacement, to build a decision tree. At each node only a random sample of explanatory variables are available to split the data. Each tree is grown as in CART, except they are grown to their full extent without pruning. The remaining un-sampled observations, known as *out-of-bag* data, are then used to determine the decision tree’s classification error. Each tree classifies the response and contributes their “vote” to the final prediction. The ensemble of trees can either select the response which most trees voted for or report the percent of trees which voted for each class (Breiman 2001).

By growing a large number of trees RF can greatly increase classification accuracy. The algorithm is less sensitive to noisy data or outliers and is far less likely to overfit the model, especially as more trees are added (Breiman 2001). RF and CART are both appropriate modeling techniques when there may be complex interactions between explanatory variables. Unlike linear techniques that remove highly correlated predictors, even if they are good predictors, classification trees distribute importance and maintain all contributing variables (Cutler et al. 2007). RF has been shown to be more effective than most commonly used methods (i.e. linear discriminant analysis, logistic regression and classification trees) to model presence/absence data (Cutler et al. 2007).

Unlike CART, which provides a straightforward representation of decision rules, RF consists of many decision trees, which cannot be simultaneously displayed or assessed. While the exact relationship between response and predictor variables is unknown, variable importance measures and partial dependence plots can help characterize variable relationships (Breiman 2001; Cutler et al. 2007). Mean decrease in accuracy, a common variable importance measure, compares the classification accuracies of *out-of-bag* data between true values and randomly permuted values for each predictor variable. The reported importance value is the normalized difference between the misclassification rates (Cutler et al. 2007).

RF minimizes overall error, and therefore may not classify rare cases well if the data are extremely imbalanced. In such cases either a weighted or balanced RF can be applied to solve data imbalance issues. Balanced RF down-samples the majority class and weighted RF increases the penalty of misclassifying the rare class. Both methods work similarly well, however balanced RF (bRF) is more computationally efficient than weighted RF (Chen et al. 2004).

It has recently been shown that RF and CART have a selection bias against categorical variables. Since explanatory variables are selected and split in the same step, continuous variables or categorical variables with many classes provide a greater number of splitting options, thereby increasing their change of being selected for the decision tree (Hothorn et al. 2006; Strobl et al. 2007). Mean decrease in accuracy variable importance measures may also overestimate the importance of highly correlated variables. Since RF and CART are often used to infer variable relationships and select important variables, it is important to consider if variable selection is biased (Strobl et al. 2007).

Conditional Inference Trees

In response to concerns about CART and RF selection bias, Horthern et al. (2006) developed an unbiased variable selection process known as conditional inference trees (cTREE). cTREE follows the same general process described for CART, except cTREE separates variable selection and splitting into 2 distinct steps. First, the split variable is selected based on chi square significance tests between the response and explanatory variables. Once the predictor with the strongest association to the response is selected, an optimal split is chosen for that variable. Any split method, such as those used by CART and RF, can be applied. cTREE applies predetermined statistical rules to stop tree growth. Since split variables are only selected as long as they have a minimum association to the response variable, unlike CART, cTREE does not need to be pruned in order to avoid overfitting (Hothern et al. 2006).

Just as RF is the ensemble analog of CART, conditional random Forest (cRF) is the ensemble analog of cTREE. Despite structural differences in data partitioning, conditional inference methods have shown to provide equally effective predictions to CART and RF. The main difference is likely to appear in assessments of variable importance (Strobl et al. 2007).

Logistic Regression

Logistic regression (LR) is a regression technique for predicting binary response variables. Similar to common linear regression modeling, LR can be used to model a presence/absence response from a collection of continuous and or categorical variables. The LR model for p independent variables is

$$P(Y = 1) = \frac{1}{1+e^{(\beta 0+ \beta 1x1+ \beta 2x2+\ldots+\beta pxp)}}$$

where P(Y = 1) is the probability of presence and β1, β2,. . . βp are the regression coefficients (Hosmer and Lemeshow 2000).

Model Assessment

Threshold Selection

Perhaps the most important step of model building is the evaluation of model performance (Pearce and Ferrier 2000). Although it is common for presence/absence models to predict the relative probability of presence, many accuracy metrics require a discrete presence/absence response (Manel et al. 2001). Additionally, natural resource managers and decision makers are often interested in a simple presence/absence prediction requiring discretized model output. Therefore, it is important to select a reasonable threshold with which to classify the response as presence or absence.

There are a variety of subjective and objective approaches for selecting classification thresholds. When the response variable is balanced (frequency of presence in the sample is near 50 percent) there is little difference between threshold selection methods. However, in the case of imbalanced response data, the predicted response is weighted towards the larger of the two groups (Fielding and Bell 1997; Cramer 1999). One of the most commonly used thresholds, 0.5, is only effective when the data are balanced. This threshold can often result in illogical classifications, especially if prevalence (frequency of presence) is small. In a comparison of different threshold selection methods, Liu et al. (2005) found that the 0.5 threshold performs poorly compared to many other threshold selection techniques.

Whereas 0.5 is a subjective threshold, there are many objective approaches, which select a threshold to optimize the agreement between observed and predicted/classified responses. Prevalence (the proportion of presences in the model-building data) is one of the most robust thresholds. It often performs very well and, even at its worst, still performs as well as or better than many other thresholds (Liu et al. 2005; Jimenez-Valverde and Lobo 2007). Furthermore, prevalence has also been recognized by statisticians as a theoretically appropriate classification threshold (Cramer 1999). Objective thresholds can also be determined by maximizing some relevant accuracy metric and/or the trade-off between conflicting metrics. Threshold selection methods, which consider both sensitivity and specificity (i.e. sensitivity-specificity sum maximization and difference minimization) often perform well, similar to prevalence (Liu et al. 2005).

Accuracy Measures - *Threshold Dependent*

Once the response has been classified as presence/absence, model performance can be summarized by comparing observed and predicted presences/absences in a confusion matrix (Table 2) (Fielding and Bell 1997; Pearce and Ferrier 2000). Accuracy assessment metrics, which require a presence/absence response, can be derived from the values provided in the confusion matrix (Table 2).

Table 2. Confusion matrix describing the agreement between predicted and observed responses. Each value (A, B, C, D) denotes the number of observations such that A + B + C + D = N.

		Observed	
		Presence	**Absence**
Predicted	**Presence**	A	B
	Absence	C	D

Sensitivity

$$= \frac{\text{Number of correctly predicted presences}}{\text{Total number of presences in the sample}} = \frac{\text{A}}{\text{A} + \text{C}}$$

Specificity

$$= \frac{\text{Number of correctly predicted absences}}{\text{Total number of absences in the sample}} = \frac{\text{D}}{\text{B} + \text{D}}$$

Percent Correctly Classified (PCC)

$$= \frac{\text{Number of correctly predicted presences and absences}}{\text{Total number of observations in the sample}} = \frac{\text{A} + \text{D}}{\text{A} + \text{B} + \text{C} + \text{D}}$$

Bias

$$= \frac{\text{Number of predicted presences}}{\text{Number of observed presences}} = \frac{\text{A} + \text{B}}{\text{A} + \text{C}}$$

Sensitivity measures the percentage of observed presences correctly predicted while specificity measures the percentage of observed absences correctly predicted (Fielding and Bell 1997; Pearce and Ferrier 2000). Sensitivity and specificity are true accuracy measures, because they are not sensitive to prevalence (Pearce and Ferrier 2000).

Percent correctly classified (PCC), also sometimes referred to as accuracy, is another commonly used assessment metric (Liu et al. 2005). Percent correctly classified (PCC) is the percentage of correctly predicted absences and presences. If prevalence is low, PCC could be artificially high due to a large number of true absences (Fielding and Bell 1997). Although often used, PCC can be misleading without knowing the prevalence. For example, a data set with only 5 percent presence could incorrectly classify all presences as absences, but still have 95 percent correctly classified. Without considering additional metrics, PCC can be very misleading (Pearce and Ferrier 2000).

Bias measures whether or not a model tends to over or underestimate presence. Reliable models should predict the correct proportion of presences (e.g. little bias). Biased models will inflate either false positive or false negatives (Pearce and Ferrier 2000; Vaughan and Ormerod 2005).

There are many different model accuracy measures, all with their own strengths and weaknesses. During model assessment it is best to report and evaluate a variety of carefully selected accuracy metrics. It is important to consider the distribution of the response data and the intended use of the model when selecting appropriate accuracy metrics (Fielding and Bell 1997; Cutler et al. 2007).

Accuracy Measures – Threshold Independent

Unlike the previously described accuracy metrics, which require the response to be classified as presence/absence, the receiver operating characteristic (ROC) curve is a threshold independent model assessment technique (Fielding and Bell 1997). The ROC graph compares the true positive rate (sensitivity) plotted on the y-axis to the false positive rate (1 – specificity) plotted on the x-axis. The upper left hand corner of the ROC graph indicates perfect classification (i.e. true positive rate = 1 and false positive rate = 0). To create the ROC curve, the true positive and false positive rates are calculated and plotted for many different classification thresholds (Fawcett 2006). ROC curves represent the ability of a model to discriminate between presences and absences and have been widely applied to compare and assess the performance of presence/absence models. Fielding and Bell (1997) suggest using the ROC curve to compare multiple models because it is not dependent on the confusion matrix and corresponding classification threshold.

Area under the curve (AUC) can be derived from the ROC curve as a threshold independent single unit accuracy measure (Fawcett 2006). Although AUC is a widely reported and standard metric for model comparison, its reliability has be questioned for a variety of reasons (Lobo et al. 2008). In addition to AUC, it is important to interpret the full ROC curve. The shape of the curve describes how predicted probabilities relate to observed presence/absence as well as the tradeoffs between sensitivity and specificity at different thresholds (Pearce and Ferrier 2000; Pontius and Parmentier 2014).

Chapter 3. METHODS

Harvest presence/absence was identified for forested FIA plot conditions and related to social and biophysical characteristics measured on the same plots. Additional variables were derived through overlay analysis with GIS layers including the National Hydrography Dataset (NHD), ESRI StreetMap Premium, Timber Product Output (TPO) reports, the National Land Cover Database (NLCD), census data, the Protected Areas Database of the United State (PADUS) and the National Conservation Easement Database (NCED). Harvest presence/absence was modeled as a function of these variables using classification and regression trees (CART), random forest (RF), balanced random forest (bRF), conditional decision trees (cTREE), conditional random forest (cRF) and logistic regression (LR). These models were used to identify variables with a significant impact on harvest probability. The best models were then selected to predict future harvests and estimate the available wood supply in Virginia.

Data Processing

Variable Selection

Harvesting at Forest Inventory and Analysis (FIA) plots in Virginia was modeled as a function of social and biophysical harvesting constraints. Predictor variables were selected based on previous research regarding harvesting behavior as well as data availability. Special care was taken to select widely available data so this assessment could be easily repeated in other states. After preliminary analyses, 43 variables were selected for further investigation (Table 3). A description of data processing to obtain these variables follows.

Table 3. General description of explanatory variables used in analysis. Condition level variables are denoted by an *, all other variables are based on the FIA plot location. See Appendix A for a more complete description of variables.

	Variable	Description	Data Type	Units/ Basic Description
FIA	RdDist	Distance to closest road	Ord.	1 – 9, 1 = closer
	Water	Water on plot	Cat.	0 = no water
	Reserv	Reserved status	Cat.	0 = not reserved, 1 = reserved
	Own	Ownership type*	Cat.	
	StdAge	Stand age *	Cont.	Years
	StdSz	Stand size*	Ord.	1 – 4, 1 is greater
	SiteCl	Site productivity*	Ord.	1 – 7, 1 is greater
	StdOrg	Type of stand regeneration*	Cat.	0 = natural, 1 = artificial
	Slope	Slope*	Cont.	Percent
	PhysCl	Physiographic class*	Cat.	
	AlStk	Stocking by live trees*	Ord.	1 – 5, 1 is greater
	Operability	Ability to operate logging equipment*	Cat.	0 = none
	NetVol	Net tree volume	Cont.	Cubic feet/ acre
	Dstrb_any	Any disturbances recorded*	Cat.	0 = none
	Dstrb_tot	Total disturbances recorded*	Count	
	Mgmt_harv	Harvest*	Cat.	0 = none
	Mgmt_any	Any management*	Cat.	0 = none
NHD	StrmDist_per	Distance to closest perennial stream	Cont.	Feet
	StrmDist_any	Distance to closest stream	Cont.	Feet

Table 3 continued.

	Variable	Description	Data Type	Units/ Basic Description
StreetMap /TPO	RdDist_any	Distance to closest road	Cont.	Feet
	RdDist_hw	Distance to closest highway	Cont.	Feet
	SA_25mi	Mills in the service area	Count	Total within 25 miles
	SA_50mi	Mills in the service area	Count	Total within 50 miles
	SA_75mi	Mills in the service area	Count	Total within 75 miles
	SA_100mi	Mills in the service area	Count	Total within 100 miles
NLCD	CoverType	Land cover type at plot	Cat.	
	PctFor_dec	Deciduous forest	Cont.	Percent within 10 miles
	PctFor_ev	Evergreen forest	Cont.	Percent within 10 miles
	PctFor_mix	Mixed forest	Cont.	Percent within 10 miles
	PctFor_all	Forest (any type)	Cont.	Percent within 10 miles
	PctDev_open	Developed, open space	Cont.	Percent within 10 miles
	PctDev_low	Developed, low intensity	Cont.	Percent within 10 miles
	PctDev_med	Developed, med. intensity	Cont.	Percent within 10 miles
	PctDev_high	Developed, high intensity	Cont.	Percent within 10 miles
	PctDev_all	Developed (any type)	Cont.	Percent within 10 miles
	PctWater	Open water	Cont.	Percent within 10 miles
Census	PopDens	Population density of census block	Cont.	People / ha
	HousDens	Housing density of census block	Cont.	Housing unit / ha
	UrbanDist	Distance to closest urban area or cluster	Cont.	Miles
	UrbanPGI	Population gravity index	Cont.	Population/ Distance (mi) ^{2}
	UrbanDist_A	Distance to closest urbanized area	Cont.	Miles
Other	PAD	Protected area	Cat.	
	NCED	National conservation easement	Cat.	

FIA Data

The response variable, as well as many predictor variables, was obtained from the FIA Database (FIADB). Virginia FIA data were downloaded from the FIA DataMart (Forest Inventory and Analysis 2013a). All plot conditions that satisfied the following requirements were included in this analysis: 1) Plot sampled with the national plot design, 2) Forestland on the condition, 3) Re-measured at least once. The response variable was derived from the FIA condition treatment codes. The possible treatment codes include cutting, site preparation, artificial regeneration, natural regeneration or other silvicultural treatment. Cutting is defined as "the removal of one or more trees from a stand" (Forest Inventory and Analysis 2013b, p. 66). Up to 3 treatments per condition are reported; if at least one condition was classified as cutting, then that condition was coded as a harvest. If no cutting or treatment was reported, then that condition was coded as no harvest.

Many of the FIA derived predictors investigated in this analysis are directly impacted by timber harvest. For example, if harvest occurs, it is likely that many attributes such as stand age and stand volume will be reduced and will reflect that harvest. Since we are interested in predicting future harvests, rather than describing post-harvest conditions, it is necessary to select predictors from measurements prior to the observed response. The Subplot Condition Change Matrix from the FIADB was used to link the condition response to the plot and condition attributes from the previous measurement. All FIA derived predictors were selected from the measurement prior to the linked response. The average time between measurements was 4.9 years. Predictor variables obtained from the FIADB include distance to road, water conditions, ownership, stand age, stand size, site productivity, stand regeneration, slope, stocking, operability constraints, net tree volume, disturbances, previous harvest, and previous management. The FIADB table and source column for these variables as well as any necessary processing is described in Table 4.

Supplementary Data

In addition to FIA data, predictor variables were obtained from supplementary datasets including Census Population (U.S. Census Bureau 2010b) and Urban Areas data (U.S. Census Bureau, 2010a), the National Land Cover Database (NLCD) (Fry, J. et al. 2011), the National Hydrography Dataset (NHD) (U.S. Environmental Protection Agency 2006), ESRI StreetMap Premium Data (ESRI 2013), Timber Product Output (TPO) Reports (Johnson et al. 2009), the

Protected Areas Database of the U.S. (PADUS) (U.S. Geological Survey, 2010) and the National Conservation Easement Database (NCED) (National Conservation Easement Database 2011). Variables were extracted from these data sets in ArcGIS based the FIA plot locations. This was completed at the USDA Forest Service Southern Research Station using the true plot coordinates for all variables except for those related to percent land cover. Data processing for all supplementary variables is described in Tables 5 – 8.

Table 4. Derivation of predictor variables from FIA data. Table and column names can be found in the FIADB documentation (Forest Inventory and Analysis 2013b).

Variable	FIA Table	FIA Column name	Processing (if needed)
RdDist	Plot	RDDISTCD	n/a
Water	Plot	WATERCD	n/a
Own	Condition	OWNCD	n/a
StdAge	Condition	STDAGE	n/a
StdSz	Condition	STDSZCD	n/a
SiteCl	Condition	SITECLCD	n/a
StdOrg	Condition	STDORGCD	n/a
Slope	Condition	SLOPE	n/a
Physiog	Condition	PHYSCLCD	n/a
AlStk	Condition	ALSTKCD	n/a
Operable	Condition	OPERABILITY_SRS	n/a
NetVol	Tree	VOLCFNET, TPA_UNADJ	sum by plot (TREE.VOLCFNET * TREE.TPA_UNADJ)
Dstrb_any	Condition	DSTRBCD	if (COND.DSTRBCD1 > 0 | COND.DSTRBCD2 > 0 | COND.DSTRBCD3 > 0, 1, 0)
Dstrb_tot	Condition	DSTRBCD	sum (if (COND.DSTRBCD1 > 0, 1 COND.DSTRBCD2 > 0, 1 COND.DSTRBCD3 > 0, 1))
Mgmt_harv	Condition	TRTCD	if (COND.TRTCD1 = 10 | COND.TRTCD2 = 10 | COND.TRTCD3 = 10, 1, 0)
Mgmt_any	Condition	TRTCD	if (COND.TRTCD1 > 0 | COND.TRTCD2 > 0 | COND.TRTCD3 > 0, 1, 0)

Table 5. Derivation of predictor variables from census data.

Variable	Derivation in ArcGIS	Data Source
PopDens	Block Population /Block Area (ha)	2010 Census Population & Housing Unit Counts – Blocks
HousDens	Block Housing Unit Counts /Block Area (ha)	
UrbanDist_A	Straight line distance (miles) from plot location to closest urbanized area (>50,000 people)	2010 Census Urban Areas
UrbanDist_C	Straight line distance (miles) from plot location to closest urban cluster (2,500 – 50,000 people)	
UrbanDist	Straight line distance (miles) from plot location to closest urbanized area or urban cluster (>2,500 people)	
UrbanPGI	$\frac{\text{Population of closest urbanized area or urban cluster}}{(\text{UrbanDist})^2}$	

Table 6. Derivation of predictor variables from StreetMap premium data.

Variable	Derivation in ArcGIS	Data Source
RdDist_any	Straight line distance from plot location to closest road of any type	StreetMap Premium Roads
RdDist_hw	Straight line distance from plot location to closest highway	
SA_25mi	Total number of 25 mile mill service areas that intersect at the plot location	- Mill locations determined from Timber Product Output (TPO) reports. - Service areas created for every mill within 100 miles of Virginia using ArcGIS Network Analyst with StreetMap Premium Roads.
SA_50mi	Total number of 50 mile mill service areas that intersect at the plot location	
SA_75mi	Total number of 75 mile mill service areas that intersect at the plot location	
SA_100mi	Total number of 100 mile mill service areas that intersect at the plot location	

Table 7. Derivation of predictor variables from NLCD.

Variable	Derivation in ArcGIS
CoverType	Extracted pixel value at each plot location.
Percent Cover	Created a 10 mile buffer for each plot and then determined the percent cover for the relevant NLCD pixel class. Note: Percent cover variables based on publicly available *fuzzed and swapped* plot locations. Since percent land cover is based on a 10 mile plot radius, the fuzzed plot locations should not have a major impact on the derived variable.
	NLCD Class/Value
PctFor_dec	41: Deciduous Forest
PctFor_ev	42: Evergreen Forest
PctFor_mix	43: Mixed Forest
PctFor_all	41 + 42 + 43
PctDev_open	21: Developed, Open Space
PctDev_low	22: Developed, Low Intensity
PctDev_med	23: Developed, Medium Intensity
PctDev_high	24: Developed, High Intensity
PctDev_all	21 + 22 + 23 + 24
PctWater	11: Open Water

Table 8. Derivation of additional (miscellaneous) predictor variables.

Variable	Derivation in ArcGIS	Data Source
PAD	Intersection with protected areas	PADUS
NCED	Intersection with conservation easement	NCED
StrmDist_per	Straight line distance to closest perennial stream	National Hydrography Dataset (NHD)
StrmDist_any	Straight line distance to closest stream of any type	National Hydrography Dataset (NHD)

Model Development

Any observations with missing data were removed before beginning model development. Missing data were removed to ensure consistency across all models because not all of the modeling techniques investigated in this analysis can handle missing data. Seventy percent of the plot conditions were randomly selected as a training data set for model building (n = 5148), while the remaining 30 percent were reserved for model testing (n=2205). All models were trained and tested with the same data using R statistical software (R Development Core Team 2013). Six different presence/absence models to predict harvest occurrence were built using classification and regression trees (CART), random forest (RF), balanced random forest (bRF), conditional decision trees (cTREE), conditional random forest (cRF) and logistic regression (LR). Except for where otherwise indicated, models were developed using the default package parameters.

First, a classification tree model was built with CART (Breiman et al. 1984; R package rpart; http://cran.r-project.org/web/packages/rpart/index.html). Various model parameters can be set to minimize processing time by avoiding nodes that will likely be pruned. The minimum number of samples required to attempt a split was set to 15. Additionally, the complexity parameter (cp) was set to 0.001, requiring all splits to decrease the overall lack of fit by the cp. The CART model was pruned using the 1-SE rule (as recommended by Breiman et al. 1984), which applies a 10-fold cross validation to select the simplest tree (i.e. largest cp) with a cross-validation error ≤ 1 standard deviation of the smallest cp.

The next model was developed with CART's ensemble equivalent, RF (Breiman 2001). The default forest size in randomForest is 500 trees. Classification error stabilizes as the number of trees increases. A sufficient number of trees must be selected to ensure error has stabilized; more trees are needed for a greater number of variables. Since RF is by nature random, variable importance should also be verified from multiple runs to ensure the variable rank is generally consistent, if not the number of trees should be increased. Based on classification error (Appendix B) and a comparison of variable importance over multiple runs (Appendix C), the RF forest size was set to 1,000 trees.

Due to small number of harvested conditions relative to unharvested conditions in the data, a balanced random forest (bRF) model was also developed. bRF selects observations equally from each class depending on the number of observations from the minority class (Chen et al. 2004). bRf trees were built with 452 harvested conditions and 425 unharvested conditions selected with replacement. Due to high variability in the classification error (Appendix B) and variable importance rank (Appendix C), 4,000 trees were needed to stabilize the bRF model.

The next two models, cTREE (R package party, cTREE function; http://cran.rproject.org/web/packages/part/indexhtml), and its ensemble equivalent cRF (R package party, cforest function) are based on conditional inference trees (Hothorn et al. 2006). cTREE and cRF models were developed using the default p-value of 0.05. Trees were grown until there were no more significant splits. In cRF the default number of variables randomly selected at each node is 5. For consistency with the RF and bRF models, this was set to the RF default (square root of the number of parameters). cRF does not provide an error plot, but variable importance plots were compared for multiple runs to check model stability. Based on consistent variable importance rank (Appendix C), the forest size was set to 1,000 trees.

The CART and cTREE models are reported as dichotomous trees, which provide a graphical representation between the response and predictor variables. The decision trees were plotted with the R package partykit (http://cran.r-project.org/web/packages/partykit/index.html). In the decision tree diagram, the observations are split at nodes, connected by branches, which eventually reach the final classification at the terminal nodes. The individual decision trees cannot be provided for the ensemble methods (RF, bRF and cRF models). Instead, variable importance plots are provided to characterize the relationships between response and explanatory variables. Mean decrease in accuracy is plotted for each variable; variables with a greater mean decrease in accuracy have a greater impact on the model. Variable importance is interpreted as a relative ranking, actual importance values cannot be compared between models.

The final model was developed with LR. Unlike decision tree methods, LR is not well suited to model many explanatory variables. Therefore some sort of initial variable selection was needed to reduce the number of input parameters. Variables were selected for further analysis based on the cRF variable importance measure. The 14 variables with the greatest mean decrease in accuracy were selected (CoverType, NetVol, StdOrg, PctFor_ev, StdAge, UrbanDist_A, PctDev_low, StdSz, PctDev_all, PctFor_dec, PopDens, PctDev_high, HousDens, and SiteProd. Two additional variables (Own and SA_75mi) were also included due to previous research and preliminary analyses, which indicated their potential significance. All possible regressions were fit using a generalized linear model, with a binomial distribution (link = logit) (R Development

Core Team 2013; glm function). From all possible regressions the best model was selected using the Akaike information criterion (AIC), which assesses the tradeoff between model fit and complexity. Goodness of fit was assessed with the Hosmer-Lemeshow test (R Resource Selection package, hoslem.test function; http://cran.r-project.org/web/packages/ResourceSelection/index.html).

Model Accuracy

Model performance was assessed with 5 different accuracy metrics: percent correctly classified (PCC), sensitivity, specificity and area under the curve (AUC). Bias was also reported to indicate whether the model tended to over- or underpredict harvest occurrences.

PCC, sensitivity, specificity and bias were all derived from the confusion matrix, which compares observed with predicted responses. Since all of the models predict the relative probability of harvest, a threshold must be selected to classify model output. A variety of threshold selection techniques are available, which can impact model interpretation, especially if data are imbalanced. The prevalence threshold and minimum difference threshold (MDT) were both examined. The prevalence threshold is determined from the model building data based on the percent of total presences (i.e. harvests) in the data set. Prevalence is the same for all models (threshold = 0.088) except for bRF (threshold = 0.5). MDT is different for each model because it optimizes the classification of predicted probabilities by minimizing the difference between sensitivity and specificity. The MDT for each model is reported with the model accuracy measures.

AUC is a threshold independent measure derived from the receiver operative characteristic (ROC) curve (R package pROC, http://cran.r-project.org/web/packages/pROC/index.html). AUC ignores much of the important information provided by the ROC curve and should not be interpreted in isolation. Therefore in addition to reporting AUC, the full ROC curve was also plotted and interpreted. All model accuracy measures are reported for training and testing data. However, since we are more interested in how well the models will predict future data, our interpretation focused on the model accuracy when applied to the testing data.

Availability Estimate

The best models were selected and recalibrated with all of the available data (n = 7353) to provide the most robust harvest predictions possible (Fielding and Bell 1997). The final models were applied to predict the relative probability of future harvests at FIA plots in Virginia from the 2012 evaluation group. Given an average of 5 years between measurements in the training data, the harvest predictions represent the probability of harvest between 2012 and 2017. The probability of harvest was classified as harvest presence/absence based on 1) the prevalence threshold and 2) the percentage of conditions expected to be harvested. Plot conditions where harvest presence was predicted were considered available and tree volume was estimated for those plots. Following the general methods described by Prestemon and Wear (2000), area expansion factors were applied to estimate the volume of wood available for harvesting from timberland in Virginia.

Chapter 4. RESULTS

Model Development

Single Tree Models

The pruned CART model produced a decision tree with 12 splits and 13 terminal nodes (Figure 3, Table 9), while the cTREE model split the data 11 times into 12 terminal nodes (Figure 4, Table 10). Of the 13 terminal nodes in the CART model, 9 nodes predicted harvest absence, 3 of which did so with less than 5 percent error. Harvest was predicted by 4 nodes with error ranging from 4 – 9 percent. The cTREE model only predicted harvest in 1 node with 41 percent error. All but 2 of the terminal nodes, which predicted harvest absence, had less than 15% error.

Both models (Figures 3, 4) selected natural vs. artificial stand origin (StdOrg) for the first split. For natural regeneration (StdOrg = 0), the CART model split the data by percent evergreen forest cover (PctFor_ev = 10.23) and the cTREE model split the data by net volume per acre (NetVol = 1883.35). For greater percent cover and net volume, both models selected land cover type at the plot location (CoverType) for the next split. For the CART model, plots with shrub/scrub, grassland/herbaceous, or cultivated crops were split again by net volume per acre (NetVol = 2685.1). Observations with greater volume were classified as harvests. For the cTREE model, given greater net volume, shrub/scrub, grassland/herbaceous, cultivated crops, or barren land cover were classified as no harvest, but with nearly 40% classification error.

For artificial regeneration (StdOrg = 1), the CART model split the data by stand age (StdAge = 14.5), and classified younger stands as harvest absence. Older stands with shrub/scrub, grassland/herbaceous, cultivated crops, or barren land cover was classified as harvest presence. For other land cover types, given a minimum percent evergreen forest cover (PctFor_ev ≥ 17.06) and distance to highway (RdDist_hw ≥ 254.49), 2 of the 5 resulting nodes predicted harvest. The cTREE model split artificially regenerated stands by volume (NetVol = 1316.28); greater volumes lead to terminal nodes with a greater proportion of harvests. However, only observations with shrub/scrub, grassland/herbaceous, cultivated crops or barren land cover were classified as harvests.

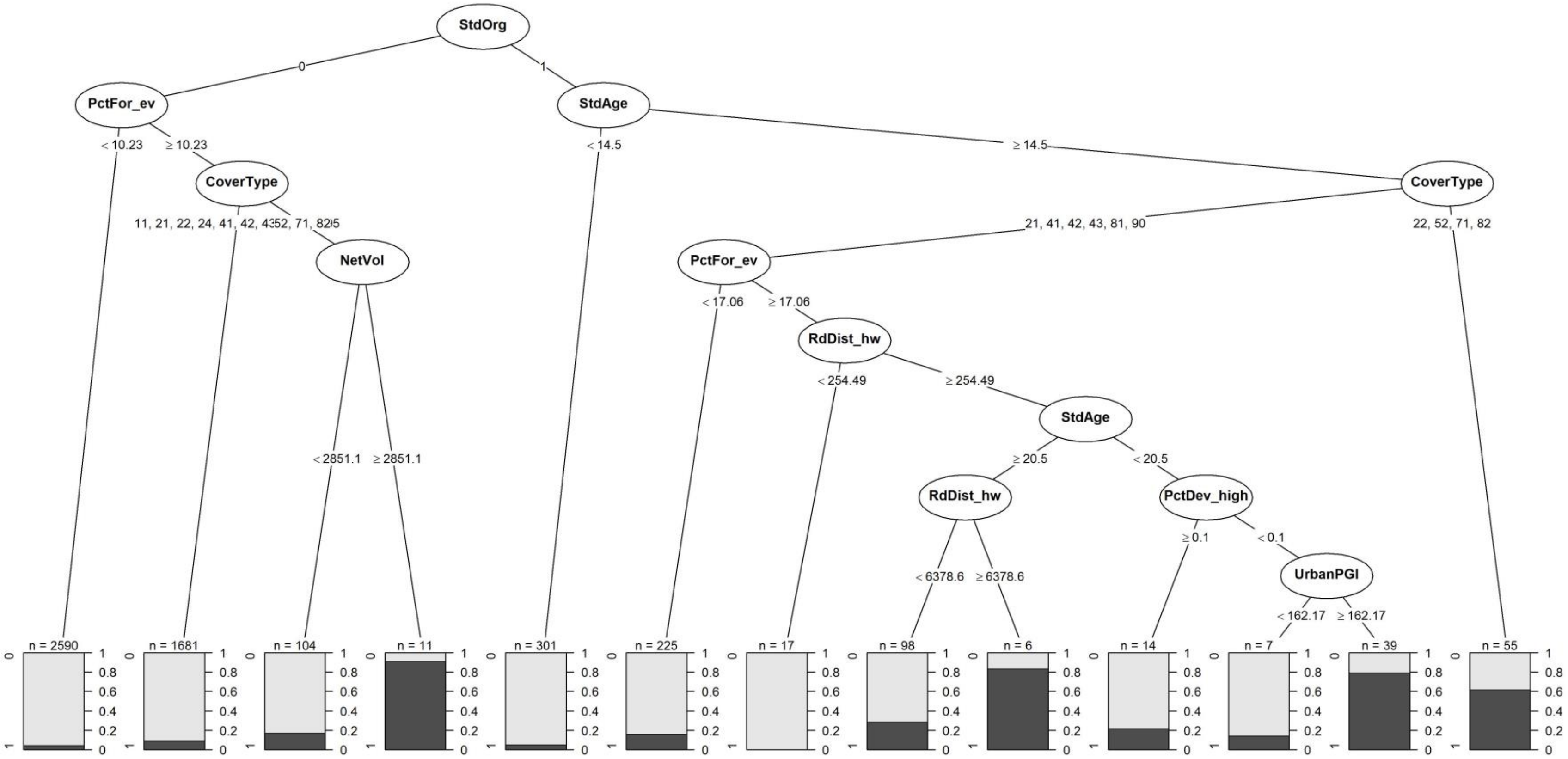

Figure 3. Decision tree for the Classification and Regression Tree (CART) model. Variables selected for splits are listed in ovals and their splitting criteria are listed on the lines connecting variables. For example, if an observation had StdOrg = 1, then it was grouped with data from the branch on the right. Bar graphs in the terminal node describe the proportion of harvest and no harvest responses for each final classification group; dark grey indicates harvest and light grey indicates no harvest. Further details about the CART model are provided in Table 9. See Table 1 for a general description of variables.

Table 9. CART model details. The splitting variables and terminal node predictions are bolded. Any terminal nodes that predicted harvest are underlined.

```
[1] Root
|  [2] StdOrg in 0
|  |  [3] PctFor_ev < 10.23277: 0 (n = 2590, err = 4.4%)
|  |  [4] PctFor_ev >= 10.23277
|  |  |  [5] CoverType in 11, 21, 22, 24, 41, 42, 43, 81, 90, 95: 0 (n = 1681, err = 9.3%)
|  |  |  [6] CoverType in 52, 71, 82
|  |  |  |  [7] NetVol < 2851.10119: 0 (n = 104, err = 17.3%)
|  |  |  |  [8] NetVol >= 2851.10119: 1 (n = 11, err = 9.1%)
|  [9] StdOrg in 1
|  |  [10] StdAge < 14.5: 0 (n = 301, err = 5.0%)
|  |  [11] StdAge >= 14.5
|  |  |  [12] CoverType in 21, 41, 42, 43, 81, 90
|  |  |  |  [13] PctFor_ev < 17.05607: 0 (n = 225, err = 16.0%)
|  |  |  |  [14] PctFor_ev >= 17.05607
|  |  |  |  |  [15] RdDist_hw < 254.48831: 0 (n = 17, err = 0.0%)
|  |  |  |  |  [16] RdDist_hw >= 254.48831
|  |  |  |  |  |  [17] StdAge >= 20.5
|  |  |  |  |  |  |  [18] RdDist_hw < 6378.59999: 0 (n = 98, err = 28.6%)
|  |  |  |  |  |  |  [19] RdDist_hw >= 6378.59999: 1 (n = 6, err = 16.7%)
|  |  |  |  |  |  [20] StdAge < 20.5
|  |  |  |  |  |  |  [21] PctDev_high >= 0.09673: 0 (n = 14, err = 21.4%)
|  |  |  |  |  |  |  [22] PctDev_high < 0.09673
|  |  |  |  |  |  |  |  [23] UrbanPGI < 162.17258: 0 (n = 7, err = 14.3%)
|  |  |  |  |  |  |  |  [24] UrbanPGI >= 162.17258: 1 (n = 39, err = 20.5%)
|  |  |  [25] CoverType in 22, 52, 71, 82: 1 (n = 55, err = 38.2%)

Number of inner nodes: 12
Number of terminal nodes: 13
```

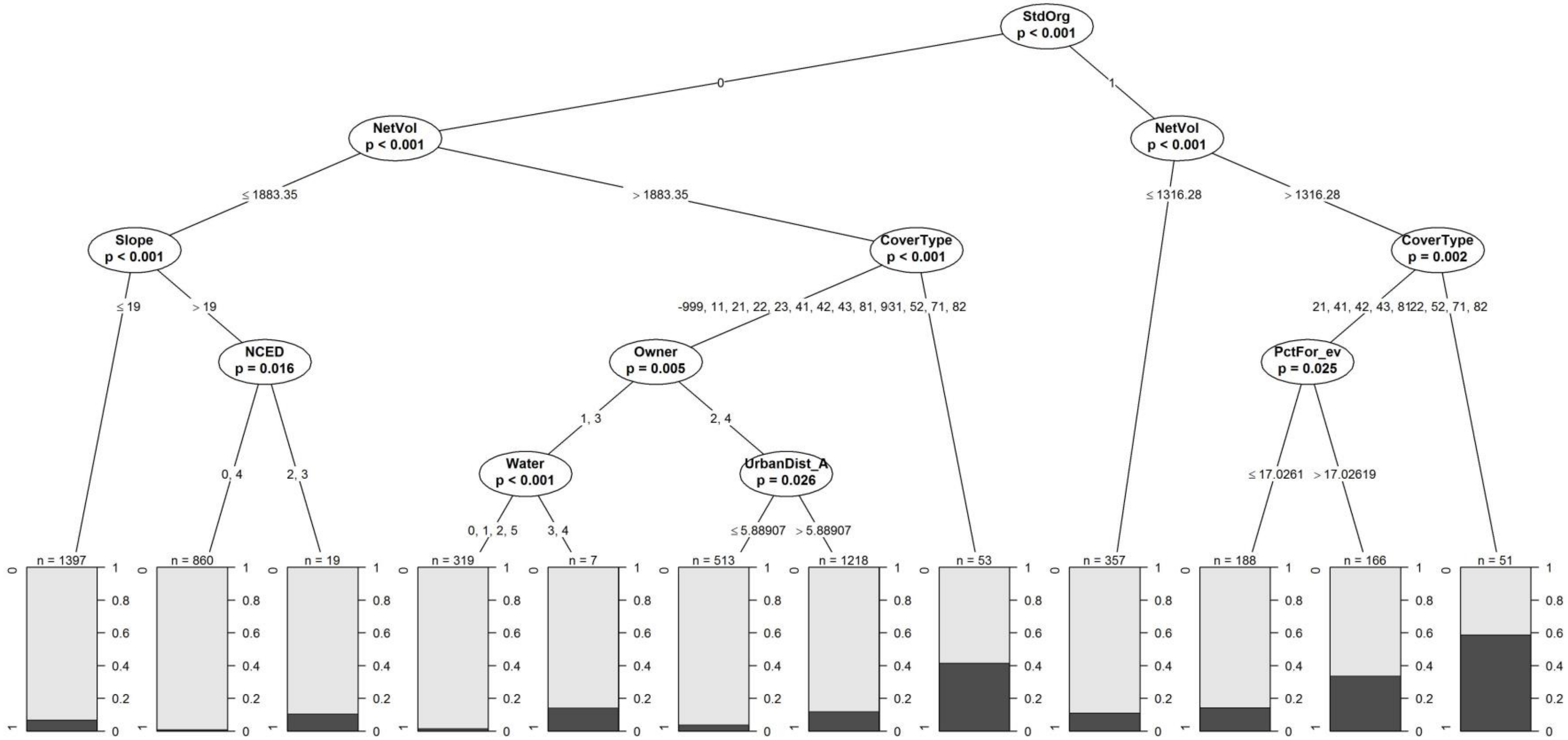

Figure 4. Decision tree for the conditional inference tree (cTREE) model. See Figure 3 for decision tree explanation. The significance of each split variable is listed below the variable names. See Table 10 for further details about the cTREE model.

Table 10. cTREE model details. The splitting variables and terminal node predictions are bolded. Any terminal nodes that predicted harvest are underlined.

```
[1] Root
|  [2] StdOrg in 0
|  |  [3] NetVol <= 1883.35046
|  |  |  [4] Slope <= 19: 0 (n = 1397, err = 6.8%)
|  |  |  [5] Slope > 19
|  |  |  |  [6] NCED in 0, 4: 0 (n = 860, err = 1.0%)
|  |  |  |  [7] NCED in 2, 3: 0 (n = 19, err = 10.5%)
|  |  [8] NetVol > 1883.35046
|  |  |  [9] CoverType in -999, 11, 21, 22, 23, 41, 42, 43, 81, 90
|  |  |  |  [10] Owner in 1, 3
|  |  |  |  |  [11] Water in 0, 1, 2, 5: 0 (n = 319, err = 1.6%)
|  |  |  |  |  [12] Water in 3, 4: 0 (n = 7, err = 14.3%)
|  |  |  |  [13] Owner in 2, 4
|  |  |  |  |  [14] UrbanDist_A <= 5.88907: 0 (n = 513, err = 3.9%)
|  |  |  |  |  [15] UrbanDist_A > 5.88907: 0 (n = 1218, err = 11.9%)
|  |  |  [16] CoverType in 31, 52, 71, 82: 0 (n = 53, err = 41.5%)
|  [17] StdOrg in 1
|  |  [18] NetVol <= 1316.27994: 0 (n = 357, err = 11.2%)
|  |  [19] NetVol > 1316.27994
|  |  |  [20] CoverType in 21, 41, 42, 43, 81, 90
|  |  |  |  [21] PctFor_ev <= 17.02619: 0 (n = 188, err = 14.4%)
|  |  |  |  [22] PctFor_ev > 17.02619: 0 (n = 166, err = 33.7%)
|  |  |  [23] CoverType in 22, 52, 71, 82: 1 (n = 51, err = 41.2%)

Number of inner nodes: 11
Number of terminal nodes: 12
```

Ensemble Models

Net volume (NetVol) was the most important variable identified by the RF (Figure 5), bRF (Figure 6), and cRF (Figure 7) models. Distance to urbanized area (UrbanDist_A), and stand age (StdAge) were also identified among the top 10 most important variables for all 3 of the ensemble models. These 3 common variables were ranked among the top 5 most important variables for all of the models (Table 12). The RF and cRF models also both identified land cover (CoverType), developed – low intensity percent cover (PctDev_low) and evergreen forest percent cover (PctFor_ev) as important variables (Table 12). Cover type was the second most important variable in both models, and except for distance to urban area, the rest of the common important variables ranked similarly. The RF model placed greater emphasis on distance to urban areas than the cRF model.

Other important variables selected by the RF model include mixed forest percent cover (PctFor_Mix), water percent cover (PctWater), population gravity index (UrbanPGI) and total forest percent cover (PctFor_all). In addition to stand origin, the cRF model also selected stand size (StdSz) and percent developed – all types (PctDev_all) as very important variables. None of the other variables from the bRF model were also identified as most important by either the RF or cRF model. In addition to the 3 common variables, the bRF model also selected ownership (Own), protected areas (PAD), road distance to highways (RdDist_hw), distance to stream (StrmDist_any), distance to urbanized area or cluster (UrbanDist), and water on plot (Water) as then 10 most important variables.

As expected, given the RF selection bias towards continuous variables and categorical variables with many groups, 9 of the 10 most important variables identified by the RF model are continuous and the one categorical variable (CoverType) has many categories. In the cRF model, 3 of the top 10 variables were categorical (StdOrg, StdSz, CoverType). The bRF did not appear to suffer from any selection bias. Three of the top 10 variables were categorical, none of which have as many categories as CoverType.

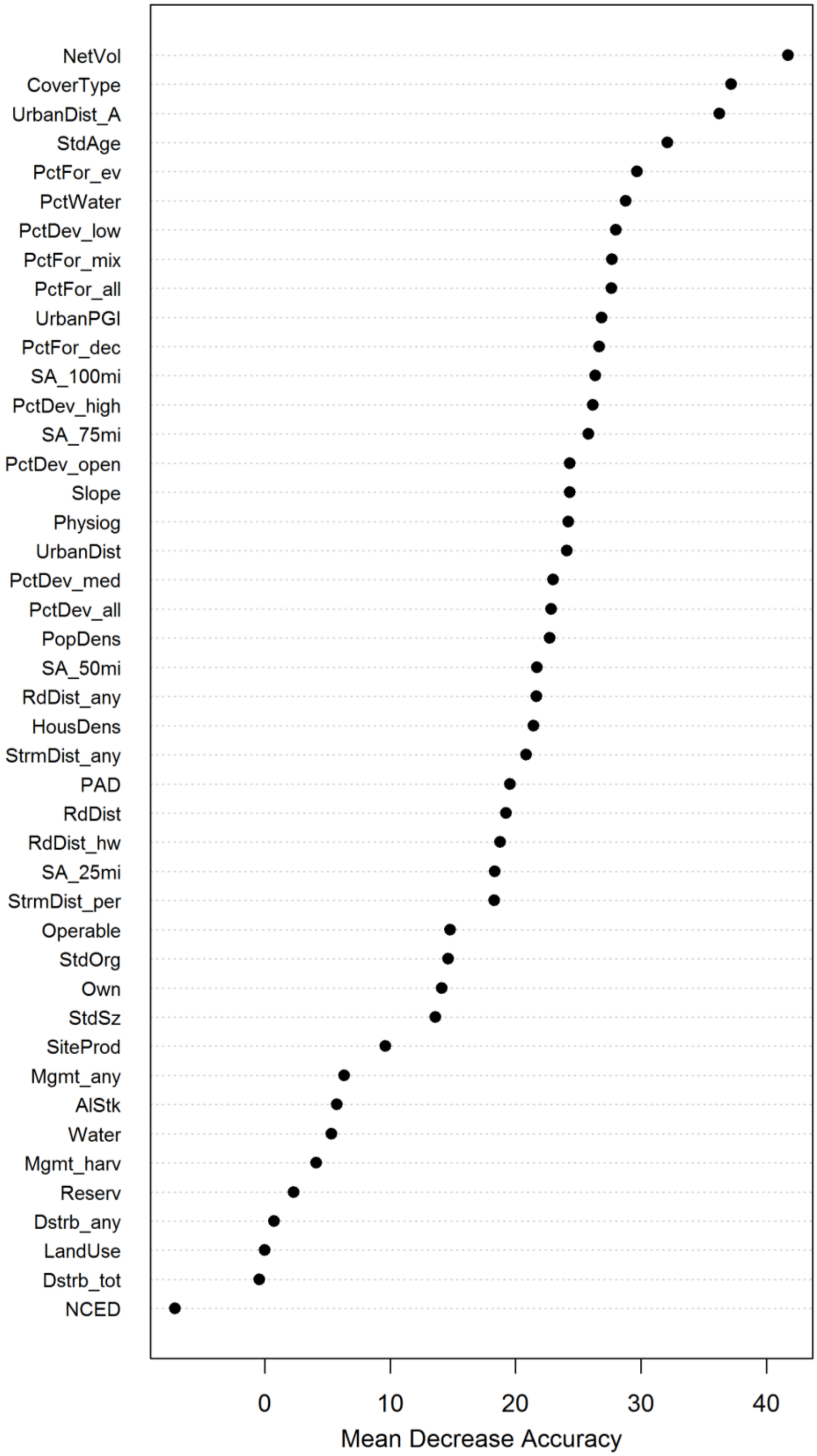

Figure 5. Random forest (RF) model variable importance.

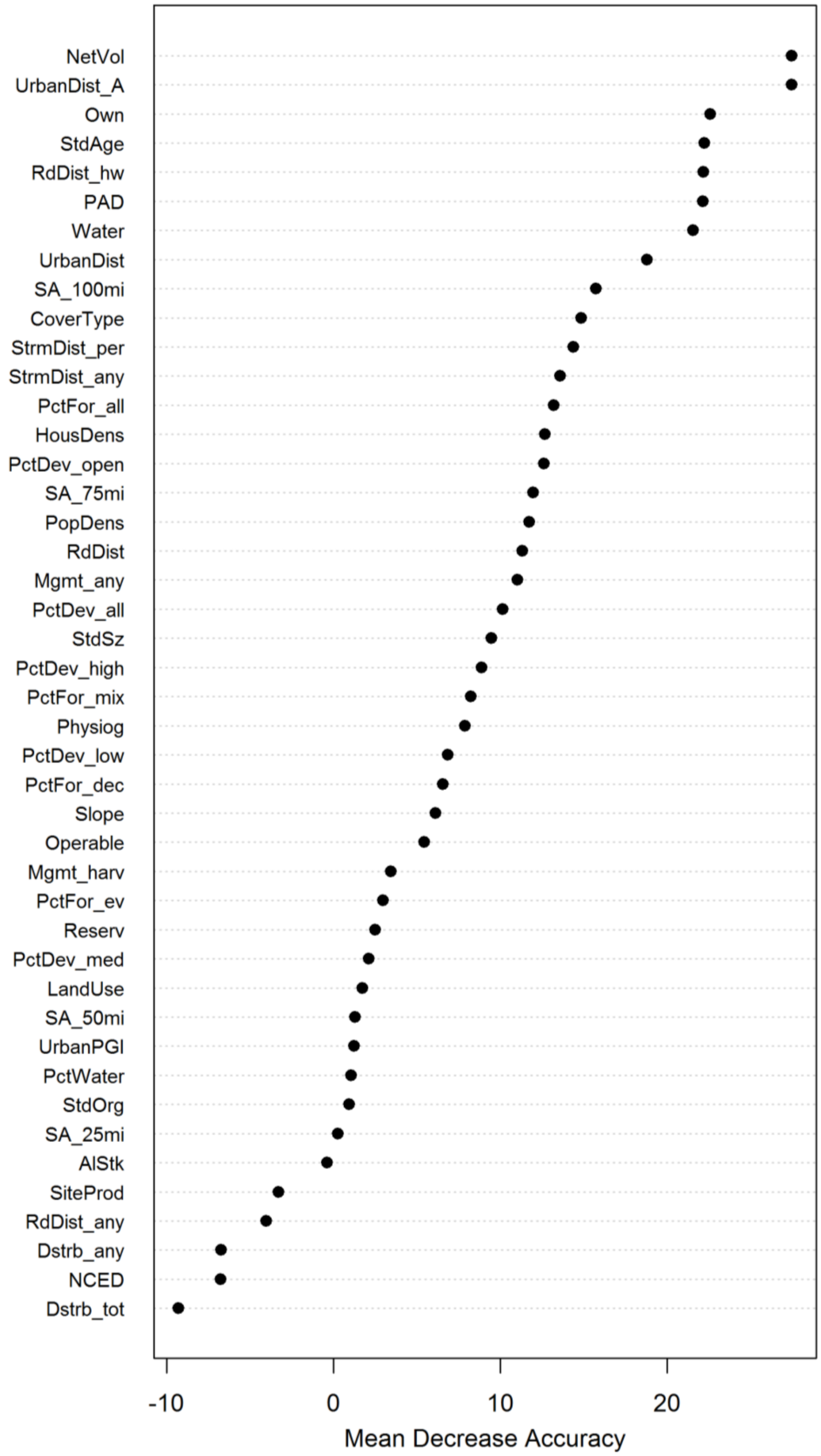

Figure 6. Balanced random forest (bRF) model variable importance.

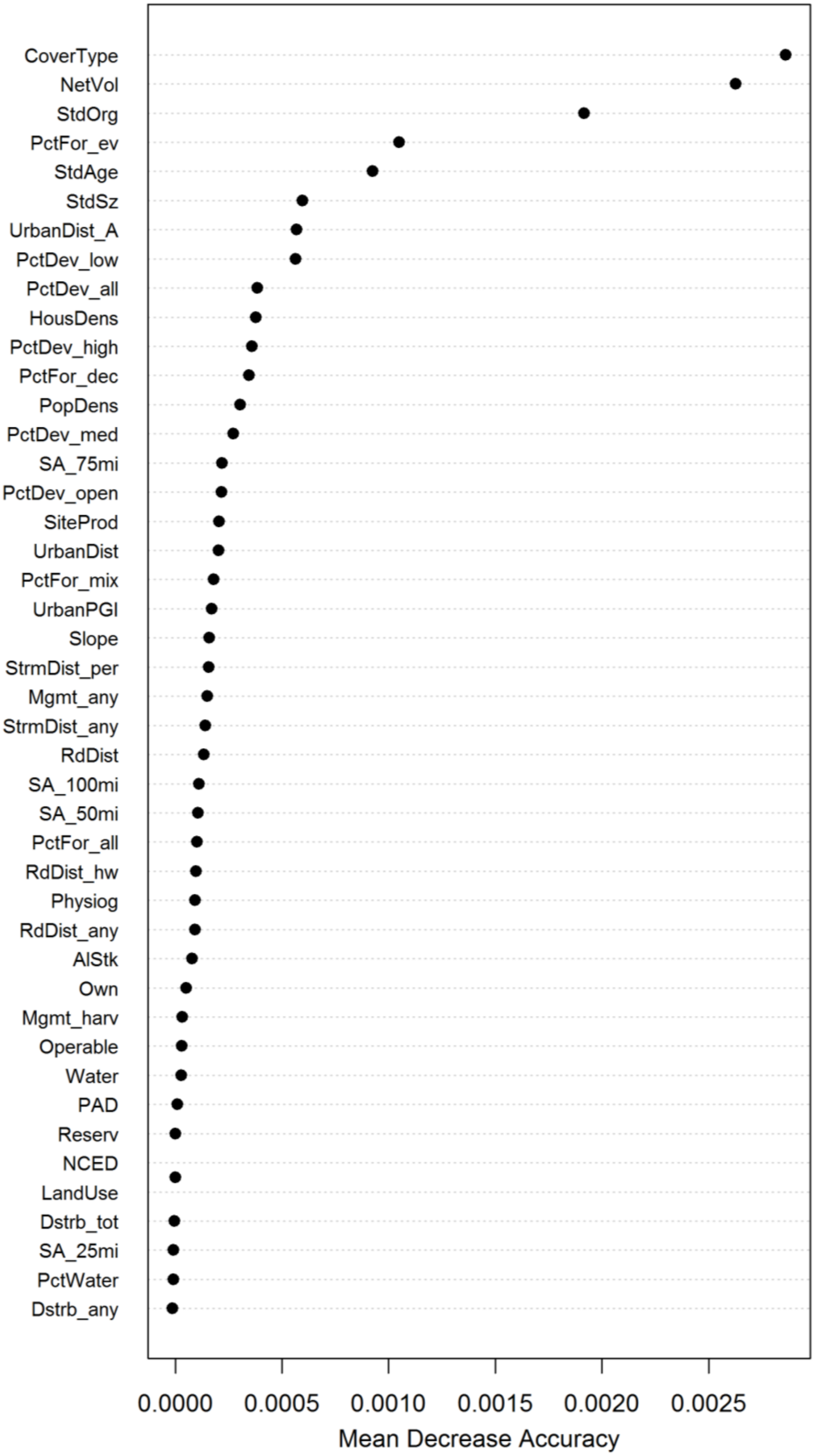

Figure 7. Conditional random forest (cRF) model variable importance.

Logistic Regression Model

AIC was used to select the best LR model from all possible regressions of the initially selected variables. The variables and coefficients of the final model are shown in Table 11. The Hosmer-Lemeshow goodness of fit statistic ($P > 0.05$) showed good model fit. In the selected LR model, the log odds of harvest decreased significantly ($p < 0.05$) with developed - low intensity land cover (CoverType 21) and population density (PopDens) and increased significantly with net volume (NetVol) and medium diameter trees (StdSz 2).

Table 11. Logistic regression model coefficients and significance.

Variable	Estimate	Std. Error	p-value
CoverType 21	-5.39	1.13	0.00
CoverType 41	0.31	1.11	0.78
CoverType 42	0.35	1.05	0.74
CoverType 43	0.49	1.05	0.64
CoverType 52	0.60	1.07	0.57
CoverType 71	1.62	1.07	0.13
CoverType 81	2.04	1.09	0.62
CoverType 82	0.54	1.10	0.62
CoverType 90	1.69	1.12	0.13
StdOrg 1	-0.05	1.07	0.96
StdSz 2	0.81	0.14	0.00
StdSz 3	-0.48	0.51	0.35
StdSz 5	0.01	0.39	0.98
Own 31	0.71	0.22	0.00
Own 46	0.93	0.54	0.09
NetVol	1.84	0.35	0.00
PctFor_ev	0.00	0.00	0.00
PctDev_all	0.03	0.01	0.00
PopDens	-0.10	0.03	0.00
PctDev_high	0.00	0.00	0.01

Table 12. Comparison of important variables from all models. All variables from the CART, cTREE and LR models are included. Only the10 most important variables from the RF, bRF and cRF models are included. An X indicates that variable was selected/important for that model.

	Model					
	CART	cTREE	RF	bRF	cRF	LR
CoverType	X	X	X		X	X
NCED		X				
NetVol	X	X	X	X	X	X
Own		X		X		X
PAD				X		
PctDev_all					X	X
PctDev_high	X					X
PctDev_low			X		X	
PctFor_all			X			
PctFor_dec					X	
PctFor_ev	X	X	X		X	X
PctFor_mix			X			
PctWater			X			
PopDens						X
RdDist_hw	X			X		
Slope		X				
StdAge	X		X	X	X	
StdOrg	X	X			X	X
StdSz					X	X
StrmDist_any				X		
StrmDist_per				X		
UrbanDist				X		
UrbanDist_A		X	X	X	X	
UrbanPGI	X		X			
Water		X		X		

Model Accuracy

Model predictions were classified with the prevalence (Table 13) and minimum differences thresholds (Table 14) to calculate accuracy metrics (Tables 15,16). The prevalence threshold was 0.088 for all models except for bRF, which was classified with a threshold of 0.05. The minimum difference threshold was much higher than the prevalence threshold for CART, (MDT = 0.382), RF (MDT = 0.468), and bRF (MDT = 0.694) models (Table 14). This ultimately resulted in reducing the bias for those models and increasing the specificity, but at the cost of greatly lowering the sensitivity (Table 16). Since this study aims to predict harvest presence, the prevalence threshold was chosen instead of MDT for further model assessments and predictions.

Using the prevalence threshold (0.088), all of the models classified more than 60% of the testing data correctly (PCC > 60%; Table 15). The cRF and bRF had the greatest PCC accuracy (PCC > 80%). These models performed similarly in all aspects. Both the cRF and bRF models have low prediction bias (1.538, 2.188 respectively) and according to AUC discriminate well between harvest presence and absence.

The CART model had the greatest bias, but still performed somewhat better than chance alone (AUC = 0.656) and similar to the cTREE model (AUC = 0.685). However, upon closer inspection of the ROC curves, it is evident that the cTREE and CART models behave differently at various thresholds (Figure 8). The cTREE model has a lower prediction bias and classifies fewer harvests than the CART model.

All of the ensemble tree models, as well as the LR model had similar AUC values (.738 - .773), and similar ROC curves for the testing data. They intersect frequently making it difficult to identify one curve as absolutely better than the others. Although all of the ensemble tree models perform better than the single-tree models, the bRF model seems to be the best classifier over all regions of the ROC graph (Figure 8).

The cRF model correctly predicted the fewest harvests (Sens = 0.419), however given its high specificity (Spec = 0.897) it also rarely falsely predicts harvest. The bRF model performed similarly although with slightly higher sensitivity (0.559) and lower specificity (0.850). The LR model performs similarly to the RF model according to all measures. The RF and LR models have fairly high selection bias (Bias = 3.844,3.796 respectively) but AUC values similar to the cRF and bRF models. Perhaps due to the greater prediction bias, the LR and RF models have higher sensitivity and lower specificity than the bRF and cRF models.

Table 13. Confusion matrix with responses classified by the prevalence threshold.

			Observed			
			Training		Testing	
Model	Thresh		Harvest	No Harvest	Harvest	No Harvest
Classification tree	0.088	Harvest	323	1917	122	834
		No Harvest	129	2779	64	1185
Conditional classification tree	0.088	Harvest	259	1123	94	500
		No Harvest	193	3573	92	1519
Random forest	0.088	Harvest	452	388	122	593
		No Harvest	0	4308	64	1426
Balanced random forest	0.5	Harvest	452	0	104	303
		No Harvest	504	4192	82	1716
Conditional random forest	0.088	Harvest	376	391	78	208
		No Harvest	76	4305	108	1811
Logistic regression	0.088	Harvest	297	1419	120	612
		No Harvest	155	3277	66	1407

Table 14. Confusion matrix with responses classified by the minimum difference threshold.

			Observed			
			Training		Testing	
Model	Thresh		Harvest	No Harvest	Harvest	No Harvest
Classification tree	0.382	Harvest	80	31	26	31
		No Harvest	372	4665	160	1988
Conditional classification tree	0.094	Harvest	259	1123	94	500
		No Harvest	193	3573	92	1519
Random forest	0.468	Harvest	449	49	26	54
		No Harvest	3	4647	160	1965
Balanced random forest	0.694	Harvest	441	114	45	76
		No Harvest	11	4582	141	1943
Conditional random forest	0.071	Harvest	402	519	95	270
		No Harvest	50	4177	91	1749
Logistic regression	0.083	Harvest	304	1537	126	661
		No Harvest	148	3159	60	1358

Table 15. Model accuracy measures derived for the prevalence threshold. See Table 13 for confusion matrix used to derive measures.

		Accuracy Measure				
Classification method		PCC	Bias	Sens	Spec	AUC
Classification tree	Test	0.605	5.140	**0.645**	0.601	**0.656**
	Train	0.617	4.956	0.706	0.609	0.714
Conditional classification tree	Test	0.732	3.194	**0.505**	0.752	**0.685**
	Train	0.744	3.194	0.573	0.761	0.732
Random forest	Test	0.702	3.844	**0.656**	0.706	**0.739**
	Train	0.925	1.858	1.000	0.917	0.723
Balanced random forest	Test	0.826	2.188	**0.559**	0.850	**0.773**
	Train	0.902	0.473	1.000	0.893	0.769
Conditional random forest	Test	0.857	1.538	**0.419**	0.897	**0.761**
	Train	0.925	1.858	1.000	0.917	0.757
Logistic regression	Test	0.693	3.935	**0.645**	0.697	**0.738**
	Train	0.694	3.796	0.657	0.698	0.752

Table 16. Model accuracy measures derived for the minimum difference threshold. See Table 14 for the confusion matrix used to derive measures.

			Accuracy Measure				
Classification method		Threshold	PCC	Bias	Sens	Spec	AUC
Classification tree	Test		0.913	0.306	**0.124**	0.986	**0.656**
	Train	0.382	0.925	0.246	0.181	0.997	0.714
Conditional classification tree	Test		0.732	3.194	**0.505**	0.752	**0.685**
	Train	0.094	0.744	3.058	0.573	0.761	0.732
Random forest	Test		0.903	0.430	**0.140**	0.973	**0.739**
	Train	0.468	0.990	1.102	0.993	0.990	0.723
Balanced random forest	Test		0.902	0.651	**0.242**	0.962	**0.773**
	Train	0.694	0.976	1.228	0.976	0.976	0.769
Conditional random forest	Test		0.836	1.962	**0.511**	0.866	**0.761**
	Train	0.071	0.889	2.038	0.889	0.889	0.757
Logistic regression	Test		0.673	4.231	**0.677**	0.673	**0.738**
	Train	0.083	0.673	4.073	0.673	0.673	0.752

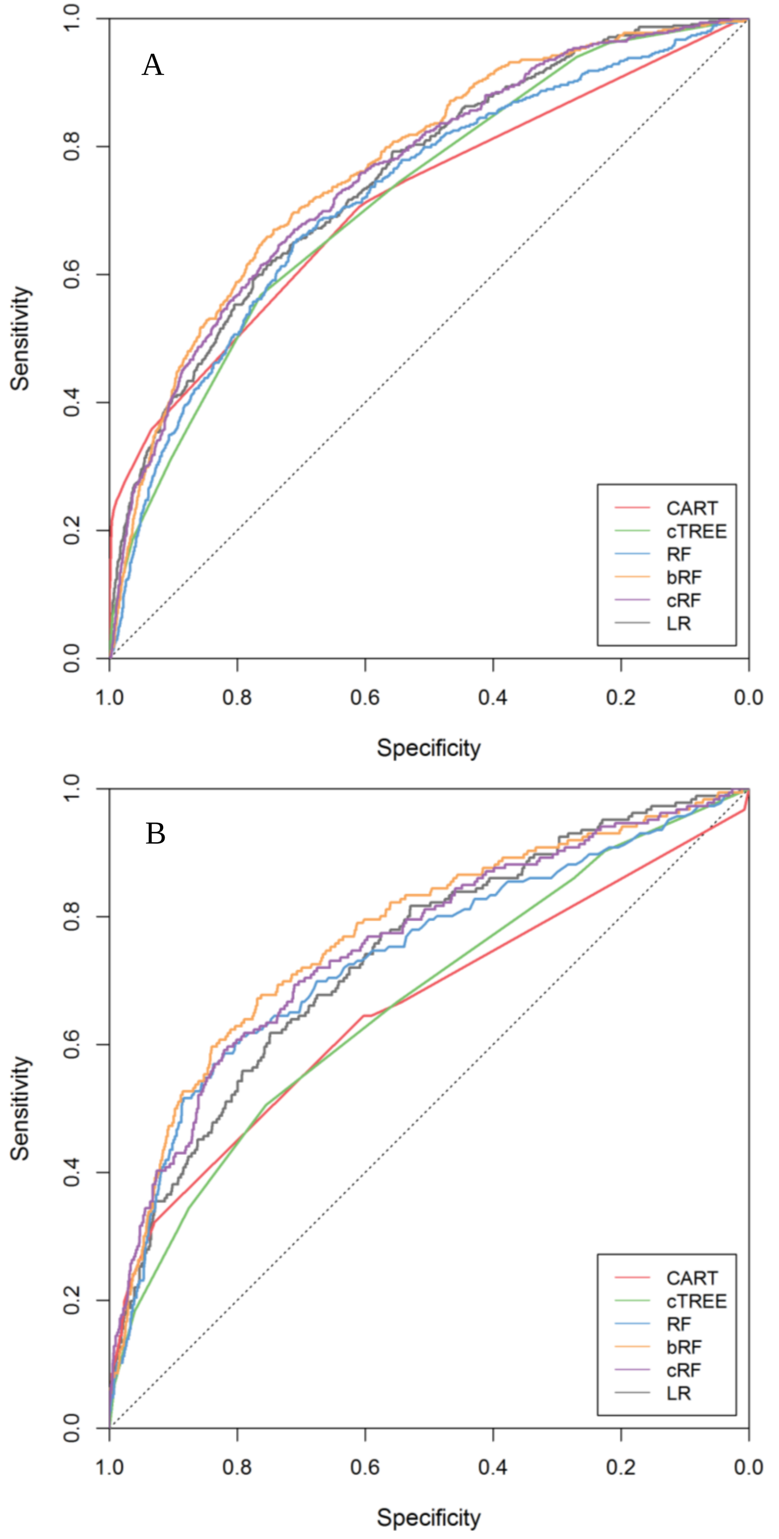

Figure 8. ROC curves for training data (A) and testing data (B).

Model Predictions

The recalibrated bRF and LR models were applied to predict future harvest occurrences at FIA plot conditions in Virginia. The bRF model was selected because provided the greatest balance between bias and sensitivity with the prevalence threshold accuracy metrics and performed slightly better than all other models over most regions of the ROC curve. The LR model was selected because it performed almost as well as the ensemble tree methods and can be much more easily shared and repeated. After predicting the probability of harvest with both the bRF and LR model (Figures 9,10), harvest presence/absence was classified based on the newly estimated prevalence threshold from all of the available data (bRF = 0.5, LR = 0.095). The bRF model predicted 21.8 percent of plot conditions would be harvested (Figure 11), while the LR model predicted 34.6 % of plot conditions would be harvested (Figure 12). The bRF and LR models predicted far more plot conditions available for harvest then had been observed in the model building data (9.5%). Therefore harvest presence/absence was also classified for the 9.5% of plots with the greatest probability of harvest (Figures 13, 14).

In both models, plot conditions with high and low probabilities of harvest were concentrated in similar areas of Virginia (Figures 9,10). FIA plot conditions with a higher probability of harvest are concentrated in the Piedmont and Coastal Plain Regions. Both models predicted low probabilities of harvest along the northwestern boundary of Virginia essentially outlining the George Washington and Jefferson National Forests. Additionally, in both models, the probability of harvest decreases in proximity to urban area, Richmond in particular. Harvest probability also seems to decrease near Norfolk/ Virginia Beach and Washington DC. It is evident from all of the harvest presence/absence classification maps (Figures 11,12,13,14), that most harvests are predicted to occur surrounding, but not directly adjacent to urban areas, specifically Richmond.

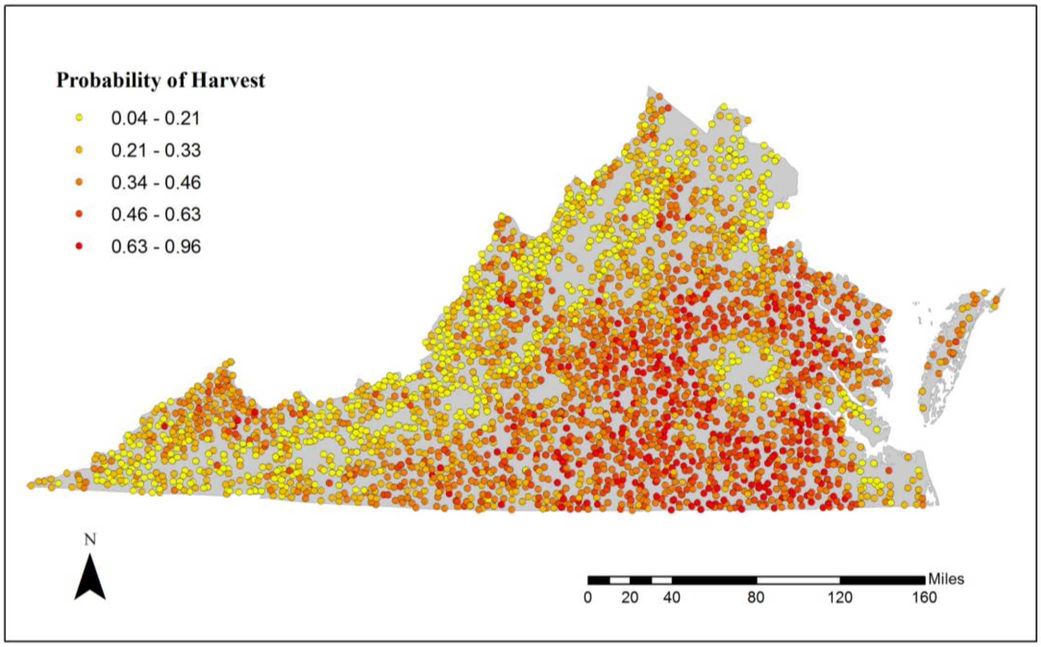

Figure 9. Probability of harvest predicted by the bRF model.

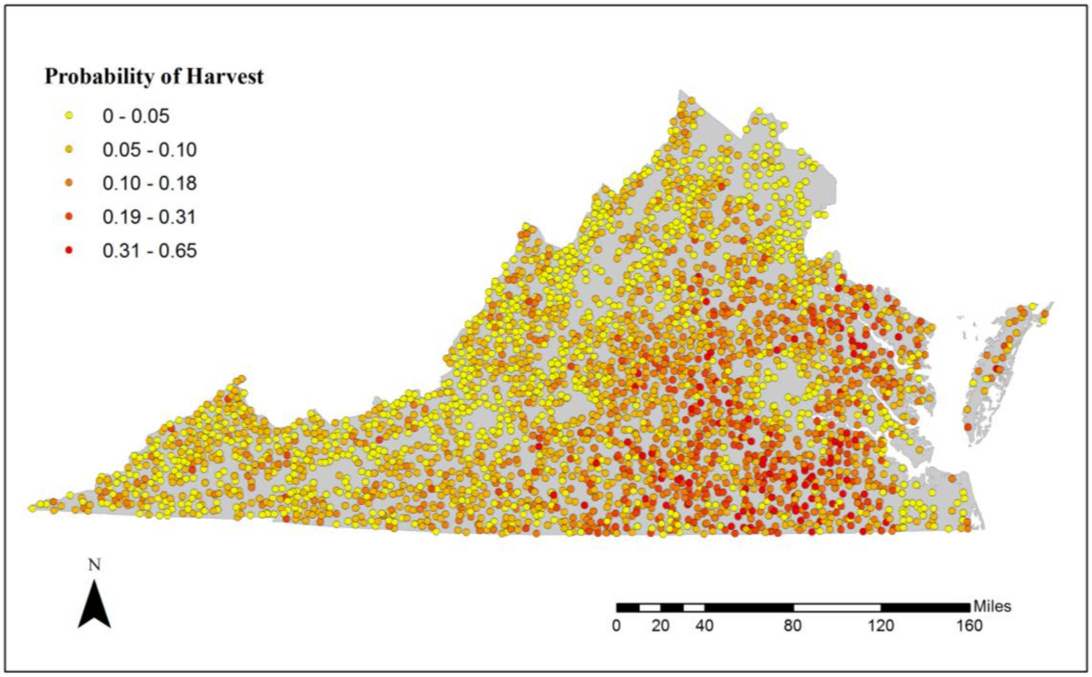

Figure 10. Probability of harvest predicted by the LR model.

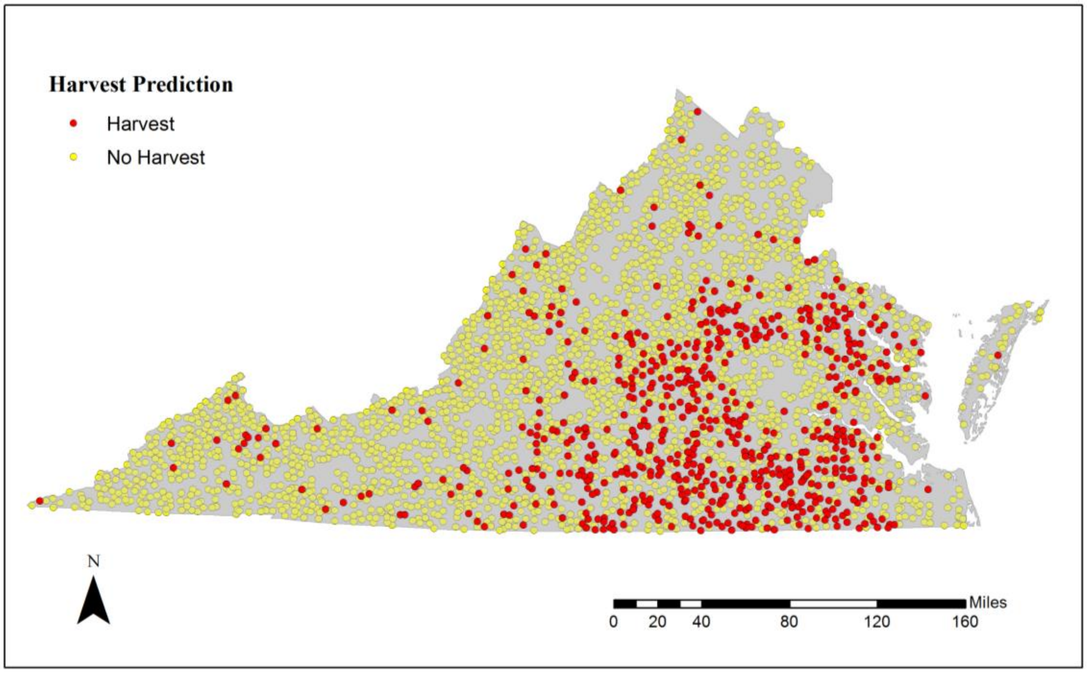

Figure 11. Harvest presence/absence predicted by the bRF model with a 0.5 classification threshold.

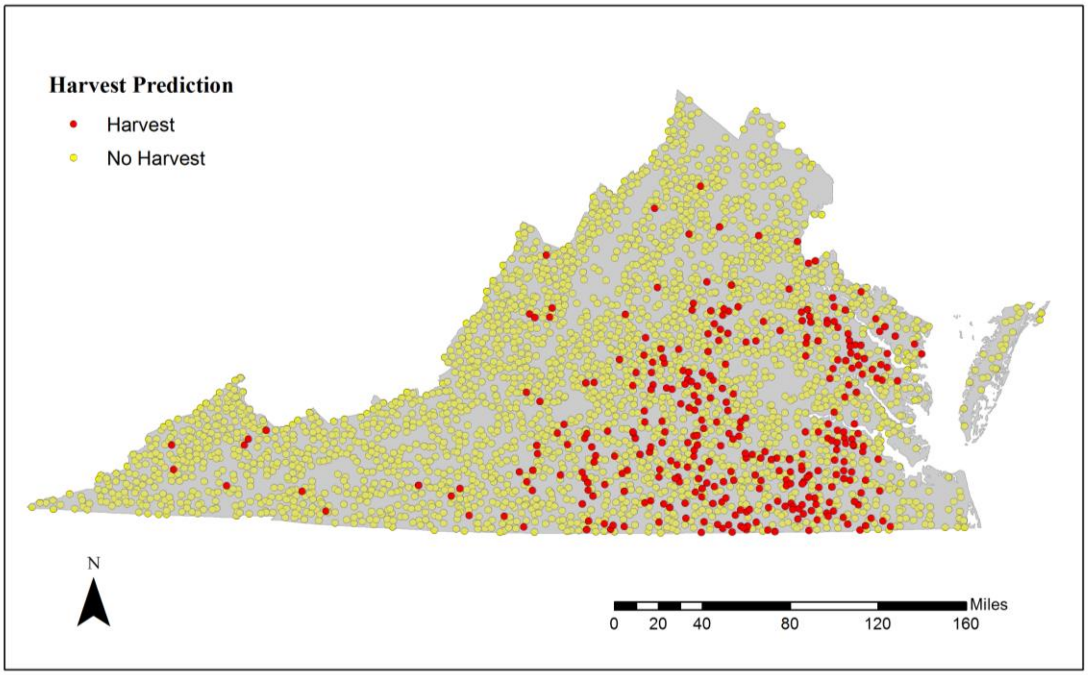

Figure 12. Harvest presence/absence predicted by the LR model with a 0.095 classification threshold.

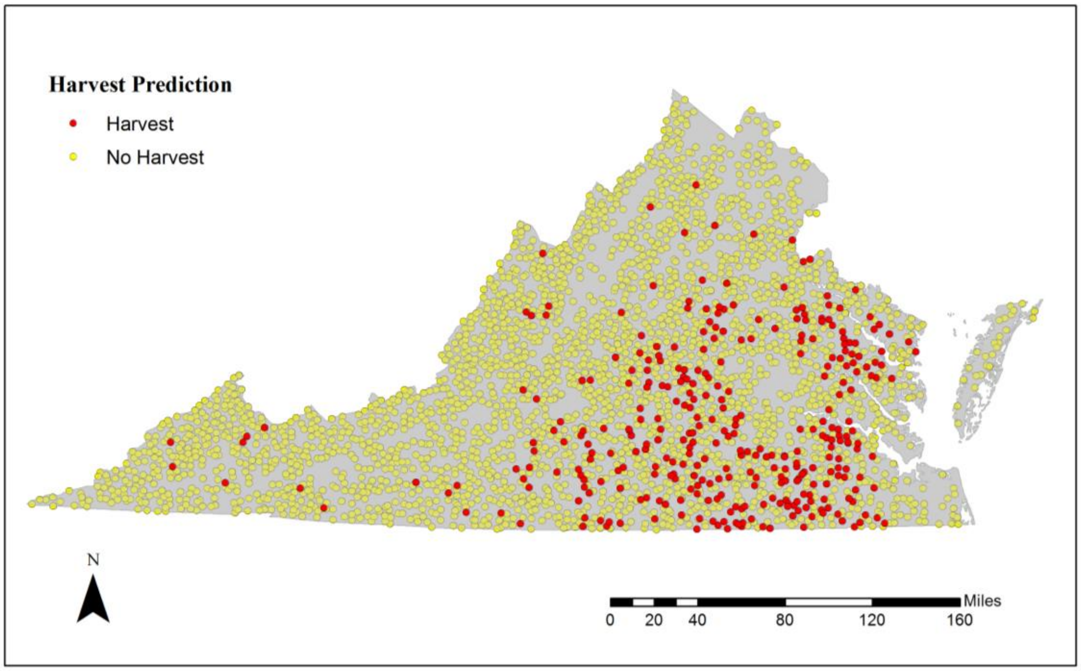

Figure 13. Harvest presence predicted for the 9.5% of conditions with the highest probability of harvest based on the bRF model.

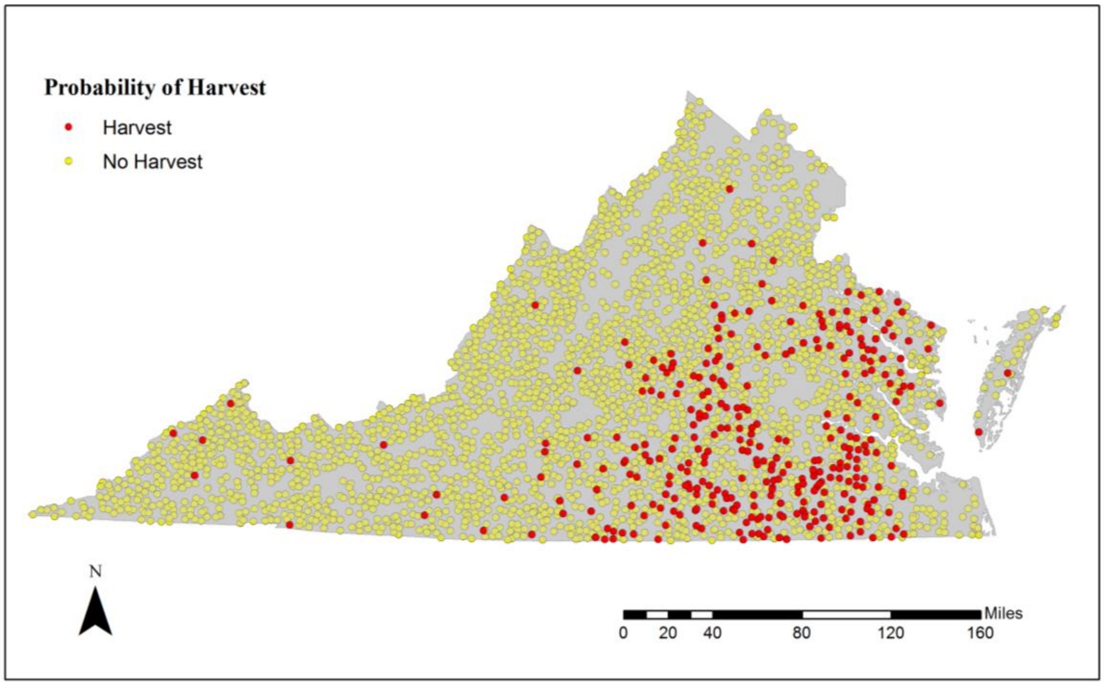

Figure 14. Harvest presence predicted for the 9.5% of conditions with the highest probability of harvest based on the LR model.

Depending on the selected model and classification threshold, the estimated available wood volume from plot conditions likely to be harvested ranged from about 10 – 43 percent of total wood volume (34.09 billion cubic feet) on timberland (Table 17). The availability estimate from the 9.5% of conditions with the highest probability of harvest is very similar for the bRF model (3.55 billion cubic feet, 10% of total volume) and LR model (4.53 billion cubic feet, 13% of total volume). There is a lot more variability in the availability estimate using the prevalence threshold. The LR model predicts twice as much wood is available (14.7 billion cubic feet, 43% of total volume) than the bRF model predicts is available (4.53 billion cubic feet, 18% of total volume).

Table 17. Estimated wood availability on timberland. Available volume was calculated for all conditions with a probability of harvest greater than the prevalence threshold and for the 9.5% of plots with the highest probability of harvest. Volume is estimated from the FIA Virginia 2012 evaluation group.

Harvest Classification		Total Volume	Estimated Available Volume	
Model	Threshold	(Billion Cubic Feet)	(Billion Cubic Feet)	(Percent of Total)
bRF	Prevalence (0.5)	34.09	6.15	18.05
bRF	Top 9.5%		3.55	10.42
LR	Prevalence (0.095)		14.70	43.14
LR	Top 9.5%		4.53	13.28

Chapter 5. DISCUSSION

Model Performance

Six modeling techniques were applied to predict the presence/absence of harvest at FIA plot conditions in Virginia. The single-tree models, CART and cTREE, provide simple diagrams that are easy to understand with minimal instruction. However, such methods are often unstable and are sensitive to variations in the training the data. While the CART and cTREE models predicted harvest presence better than random, they did not predict harvests nearly as well as the logistic regression and ensemble tree models (RF, bRF, and cRF). This is consistent with previous research showing that both RF (Cutler et al. 2007; Maloney et al. 2009) and cRF (Maloney et al. 2009) are better classifiers than CART and cTREE for presence/absence data.

Although they provide more accurate predictions, ensemble methods are not as intuitive to interpret as CART and cTREE models. Variable importance plots help characterize the relationship between response and predictor variables, but the effect of the variable on the response is not evident without providing partial dependence plots. There were a few notable differences between the ensemble methods. Although the cRF model had the lowest sensitivity (proportion of actual harvest presences correctly predicted), it also had the greatest specificity (proportion of actual harvest absences correctly predicted) and the smallest bias. This indicates that while the cRF model doesn't predict harvests as often as other models and fails to predict some actual harvests, the harvest presences predictions may be more reliable.

Conversely, the RF model had the greatest sensitivity and largest bias of all of the ensemble methods. The RF model correctly predicts harvest presence more often than cRF, but with false positive harvest predictions as well. The bRF model seems to make the best compromise between the cRF and RF models. The bRF model increases the sensitivity of predictions without sacrificing much reliability - bias is still fairly low and specificity is not much lower than the cRF model (Table 15). An added benefit of bRF is greatly reduced processing time. The bRF model down-samples the larger response class, thereby reducing the number observations that need to be to fit at each tree.

The LR model performed similarly to RF for all accuracy measures on the testing data. Given LR's weaknesses modeling complex and highly correlated data we expected the LR model to perform much worse than the ensemble tree methods. However, even though the LR model does not predict harvest quite as well as the bRF model, it still predicts harvest relatively well. And unlike the bRF (or any other ensemble tree) model, LR models can be easily shared with other users. In order to share a bRF model, it is also necessary to supply the underlying data, which generates the decision trees. Although log odds coefficients can be confusing to interpret, no other information is needed to replicate a LR model. The bRF model is the most accurate while the LR model is the most easily reproduced, without losing too much accuracy.

Unlike decision tree models, logistic regression is not well suited to model responses with many predictors. LR modeling can be complicated and time consuming if the user must sort through many candidate variables. By using cRF variable importance measures for initial variable selection, the LR modeling process was greatly simplified. cRF was selected for variable selection since both RF and bRF variables importance measures are known to have a bias towards continuous and correlated variables (Strobl et al. 2007).

Depending on the intended application, either the bRF or LR model may be appropriate. The main benefit of the LR model is that it can be quickly and easily shared. Although not quite as accurate as the bRF model, it provides a good approximation of harvest probability. If a user is simply interested in a quick assessment of harvest probability, the LR model may be sufficient. However, if a user wants more detail about variable importance or more accurate predictions, the bRF model is more suitable. While the LR model can provide a quick estimate, it is important to note that both the LR and BRF models are calibrated for Virginia. For accurate predictions in new locations, especially areas that may be very different from Virginia, new models should be developed.

Variable Importance

In terms of variable importance, the RF (Figure 5) and cRF (Figure 7) models were more similar than the bRF model (Figure 6). As expected, given RF's selection bias against categorical variables (Strobl et al. 2007), the cRF model identified more categorical variables as important than the RF model. However, this selection bias was not apparent in the bRF model, which identified 3 categorical variables as very important. Both the RF and cRF models ranked variables related to percent cover much higher than the bRF model. Such variables have less significance when the data are balanced, explaining why the bRF model ranked other categorical variables higher than the RF model.

None of the ensemble methods performed drastically better or worse than another, thus the differences in variable importance don't seem to have a huge impact on model accuracy. Net volume, cover type, distance to urbanized areas, stand age and percent forest cover were some of the most important variables selected by all of the decision tree models (Table 12). The final LR model also indicated that harvest presence is significantly correlated to net volume and land cover type. Although distance to urban area wasn't included in the LR model, percent developed, which likely represents similar population influence, had a significant relationship to harvest. Similarly, stand age was not included in the LR model, but stand size (predominant tree diameter) was a significant variable. The most important variables common to all of the models can be broadly characterized as either a stand condition or population pressure. Biophysical constraints related to accessibility/operability (e.g. slope) were rarely identified as important harvest constraints.

Butler et al. (2010) recently estimated the impact of social and biophysical constraints on wood availability in the Northern US. In their study they constrained harvest by slope, physiographic class, site productivity and stand size as determined by FIA data. The exact same variables were included in this analysis. Based on our results, operability factors including slope and physiographic class may not constrain harvest as much as Butler et al. (2010) estimated. Other stand related conditions including site productivity and stand size seem to be more important in predicting availability. Further analysis is needed to determine to what extent operability constraints may or may not impact the probability of harvest and ultimately wood availability.

Butler et al. (2010) also included a variety of social constraints, some of which were also included in this study such as distance to road and population metrics. Our models indicate that variables related to population are important availability constraints and distance to road may also be important. Butler et al. (2010) also constrained wood availability based on the size of forest holdings. Although we were unable to incorporate this variable, there is significant evidence from previous research indicating that the size of forest holdings is significantly correlated to harvest (Table 1). Future research should incorporate this variable into harvest models, as it may be an important harvest predictor.

Predicting Future Availability

The bRF and LR models were selected as the best models and applied to predict the future probability of harvest at FIA plots in Virginia from 2012 - 2017. This is first application of decision tree/random forest methods to predict harvest presence/absence. Most previous studies have applied logistic regression, but at much smaller scales and without widely available data. This research shows that using random forest, specifically bRF, future harvests can be predicted over large regions. Since widely available datasets were identified to represent harvesting constraints, new harvest models can be easily developed for additional locations.

Since harvest was predicted across the entire state of Virginia, we can evaluate the spatial distribution of wood supply availability. Harvest predictions are distributed similarly throughout

Virginia for both the LR and bRF models. As described in the results, decreased probability of harvest is evident near urban areas and National Forest in both models. Such predictions around urban areas are reasonable since the LR model included variables for population density as well as percent of developed land cover and the bRF model ranked distance to urban areas as very important. This reaffirms that variables correlated to population and development may be very important in predicting future harvests. Additionally, lower probabilities of harvest for national forest make sense for the models given that the LR model included an ownership variable, while the bRF model ranked protected area designation and ownership as important predictors.

Harvest presence was predicted from both the LR and bRF models for (1) all observations with a probability of harvest was greater than the prevalence threshold and (2) the 9.5% of plots with the highest probability of harvest. The prevalence threshold prediction (1) represents an upper estimate of available volume while the high probability prediction (2) represents a lower estimate of available volume. The bRF model predicted that 3.55 – 6.15 billion cubic feet, 10 – 21 percent of the total tree volume on timberland is available for harvest or likely to be harvested between 2012 and 2017. The LR model predicted that between 4.53 and 14.7 billion cubic feet, 12 – 44 percent of the total tree volume on timberland would be available wood supply.

Between 2002 and 2007, 4.14 billion cubic feet of live trees were removed from timberland (Rose, 2009). The bRF estimate of 3.55 – 6.14 billion cubic feet of available wood supply is most similar to previously reported removals within a similar 5-year time frame (Table 18). The LR estimate of 4.53 – 14.70 billion cubic feet of available wood supply has a much larger range and greater upper estimate of availability, around 255% more than previously observed removals (Table 18). Given the bRF model’s greater accuracy (see discussion of model performance) and its similarity to previous removals, the bRF model’s wood availability estimate is likely more reliable than the LR model’s wood availability estimate.

Table 18. Comparison of estimated availability to previously reported removals (Rose, 2009).

Harvest Classification		2002 - 2007 Removals	Estimated Available Volume	
Model	Threshold	(Billion Cubic Feet)	(Billion Cubic Feet)	(Percent Change)
bRF	Prevalence (0.5)	4.14	6.15	48.68
bRF	Top 9.5%		3.55	-14.19
LR	Prevalence (0.095)		14.70	255.40
LR	Top 9.5%		4.53	9.41

Whether the LR or bRF model is applied, if the available volume is estimated with the top 9.5% of plots, the estimated removal volumes are similar and within 15% of previous removals (Table 18). This threshold provides a conservative estimate of availability, applicable to forecast the minimum available supply. Either the LR or bRF model, with this threshold could be used to estimate minimum availability. In terms of minimum availability, it is likely that at least 10% of total wood volume, but perhaps even more of the total wood volume, will be available for removals from 2012 – 2017 (Table 17).

Although the bRF model provides a more realistic estimate of availability, with the prevalence threshold, both models predicted harvest presence much more often than was previously observed in the training data. This upper limit reflects what could become available, but is likely an overestimation of availability due to constraints, which are unknown or difficult to model. Based on the bRF model, it is possible that 21% of the total wood supply on timberland could be available, but this estimate probably includes some plot conditions that may not be harvested for a variety of reasons, in particular landowner preference. It is likely that only a portion of the upper availability estimate (21%) will actually be available due to landowners who simply decide not to harvest. If more information were known about ownership, specifically private landowners, a more robust estimate of availability would be possible. Future studies may be able to incorporate landowner preferences through the National Woodland Owner Survey (NWOS), similar to Butler et al. 2010.

Depending on the application, either the lower or upper availability estimate may be preferred. For example, if a user is interesting in establishing a new mill, a more conservative, but reliable minimum estimate may be preferred. However, a policy maker aiming to regulate any potentially available wood supply may be more interested in a slight overestimate, which is more likely to include all future harvests. For a lower estimate, either the bRF or LR model may be applied. But only the BRF model should be used for a reliable upper availability estimate.

Although many improvements are certainly possible, this study provides an initial estimate of wood availability in Virginia. This research builds upon previous findings to describe how harvest occurrences can be predicted at a statewide level. The models developed are best suited to answer questions regarding the probability of harvest (i.e. what is most likely to be harvested) rather than predicting exact harvest presence. Observations with a high probability of harvest can be used to estimate how much wood is available to be extracted.

Effective forest conservation and management requires realistic estimates of wood supply and distribution over large geographic areas. These results describe the current status and distribution of wood availability in in Virginia. It is evident that not all wood is equally available to be harvested and that the total wood supply is substantially different from the available wood supply. Estimates of future wood availability can help natural resource professionals, forest industry, government and policy-makers develop management plans that are relevant to current and future forest conditions.

REFERENCES

Alig, R. J., and D. N. Wear. 1992. Changes in Private Timberland: Statistics and Projections for 1952 to 2040. *Journal of Forestry*. 90(5):31–36.

Barlow, S. A., Munn, D. Cleaves, and D. L. Evans. 1998. The Effect of Urban Sprawl on Timber Harvesting: A Look at Two Southern States. *Journal of Forestry*. 12(5):10–14.

Bechtold, W.A., P.L. Patterson., Editors. 2005. *The enhanced Forest Inventory and Analysis Program – national sampling design and estimation procedures*. USDA Forest Service. Gen. Tech. Rep. SRS-80. 85 p.

Boyd, R. 19984. Government support of non-industrial production: the case of private forests. *Southern Economic Journal*. 51(1): 89 – 107.

Breiman, L. 2001. Random Forests. *Machine Learning*. 45(1):5–32.

Breiman, L., J.H. Olshen, and C.J. Stone. 1984. *Classification and regression trees*. Wadsworth and Brooks, Monterey, California, USA. 368 p.

Brinckman, M. D., and J. F. Munsell. 2012. Disproportionality, Social Marketing, and Biomass Availability: A Case Study of Virginia and North Carolina Family Forests. *Southern Journal of Applied Forestry*. 36(2):85–91.

Butler, B. J., Z. Ma, D. B. Kittredge, and P. P. Catanzaro. 2010. Social versus biophysical availability of wood in the northern United States. *Northern Journal of Applied Forestry.* 27(4):151–159.

Chen, C., A. Liaw, L. Breiman. 2004. *Using Random Forest to Learn Imbalanced Data.* UC Berkeley. Technical Report 666. 12 p.

Cleaves, D., and M. Bennett. 1995. Timber Harvesting by Nonindustrial Private Forest Landowners in Western Oregon. *WJAF*. 10(2):66–71.

Cleaves, D., Munn, S. A. Barlow, and D. L. Evans. 2002. Urbanization's impact on timber harvesting in the south central United States. *Journal of Environmental Management*. 64(1):65–76.

Conway, M.C., S. Chapman, G.S. Amacher, J. Sulliavan. 2000. Differences in non-industrial landowner behavior between Hardwood and Pine regions of Virginia: implications for timber supply. *SOFAC* Report No. 19.

Cramer, J. S. 1999. Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society: Series D (The Statistician)*. 48(1):85–94.

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

De'Ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*. 81(11):3178–3192.

Dennis, D. F. 1990. A probit analysis of the harvest decision using pooled time-series and cross-sectional data. *Journal of Environmental Economics and Management*. 18:176–187.

Dennis, D. F. 1989. An Economic Analysis of Harvest Behavior: Integrating Forest and Ownership Characteristics. *Forest Science*. 35(4):1088–1104.

ESRI. 2013. *StreetMap Premium for ArcGIS.*

Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters*. 27(8):861–874.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*. 24(1):38–49.

Forest Inventory and Analysis. 2013a. *FIA DataMart FIADB version 5.1*. Available online at http://apps.fs.fed.us/fiadb-downloads/datamart.html; last accessed May 30, 2013.

Forest Inventory and Analysis. 2013b. *The Forest Inventory and Analysis Database: Database*

Description and Users Manual Version 5.1.6 for Phase 2. Arlington, VA. USDA Forest Service 556p.

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., and Wickham, J., 2011.Completion of the 2006 National Land Cover Database for the Conterminous United States, *PE&RS*, Vol. 77(9):858-864.

Hothorn, T., K. Hornik, and A. Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics.* 15(3):651–674.

Hosmer, Jr, D.W., and S. Lemeshow. 2004. *Applied logistic regression*. John Wiley and Sons, Hoboken, New Jersey, USA. 392 p.

Hyberg, B. T., and D. M. Holthausen. 1989. The behavior of nonindustrial private forest landowners. *Canadian Journal of Forest Research*. 19:1014–1023.

Jamnick, M. S., and D. R. Beckett. 1988. A logit analysis of private woodlot owner's harvesting decisions in New Brunswick. *Canadian Journal of Forest Research*. 18(3):330–336.

Jimenez-Valverde, A., and J. M. Lobo. 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*. 31(3):361–369.

Johnson, T.G., J.W. Bentley, M. Howell. 2009. *The South's timber industry – an assessment of timber product output and use, 2007*. Asheville, NC. U.S. Department of Agriculture Forest Service, Southern Research Station Resour. Bull. SRS-164. 52 p.

Kelty, M.J., A.W. D'Amato, P.K. Barten. 2008. *Silvicultural and Ecological Considerations of Forest Biomass Harvesting in Massachusetts.* Prepared for the Massachusetts Division of Energy and Resources and Massachusetts Department of Conservation and Recreation.

Kuuluvainen, J., H. Karppinen, and V. Ovaskainen. 1996. Landowner Objectives and Nonindustrial Private Timber Supply. *Forest Science*. 42(3):300–309.

Kuuluvainen, J., J. Salo. 1991. Timber supply and life cycle harvest of non-industrial private forest owners: an empirical analysis of the Finnish case. *Forest Science.* 37: 1011 – 1029.

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*. (28):385–393.

Lobo, J. M., A. Jimenez-Valverde, and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*. 17(2):145–151.

Maloney, K. O., D. E. Weller, M. J. Russell, and T. Hothorn. 2009. Classifying the biological condition of small streams: an example using benthic macroinvertebrates. *Journal of the North American Benthological Society*. 28(4):869–884.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*. 38(5):921–931.

Markowski-Lindsay, M., P. P. Catanzaro, D. Damery, D. B. Kittredge, B. J. Butler, and T. Stevens. 2012. Forest-based biomass supply in Massachusetts: how much is there and how much is available. *Journal of Environmental Management*. 106:1–7.

McRoberts, R. E., G. R. Holden, M. D. Nelson, G. C. Liknes, W. K. Moser, A. J. Lister, S. L. King, E. B. LaPoint, J. Coulston, and W. B. Smith. 2005. Estimating and circumventing the effects of perturbing and swapping inventory plot locations. *Journal of Forestry*. 103(6):275–279.

Munn, and Arano, K. G. 2006. Evaluating forest management intensity: a comparison among major forest landowner types. *Forest Policy and Economics*. 9:237–248.

National Conservation Easement Database. 2011. *National Conservation Easement Database (NCED).* Available online at http://www.conservationeasement.us/; last accessed July

15, 2013.

Prisley, S.P., H. Wang, P. Radtke, J. Coulston. 2009. Combining FIA Plot Data with Topographic Variables: Are Precise Locations Needed? USDA Forest Service Proceedings of *2008 Forest Inventory and Analysis (FIA) Symposium*. Park City, RMRS.

Paula, A., C. Bailey, R. Barlow, and W. Morse. 2011. Landowner Willingness to Supply Timber for Biofuel: Results of an Alabama Survey of Family Forest Landowners. *Southern Journal of Applied Forestry*. 35(2):93–97.

Pearce, J., and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*. 133:225–245.

Perlack, R., L. Wright, A. Turhollow, R. Graham, B. Stokes, D. Erbach. 2005. *Biomass as feedstock for a bioenergy and bioproducts industry: The technical feasibility of a billion-ton annual supply*. USDE and USDA. Oak Ridge National Laboratory. DOE/GO - 02005-2135. 59 p.

Pontius, R. G., and B. Parmentier. 2014. Recommendations for using the relative operating characteristic (ROC). *Landscape Ecol*. 29(3):367–382.

Prestemon, J. P., and D. N. Wear. 2000. Linking harvest choices to timber supply. *Forest Science*. 46(3).

R Core Team 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reams, G. A., F. A. Roesch, and N. D. Cost. 1999. Annual forest inventory: cornerstone of sustainability in the South. *Journal of Forestry*. 97(12):21–26.

Romm, J., R. Tuazon, and C. Washburn. 1987. Relating Forestry Investment to the Characteristics of Nonindustrial Private Forestland Owners in Northern California. *Forest Science*. 33(1):197–209.

Rose, Anita K. 2009. V*irginia's forests, 2007*. Asheville, NC. U.S. Department of Agriculture Forest Service, Southern Research Station. Resour. Bull. SRS - 159. 77p.

Smith, B. W. 2002. Forest inventory and analysis: a national inventory and monitoring program. *Environmental Pollution*. 116 (Suppl 1):233–42.

Strobl, C., A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*. 8:25.

Størdal, S., Baardsen, S., and G. Lien. 2008. Analyzing determinants of forest owners' decision-making using a sample selection framework. *Journal of Forest Economics*. 14:159–176.

U.S. Census Bureau. 2010a. *2010 Census Urban and Rural Classification and Urban Area Criteria: Lists of 2010 Census Urban Areas*. Available online at http://www.census.gov/geo/reference/ua/urban-rural-2010.html; last accessed June 15, 2013.

U.S. Census Bureau. 2010b. *TIGER/Line Shapefiles Pre-joined with Demographic Data: 2010 Census Population & Housing Unite Counts - Blocks*. Available online at http://www.census.gov/geo/maps-data/data/tiger-data.html; last accessed June 15, 2013.

U.S. Geological Survey. 2012. *Protected Areas Database of the United States (PADUS) version 1.3.* http://gapanalysis.usgs.gov/padus/data/download/; last accessed July 15, 2013.

U.S. Environmental Protection Agency. 2006. *NHDPLUS Flowlines*. Available online at http://nhd.usgs.gov/data.html; last accessed June 15, 2013.

Vaughan, I. P., and S. J. Ormerod. 2005. The continuing challenges of testing species distribution models. *Journal of Applied Ecology*. 42(4):720–730.

Vayssières, M. P., R. E. Plant, and B. H. Allen Diaz. 2000. Classification trees: An alternative non- parametric approach for predicting species distributions. *Journal of vegetation science*.

11(5):679–694.

Virginia Department of Forestry. 2012. *2012 State of the forest. Annual report on Virginia's forests.* Available online at http://dof.virginia.gov/print/aboutus/SOF-2012.pdf; last accessed January 15, 2013.

Vokoun, M., G. S. Amacher, and D. N. Wear. 2006. Scale of harvesting by non-industrial private forest landowners. *Journal of Forest Economics*. 11:223–224.

Wear, D.N., C.R. Douglas, J.Prestmon. 2007. *The U.S. South's timber sector in 2005: a prospective analysis of recent change.* USDA Forest Service Gen. Tech. Re. SRS-99 29 p.

Wear, D. N., and R. Flamm. 1993. Public and Private Forest Disturbance Regimes in the Southern Appalachians. *Natural Resource Modeling*. 7(4):379–397.

Wear, D. N., M. C. Conway, G. S. Amacher, and J. Sullivan. 2003. Decisions nonindustrial forest landowners make: an empirical examination. *Journal of Forest Economics*. 9(3):181–203.

Wear, D. N., R. Liu, J. Michael Foreman, and R. M. Sheffield. 1999. The effects of population growth on timber management and inventories in Virginia. *Forest Ecology and Management*. 118(1-3):107–115.

Woudenberg, S. W., B. L. Conkling, and B. M. O'Connell. 2010. The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2.

Zhang, D. 2004. Endangered Species and Timber Harvesting: The Case of Red-Cockaded Woodpeckers. *Economic Inquiry*. 42(1):150–165.

APPENDIX A: Variable Descriptions and Summary Statistics

Data Source	Variable	Description	Data Type	Units/Codes	Descriptive Statistics		
					Freq.	Mean	Range
FIA	RdDist	Distance to closest road	Ord.	1 - 100 ft or less	769		
				2 - 101 ft to 300 ft	1109		
				3 - 301 ft to 500 ft	953		
				4 - 501 ft t0 1000 ft	1563		
				5 - 10001 ft - 1/2 mile	2596		
				6 - 1/2 to 1 mile	809		
				7 - 1 to 3 miles	220		
				8 - 3 to 5 miles	4		
				9 - Greater than 5 miles	3		
FIA	Water	Water on plot	Cat.	0 - none	7204		
				1 - Permanent streams or ponds too small to qualify as non-census water	498		
				2 - Permanent water in the form of deep swamps, bogs, marshes without standing trees	100		
				3 - Ditch/canal - human-made channels used a means of moving water	21		
				4 - temporary streams	155		
				5 - Flood zones	36		
				9 - Other temporary water	39		
FIA	Reserv	Reserved status, land management for wood products prohibited		0 = Not reserved	7863		
				1 = Reserved	190		

Data	Variable	Description	Data	Units/Codes	Descriptive Statistics		
Source			Type		Freq.	Mean	Range
FIA	Own	Ownership type*	Cat.	11 - National Forest	649		
				21 - National Park Service	78		
				24 - Departments of Defense/Energy	110		
				25- Other Federal	5		
				31 - State	164		
				32 - Local	118		
				46 - Private	6898		
FIA	StdAge	Stand age *	Cont.	Years		47.77	0 - 173
FIA	StdSz	Stand size*	Ord.	1 - Large diameter trees	4715		
				2 - Medium diameter trees	2049		
				3 - Small diameter trees	1233		
				4 - Nonstocked	56		
FIA	SiteProd	Site productivity*	Ord.	1 = 225 + (ft^3/ac per year)	7		
				2 = 165 - 224	253		
				3 = 120 - 164	840		
				4 = 85 - 119	2303		
				5 = 50 - 84	3472		
				6 = 20 - 49	1162		
				7 = 0 - 19	16		
FIA	StdOrg	Type of stand regeneration*	Cat.	0 = Natural stands	6896		
				1 = Artificial regeneration	1157		
FIA	Slope	Slope*	Cont.	Percent			0 - 150
FIA	Physiog	Physiographic class*	Cat.	1 = xeric	978		
				2 = mesic	6906	18.46	
				3 = hydric	92		

Data Source	Variable	Description	Data Type	Units/Codes	Descriptive Statistics		
					Freq.	Mean	Range
FIA	AlStk	Stocking by live trees*	Ord.	1 = Overstocked (130% +)	790		
				2 = Fully stocked (100 - 129.9%)	3711		
				3 = Medium Stocked (60 - 99.9%)	2753		
				4 = Poorly stocked (16.7 - 59.9%)	699		
				5 = Nonstocked (<16.7%)	100		
FIA	Operable	Ability to operate logging equipment*	Cat.	0 = No problems	4516		
				1 = Seasonal access due to water conditions in wet weather	565		
				2 = Mixed wet and dry areas	90		
				3 = Broken terrain that would severely limit equipment, access or use	22		
				4 = Year-round water problems	135		
				5 = Slopes 20 - 40%	2465		
				6 = Slopes greater than 40%	260		
FIA	NetVol	Net tree volume	Cont.	Cubic feet/ acre		1743	1.50 - 8833
FIA	Dstrb_any	Any disturbances recorded*	Cat.	0 = No disturbances	7021		
				1 = At least one disturbance	1032		
FIA	Dstrb_tot	Total disturbances recorded*	Count				0 - 3
FIA	Mgmt_harv	Harvest*	Cat.	0 = Not harvested (previous)	7223		
				1= Harvested (previous)	830		
FIA	Mgmt_any	Any management*	Cat.	0 = Not managed (previous)	7152		
				1 = Managed (previous)	901		

Data Source	Variable	Description	Data Type	Units/Codes	Descriptive Statistics		
					Freq.	Mean	Range
NHD	StrmDist_per	Distance to perennial stream	Ord.	Feet	624.6		0.21 - 64999
NHD	StrmDist_any	Distance to closest stream	Cont.	Feet	220.2		0.21 - 64999
StreetMap	RdDist_any	Distance to closest road	Cont.	Feet		214.2	0.128 - 3388
StreetMap	RdDist_hw	Distance to closest highway	Cont.	Feet		2471	0.23 – 21360
TPO	SA_25mi	Mills in the service area	Count	Total within 25 miles			0 - 13
TPO	SA_50mi	Mills in the service area	Count	Total within 50 miles			0 - 40
TPO	SA_100mi	Mills in the service area	Count	Total within 100 miles			1 - 114
NLCD	CoverType	Land cover type at plot	Cat.	11 = Open water	9		
				21 = Developed, open space	251		
				22 = Developed, low intensity	33		
				23 = Developed, med. intensity	12		
				24 = Developed, high intensity	3		
				31 = Barren land	14		
				41 = Deciduous forest	4935		
				42 = Evergreen forest	1084		
				21 = Developed, open space	251		
				43 = Mixed forest	345		
				52 = Shrub/scrub	263		
				71 = Grassland/herbaceous	130		
				81 = Pasture/hay	270		
				82 = Cultivated crops	73		
				90 = Woody wetlands	602		
				95 = Emergent herb. wetlands	16		

Data Source	Variable	Description	Data Type	Units/Codes	Descriptive Statistics		
					Freq.	Mean	Range
NLCD	PctFor_dec	Deciduous forest	Cont.	Percent within 10 miles		44.54	0.59 - 86.86
NLCD	PctFor_ev	Evergreen forest	Cont.	Percent within 10 miles		11.17	0.09 - 38.33
NLCD	PctFor_mix	Mixed forest	Cont.	Percent within 10 miles		3.457	0.16 - 10.36
NLCD	PctFor_all	Forest (any type)	Cont.	Percent within 10 miles		59.29	2.52 - 92.01
NLCD	PctDev_open	Developed, open space	Cont.	Percent within 10 miles		5.04	1.40 - 25.48
NLCD	PctDev_low	Developed, low intensity	Cont.	Percent within 10 miles		1.61	0 – 19.2
NLCD	PctDev_high	Developed, high intensity	Cont.	Percent within 10 miles		0.15	0 - 3.30
NLCD	PctDev_all	Developed (any type)	Cont.	Percent within 10 miles		7.306	2.02 - 52.57
NLCD	PctWater	Open water	Cont.	Percent within 10 miles		2.90	0.01 - 70.90
Census	PopDens	Population density of census block	Cont.	People / ha		71.87	0 - 9191
Census	HousDens	Housing density of census block	Cont.	Housing unit / ha		11.04	0 - 3120
Census	UrbanDist	Distance to closest urban area or cluster	Cont.	Miles		3.70	0.168 - 11.67
Census	UrbanPGI	Population gravity index	Cont.	Population/ Distance (mi) ^{2}		11580	18.6 - 2060000
Census	UrbanDist_A	Distance to closest urbanized area	Cont.	Miles		8.70	0.47 - 19.14

Data Source	Variable	Description	Data Type	Units/Codes	Descriptive Statistics		
					Freq.	Mean	Range
PADUS	PAD	Protected area	Cat.	1 = Inholding	13		
				2 = National Park	53		
				3 = National Wildlife Refuge	31		
				4 = Parkway	4		
				5 = Reservoir	33		
				6 = State Forest	31		
				7 = State Park	20		
				8 = Wildlife Management Area	95		
				9 = Easement	2		
				10 = Military Installation	60		
				11 = National Forest	560		
				12 = National Scenic Trail	4		
				13 = Private Land	7022		
				14 = Special Biological Area	30		
				15 = State Natural Area Preserve	6		
				16 = Wilderness Area	40		
NCED	NCED	Conservation easement	Cat.	0 = n/a	738		
				1,2 = Managed for biodiversity	23		
				3 = Managed for multiple uses	152		
				4 = No known mandate	39		

APPENDIX B: Model Classification Error

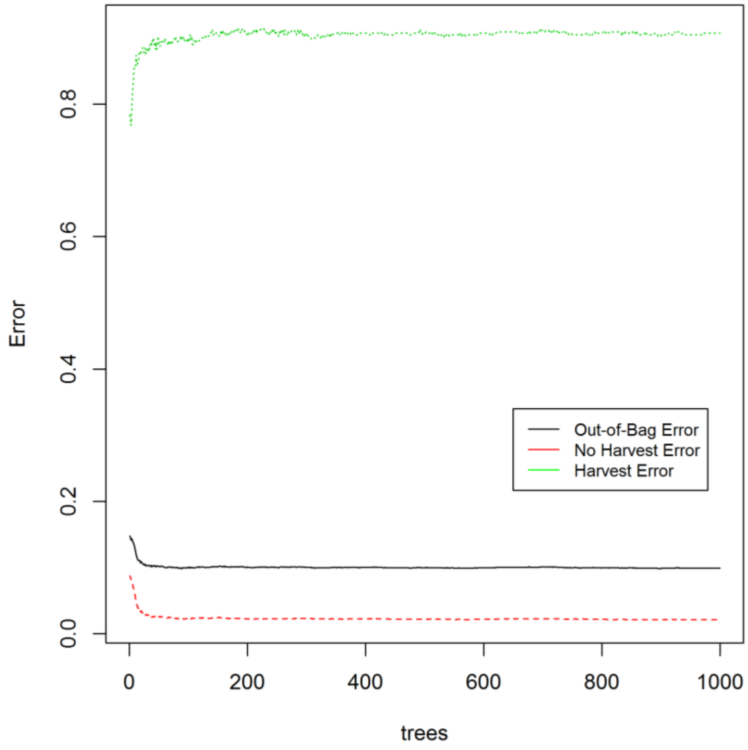

Figure B.1. RF model classification error.

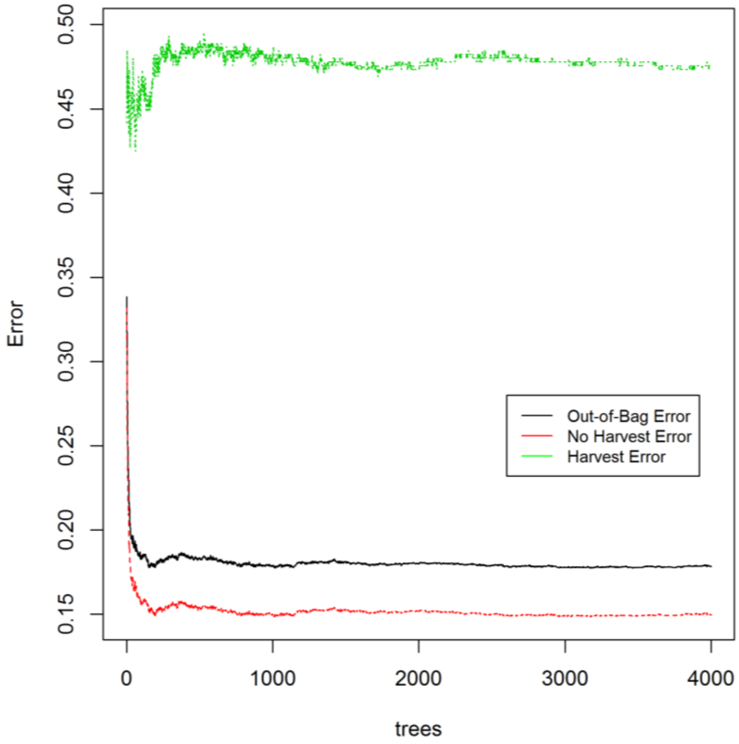

Figure B.2. bRF model classification error.

APPENDIX C: Additional variable importance plots

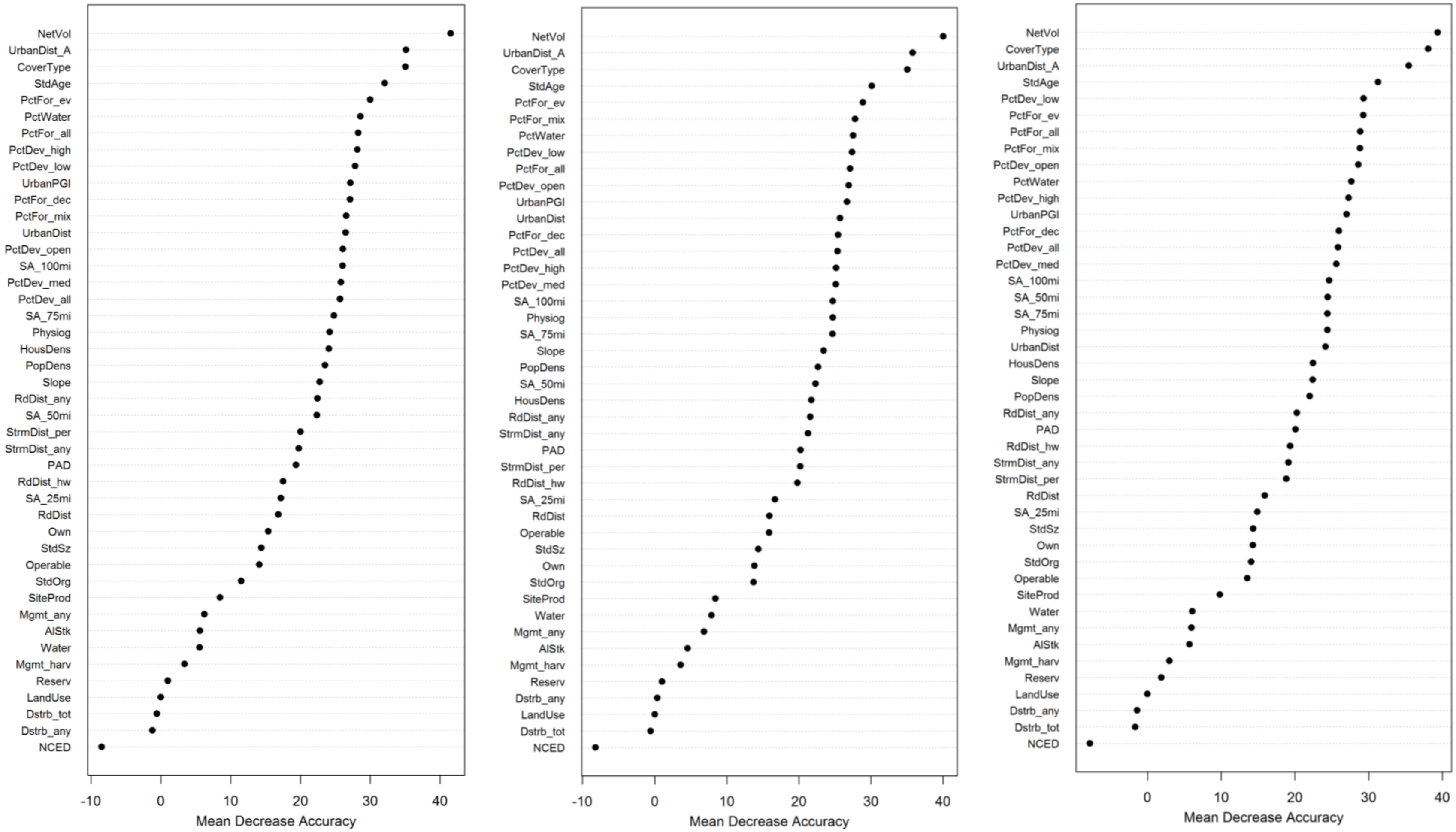

Figure C.1 RF model variable importance from additional runs.

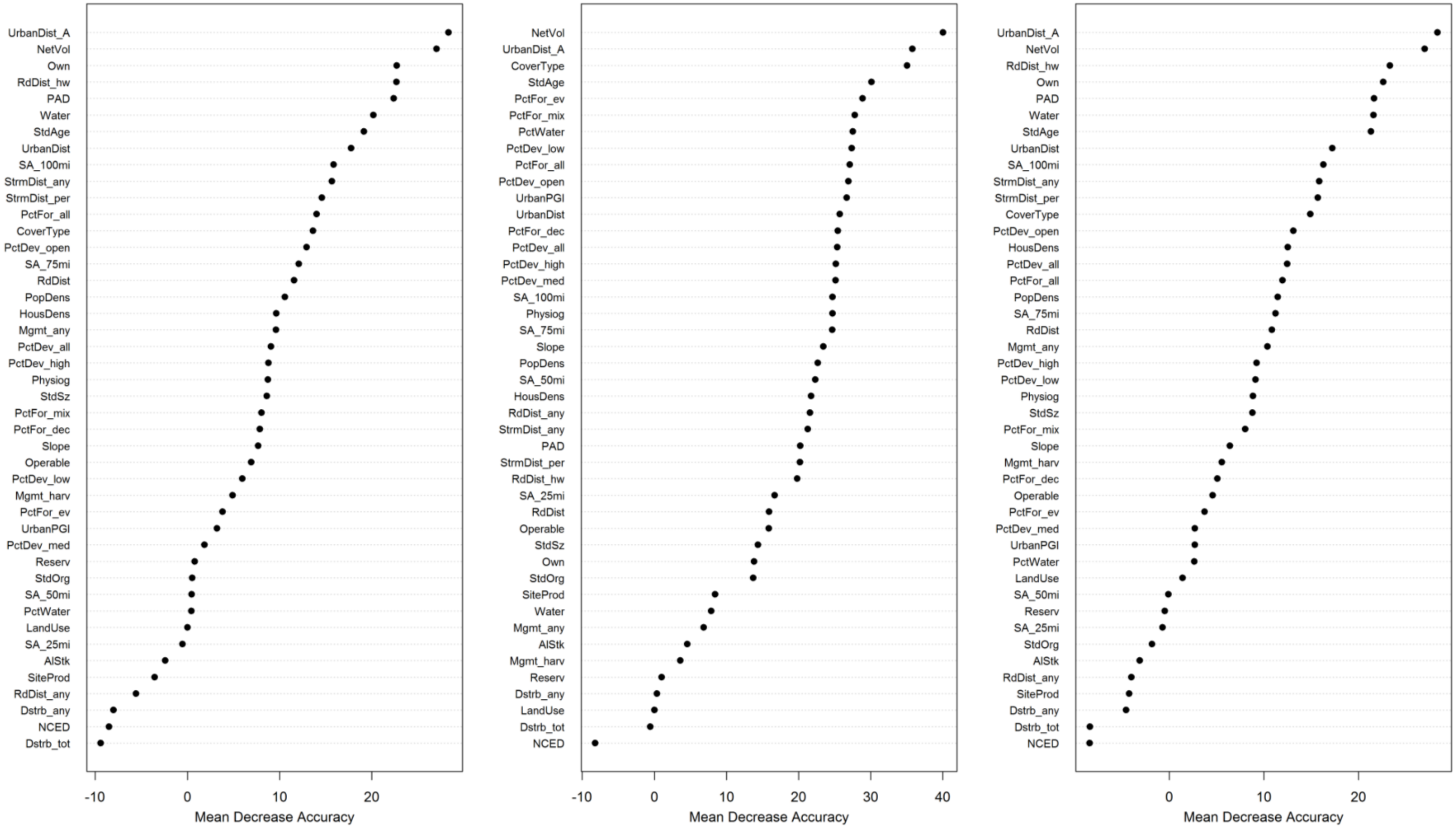

Figure C.2 bRF model variable importance from additional runs.

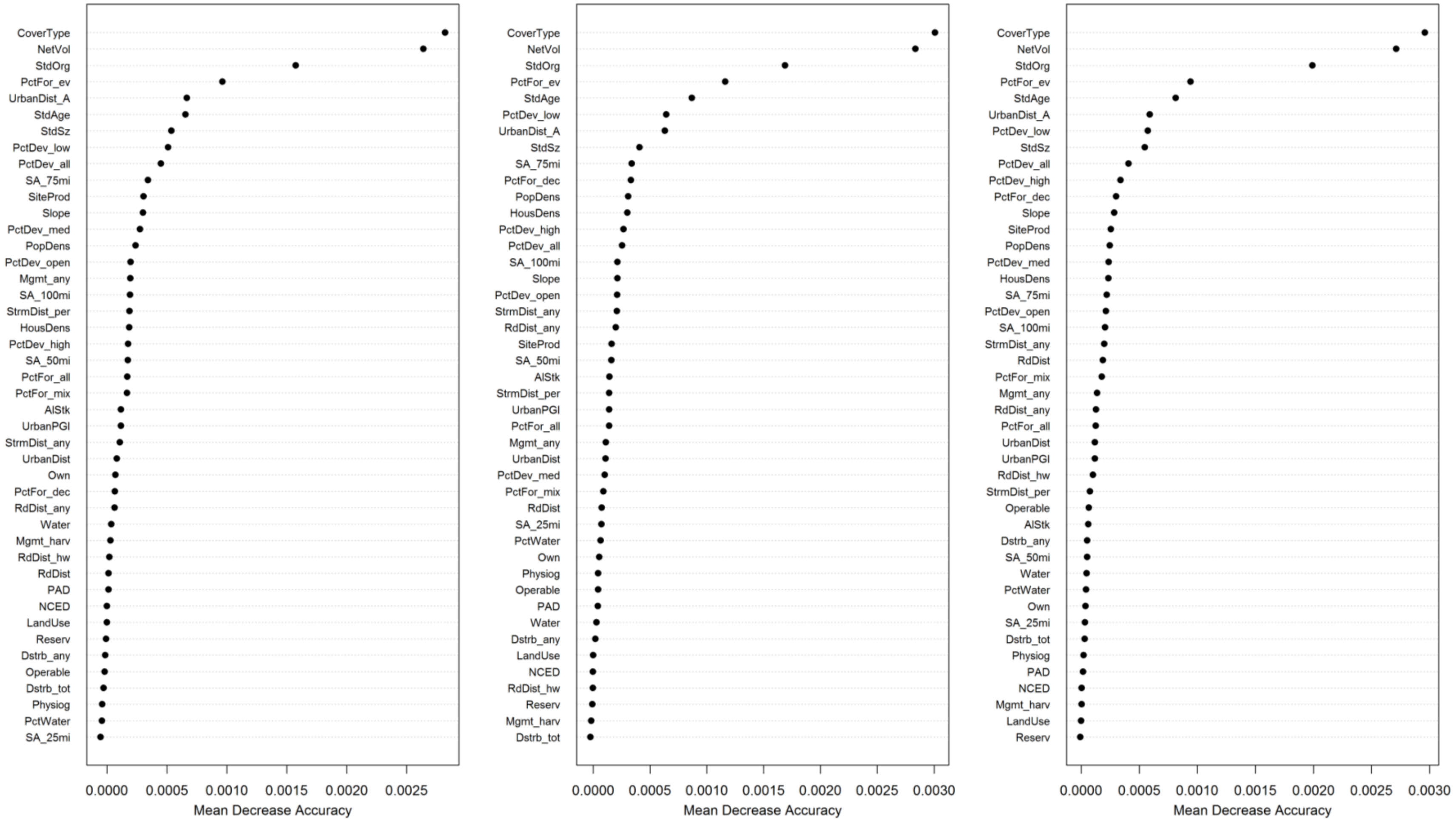

Figure C.3. cRF model variable importance from additional runs.