Methods for the spatial modeling and evaluation of tree canopy cover

Jill M. Datsko

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

Doctor of Philosophy
in
Forestry

Randolph H. Wynne, Co-chair
Valerie A. Thomas, Co-chair
John W. Coulston
Kurt Luther

February 21, 2022
Blacksburg, Virginia

Keywords: Time-series, crowdsourcing, sampling, ensemble methods.
Copyright 2022, Jill M. Datsko

Methods for the spatial modeling and evaluation of tree canopy cover

Jill M. Datsko

(ABSTRACT)

Tree canopy cover is an essential measure of forest health and productivity, which is widely studied due to its relevance to many disciplines. For example, declining tree canopy cover can be an indicator of forest health, insect infestation, or disease. This dissertation consists of three studies, focused on the spatial modeling and evaluation of tree canopy cover, drawing on recent developments and best practices in the fields of remote sensing, data collection, and statistical analysis.

The first study evaluates how well harmonic regression variables derived at the pixel-level using a time-series of all available Landsat images predict values of tree canopy cover. Harmonic regression works to approximate the reflectance curve of a given band across time. Therefore the coefficients that result from the harmonic regression model estimate relate to the phenology of the area of each pixel. We use a time-series of all available cloud-free observations in each Landsat pixel for NDVI, SWIR1 and SWIR2 bands to obtain harmonic regression coefficients for each variable and then use those coefficients to estimate tree canopy cover at two discrete points in time. This study compares models estimated using these harmonic regression coefficients to those estimated using Landsat median composite imagery, and combined models. We show that (1) harmonic regression coefficients that use a single harmonic coefficient provided the best quality models, (2) harmonic regression coefficients from Landsat-derived NDVI, SWIR1, and SWIR2 bands improve the quality of tree canopy cover models when added to the full suite of median composite variables, (3) the harmonic regression constant for the NDVI time-series is an important variable across models, and (4) there is little to no additional information in the full suite of predictors compared to the harmonic regression coefficients alone based on the information criterion provided by principal components analysis.

The second study presented evaluates the use of crowdsourcing with Amazon's Mechanical Turk platform to obtain photointerpretated tree canopy cover data. We collected multiple interpretations at each plot from both crowd and expert interpreters, and sampled these data using a Monte Carlo framework to estimate a classification model predicting the "reliability" of each crowd interpretation using expert interpretations as a benchmark, and identified the most important variables in estimating this reliability. The results show low agreement between crowd and expert groups, as well as between individual experts. We found that variables related to fatigue had the most bearing on the "reliability" of crowd interpretations followed by whether the interpreter used false color or natural color composite imagery during interpretation. Recommendations for further study and future implementations of crowdsourced photointerpretation are also provided.

In the final study, we explored sampling methods for the purpose of model validation. We evaluated a method of stratified random sampling with optimal allocation using measures of prediction uncertainty derived from random forest regression models by comparing the accuracy and precision of estimates from samples drawn using this method to estimates from samples drawn using other common sampling protocols using three large, simulated datasets as case studies. We further tested the effect of reduced sample sizes on one of these datasets and demonstrated a method to report the accuracy of continuous models for domains that are either regionally constrained or numerically defined based on other variables or the modeled quantity itself. We show that stratified random sampling with optimal allocation provides the most precise estimates of the mean of the reference Y and the RMSE of the population. We also demonstrate that all sampling methods provide reasonably accurate estimates on average. Additionally we show that, as sample sizes are increased with each sampling method, the precision generally increases, eventually reaching a level of convergence where gains in estimate precision from adding additional samples would be marginal.

Methods for the spatial modeling and evaluation of tree canopy cover

Jill M. Datsko

(GENERAL AUDIENCE ABSTRACT)

Tree canopy cover is an essential measure of forest health, which is widely studied due to its relevance to many disciplines. For example, declining tree canopy cover can be an indicator of forest health, insect infestation, or disease. This dissertation consists of three studies, focused on the spatial modeling and evaluation of tree canopy cover, drawing on recent developments and best practices in the fields of remote sensing, data collection, and statistical analysis.

The first study is an evaluation of the utility of variables derived from time-series satellite imagery, which describe the timing and magnitude of green-up and leaf loss at each location, to estimate tree canopy cover. This study compares models estimated using these time-series variables to those estimated using image composites, which are compiled using the median reflectance value from multiple dates at each location, and models which used both types of variables. Our results indicate that time-series variables are useful for modeling tree canopy cover, and that they contain more information and are less influenced by extreme observations compared to commonly used multi-temporal variables, suggesting that they may be useful in other applications as well.

The second study presented, evaluates the use of crowdsourcing, a method of data collection where tasks are outsourced to an anonymous online workforce (the crowd) who may or may not have experience with the subject matter. We used Amazon's Mechanical Turk platform to obtain tree canopy cover data, as interpreted from aerial images. Crowd workers view aerial imagery through our online interface and determine whether a series of 100 dots organized in a grid land on trees. These photointerpretations are then aggregated to obtain % tree canopy cover measurements across that grid. We used these photointerpretations and ancillary data collected during interpretation to estimate a classification model predicting the "reliability" of each crowd interpretation using expert interpretations as a benchmark, and identified the most important variables in estimating this "reliability". The results show low agreement between crowd and expert groups, as well as between individual experts. We found that variables related to worker fatigue had the most bearing on the reliability of crowd interpretations followed by variables related to the display settings used to view imagery during interpretation, suggesting that future endeavors may be more successful if crowd workers and experts used the same display settings and if the number of assignments a crowd worker could complete in a specified amount of time was limited.

The final study in this dissertation explored sampling methods for the purpose of model

validation. We evaluated a method called 'stratified random sampling with optimal allocation', a sampling method that is specifically designed to improve the precision of sample estimates, using measures of the variability in predictions from different models. We compared the accuracy and precision of estimates from samples drawn using this method to estimates from samples drawn using other common sampling protocols. We further tested the effect of smaller sample sizes on one of these data products. We show that the method of stratified random sampling with optimal allocation could provide more precise results and could require smaller sample sizes compared to other common methods, therefore providing higher quality results and cost savings associated with data collection.

Dedication

To L.B.D. and M.P.D.

Acknowledgments

First, I would like to acknowledge my co-advisers, Dr. Randy Wynne and Dr. Val Thomas, for their guidance, advice, and expertise and their immense patience in addressing all of my questions while also making me feel like a respected member of a team. Their support over the years has been humbling, and I am tremendously grateful.

This work would also not be possible without the input and guidance of my other committee members, Dr. Kurt Luther, whose expertise about crowdsourcing has been instrumental to this dissertation, and Dr. John Coulston, whose mentorship has been invaluable in demystifying some of the more complex concepts herein, and who has been very generous in taking the time to work through questions and address issues related to each part of the studies that follow.

I am indebted to our collaborators at the US Forest Service, including Greg Likness, Dr. Karen Schleeweis, Stacie Bender, and others, whose guidance and dedication was unyielding over many meetings and presentations and whose continued interest has vastly improved this work.

I would not have been able to compile this work without the help of the greater Virginia Tech community. Specifically, I would like to thank Dr. Evan Brooks, who worked directly with me to apply components of his work to these studies, Seth Peery, who patiently lent his expertise through various iterations in the development and creation of our crowd sourcing interface, Dr. Susmita Sen, who dedicated her time and expertise in photointerpretation of tree canopy cover to enrich this work, and Muna Khatiwada who volunteered many hours in support of data collection efforts for these studies.

I would like to thank Dr. Carol Franco for her help in navigating this process and for keeping me apprised of various research activities at every possible opportunity to the extent that time allowed. I would additionally like to acknowledge the support of other faculty and staff members. In particular, Dr. Christine Blinn, Dr. Stella Schons, Dr. John McGee, Dr. Leandro Castello, and Stacey Kuhar who have all provided insight and encouragement throughout this degree.

Additionally, I would like to thank my Remote Sensing IGEP colleagues, who have always been collaborative and supportive during the various struggles associated with pursuing a graduate degree, and who have pursued interdisciplinary study with openness and interest. In particular, I would like to thank Cameron Houser who has always been available to chat about coding problems or participate in study and writing sessions over the years, as well

as Dr. Ranjith Gopalakrishnan, Dr. Sherin Ghanam, Dr. Samantha Fried, Dr. Iris Fynn, Adam Oliphant, Matt House, Snehal Moore, and Paige Williams, among others.

I have been fortunate to have had support from family and friends during this endeavor as well. In particular, I would like to thank my husband and partner Scott, who has been incredibly supportive through every step of this process.

Contents

List of Figures xiii

List of Tables xvii

1 Introduction and Objectives 1

1.1 Introduction and Justification . . . 1

1.2 Focus and Objectives of by Chapter . . . 1

1.2.1 Chapter 2 Objectives . . . 1

1.2.2 Chapter 3 Objectives . . . 2

1.2.3 Chapter 4 Objectives . . . 2

1.3 Selected Background . . . 3

1.3.1 Multi-temporal and Time-Series Data . . . 3

1.3.2 Photointerpretation . . . 6

1.3.3 Crowdsourcing . . . 9

1.3.4 General Considerations for Selecting a Sampling Protocol . . . 11

Bibliography 13

2 Estimating tree canopy cover using harmonic regression coefficients derived from multi-temporal Landsat data 22

Abstract . . . 22

2.1 Introduction . . . 22

2.2 Data and Methods . . . 26

2.2.1 Study area . . . 26

2.2.2 Training (response) data . . . 26

2.2.3 Training (response) data characteristics . . . 26

2.2.4 Development of predictor variables . . . 27

2.2.5 Predictive modeling with random forests . . . 29
2.2.6 Model comparison . . . 29
2.3 Results . . . 31
2.3.1 Variable importance and correlations . . . 34
2.3.2 Inherent dimensionality . . . 34
2.3.3 Mapping TCC across the landscape . . . 36
2.3.4 Variance of predictions across land cover types and TCC values . . . 36
2.4 Discussion . . . 36
2.5 Conclusions . . . 38
2.6 References . . . 38

3 Consistency in expert and crowdsourced interpretations of aerial photographs to measure tree canopy cover 40

Abstract . . . 40
3.1 Introduction . . . 42
3.1.1 Background: Measurement of Tree Canopy Cover . . . 42
3.1.2 Objectives . . . 44
3.2 Data and Methods . . . 45
3.2.1 Study area and sample site selection . . . 45
3.2.2 Mechanical Turk HIT interface development . . . 46
3.2.3 Training and advertising . . . 48
3.2.4 HIT release and quality control . . . 49
3.2.5 Data analysis . . . 49
3.2.6 Reliability model . . . 51
3.3 Results . . . 54
3.3.1 Interface technical issues and feedback . . . 54
3.3.2 All interpretations . . . 55
3.3.3 Expert interpreters and inter-expert agreement . . . 60
3.3.4 Reliability . . . 62

3.3.5 Important variables . . . 66
3.4 Discussion . . . 69
3.4.1 The wisdom of crowds . . . 69
3.4.2 Interpretation consistency and reliability . . . 71
3.4.3 Interpreter fatigue . . . 72
3.4.4 Display . . . 72
3.4.5 Potential earnings . . . 73
3.4.6 Interpreted TCC . . . 74
3.4.7 Recommended improvements . . . 74
3.5 Conclusion . . . 76
3.6 Acknowledgements . . . 77
3.7 References . . . 78

4 Sampling methods for the accuracy assessment of continuous variables derived from remotely sensed data 84

Abstract . . . 84
4.1 Introduction . . . 86
4.2 Data and methods . . . 87
4.2.1 Simulation of predictors and response variable . . . 87
4.2.2 Model prediction . . . 92
4.2.3 Sampling . . . 92
4.2.4 Statistical analysis . . . 96
4.2.5 Domain estimation . . . 97
4.3 Results . . . 98
4.3.1 Sampling protocol . . . 98
4.3.2 Number of samples . . . 105
4.3.3 Domain estimation . . . 108
4.4 Discussion . . . 108
4.4.1 Sampling protocol . . . 108

4.4.2 Sample size 109
4.4.3 Domain estimation 110
4.5 Conclusion 110
4.6 Acknowledgements 111
4.7 References 112

5 Conclusions **115**
5.1 Summary and conclusion 115
5.2 Future work 116

List of Figures

1.1 The effect of changes in the harmonic coefficients to the resulting curve. Each graph shows a variation in a different combination of coefficients: B_0 varying alone (top-left), B_1 varying along (top-center),B_2 varying alone (top-right), B_1 and B_2 varying concurrently (bottom-left), B_1 and B_2 varying inversely with a smaller difference in coefficients' values (bottom-center),B_1 and B_2 varying inversely with a larger difference in coefficients' values (bottom-right), 5

2.1: Median composite and harmonic regression. 25

2.2: Harmonic regression coefficients for the Landsat 8 OLI panchromatic band from 2014–2016. 26

2.3: Western and Southern study areas with 2011 National Land Cover Database classification. 27

2.4: Histograms of photo interpreted TCC observations for time 1 time 2 in the South and West. 28

2.5: Scatter plots comparing time 1 and time 2 photo interpreted TCC observations for the South and West. 28

2.6: Comparison of models using harmonic regression coefficients for NDVI, SWIR1 and SWIR2 to the median composite values for those same three bands. . . 30

2.7: Comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using the full suite of median composite and terrain variables used in the 2011 TCC product plus the harmonic regression coefficients from NDVI, SWIR1, and SWIR2. 31

2.8: Comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using the harmonic regression coefficients for NDVI, SWIR1, and SWIR2. 32

2.9: Observed vs. predicted TCC for mean (with respect to pseudo R2 out of 500) Time 2 models for the southern study area for the two models using harmonic regression coefficients as part of the predictor variable suite, colored by 2011 NLCD land cover at sample locations. 32

2.10: Top 11 variable importances from full random forests model for each region/-time. 33

2.11: Eigen analysis of predictor variable sets. 34

2.12: TCC maps for the southern study area for the two models that included harmonic regression coefficients as part of the predictor variable suite. 34

2.13: Mean variance by NLCD land cover class (water excluded) for South, all models. 35

2.14: Mean variance by NLCD land cover class (water excluded) for West, all models. 35

2.15: Observations from two similar South Time 1 plots with fitted harmonic regression line, harmonic regression constant, and median composite value shown. 36

3.1: Mechanical Turk interface. 47

3.2: Distribution of expert and crowd interpretation values and disagreement for 2016. 56

3.3: All combinations of expert vs crowd interpretations from 2016 and all combinations of expert vs crowd interpretations from 2016 split into "unreliable" and "reliable" classifications based on model predicitons. 56

3.4: Plots from an individual crowd worker's interpretations showing the experts' interpretations vs the worker's interpretations, the sequence of the worker's interpretations vs the interpretation's value and the sequence of the worker's interpretations vs the time it took to complete the interpretation. 57

3.5: Plots comparing the median and mean of all assignments submitted for each HIT to the individual expert interpretations evaluated at the same plot locations. 57

3.6: Each image shows a panel with a crowd interpretation and the underlying NAIP imagery in the display used during the majority of that interpretation, an expert interpretation and the underlying NAIP imagery in the display used during the majority of that interpretation, and a context image. 58

3.7: Scatter plot between individual experts for 2011 and 2016. 61

3.8: Correlation matrix showing a subset of variables which have a correlation greater than $|0.5|$ with at least one other variable in the subset. 63

3.9: Variable importance scores, calculated using a rank-score method which assigns points to the top ten most important variables in each of the 5000 models and adds them together (2016 only). Dashed line indicates the top five most important variables across all models. These scores were derived from the mean decrease in accuracy metric supplied by the randomForest function in R. 65

3.10: Summary plots for important variables, including a bar plot showing percent of interpretations which use a single display vs a combination of displays, and box plots showing percent of dots in false color display for each interpretation, percent of dots in natural color display for each interpretation, interpreted TCC, number of plots interpreted,and number of HITs available at the time of interpretation (bottom right) by interpretation class (all crowd, expert,"unreliable" crowd, and "reliable" crowd). 67

3.11: Scatter plots demonstrating an increase in error and decrease in TCC value with additional plots as interpreted by crowd workers. 68

4.1: Raster maps of predictor variables, response variable, as well as the mean of random forest regression tree predictions and the standard deviation of predictions for case 1. 88

4.2: Raster maps of predictor variables, response variable, as well as the mean of random forest regression tree predictions and the standard deviation of predictions for case 2. 89

4.3: Raster maps of normally distributed Y for case 1, non-parametric Y for case 2, and the USFS Analytical TCC product for case 3. 90

4.4: Distribution of Y values for all three cases as well as the distribution of the USFS Analaytical TCC map product. 91

4.5: USFS Analytical TCC layer and Standard Error layer. 93

4.6: Correlation matrix for X and Y variables for cases 1 and 2. 94

4.7: Scatter plots showing the correlation between the variable to be estimated and the stratification variables used and error vs the standard deviation across random forest regression tree predictions. 98

4.8: Histograms showing the distribution of ensemble estimates of the RMSE and mean Y across the population, calculated from 250 sets of samples using each sampling protocol. 99

4.9: Scatter plots showing the correlation between the variable to be estimated and the stratification variables used. Error vs the standard deviation across random forest regression tree predictions. Reference Y vs the mean of regression tree predictions. Only one 1416 x 1770 pixel tile (out of 30) is shown, however all tiles are similar. 100

4.10: Histograms showing the distribution of ensemble estimates of the RMSE and mean Y across the population, calculated from 250 sets of samples using each sampling protocol. 100

4.11: Scatter plots showing the correlation between the variable to be estimated and the stratification variables used. Error vs the standard error layer value from the 2016 Analytical TCC product. Reference Y vs the 2016 Analytical TCC product. 103

4.12: Histograms showing the distribution of ensemble estimates of the RMSE and mean Y across the population, calculated from 250 sets of samples using each sampling protocol. 103

4.13: Width of the confidence intervals of samples estimates of the average reference Y value and the average RMSE value by sampling protocol and sample size. 106

4.14: Confusion matrix showing the number of samples in target and prediction classes. 107

List of Tables

2.1: Phenology metrics used to map tree canopy cover using satellite data. 24

2.2: US National Agricultural Imagery Program (NAIP) image acquisition dates. . 27

2.3: Number of photo interpretation (response data) samples per region and time period. 27

2.4: Parameters used in the EWMACD algorithm. 29

2.5: Parameters used in the randomForest algorithm. 29

2.6: Additional R libraries used in computing. 29

2.7: Model-specific predictor variables. 36

2.8: Correlation matrix of seven-most important predictor variables and tree canopy cover from the South, time 2 model. 37

2.9: Correlations between TCC in each region/time and the two NDVI predictor variables. 37

3.1: Schedule and characteristics of each HIT release including phase, description, number of plots, dates, and timing. 49

3.2: Dependant variable explanation for reliability model. 51

3.3: Parameter settings for the randomForest classification model in R. 51

3.4: Predictor variable explanation for reliability model. 52

3.5: Number of plots by number of assignments completed by crowd workers. . . . 54

3.6: Number of plots by number of assignments completed by expert workers. . . 54

3.7: Summary statistics for the number of plots interpreted by individual crowd workers. 55

3.8: Summary values for interpretations as collected. Disagreement is the standard deviation of all interpretations of each plot by worker type The weighted mean confidence interval width refer to the confidence interval around the mean for all group-specific interpretations at each plot. 55

3.9: Number of interpretations by expert. 60

3.10: Expert-to-expert comparison showing number of plots interpreted between experts, proportion of interpretations within 10%, weighted Kolmogrov-Smirnov index, and weighted Pearson's correlation coefficient. 60

3.11: Agreement between crowd and expert interpreters as demonstrated by the summary metrics for the crowd workers' proportion of plots within 10% TCC of experts summary metrics for the Kolmogrov-Smirnov test results between each crowd worker's interpretations and expert interpretations, as well as the number of workers whose interpretations are likely to have come from the same distribution as the expert interpretations. 62

3.12: Average number of interpretations predicted and/or observed to be "reliable" and "unreliable". Average Overall % accuracy across all 5000 models is in the bottom right cell. 64

3.13: Weighted Pearson's correlation coefficients and weighted Kolmogrov-Smirnov indices between all crowd and experts for 2011 and 2016 and all dates, and between "reliable" crowd and experts and "unreliable" crowd and experts for 2016. 66

3.14: Weighted ANOVA between all crowd and experts for 2011 and 2016 and all dates, and between "reliable" crowd and experts and "unreliable" crowd and experts for 2016. 66

3.15: Weighted mean values of the top five most important variables by interpreter class for 2016 only. 68

4.1: Population variance and mean of the reference Y and population variance of squared residuals and RMSE for all cases. 98

4.2: Average of estimated population mean, standard deviation of population mean estimates, width of the 95% confidence interval of population mean estimate, and mean difference from actual population mean for 250 iterative sample estimates. Stratification is based on the mean of model predictions at each pixel. 101

4.3: Average of estimated RMSE, standard deviation of RMSE estimates, width of the 95% confidence interval of RMSE estimate, and mean difference from actual population RMSE for 250 iterative sample estimates. Stratification is based on the standard deviation of model predictions at each pixel. 101

4.4: Average of estimated population mean, standard deviation of population mean estimates, width of the 95% confidence interval of population mean estimate, and mean difference from actual population mean for 250 iterative sample estimates. Stratification is based on the mean of model predictions at each pixel. 102

4.5: Average of estimated RMSE, standard deviation of RMSE estimates, width of the 95% confidence interval of RMSE estimate, and mean difference from actual population RMSE for 250 iterative sample estimates. Stratification is based on the standard deviation of model predictions at each pixel. 102

4.6: Average of estimated population mean, standard deviation of population mean estimates, width of the 95% confidence interval of population mean estimate, and mean difference from actual population mean for 250 iterative sample estimates. Stratification is based on the mean of model predictions at each pixel. 104

4.7: Average of estimated RMSE, standard deviation of RMSE estimates, width of the 95% confidence interval of RMSE estimate, and mean difference from actual population RMSE for 250 iterative sample estimates. Stratification is based on the standard error of model predictions at each pixel. 104

4.8: Width of the 95% confidence interval of estimates of the mean and estimates of RMSE for each sampling protocol at various sample sizes using case 3 data. 105

4.9: Within-domain estimates of RMSE, variance in pixel-wise errors, standard error of pixel-wise errors, and sample size for one set of samples. The samples are drawn with stratified random sampling using optimal allocation. The strata are defined by the prediction uncertainty metric (the standard error values extracted from the USFS Analytical TCC Standard Error product for 2016) and the domains are defined by 10% intervals of the simulated TCC reference data. 108

4.10: Within-domain estimates of RMSE, variance in pixel-wise errors, standard error of pixel-wise errors, and sample size for one set of samples. The samples are drawn with stratified random sampling using optimal allocation. The strata are defined by the prediction uncertainty metric (the standard error values extracted from the USFS Analytical TCC Standard Error product for 2016) and the domains are defined by the NLCD land cover product for 2016. 109

List of Abbreviations

ANOVA Analysis of Variance

EWMACD Equally Weighted Moving Average Change Detection

FIA Forest Inventory and Analysis

GTAC Geospatial Technology and Applications Center

HIT Human Intelligence Task

HR Harmonic Regression

LiDAR Light Detection and Ranging

MC Median Composite

MSE Mean Squared Error

MSR Mean Squared Residuals

NAIP National Agriculture Imagery Program

NDMI Normalized Difference Moisture Index

NDVI Normalized Difference Vegetation Index

NLCD National Land Cover Database

PCA Principal Component Analysis

RMSE Root Mean Squared Error

SSR Sum of Squared Residuals

SWIR Short-wave Infrared

TCC Tree Canopy Cover

Chapter 1

Introduction and Objectives

1.1 Introduction and Justification

Tree canopy cover (TCC) is defined as the measure of the proportionate area covered by the vertical components of tree crowns [53]. This important metric is widely studied and modeled at various scales due to its implications across many scientific disciplines, such as the study of climate change [60], forest management, and global health, as disease outbreaks such as the Ebola Virus epidemic of 2014 have been linked to deforestation [70].

Knowledge of the extent of forest across the globe and the changes in forest extent over time, especially forest loss, is critical to these areas of study. According to Sexton, et al. (2016) there are over 800 official definitions of forest, many of which are based on models of TCC, however these model estimates can vary substantially. Robust estimates and detailed evaluation of TCC models is essential to maintaining a reliable standard for these products and subsequently linking national and global forest extent to other systems [86]. As such, this dissertation aims to test methods for the modeling and evaluation of TCC, exploring state of the art techniques enabled by advances in big data analysis and high performance computing including ensemble machine learning algorithms and open source data products.

1.2 Focus and Objectives of by Chapter

1.2.1 Chapter 2 Objectives

The primary objective of this chapter, entitled 'Estimating tree canopy cover using harmonic regression coefficients derived from multi-temporal Landsat data' is to evaluate the utility of harmonic regression coefficients derived from Landsat time-series as predictors of TCC alone and in conjunction with Landsat median-composite variables. We answered the questions:

(i) Are harmonic regression coefficients better predictors of TCC than median composite variables for the same vegetation index and wavelength bands?

(ii) When added to a full suite of median composite and terrain variables, can harmonic regression coefficients improve the quality of TCC models?

1.2.2 Chapter 3 Objectives

The goal of this study, titled 'Consistency in expert and crowdsourced interpretations of aerial photographs to measure tree canopy cover' is to determine the how consistent photointerpreted TCC values can be and what factors influence that consistency. To accomplish this, we did the following:

(i) Compared interpretations between each expert and between groups of crowd and expert interpreters.

(ii) Explored factors that influence the "reliability" of crowd interpretations as they compare to expert interpretations.

(iii) Proposed recommendations for the design of future studies in order to improve consistency and agreement in photointerpretation of TCC.

1.2.3 Chapter 4 Objectives

In Chapter 4, we explored sampling methods for the purpose of model validation. In this study, titled 'Sampling methods for validating continuous variables estimated using remotely sensed data', we evaluated methods for selecting data points in simulated continuous data products that represent large spatial extents using estimates of prediction uncertainty to stratify the sampling frame. In particular, we:

(i) Compared various sample allocation methods in terms of the precision of estimates derived from those samples and demonstrated a method of stratified random sampling with optimal allocation using measures of prediction uncertainty derived from random forest regression models.

(ii) Compared the precision of estimates derived from samples drawn using different sampling protocols with different sample sizes.

(iii) Demonstrated a method to report the accuracy of continuous models for domains that are either regionally constrained or numerically defined based on other variables or the modeled quantity itself.

1.3 Selected Background

1.3.1 Multi-temporal and Time-Series Data

The temporal dimension of remote sensing imagery is one of the things that sets it apart from other types of data collection. In Chapter 2, we explore the use of multi-temporal image composites and time-series derived harmonic regression coefficients in modeling tree canopy cover. Multi-temporal data serves many purposes in these models. First, the use of multi-temporal composite imagery for modeling is desirable because when pixel-reflectance values associated with an individual date are occluded by cloud cover or, in the case of Landsat 7, uncorrected scan lines, other dates and values are available to replace them. Using many dates to obtain an average or median value is desirable because the resulting image is less likely to be affected by anomalies that are not representative of the land in reality (such as snow cover or clouds that evaded the cloud mask).

Multi-temporal composites are made up of several images, sometimes spanning part of a season or representing values within a range compared to a target value (such as within 10 % of peak green-up). Time-series data, by contrast usually utilize a chronological record of images from all dates spanning a period of time. The complete catalog of reflectances for a pixel can provide information about the patterns periodic changes that occur in the land each year and they can identify areas that have diverged from those patterns [10]. Our study demonstrates that this added information is valuable in the modeling of tree canopy cover, and possibly for other applications as well.

In contrast to other disciplines, the temporal nature of satellite imagery makes the longevity of programs like Landsat a critical priority. Wulder et al. (2012) commented on this when they considered that we've entered an age of satellite remote sensing data compatibility rather than competition [99]. Woodcock et al (2001) also cited the importance of legacy data when they called for the improved availability and cost of Landsat imagery as well as continuity in the Landsat Program [98].

To support the need for historic data, the Landsat archive was opened and made freely available to the public in 2008 through USGS' No-Cost Data Policy [62, 78, 79, 93, 97]. This policy, as well as recent developments in high performance and cloud-based computing servers, such as Google Earth Engine, have provided the opportunity for increasingly complex and sophisticated approaches to modeling with remote sensing data. These technologies, harness new methods for handling big data to produce meaningful models that span large spatial scales [6, 7, 36, 61, 61, 99].

Multi-temporal data has been shown to be highly important in the field of remote sensing, especially for environmental and natural resource applications. Multiple dates of imagery have been demonstrated to improve static models such as land cover classification using both clustering [38] and transformation [28, 41, 67] as well as regression analysis utilizing

multiple dates and spectra of data [39, 40, 100]. For example, the modeling of biomass can be limited by optical saturation and temporal instability of vegetation. These phenomena can be captured and corrected when multiple time steps are available [71]. Multi-temporal analysis has also been used in fusion algorithms [30, 43, 44, 77] to produce more frequent, higher resolution imagery, and to remove anomalies from existing images, such as clouds [31, 54], striping [14, 54, 89], and more. Nemani and Running (1997) found multi-temporal data to be useful in land cover classification with AVHRR data and Hansen, et al. (2000) used multi-temporal AVHRR data in a regression tree model to create a continuous 1 km global land cover classification [35, 69].

The most common application of multi-temporal images has been to estimate broad-scale change over time [36, 37, 39, 40, 56, 59, 98]. For example, Hansen and Defries (2004) also used multi-temporal AVHRR data to measure and detect decadal change in proportional forest cover [34]. Subtle changes, such as disturbances like canopy defoliation due to Gypsy Moth infestation can also be observed from multi-temporal data [68].

In contrast to multi-temporal analysis, time-series analysis utilizes multiple repeated measurements to interpret trends over time based on a chronological sequence of observations [64]. Building on literature using AVHRR time-series [5, 51, 52, 65, 76], Landsat time-series analysis uses all available observations from Landsat images between the dates of interest in order to incorporate information about regular and anomalous changes in the landscape. The global archive of normalized images from Landsat lends itself to the idea of time-series analysis, as one can easily derive vectors of temporal information from each pixel-based location in a stack of spatially registered images. These data are shown to detect recent changes occurring across landscapes [10, 12, 47] or observations related to seasonal phenology that may not be visible using static observations. This information has proved useful in land classification [50, 100], cloud detection [11] and modeling of phenological characteristics [55].

Many time-series analysis techniques utilize a pixel-level vector of temporal observations to obtain extract detailed information about trends in reflectance over time and summarize those trends in a small number of output bands. One approach 'the shapes algorithm' takes a pixel's time-series and selects the best-fitting shape from a library of pre-defined trajectories. This algorithm provides information on forest dynamics and monitoring and can also be used to predict dynamics in tree canopy cover [63] and disturbance causality [84].

Harmonic regression can also be used to summarize attributes of a time series. Immerzeel (2005) used harmonic analysis of a landsat Landsat time series to determine the relationship between rainfall and reflectance [48]. Using only cosine terms, they found the constant term, amplitude of the first cosine term, amplitude of the second cosine term, the sum of those amplitudes and the difference between minimum and maximum NDVI to be useful in this regard.

Changes in the harmonic equation relate to changes in the phenology of a pixel. In Figure 1.1 we demonstrate how changes to each coefficient in the harmonic equation affects the shape and timing of the reflectance curve it represents. The constant term (B_0) equates to the

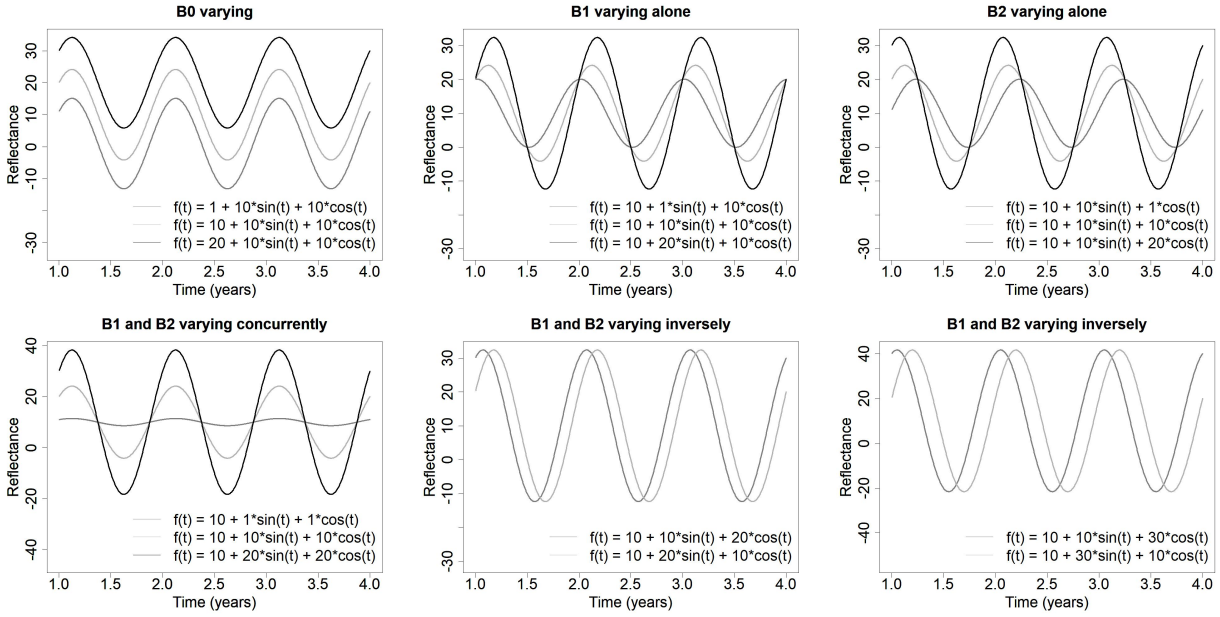

Figure 1.1: The effect of changes in the harmonic coefficients to the resulting curve. Each graph shows a variation in a different combination of coefficients: B_0 varying alone (top-left), B_1 varying along (top-center),B_2 varying alone (top-right), B_1 and B_2 varying concurrently (bottom-left), B_1 and B_2 varying inversely with a smaller difference in coefficients' values (bottom-center),B_1 and B_2 varying inversely with a larger difference in coefficients' values (bottom-right),

average value for the modeled curve over time. Wile adjustments to B_0 increase or decrease the mean reflectance value over time, the amplitude and time of peak reflectance remain the same. If the Y axis represented NDVI, if the harmonic regression equation for two pixels only differed in terms of B_0, the pixels would both have the same magnitude of seasonal increase in NDVI between leaf-off and peak green-up, and the timing of peak green-up and leaf-off conditions would be the same during the season, but the pixel with the lower B_0 would have a lower NDVI in leaf-off condition and would reach a lower NDVI during peak green-up compared to the pixel with a higher B_0.

Varying either the sine (B_1) or cosine (B_2) coefficients alone changes both the time of peak reflectance, as well as the amplitude of the curve, with a greater time shift and change in amplitude shown by a larger difference in sine and cosine terms. Increasing either B_1 or B_2 alone increases the amplitude of the curve, but increasing B_1 causes a time shift to the later part of the series, while increasing B_2 causes a time shift to the earlier part of the series. If all other coefficients were the same, a pixel with a higher B_1 might represent an area where vegetation has a greater change in NDVI between leaf-off and peak green-up conditions, with the seasonal period of leaf-off to green-up happening later than a pixel with a lower B_1. If all other coefficients were the same, a pixel with a higher B_2 would representa an area where vegetation has a smaller change in NDVI between leaf-off and peak green-up conditions, with the seasonal period of leaf-off to green-up happening earlier than a pixel with a lower B_2. In both of these cases, the seasonal period represents the same amount of time, but the changes in the coefficients shifts the peaks to be earlier or later.

Adjusting B_1 and B_2 concurrently maintains the time of peak reflectance, but adjusts the amplitude of the curve, proportionately to the difference between the constant term and the other coefficients. If two pixels had the same B_0 coefficient, the pixel with higher B_1 and B_2 coefficients (assuming B_1 and B_2 were equal) would represent land with a larger difference in NDVI between peak green-up and leaf-off conditions. Adjusting B_1 and B_2 inversely changes the time of peak reflectance, with a greater difference between the coefficients yielding a greater shift. This would reflect two pixels that experienced peak green-up and leaf-off conditions at different times during the growing season, however the length of the growing season remains the same.

When the harmonic curves estimated for each pixel are compared across gradients of TCC%, or grouped by type of evergreen or deciduous forest type, the curves take on very different characteristics.

1.3.2 Photointerpretation

Tree canopy cover is measured in many ways. In the field, labor-intensive methods such as grid sampling using a Cajanus tube [58] can be used as well as line-intersect sampling [23, 92]. There are also more efficient techniques such as below-canopy digital photography or spherical densiometer measurements (which measure canopy closure, a related, but different

metric) and ocular estimates [19]. Additionally, Toney et al. (2009) proposed a stem-map model to measure tree canopy cover using the locations of tree stems and modeling tree canopies using crown width estimates by species[92].

Photointerpretation, defined as "the act of examining the photographic images of objects for the purpose of identifying those objects and deducing their significance" (Colwell et al., 1952, page 375), has also been used to collect training data for the NLCD Tree Canopy Cover products [17, 21, 91]. The third chapter of this dissertation focuses on the ability of experts and non-experts to consistently collect photointerpreted tree canopy cover measurements.

Photointerpretation provides a cost-saving solution for some applications. This data collection method utilizes higher resolution imagery or sensor information to measure field characteristics where possible. This method can be completed without transit time between field sites, and does not require lengthy field visits or travel costs. Since the cost of travel between plot locations often represents the largest proportion of the cost of an inventory (especially at national scales), photointerpretation has become an essential tool as it offers the ability to obtain training and testing data across large study areas at relatively low cost [61]. For example, Esseen et al (2006) show that photointerpretation of forest edge lengths was found to be eight times faster and five times more cost effective than field measurements [23]. Additionally, photointerpretation also has the potential to reduce constraints against physical access to sites which could potentially cause bias in sampling designs (though this can also be remedied through sampling scheme selection, [18]).

Photointerpretation methods are widely used in the field of remote sensing of natural resources. For example, this method is frequently used to interpret and classify features on the landscape from aerial imagery. Multispectral aerial imagery is particularly valuable to the remote sensing of natural resources as it can provide additional perspectives outside of the visual spectrum in order to better classify soil, vegetation, and water [3, 87]. Additionally, the Forest Inventory and Analysis (FIA) program, a program run by the USDA Forest service which collects, analyzes, and reports measurements and trends for America's forests in an annual survey [95], has been using photointerpretation since the 1950s to assign plots to strata [74].

The availability of high resolution digital imagery from satellites, instruments like unmanned aerial systems (UAS), or programs like the National Agricultural Imagery Program (NAIP) and open curation in software like Google Earth has created a resurgance of interest in photointerpretation for new applications. As research expands to support international initiatives to manage natural resources, this method of data collection has been employed to collect globally distributed datasets using consistent measurements and interpreters [8, 66, 81, 90]. However, equal access to high quality fine resolution imagery has yet to be achieved across all nations [82].

The flexibility in the types of data that can be collected has resulted in the use of photointerpretation for a variety of applications. There are many examples in the literature that demonstrate that these methods can be used to obtain categorical measurements such as

those used in landcover classification [15, 66, 81], urban tree mapping [72], urban land cover classification for use in classifying urban structure types [42], coastal change analysis [20], and to test consistency in existing landcover datasets [9]. Photointerpretation is also used to derive continuous measurements from high resolution images (commonly from categorical scale references [3, 16, 17, 87], intercept-sampling methods [23], or indirect measurement). These methods have been used to determine seagrass extent in North Carolina [94], observe landscape fragmentation [57], improve regional catch estimates from fishing weirs in the Persian Gulf [1], and to calculate biotic score in the Delaware River Basin to evaluate factors contributing to stream health [75].

Timing of data collection is an important consideration when modeling landscape features such as tree canopy cover using remotely sensed data. In particular, for studies using ground-based training data collection with large study areas, it can be a challenge to collect data in a timely manner so that conditions on the ground remain consistent with the conditions present when imagery was taken. Seasonal variations and disturbances, such as leaf-loss in winter can change the TCC values considerably in a particular location. In the absence of a very large, highly costly team of field experts, it would be very challenging to perfectly coordinate field activities with satellite image acquisition for national-scale modeling activities. Photointerpretation provides a solution to this problem by creating an unchanging record of the landscape that can be measured as time permits.

Photointerpretation is not always the gold standard of measurement for natural resources surveys. According to Congalton and Green, (2009), aerial photointerpretations are falsely assumed to be higher quality than the satellite imagery used in the models because the spatial resolution is often much finer and because image interpretation is a 'time-honored skill that is accepted as accurate' ([20], page 88). However errors do occur due to the complexity of the task and interpreter ability. There are a number of conditions that can affect the quality of image interpretation. Campbell and Wynne (2011) cite several factors which can affect the quality of photointerpretation including the overhead 'birds-eye' viewing angle, which can be less familiar or disorienting to inexperienced interpreters, and differences in the resolution and spatial scale of an image, which can reduce the ability to resolve certain features across the landscape [13]. Additionally the utilization of non-visible bands (a critical resource for the detection of different landscape components) can cause confusion for inexperienced interpreters. The chosen classification schemes can also affect the ability to interpret land classes through photointerpretation, especially if there is subjectivity or ambiguity in class definitions [90]. Congalton and Green (2009) also suggest that measures can be taken during the planning phase of a study to improve the consistency between field observations and photointerpretation for image classification and accuracy assessment, such as consistency in training and data collection, a rule-based process of interpretation, and ground validation at a subset of plots [20].

1.3.3 Crowdsourcing

Crowdsourcing is a form of participatory research that capitalizes on input from the global online workforce [101] in "the act of outsourcing tasks… to an undefinably large, heterogeneous mass of potential actors." [33]. Crowdsourcing processes, called Human Intelligence Tasks (HITs) utilize an identified, voluntary, often heterogeneous crowd to meet a specific goal for an identified solicitor or 'crowdsourcer'. The objective is an online participatory process and the transaction between crowdworker and crowdsourcer is clearly defined [24]. Crowdsourcing, as with other forms of participatory science, usually includes non-experts, but in contrast these crowd workers are often anonymous to the requester. Recently, crowdsourcing has been used more commonly in scientific research to generate datasets more quickly and at a reduced cost.

In the third chapter of this dissertation we use crowdsourcing to collect multiple photointerpreted measurements of tree canopy cover from non-experts. We then use an ensemble of models to identify factors which contribute to the "reliability" of those crowd interpretations as compared to experts.

The term 'crowdsourcing' was first mentioned by Howe, 2006. The article discussed the increase in the outsourcing of commercially traded tasks to an online labor network. Since then crowd sourcing has grown to be utilized for many applications, especially those that require quick response at a low cost [45]. These HITs are often difficult to automate, would involve considerable expense if performed by experts [32], and involve only simple, objective analyses in order to reduce bias [29].

Geo-wiki.org is a popular platform that has been used to generate training data for classifications using imagery from Google Earth. These data have been utilized to train and order to improve upon existing global landcover [27] classifications in regions where they disagree. Schepaschenko et al. (2015) uses these data along with several commonly used models of forest cover to create a forest probability layer and a map of percent forest cover which were then combined to create a best guess hybrid map. Their study yields a resulting model of forest classification with an overall accuracy of 93% improving upon all input models, and 0.87% R^2 which is tied for the highest correlation with one other model [83].

There are challenges to using crowdsourced data, obtained by non-experts in research. First of all, the high data quality standards in science and perception of expertise [96] have led to skepticism about the ability of crowds to produce comparable results to experts. Photointerpretation in particular requires some expertise to adequately obtain high quality results, such as knowledge of subject, geographic region, and remote sensing system [13] and crowd workers traditionally have unknown or variable levels of experience with all of these. To ensure data quality, participants should be trained to perform the task at hand, quality control measures should be considered in the experimental design and analysis, and data should be validated. [22, 85] Many studies that employ crowdsourced data compare the results results to equivalent expert interpretations to confirm their quality. Many studies

show that crowdsourced data can be at least as accurate as expert data.

See et al. (2013) utilized data from Geowiki.org to identify areas impacted by humans (five classes) and landcover (ten classes), compared interpretations from more than 60 crowd workers to a set of three expert interpreters at control points where initial analysis suggested non-experts should be able to interpret these qualities, and collected interpreter ratings of confidence in interpretation. This quality analysis was conducted at three stages during their study. Their analysis shows that volunteers underestimated the degree of human impact in most land cover types, but they were more consistent than experts (3% lower average deviation) in the estimates they did provide. Experts were more adept at classifying landcover than non-experts (69% vs 62% at the beginning of the study increasing to 76.2% vs 65.9% at the end) and they were also slightly (0.4%) more consistent than experts. They found that, for all interpreters, decisions were more consistent in locations where they were highly confident in their interpretations. Time on task in identifying landcover type resulted in a decrease in accuracy in this study and accuracy of both human impact and landcover classification increased in the final stages of competition as compared to the earlier quality assessments [85].

High quality interpretations can be ensured through multiple approaches. [73] (learning from crowds) and [49] (quality management algorithms) use algorithms to identify interpreter quality and then weight responses, while [46] (data quality from crowdsoucing) show that multiple noisy non-expert observations can be used to generate better observations.

Voting mechanisms are common with crowdsourced observations. For example, Foody, et al (2018) also used data from Geo-Wiki.org to show that consensus approaches such as simple voting or weighted voting can improve land cover classifications over those generated from individual observations. The researchers explored several methods including basic majority voting (scenario 1), volunteer removal if the class-level accuracy was substantially lower than the maximum obtained (scenario 2), volunteer removal of the full set of contributions one (scenario 3a) or two (scenario 3b) of the lowest quality contributers in terms of overall accuracy, and weighting using class-weighted accuracy by magnification of contributions from high quality workers (scenario 4a), the magnification of most accurate and removal of least accurate workers (scenario 4b), and the removal of contributions from the least accurate worker (scenario 4c) . They found for their study, that removal of inaccurate data improved results more than magnification of high quality interpretations, improving results between scenario 4a and scenario 4c from 78.59% to 80.60%. Removal based on lowest quality interpretations based on unweighted class accuracies (scenario 3) were higher than removal of a subset of interpretations based on class-level accuracies (scenario 2) and both of these scanerios were more accurate than simple voting (scenario 1), however simple (unweighted) voting provided higher accuracy (76.58%) than the most accurate worker (73.91% accuracy) [25].

Salk et al. (2016) also explore a majority rule procedure to improve voluntary observations for use in cropland identification, but they suggest this method may not be a substitute for

expert validation especially in difficult to interpret images. Observations for this study were provided by volunteers using the Cropland Capture game. They were validated using a pair of two experts, and additional assessment was performed to determine agreement. Volunteer classifications usually agreed with expert validations for both categories in easy to interpret images (greater than 90% agreement in over half of the 'easy' images). Additionally, they found that agreement among volunteers across the full sample was greater than 90% for approximately 80% of the images and was related (significantly, but weakly) to interpreter confidence. By contrast, experts agreed 66.5% of the time. Inter-volunteer disagreement was also correlated with inter-expert disagreement. While other studies suggest that the use of majority voting rules with multiple crowd interpretations can achieve higher accuracy than individual interpretations [25], in this study 22% of images were mis-categorized using a majority voting technique, with in increase in mis-classification for more difficult images. This finding highlights the notion that voting and marjority rule procedures should be used with caution (especially when synthesizing a dataset for model validation) as they result in a model of the original data and error is possibly introduced through aggregation, especially where systematic biases occur among interpreters [80].

1.3.4 General Considerations for Selecting a Sampling Protocol

The fourth chapter of this dissertation demonstrates a method for sample allocation across continuous data products created with ensemble models such as random forest regression. The sampling protocols tested in this chapter are considered to be statistically valid, and were chosen with the consideration of the following criteria found in the literature.

When planning a data collection for the purpose of accuracy assessment, three main considerations are common. First, the activity must yield information that is useful to the user. If accuracy is assessed under conditions that are only relevant to a very specific use of the data, that must be made clear upon reporting. Secondly, budget is often an unavoidable consideration in accuracy assessment, especially when choosing a sample size to meet an intended level of confidence. Under funding constraints, selecting a sample design that maximizes the utility of the sample is all the more important. To address these issues, it is important to include additional statistical information related to the quality of the accuracy assessment activity when reporting results. Foody (2009) recommends careful planning and that the interpreter specify effect size, significance level, statistical power, and reports the confidence intervals along with accuracy measures [26].

Finally, the interpreter must consider whether the sampling design selected is scientifically defensible. To select a sample and perform a statistically valid accuracy assessment, samples must be independent, the sampling protocol must be based on each sample's inclusion probability, and the sample size must be based on the desired precision of the accuracy assessment.

The sampling protocol is the method used to select samples from the sampling frame. In

order to have a statistically valid sample, the sampling protocol should utilize the inclusion probability of each potential sample, and the probability of inclusion should be approximately equal for each sample. Known sampling protocols, such as random, stratified, cluster, and systematic sampling do not require that these inclusion probabilities be explicitly calculated, however Stehman and Czaplewski (1998) state that 'if the sampling protocol cannot be identified as a standard probability sampling design and the project planners are unable to specify the nonzero inclusion probabilities, the proposed design should be discarded.' [88]

Further, a sampling protocol should be selected that enables sample independence, which is particularly important for spatial data due to spatial autocorrelation. Spatial autocorrelation is the increase in similarity between locations that are closer to one-another on a landscape. This means that selecting close-by samples may reduce sample independence. For example, a simple random sampling protocol, which randomly select samples from the sampling frame is particularly susceptible to spatial autocorrelation despite having equal inclusion probabilities across samples, since it is likely that some samples will be closer together than others [2, 20]. Cluster sampling, which divides the population into clusters and then selects a random sample of those clusters is also likely to be impacted by spatial autocorrelation, resulting in samples which are not independent from one another.

The corollary to closer samples being similar is that samples that are too far away may lead to gaps in coverage [2]. This can happen with simple random sampling, but is also a risk associated with systematic sampling, which sets up evenly-spaced samples across the study area. If the landscape in question has periodic characteristics, such as terrain, that coincide at any point with sample spacing, parts of the population will be missed leading to bias [4, 20]. Additionally, in systematic sampling, each sample unit does not have an equal probability of being selected [20].

The wide distribution of samples across the study area can reduce the impact of spatial autocorrelation, however representation of rare or underrepresented values in the model in question may be very important for accuracy assessment. Stratified sampling schemes can be used to improve coverage across the feature space of the predictor variables, but may result in samples that are spatially imbalanced across the study area. Further, the inclusion probabilities of each strata may vary depending on the sample allocation method. For example, for samples that distributed with optimal allocation, which weights the number of samples per strata by a combination of within-strata variance and strata area, a sample in a more variable strata may be more likely to be included than a sample in a less-prevalent strata. This can be remedies by re-balancing the resulting accuracy metrics by the strata allocation weights [20].

Bibliography

[1] Al-Abdulrazzak, D. and Pauly, D. (2014). Managing fisheries from space: Google earth improves estimates of distant fish catches. *ICES Journal of Marine Science*, 71(3):450–454.

[2] Atkinson, P. M. (1991). Optimal ground-based sampling for remote sensing investigations: estimating the regional meant. *International Journal of Remote Sensing*, 12(3):559–567.

[3] Avery., T. E. (1969). *Forester's Guide to Aerial Photo Interpretation.* Number 308 in Agriculture Handbook. U.S. Department of Agriculture Forest Service., Washington, D.C.

[4] Avery, T. E. and Burkhart, H. E. (1994). *Forest Measurements*. McGraw-Hill, 4 edition.

[5] Azzali, S. and Menenti, M. (2000). Mapping vegetation-soil-climate complexes in southern africa using temporal fourier analysis of noaa-avhrr ndvi data. *International Journal of Remote Sensing*, 21(5):973–996.

[6] Baccini, A., Goetz, S. J., Walker, W. S., Laporte, N. T., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P. S. A., Dubayah, R., Friedl, M. A., Samanta, S., and Houghton, R. A. (2012). Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change*, 2(3):182–185.

[7] Baccini, A., Laporte, N., Goetz, S. J., Sun, M., and Dong, H. (2008). A first map of tropical africa's above-ground biomass derived from satellite imagery. *Environmental Research Letters*, 3(4):45011.

[8] Bastin, J.-F., Berrahmouni, N., Grainger, A., Maniatis, D., Mollicone, D., Moore, R., Patriarca, C., Picard, N., Sparrow, B., Abraham, E. M., Aloui, K., Atesoglu, A., Attore, F., Bassüllü, Ç., Bey, A., Garzuglia, M., García-Montero, L. G., Groot, N., Guerin, G., Laestadius, L., Lowe, A. J., Mamane, B., Marchi, G., Patterson, P., Rezende, M., Ricci, S., Salcedo, I., Diaz, A. S.-P., Stolle, F., Surappaeva, V., and Castro, R. (2017). The extent of forest in dryland biomes. *Science*, 356(6338):635–638.

[9] Beuchle, R., Grecchi, R. C., Shimabukuro, Y. E., Seliger, R., Eva, H. D., Sano, E., and Achard, F. (2015). Land cover changes in the brazilian cerrado and caatinga biomes from 1990 to 2010 based on a systematic remote sensing sampling approach. *Applied Geography*, 58:116–127.

[10] Brooks, E., Wynne, R., Thomas, V., Blinn, C., and Coulston, J. (2013). On-the-fly massively multitemporal change detection using statistical quality control charts and landsat data. *IEEE Transactions on Geoscience and Remote Sensing*, 52.

[11] Brooks, E. B., Thomas, V. A., Wynne, R. H., and Coulston, J. W. (2012). Fitting the multitemporal curve: A Fourier series approach to the missing data problem in remote sensing analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 50(9):3340–3353.

[12] Brooks, E. B., Yang, Z., Thomas, V. A., and Wynne, R. H. (2017). Edyn: Dynamic signaling of changes to forests using exponentially weighted moving average charts. *Forests*, 8(9).

[13] Campbell, J. B. and Wynne, R. H. (2011). *Image Interpretation*, chapter 5, pages 130–153. Guilford Press, New York, NY, 5 edition.

[14] Chen, J., Zhu, X., Vogelmann, J. E., Gao, F., and Jin, S. (2011). A simple and effective method for filling gaps in landsat etm+ slc-off images. *Remote Sensing of Environment*, 115(4):1053–1064.

[15] Chuvieco, E. and Vega, J. M. (1990). Visual versus digital analysis for vegetation mapping: Some examples on central spain. *Geocarto International*, 5(3):21–30.

[16] Clark, J. T., Finco, M. V., Warbington, R., and Schwind, B. (2004). Digital mylar: A tool to attribute vegetation polygon features over high-resolution imagery. *Remote Sensing for Field Users, Proceedings of the Tenth Forest Service Remote Sensing Applications Conference*, April 5-9 2004.

[17] Colwell, R. N., Bradshaw, K. E., Smith, H. T. U., Thoren, R., and von Drabbe, C. A. J. (1952). Report of the commission vii (photographic interpretation) to the international society of photogrammetry. *Photogrammetric Engineering*, 18:375—451.

[18] Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37:35 – 46.

[19] Congalton, R. G. and Biging, G. S. (1992). A pilot study evaluating ground reference data collection efforts for use in forest inventory. *Photogrammetric Engineering & Remote Sensing*, 58(12):1669 – 1671.

[20] Congalton, R. G. and Green., K. (2009). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC Press/Taylor & Francis Group, Boca Raton, FL, 2 edition.

[21] Coulston, J. W., Moisen, G. G., Wilson, B. T., Finco, M. V., Cohen, W. B., and Brewer, C. K. (2012). Modeling percent tree canopy cover–a pilot study. *Photogrammetric Engineering and Remote Sensing*, 78ID - 22(7):715–727.

[22] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. (2018). Quality control in crowdsourcing. *ACM Computing Surveys*, 51(1):1–40.

[23] Esseen, P.-A., Jansson, K. U., and Nilsson, M. (2006). Forest edge quantification by line intersect sampling in aerial photographs. *Forest Ecology and Management*, 230(1):32 – 42.

[24] Estellés-Arolas, E. and L. Guevara, F. G. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38:1–14.

[25] Foody, G., See, L., Fritz, S., Moorthy, I. P., Perger, C., Schill, C., and Boyd, D. (2018). Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data. *ISPRS International Journal of Geo-Information*, 7(3):80.

[26] Foody, G. M. (2009). Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, 30(20):5273–5291.

[27] Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., and Obersteiner, M. (2009). Geo-Wiki.org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, 1(3):345–354.

[28] Fung, T. and LeDrew, E. (1987). Application of principal components analysis to change detection. *Photogrammetric engineering and remote sensing*, 53(12):1649–1658.

[29] Galloway, A. W. E., Tudor, M. T., and Haegen., W. M. V. (2006). The reliability of citizen science: A case study of oregon white oak stand surveys. *Wildlife Society Bulletin (1973-2006)*, 34(5):1425–1429.

[30] Gao, F., Masek, J., Schwaller, M., and Hall, F. (2006). On the blending of the landsat and modis surface reflectance: predicting daily landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2207–2218.

[31] Hagolle, O., Huc, M., Pascual, D. V., and Dedieu, G. (2010). A multi-temporal method for cloud detection, applied to formosat-2, venÂµs, landsat and sentinel-2 images. *Remote Sensing of Environment*, 114(8):1747–1755.

[32] Haklay, M. and Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

[33] Hammon, L. and Hippner, H. (2012). Crowdsourcing. *Business and Information Systems Engineering*, 3.

[34] Hansen, M. C. and DeFries, R. S. (2004). Detecting long-term global forest change using continuous fields of tree-cover maps from 8-km Advanced Very High Resolution Radiometer (AVHRR) data for the years 1982-99. *Ecosystems*, 7(7):695–716.

[35] Hansen, M. C., Defries, R. S., Townshend, J. R. G., and Sohlberg, R. (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21(6-7):1331–1364.

[36] Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853.

[37] Hansen, M. C., Stehman, S. V., and Potapov, P. V. (2010). Quantification of global gross forest cover loss. *Proceedings of the National Academy of Sciences*, 107(19):8650–8655.

[38] Hayes, D. and Sader, S. (2001). Comparison of change-detection techniques for monitoring tropical forest clearing and vegetation regrowth in a time series. *Photogrammetric Engineering and Remote Sensing*, 67:1067–1075.

[39] Hayes, D. J. and Cohen, W. B. (2007). Spatial, spectral and temporal patterns of tropical forest cover change as observed with multiple scales of optical satellite data. *Remote Sensing of Environment*, 106(1):1–16.

[40] Hayes, D. J., Cohen, W. B., Sader, S. A., and Irwin, D. E. (2008). Estimating proportional change in forest cover as a continuous variable from multi-year modis data. *Remote Sensing of Environment*, 112(3):735–749.

[41] Healey, S. P., Cohen, W. B., Zhiqiang, Y., and Krankina, O. N. (2005). Comparison of tasseled cap-based landsat data structures for use in forest disturbance detection. *Remote Sensing of Environment*, 97(3):301–310.

[42] Hernández, H., Gutiérrez, M., and Acuña, M. P. (2016). Urban morphologic dynamics in santiago (chile): Proposing sustainable indicators from remote sensing. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B8:873–877.

[43] Hilker, T., Wulder, M. A., Coops, N. C., Linke, J., McDermid, G., Masek, J. G., Gao, F., and White, J. C. (2009a). A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on landsat and modis. *Remote Sensing of Environment*, 113(8):1613–1627.

[44] Hilker, T., Wulder, M. A., Coops, N. C., Seitz, N., White, J. C., Gao, F., Masek, J. G., and Stenhouse, G. (2009b). Generation of dense time series synthetic landsat data through data blending with modis using a spatial and temporal adaptive reflectance fusion model. *Remote Sensing of Environment*, 113(9):1988–1999.

[45] Howe, J. (2006). The rise of crowdsourcing. *Wired*.

[46] Hsueh, P.-Y., Melville, P., and Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT Workshop on Active Learning for National Language Processing*, pages 27–35. Association for Computational Linguistics; Boulder, Colorado.

[47] Huang, C., Goward, S. N., Masek, J. G., Thomas, N., Zhu, Z., and Vogelmann, J. E. (2010). An automated approach for reconstructing recent forest disturbance history using dense landsat time series stacks. *Remote Sensing of Environment*, 114(1):183–198.

[48] Immerzeel, W., Quiroz, R., and De Jong, S. (2005). Understanding precipitation patterns and land use interaction in Tibet using harmonic analysis of SPOT VGT-S10 NDVI time series. *International Journal of Remote Sensing*, 26(11):2281–2296.

[49] Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. *KDD-HCOMP'10, July 25 2010*.

[50] Jakubauskas, M. E. and Legates, D. R. (2000). Harmonic analysis of time-series avhrr ndvi data for characterizing us great plains land use/land cover. *International Archives of Photogrammetry and Remote Sensing*, 33(B4/1; PART 4):384–389.

[51] Jakubauskas, M. E., Legates, D. R., and Kastens, J. H. (2001). Harmonic analysis of time-series avhrr ndvi data. *Photogrammetric engineering and remote sensing*, 67(4):461–470.

[52] Jakubauskas, M. E., Peterson, D. L., and Legates, D. R. (2002). Fourier decomposition of an avhrr ndvi time series for seasonal and interannual land cover change detection. In *Analysis of Multi-Temporal Remote Sensing Images*, pages 387–394. World Scientific.

[53] Jennings, S. (1999). Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. *Forestry*, 72(1):59–74.

[54] Jin, S., Homer, C., Yang, L., Xian, G., Fry, J., Danielson, P., and Townsend, P. A. (2013). Automated cloud and shadow detection and filling using two-date landsat imagery in the usa. *International Journal of Remote Sensing*, 34(5):1540–1560.

[55] Jonsson, P. and Eklundh, L. (2002). Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE transactions on Geoscience and Remote Sensing*, 40(8):1824–1832.

[56] Kennedy, R. E., Townsend, P. A., Gross, J. E., Cohen, W. B., Bolstad, P., Wang, Y., and Adams, P. (2009). Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote Sensing of Environment*, 113(7):1382–1396. Monitoring Protected Areas.

[57] Lister, A., Lister, T., and Doyle, J. A. (2009). Use of a simple photointerpretation method with free, online imagery to assess landscape fragmentation. In *Proceedings, 2009 Society of American Foresters national convention, Opportunities in a forested world; 2009 Sept. 30-Oct. 4;*. Bethesda MD: Society of American Foresters.

[58] Ma, Q., Su, Y., and Guo, Q. (2017). Comparison of canopy cover estimations from airborne lidar, aerial imagery, and satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(9):4225–4236.

[59] Main-Knorn, M., Cohen, W. B., Kennedy, R. E., Grodzki, W., Pflugmacher, D., Griffiths, P., and Hostert, P. (2013). Monitoring coniferous forest biomass change using a landsat trajectory-based approach. *Remote Sensing of Environment*, 139:277–290.

[60] Malhi, Y., Roberts, J. T., Betts, R. A., Killeen, T. J., Li, W., and Nobre, C. A. (2008). Climate change, deforestation, and the fate of the amazon. *Science*, 319(5860):169–172.

[61] McRoberts, R. and Tomppo, E. (2007). Remote sensing support for national forest inventories. *Remote Sensing of Environment*, 110:412–419.

[62] Miller, H. M., Richardson, L. A., Koontz, S. R., Loomis, J., and Koontz, L. (2013). Users, uses, and value of landsat satellite imagery: results from the 2012 survey of users. *USGS Publications Warehouse*.

[63] Moisen, G. G., Meyer, M. C., Schroeder, T. A., Liao, X., Schleeweis, K. G., Freeman, E. A., and Toney, C. (2016). Shape selection in landsat time series: A tool for monitoring forest dynamics. *Global change biology*, 22(10):3518–3528.

[64] Montgomery, D., Jennings, C., and Kulahci, M. (2015). *Introduction to Time Series Analysis and Forecasting*. Wiley Series in Probability and Statistics. Wiley.

[65] Moody, A. and Johnson, D. M. (2001). Land-surface phenologies from avhrr using the discrete fourier transform. *Remote Sensing of Environment*, 75(3):305–323.

[66] Muchoney, D., Strahler, A., Hodges, J., and LoCastro, J. (1999). The igbp discover confidence sites and the system for terrestrial ecosystem parameterization: Tools for validating global land-cover data. *Photogrammetric Engineering & Remote Sensing*, 65(9):1061–1067.

[67] Muchoney, D. M. and Haack, B. N. (1994). Change detection for monitoring forest defoliation. *Photogrammetric engineering and remote sensing*, 60(10):1243–1252.

[68] Nelson, R. F. (1983). Detecting forest canopy change due to insect activity using landsat mss. *Photogrammetric Engineering and Remote Sensing*, 49:1303–1314.

[69] Nemani, R. and Running, S. (1997). Land cover characterization using multitemporal red, near-ir, and thermal-ir data from noaa/avhrr. *Ecological Applications*, 7(1):79–90.

[70] Olivero, J., Fa, J. E., Real, R., Márquez, A. L., Farfán, M. A., Vargas, J. M., Gaveau, D., Salim, M. A., Park, D., Suter, J., King, S., Leendertz, S. A., Sheil, D., and Nasi., R. (2017). Recent loss of closed forests is associated with ebola virus disease outbreaks. *Scientific Reports*, 7(14291).

[71] Powell, S. L., Cohen, W. B., Healey, S. P., Kennedy, R. E., Moisen, G. G., Pierce, K. B., and Ohmann, J. L. (2010). Quantification of live aboveground forest biomass dynamics with landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sensing of Environment*, 114(5):1053–1068.

[72] Pu, R. (2010). Mapping urban forest tree species using ikonos imagery: Preliminary results. *Environmental monitoring and assessment*, 172:199–214.

[73] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

[74] Reams, G. A., Smith, W. D., Hansen, M. H., Bechtold, W. A., Roesch, F. A., and Moisen., G. G. (2005). The forest inventory and analysis sampling frame. In Bechtold, W. A. and Patterson, P. L., editors, *The Enhanced Forest Inventory and Analysis Program–national Sampling Design and Estimation Procedures*, Gen. Tech. Rep. SRS-80, chapter 2, page 11—26. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC.

[75] Riva-Murray, K., Riemann, R., Murdoch, P., Fischer, J. M., and Brightbill, R. (2010). Landscape characteristics affecting streams in urbanizing regions of the delaware river basin (new jersey, new york, and pennsylvania, u.s.). *Landscape Ecol*, 25:1489–1503.

[76] Roerink, G., Menenti, M., and Verhoef, W. (2000). Reconstructing cloudfree ndvi composites using fourier analysis of time series. *International Journal of Remote Sensing*, 21(9):1911–1917.

[77] Roy, D. P., Ju, J., Lewis, P., Schaaf, C., Gao, F., Hansen, M., and Lindquist, E. (2008). Multi-temporal modisâ€“landsat data fusion for relative radiometric normalization, gap filling, and prediction of landsat data. *Remote Sensing of Environment*, 112(6):3112–3130.

[78] Ryan, B. J. and Freilich, M. H. (2008). Landsat data distribution policy. *Landsat Headlines*.

[79] Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., and Ohno, H. (2005). A crop phenology detection method using time-series modis data. *Remote Sensing of Environment*, 96(3):366–374.

[80] Salk, C. F., Sturn, T., See, L., and Fritz, S. (2016). Limitations of majority agreement in crowdsourced image interpretation. *Transactions in GIS*, 21(2):207–223.

[81] Scepan, J. (1999). Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering & Remote Sensing*, 65(9):1051–1060.

[82] Schepaschenko, D., Fritz, S., See, L., Laso Bayas, J. C., Lesiv, M., Kraxner, F., and Obersteiner, M. (2017). Comment on "the extent of forest in dryland biomes". *Science*, 358(6362).

[83] Schepaschenko, D., See, L., Lesiv, M., McCallum, I., Fritz, S., Salk, C. F., Moltchanova, E., Perger, C., Shchepashchenko, M., Shvidenko, A., Kovalevskyi, S., Gilitukha, D., Albrecht, F., Kraxner, F., Bun, A., Maksyutov, S., Sokolov, A., Dürauer, M., Obersteiner, M., Karminov, V., and Ontikov, P. (2015). Development of a global hybrid forest mask

through the synergy of remote sensing, crowdsourcing and FAO statistics. *Remote Sensing of Environment*, 162:208—220.

[84] Schroeder, T. A., Schleeweis, K. G., Moisen, G. G., Toney, C., Cohen, W. B., Freeman, E. A., Yang, Z., and Huang, C. (2017). Testing a Landsat-based approach for mapping disturbance causality in U.S. forests. *Remote Sensing of Environment*, 195:230–243.

[85] See, L., Comber, A., Salk, C. F., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., and Obersteiner, M. (2013). Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLOS ONE*, 8(7):1–11.

[86] Sexton, J. O., Noojipady, P., Song, X.-P., Feng, M., Song, D.-X., Kim, D.-H., Anand, A., Huang, C., Channan, S., Pimm, S. L., and Townshend, J. R. (2016). Conservation policy and the measurement of forests. *Nature Climate Change*, 6:192—196.

[87] Smelser, R. L. and Patteson., A. W. (1975). *Photointerpretation Guide for Forest Resource Inventories.* NASA TM X-58195. U.S. Department of Commerce National Technical Information Service. National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston, TX.

[88] Stehman, S. and Czaplewski, R. (1998). Design and analysis for thematic map accuracy assessment. *Remote Sensing of Environment*, 64:331–334.

[89] Storey, J., Scaramuzza, P., and Barsi, J. (2005). Landsat 7 scan line corrector-off gap-filled product development. In *Proceeding of Pecora*, volume 16, pages 23–27.

[90] Thomlinson, J., Bolstad, P., and Cohen, W. (1999). Coordinating methodologies for scaling landcover classifications from site-specific to global: Steps toward validating global map products. *Remote Sensing of Environment*, 70(1):16–28.

[91] Toney, C., Liknes, G., Lister, A., and Meneguzzo, D. (2012). Assessing alternative measures of tree canopy cover: Photo-interpreted naip and ground-based estimates. In *Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists.*, e-Gen. Tech. Rep. SRS-157, page 209—215, Asheville, NC. U.S. Department of Agriculture, Forest Service, Southern Research Station.

[92] Toney, C., Shaw, J. D., and Nelson, M. D. (2009). A stem-map model for predicting tree canopy cover of forest inventory and analysis (FIA) plots. In *Forest Inventory and Analysis (FIA) Symposium 2008*, number January 2009 in e-Gen. Tech. Rep. RMRS-P-56CD, page 1—19, Fort Collins, CO. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

[93] Turner, W., Rondinini, C., Pettorelli, N., Mora, B., Leidner, A., Szantoi, Z., Buchanan, G., Dech, S., Dwyer, J., Herold, M., Koh, L., Leimgruber, P., Taubenboeck, H., Wegmann,

M., Wikelski, M., and Woodcock, C. (2015). Free and open-access satellite data are key to biodiversity conservation. *Biological Conservation*, 182:173–176.

[94] Uhrin, A. V. and Townsend, P. A. (2016). Improved seagrass mapping using linear spectral unmixing of aerial photographs. *Estuarine, Coastal and Shelf Science*, 171:11–22.

[95] USDA Forest Service, F. I. and Program, A. (2021). What is fia?

[96] Wiggins, A. and Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10.

[97] Woodcock, C., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S. N., Helder, D., Helmer, E. H., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P., Vermote, E., Vogelmann, J., Wulder, M., and Wynne, R. H. (2008). Free access to landsat imagery teach by the book science education. *Science*, 320:1011–10112.

[98] Woodcock, C. E., Macomber, S. A., Pax-Lenney, M., and Cohen, W. B. (2001). Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors. *Remote Sensing of Environment*, 78(1–2):194–203.

[99] Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., and Woodcock, C. E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of landsat. *Remote Sensing of Environment*, 122:2–10. Landsat Legacy Special Issue.

[100] Xavier, A. C., Rudorff, B. F. T., Shimabukuro, Y. E., Berka, L. M. S., and Moreira, M. A. (2006). Multiâ€ temporal analysis of modis data to classify sugarcane crop. *International Journal of Remote Sensing*, 27(4):755–768.

[101] Yu, L., Ball, S. B., Blinn, C. E., Moeltner, K., Peery, S., Thomas, V. A., and Wynne, R. H. (2015). Cloud-sourcing: Using an online labor force to detect clouds and cloud shadows in landsat images. *Remote Sensing*, 7(3):2334–2351.

Chapter 2

Estimating tree canopy cover using harmonic regression coefficients derived from multi-temporal Landsat data

Int J Appl Earth Obs Geoinformation 86 (2020) 101985

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

journal homepage: www.elsevier.com/locate/jag

Estimating tree canopy cover using harmonic regression coefficients derived from multitemporal Landsat data

Jill M. Derwin[a,*], Valerie A. Thomas[a], Randolph H. Wynne[a], John W. Coulston[b], Greg C. Liknes[c], Stacie Bender[d], Christine E. Blinn[a], Evan B. Brooks[a], Bonnie Ruefenacht[d], Robert Benton[d], Mark V. Finco[d], Kevin Megown[d]

[a] Virginia Tech, Forest Resources and Environmental Conservation, Blacksburg, VA 24061 USA
[b] United States Forest Service, Southern Research Station, Blacksburg, VA 24061 USA
[c] United States Forest Service, Northern Research Station, St Paul, MN 55108 USA
[d] United States Forest Service, Geospatial Technology and Applications Center, Salt Lake City, UT 84138 USA

ARTICLE INFO

Keywords:
Tree Canopy Cover
Landsat time series
Harmonic regression
Image compositing
Random forest regression models

ABSTRACT

The goal of this study was to evaluate whether harmonic regression coefficients derived using all available cloud-free observations in a given Landsat pixel for a three-year period can be used to estimate tree canopy cover (TCC), and whether models developed using harmonic regression coefficients as predictor variables are better than models developed using median composite predictor variables, the previous operational standard for the National Land Cover Database (NLCD). The two study areas in the conterminous USA were as follows: West (Oregon), bounded by Landsat Worldwide Reference System 2 (WRS-2) paths/rows 43/30, 44/30, and 45/30; and South (Georgia/South Carolina), bounded by WRS-2 paths/rows 16/37, 17/37, and 18/37. Plot-specific tree canopy cover (the response variable) was collected by experienced interpreters using a dot grid overlaid on 1 m spatial resolution National Agricultural Imagery Program (NAIP) images at two different times per region, circa 2010 and circa 2014. Random forest model comparisons (using 500 independent model runs for each comparison) revealed the following (1) harmonic regression coefficients (one harmonic) are better predictors for every time/region of TCC than median composite focal means and standard deviations (across times/regions, mean increase in pseudo R^2 of 6.7% and mean decrease in RMSE of 1.7% TCC) and (2) harmonic regression coefficients (one harmonic, from NDVI, SWIR1, and SWIR2), when added to the full suite of median composite and terrain variables used for the NLCD 2011 product, improve the quality of TCC models for every time/region (mean increase in pseudo R^2 of 3.6% and mean decrease in RMSE of 1.0% TCC). The harmonic regression NDVI constant was always one of the top four most important predictors across times/regions, and is more correlated with TCC than the NDVI median composite focal mean. Eigen analysis revealed that there is little to no additional information in the full suite of predictor variables (47 bands) when compared to the harmonic regression coefficients alone (using NDVI, SWIR1, and SWIR2; 9 bands), a finding echoed by both model fit statistics and the resulting maps. We conclude that harmonic regression coefficients derived from Landsat (or, by extension, other comparable earth resource satellite data) can be used to map TCC, either alone or in combination with other TCC-related variables.

1. Introduction

Tree canopy cover (TCC) is the proportion of the forest floor covered by the vertical projection of tree crowns (Jennings, 1999). TCC is an essential measure of forest health and productivity and is used in applications like climate change mitigation, forest management, and pest and disease monitoring. TCC influences wildlife habitat (North et al., 2017), forb yield (Muoghalu and Isichei, 1991), proportion of C3 vs. C4 grasses (Peterson et al., 2007), urban property values (Pandit et al., 2014), and rainfall partitioning (Owens et al., 2006), among others.

At the national scale, TCC is commonly studied annually. Numerous authors have shown the benefits of describing multitemporal

* Corresponding author.
E-mail addresses: jmd59@vt.edu (J.M. Derwin), thomasv@vt.edu (V.A. Thomas), wynne@vt.edu (R.H. Wynne), jcoulston@fs.fed.us (J.W. Coulston), staciebender@fs.fed.us (S. Bender), cblinn@vt.edu (C.E. Blinn), evbrooks@vt.edu (E.B. Brooks), bruefenacht@fs.fed.us (B. Ruefenacht), robertbenton@fs.fed.us (R. Benton), mfinco@fs.fed.us (M.V. Finco), kamegown@fs.fed.us (K. Megown).

https://doi.org/10.1016/j.jag.2019.101985
Received 22 February 2019; Received in revised form 16 August 2019; Accepted 1 October 2019
Available online 22 November 2019
0303-2434/ © 2019 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

Table 1
Phenology metrics used to map tree canopy cover using satellite data. Only the sensors and bands shown to be important in the resulting tree canopy cover models are listed. Abbreviations: AVHRR, Advanced Very High Resolution Radiometer; MODIS, MOderate-resolution Imaging Spectrometer; OLI, Operational Land Imager, SPOT VGT, Satellite Pour l'Observation de la Terre VEGETATION; TM, Thematic Mapper; NIR, near infrared; SWIR, shortwave infrared; NDVI, normalized difference vegetation index; NDMI, normalized difference moisture index; EVI, enhanced vegetation index; FAPAR, fraction of absorbed photosynthetically active radiation; TC B,G,W, tasseled cap brightness, greenness, and wetness; std, standard deviation; SOS, start of season; DSINT, dry season integral (annual integral minus the integral from start of season to end of season); RDERIV, right derivative (rate of decrease after growing season maximum).

Study	Sensors(s)	Bands/Indices	Curve Fitting	Phenology Metrics
Hansen and DeFries (2004)	AVHRR (monthly composite ordinal by calendar year)	red NIR NDVI BT	No	2nd 6th mean (2nd:6th) mean (2nd:4th) |6th–2nd|
Gessner et al. (2013)	MODIS	blue red SWIR (2.13 μm) NDVI EVI	Fourier	minimum* maximum mean* median amplitude
Karlson et al. (2015)	Landsat OLI	NDVI	No	minimum maximum mean median product* *std*
Brandt et al. (2016)	MODIS SPOT VGT	FAPAR	TIMESAT	DSINT* SOS* RDERIV
Ruefenacht (2016)	Landsat TM	all optical NDVI NDMI TC B,G,W	No	median composite*

*Denotes a particularly important variable.

reflectance data using phenology metrics to estimate TCC, as opposed to selecting one or only a few images throughout the growing season. DeFries et al. (1995) proposed the use of NDVI temporal dynamics to estimate (vegetation) continuous fields of canopy cover. Hansen and DeFries (2004) produced annual global TCC estimates at 8 km spatial resolution using per-pixel phenological metrics derived from Advanced Very High Resolution (AVHRR) monthly composites. Training data were developed using much higher spatial resolution imagery. Phenology metrics utilized were selected monthly composite values coupled with key descriptive statistics for the red, near infrared (NIR), normalized difference vegetation index, and brightness temperature bands.

Phenology metrics have been used as predictor variables in other, more recent, studies in which satellite imagery was used to map tree canopy cover (Table 1). Gessner et al. (2013) mapped the fraction of woody canopy cover in southern Africa using phenology metrics (Table 1) derived from Fourier-smoothed (via harmonic regression, Fig. 1) annual time series of MODIS imagery with blue, red, SWIR (2.13 μm), NDVI, and enhanced vegetation index (EVI) bands. Their resulting random forest models had an RMSE of 8.1% cover in the Kalahari and 3.1% cover in central and eastern Namibia. The three most important variables were the SWIR3 minimum in the dry season, the annual NDVI mean, and the rainy season NDVI mean. Karlson et al. (2015) mapped tree canopy cover at a 30 m pixel resolution in Burkina Faso using Landsat Operational Land Imager (OLI)-derived spectral, texture, and phenology variables. All three variable types (including one phenology metric) were necessary in the resulting random forests model that required only five variables (R^2 = 77%, RMSE = 8.9% cover) derived using backwards feature elimination to find the smallest possible model. Brandt et al. (2016) compared phenology metrics derived using TIMESAT (Jönsson and Eklundh, 2004) smoothed data from the Moderate Resolution Imaging Spectroradiometer (MODIS) and Satellite pour l'Observation de la Terre (SPOT) VEGETATION (VGT) for their utility in mapping percent woody cover at a 1 km scale across the Sahel. The most important phenology metrics, regardless of sensor, were the dry season integral and the starting point of the growing season, defined as 20% of the amplitude. The resulting multiple linear regression models had good predictive ability (R^2, MODIS = 73% and R^2, VGT = 70%) and excellent RMSEs (MODIS, 3.0% cover and VGT, 3.2% cover), especially given the challenging nature of the arid landscape.

Ruefenacht (2016) mapped tree canopy cover using 3 × 3 pixel window focal means of the growing season median composite value for each of the six multispectral bands from the Landsat Thematic Mapper (TM), plus derived indices NDVI, tasseled cap brightness, greenness, and wetness, and the normalized difference moisture index (NDMI). The growing season was defined as the period in which NDVI was within 0.1 of the growing season maximum. Tree canopy cover was mapped in Multi-Resolution Land Characteristics Consortium Mapping Zones 16, 23, 48, 54, and 59. Model quality varied across zones, with R^2s ranging from 50% to 86% and RMSEs ranging from 12.8% cover to 15.3% cover. Median composites were used operationally to map tree canopy cover as an element of the National Land Cover Database (NLCD) 2011 suite of nationwide geospatial data layers (Ruefenacht et al., 2015).

De Beurs and Henebry (2010) categorized methods by which land surface phenology can be characterized as follows: (1) thresholds, (2) derivatives, (3) smoothing functions, and (4) fitted models. Given that standardized markers of start and end of season are first order objectives of these techniques, thresholds and derivatives are less well suited to develop predictor variables not related to growing season timing or length. De Beurs and Henebry (2010) further note that complicated fitted models such as Gaussian local functions result in parameter coefficients with no clear ecological meaning. As such, simpler smoothing functions such as the autoregressive moving average or Fourier analysis can result in parsimonious models with physically meaningful coefficients with only slight loss to model fit. Given (among other issues) the subjective, heuristic nature of selecting a suitable lag time for autoregressive moving average models of land surface phenology, they play an increasingly less prominent role.

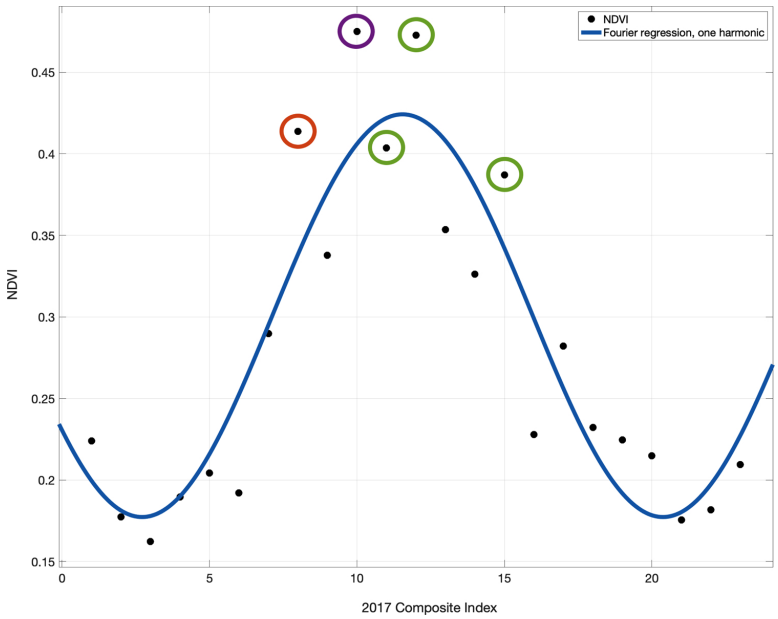

Fig. 1. Median composite and harmonic regression. This figure illustrates the median composite method using NDVI data from MODIS 250 m 16-day composites from 2017 (MYD13Q1 collection 6, coordinates 37.4378°, −80.2727°). Landsat median composite images are calculated using the median reflectance value from up to 15 images from three years preceding the target year. In this example, the maximum MODIS NDVI value is shown, circled, in violet. The Landsat median composite selects observations within 0.1 GLOVIS maximum annual NDVI value (circled in green or red for this example, using annual maximum MODIS NDVI) are used to calculate the median composite, circled in red. (Note that for non-composited Landsat data the earliest and latest value within 0.1 NDVI of the maximum serve to define the range of data used, and the pixel-specific effects of cloud cover result in a median being computed with different numbers of values from pixel to pixel.) The figure also demonstrates the harmonic regression method with MODIS NDVI data. Harmonic regression of surface reflectance observations uses all available observations (black dots) and fits a harmonic curve (blue line) following Eq. (1). Both methods exclude observations identified as cloud cover. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fourier (harmonic) analysis of annual spectral band reflectances or, more commonly, vegetation indices derived therefrom, has long been used to characterize land surface phenology (Moody and Johnson, 2001). Fourier analysis is sensitive to systematic signal changes but relatively insensitive to (nonsystematic) noise (Moody and Johnson, 2001; De Beurs and Henebry, 2010). Brooks et al. (2012) define Fourier series as "superimposed sequences, over an interval of time, of a constant with sines and cosines of increasing integer multiples of the original frequency based on the time interval" (see Eq. (1) and Fig. 1). The constant is the mean of the series, and the sine and cosine pairings at specified frequencies are called the harmonics of the series (Brooks et al., 2012) (Eq. (1)).

$$f(t) = a_0 + \underbrace{(a_1 \cos t + b_1 \sin t)}_{\text{first harmonic}} + \underbrace{(a_2 \cos 2t + b_2 \sin 2t)}_{\text{second harmonic}} + \underbrace{(a_3 \cos 3t + b_3 \sin 3t)}_{\text{third harmonic}} + \cdots \quad (1)$$

where t is the "time from some initial epoch" (Shureman, 1940), a_0 is the constant, and for each harmonic n, the coefficients a_n and b_n are the amplitudes of the cosine and sine waves, respectively (Smith, 1997). Using harmonic regression on Landsat time-series data to produce a simplified curve overcomes both noise and gaps caused by clouds, snow, or instrument problems, which is a major advantage to the approach (Brooks et al., 2012; Wilson et al., 2018).

For land surface phenology applications, the period is one year, defined in the same units as t. The first harmonic is also known as the fundamental frequency, and is the frequency that the time domain repeats itself (Smith, 1997). As Moody and Johnson (2001) noted, for vegetation indices the first harmonic is particularly associated with the overall productivity of the region, with the constant (a_0) the mean. Regional and continental applications, particularly across biomes, often use only the first harmonic even though the second (and subsequent) harmonics can improve fit. This is due to the second harmonic being variable across years and ecosystems due to its sensitivity to secondary vegetation and climate anomalies (Moody and Johnson, 2001). Fig. 1 demonstrates harmonic regression of surface reflectance observations and compares it to the median composite method, previously described.

Advantages of harmonic regression include (1) the orthogonality of the sine and cosine functions (Weisstein, 2010) and (2) the ability to store the coefficients as rasters (Brooks et al., 2012) (Fig. 2), and (3) the proven utility of the coefficients themselves in a wide variety of mapping applications, both for categorical (Brooks et al., 2016; Wilson et al., 2018) and continuous (Immerzeel et al., 2005; Wilson et al., 2018) variables. However, no prior study (Table 1) has used harmonic regression coefficients derived from moderate resolution resource satellite imagery as predictor variables to estimate tree canopy cover.

The goal of this study was to evaluate whether harmonic regression coefficients derived using all available cloud-free observations in a given Landsat pixel for a three-year period (corresponding with the timing of imagery used for training data collection) can be used to estimate tree canopy cover, and whether using harmonic regression coefficients as predictor variables is an improvement over calculating Landsat-derived predictor variables using the median composite, the current operational standard for NLCD. We explored two research questions in service of this goal, as follows: (1) Are harmonic regression coefficients better predictors of TCC than median composite variables for the same vegetation index and wavelength bands? (2) Can harmonic regression coefficients, when added to a full suite of median composite and terrain variables improve the quality of TCC models (simulating the operational framework for the original 2011 NLCD TCC product vs. that used for the 2016 NLCD TCC product)? In order to answer the first question, we compared models using harmonic regression coefficients for NDVI, SWIR1 and SWIR2, all good predictors of TCC (Table 1; Asner and Lobell, 1996; Lobell et al., 2001) to the median composite values for those same three bands. To answer the second question we compared models using (a) the full suite of median composite and terrain variables used in the original 2011 TCC product, and (b) all predictor variables used in (a) *plus* the harmonic regression coefficients from NDVI, SWIR1, and SWIR2.

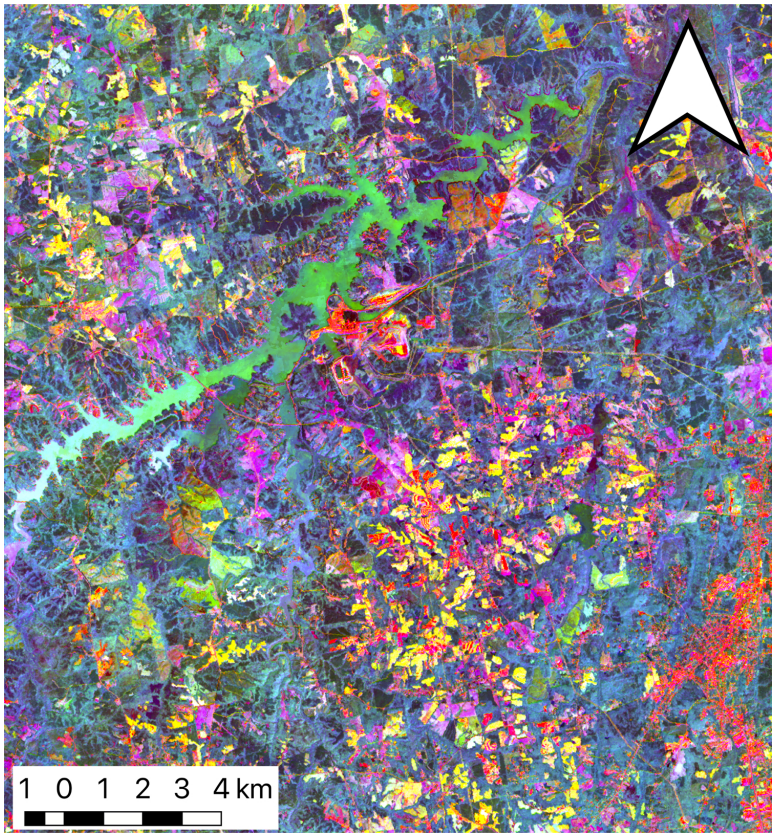

Fig. 2. Harmonic regression coefficients (one harmonic) for the Landsat 8 OLI panchromatic band (15 m spatial resolution) using all cloud-free observations for each pixel from 2014–2016. a_0 shown in red, a_1 shown in green, and b_1 shown in blue. Image boundary coordinates (x:y pairs, Conus Albers, EPSG:5070) upper left 1,485,996 m, 1,634,473 m; lower right 1,505,392 m, 1,613,539 m. Note the quality of land cover and land use discrimination that is possible even using harmonic regression coefficients from only one wavelength band. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Data and methods

2.1. Study area

There were two study areas (Fig. 3). The first was in Oregon (USA), bounded by Landsat Worldwide Reference System 2 (WRS-2) paths/rows 43/30, 44/30, and 45/30, and the second was in Georgia and South Carolina (USA) in WRS-2 paths/rows 16/37, 17/37, and 18/37. There are a variety of land cover types represented across the study areas, though the western study area has large tracts with reduced TCC, such as shrublands or pasture. The southern study area has a more diverse, dynamic landscape, characterized by managed forests, interspersed with suburban development.

2.2. Training (response) data

Training data were collected using a photo interpretation (PI) approach that superimposes an evenly spaced 90 m × 90 m dot grid, approximately the area covered by an FIA field plot, over 1 m resolution NAIP images for a sample of photo interpretation points. These data were collected by experienced interpreters as part of a joint effort between the United States Forest Service's (USFS) Forest Inventory and Analysis (FIA) program and the USFS's Geospatial Technology and Applications Center (GTAC) (Toney et al., 2009; Goeking et al., 2012). NAIP images are acquired after green up, when leaves are on trees, so the PI process assumed those conditions. Image acquisitions occurred during summer months of 2009 and 2010 for Oregon and Georgia respectively, and spring of 2011 for South Carolina for Time 1. For Time 2, acquisition occurred during the summer months of 2012 and 2013 for Oregon and Georgia respectively, and during the fall of 2013 for South Carolina (Table 2).

At every photo interpretation point, trained interpreters indicated whether or not each of 105 dots fell on tree canopy, and plot-level percent TCC was calculated by dividing the number of canopy observations by the total number of dots observed. 2626 interpreted points were used, 1360 in the southern study area and 1266 in the western study area (Table 3).

2.3. Training (response) data characteristics

For the south, both time periods have relatively consistent bimodal distributions (Fig. 4, upper half and Fig. 5, left) of observed TCC that are left-skewed. For time 1, the mean is 66.8% and the median is 86.2%. For time 2 the mean is 65.8% and the median is 83.5%. In the west, both time periods also have relatively consistent distributions, but they are not bimodal (Fig. 4, lower half and Fig. 5, right). They are skewed to the right rather than to the left. For time 1, the median is 0.9% and the mean is 18.2%. For time 2 the mean is 18.6% and median is 0%. All sets of observations (both regions/dates) have a minimum of 0% and a maximum of 100%.

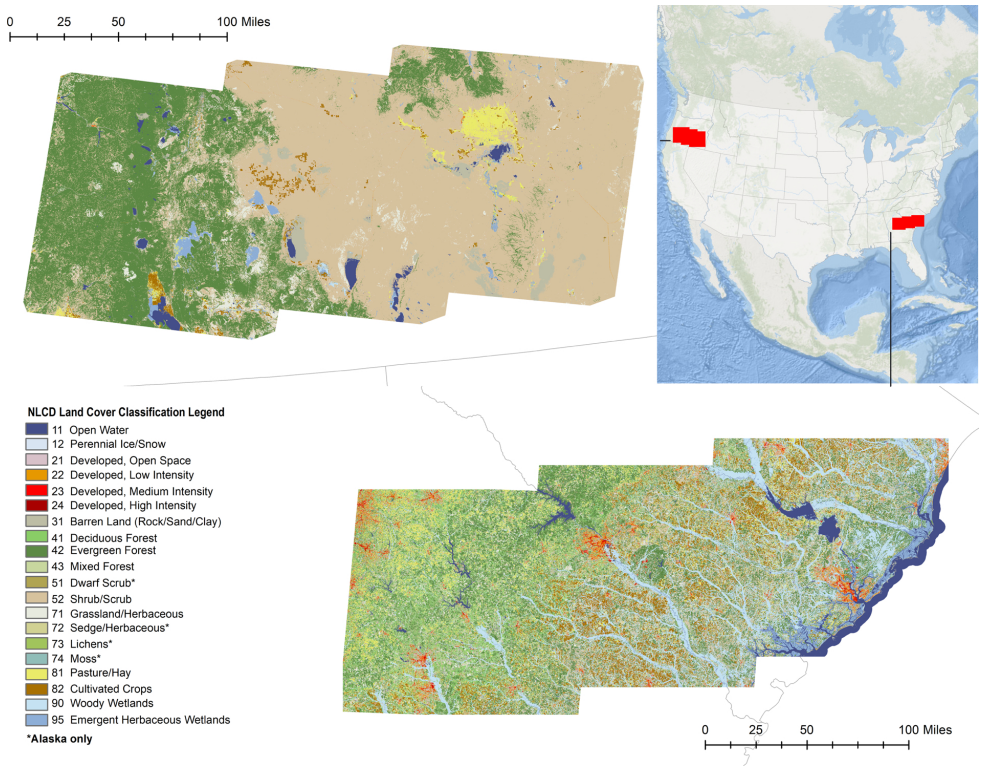

Fig. 3. Western(top) and Southern(bottom) study areas with 2011 National Land Cover Database classification.

Table 2
US National Agricultural Imagery Program (NAIP) image acquisition dates. Note that NAIP aerial images are collected with three visible bands and one near infrared band on clear days to ensure that images have less than 10 percent cloud cover per quarter quad tile. National coverage has been obtained every 3 years since 2009 (USDA Farm Service Agency, 2018).

Year	State	Date Range
2009	Oregon	June 1 to August 8
2012		June 10 to August 4
2010	Georgia	July 3 to October 6
2013		June 28 to October 26
2011	South Carolina	April 19 to May 26
2013		August 25 to October 30

Table 3
Number of photo interpretation (response data) samples per region and time period.

	Time 1	Time 2
South	1360	1360
West	1266	1266
Total	2626	2626

2.4. Development of predictor variables

2.4.1. Predictor variables for median composite models

Following Ruefenacht (2016), we developed a suite of imagery corresponding to Time 1 (2009 West, 2010–2011 South) and Time 2 (2012 West, 2013 South) of the response data by region. Landsat median composite images are a composite of observations from up to 15 dates from the three years leading up to the target year (Fig. 1). These observations were obtained during the annual periods when Global Visualization Viewer (GLOVIS) NDVI curves for forest classes (forest, evergreen forest, mixed forest, and woody wetlands) were within 10% of their annual maximum. Only images with less than 70% cloud cover were used, and clouds and cloud shadows were masked out of the remaining images using Fmask (Zhu and Woodcock, 2012). The median composite predictor variables (Table 7) include the means and standard deviations (calculated using a 90 m × 90 m focal window; Ruefenacht, 2016) of NDVI, NDMI, reflectances of Landsat 7 bands 1–5 and 7, reflectances of bands 2–7 and 9 for Landsat 8, and tasseled cap bands 1–6. In addition to median composite imagery, terrain data from a 30 m digital elevation model were derived. These included the means and standard deviations of elevation, slope, aspect, sin(aspect), and cos (aspect).

2.4.2. Harmonic regression coefficients from Landsat time series

Harmonic regression predictor variables (one harmonic, Table 7) were derived from multitemporal stacks of all available Landsat 5 and 7 images, acquired from USGS, for 3-year periods (with some pixel-specific exceptions as noted below). For Time 1, the training period started in 2008 for the western scenes and 2009 for the southern scenes. For Time 2, the period was from 2012 to 2014. This amounted to 65 dates per path/row at a minimum. To ensure a good harmonic curve fit, an R^2 threshold of 0.9 was applied for the Time 1 data using the EWMACD algorithm with the parameters listed in Table 4(Brooks et al., 2016; EWMACD v.1.8.7, Brooks, 2019). If the curve fit was less than the

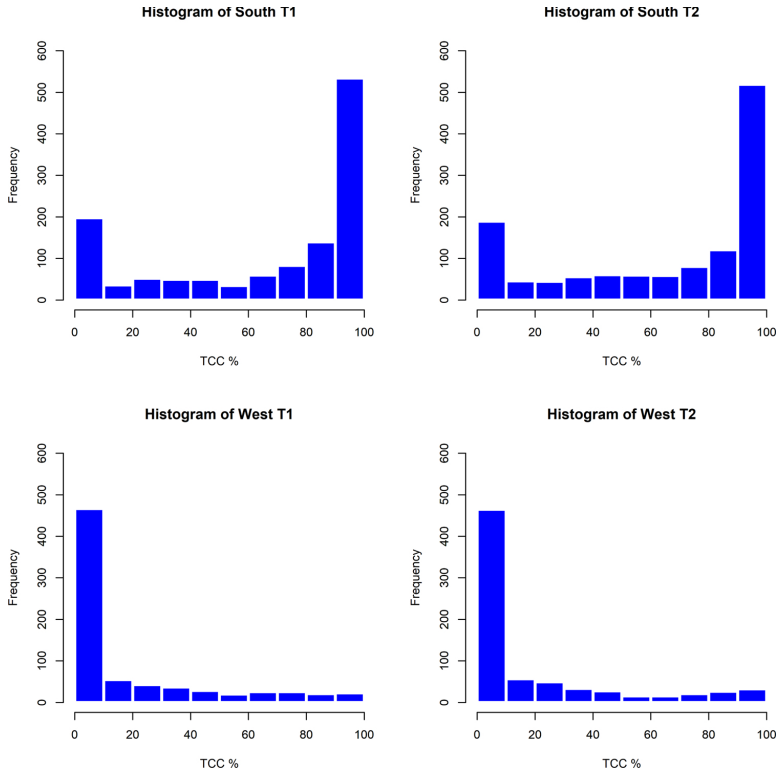

Fig. 4. Histograms of photo interpreted TCC observations for T1 (time 1) and T2 (time 2) in the South and West.

threshold, the temporal window of training observations was expanded to include more scenes from earlier dates, allowing time 1's training period to potentially extend to 2012. Therefore, the total number of input scenes (dates) varies by pixel, with the maximum number of scenes for any pixel being 192. Pre-processing for these data included the masking of clouds, cloud shadows, snow, and water from all images using cfmask provided in the LEDAPS level-1 product.

We compiled these stacks for the short wave infrared bands (SWIR 1 and SWIR 2), and NDVI (Rouse et al., 1974). For each data type, all acquisitions were stacked into multitemporal raster stacks, with each pixel representing a vector of reflectance values chronologically ordered through time. To determine the harmonic regression coefficients for the NDVI and SWIR data, a single harmonic curve (found to be superior to a two harmonic curve in preliminary analyses) was fit to each pixel in the multitemporal stack to derive the constant, sine, and cosine coefficients that describe the temporal behavior for that pixel (Brooks et al., 2012, Fig. 1, Eq. (1)).

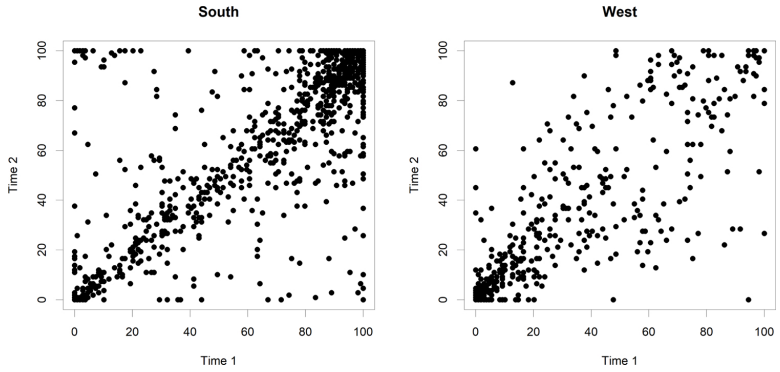

Fig. 5. Scatter plots comparing Time 1 and Time 2 photo interpreted TCC observations for the South and West.

Table 4
Parameters used in the EWMACD algorithm.

EWMACD.v.1.8.5 parameters	Inputs
trainingPeriod	'static'
numberHarmonicsSine	1
numberHarmonicsCosine	1
xBarLimit1	1.5
xBarLimit2	20
lambda	0.3
lambdaSigs	3
rounding	TRUE
persistence	10
reverseOrder	FALSE
trainingFitMinimumQuality	0.9
parallelFramework	'snow'
summaryMethod	'mean'
outputType	'coefficients only'
trainingPeriod	'static'
lowthresh	0
minTrainingLength	15
maxTrainingLength	30

Table 5
Parameters used in the randomForest algorithm.

randomForest parameters	Inputs
ntree	500
corr.bias	True
replace	True
mtry	4
nodesize	1

Table 6
Additional R libraries used in computing.

Additional libraries	Usage
raster (Hijmans, 2018)	Raster analysis
spatial.tools (Greenberg, 2018)	Spatial data analysis
foreach (Microsoft Corporation and Weston, 2017b)	Advanced for-loops
doSNOW (Microsoft Corporation and Weston, 2017a)	Parallelization with foreach

2.5. Predictive modeling with random forests

Given that the predictor variables are derived from Landsat reflectances and indices, there is multicollinearity in the predictor variables (not shown). This is somewhat mitigated among the harmonic regression coefficients, but there is still strong correlation between coefficients with the same function (i.e., the cosines for SWIR 1 and SWIR 2 are highly correlated, as are the sines). This multicollinearity causes few, if any, deleterious effects on random forest regression prediction.

To address the study objectives, four groups of variables (Table 7) were incorporated into random forest models (Breiman, 2001) using R's randomForest package (version 3.4.2; Liaw and Wiener, 2002), using the parameters listed in Table 5. Additionally, R packages listed in Table 6 were used for computing. The first two models included only variables derived from the NDVI and SWIR bands, to allow for a direct comparison between the median composite and harmonic regression approaches. The second study objective was addressed by the third and fourth models. The third model includes the same variables that were used to develop the 2011 NLCD Tree Canopy Cover product (Coulston et al., 2012; Ruefenacht, 2016). Finally, a hybrid of the median composite and harmonic regression approaches is investigated. Separate sets of models were trained to predict tree canopy cover for two time periods in 2010 and 2013 (using predictor variables and response data from appropriate date ranges) and the two study areas (South and West), discussed above.

2.5.1. Iterative modeling framework

Each model was run in R (R Core Team, 2018) with an iterative bootstrapping approach to include all available in-bag training data while also using an out-of-bag sample for model evaluation. Values from each of the input variables were extracted to plot locations and saved to a master table. For in-bag training samples, plots were randomly selected from this table with replacement, allowing for repeat plots in each bootstrapped sample. In-bag training plots were then used to estimate a random forest model based on the extracted values at each sample location. This process was repeated up to 500 times generating 500 forests of 500 decision trees. The same in-bag and out-of-bag plots were used for corresponding iterations of the four models that used different groups of predictor variables (i.e., Full model, Median Composite and Terrain, Harmonic Coefficient, and Median Composite models).

The number of points in the in-bag ample was equal to the number of extracted plots, and because of the sampling replacement, approximately 37% (Coulston et al., 2016) of each bootstrapped sample was left out. These out-of-bag samples were not used in model fitting but were used for model testing, comparing predicted and observed values for each of the 500 models (using their associated out-of-bag samples), and aggregating metrics across all models.

2.6. Model comparison

We compared models using several different measures. Variable importance for the full suite of model variables was used to assess the relative importance of included predictors, while mean correlation and error for the 500 model runs was compared for the four model types. Inherent dimensionality of predictor variables was used to determine the relative amount of information available in each model type. Models were mapped spatially in the southern study area to ensure that output maps were visually consistent. Finally, prediction variance was obtained by land cover type by calculating the mean variance of the 500 model predictions at each sample point and then obtaining the mean of those variances by land cover class.

2.6.1. Variable importance and correlations

Model strength was compared on the basis of mean RMSE (Eq. (2)) and mean pseudo R^2 (Eq. (3)) for the bootstrapped models, calculated for each model using out-of-bag samples as follows:

$$\text{RMSE} = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{x}_i - x_i)^2}{n}} \tag{2}$$

$$\text{Pseudo } R^2 = 1 - \frac{\Sigma_{i=1}^{n}(x_i - \hat{x}_i)^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

where x_i is the ith observation from the out-of-bag population with n total observations, $\hat{x}_i$ is the ith prediction, and $\bar{x}$ is the mean of the out-of-bag observations.

Both metrics were calculated by comparing observed photo interpretations to predicted values obtained for each random forest model using the 'predict' function in R's randomForest package (version 3.4.2; Liaw and Wiener, 2002).

High levels of correlation among predictor variables can reduce those variables' relative importance. Thus correlation matrices were generated and assessed in order to better inform variable importance analysis. Correlation matrices and variable importance measures were aggregated for each of the model types to determine the relative contribution of different groups of variables in the model. Once correlation was considered, mean decrease in node impurity was obtained for all variables in each combination of model type, time period, and region. This metric was obtained using the R randomForest package (Liaw and

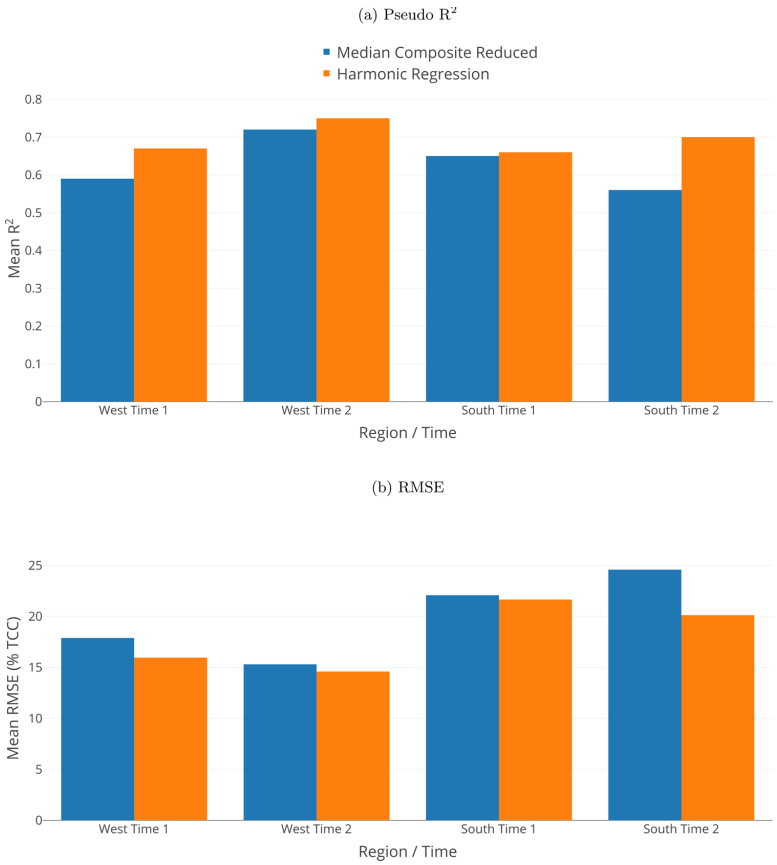

Fig. 6. Comparison of models using harmonic regression coefficients for NDVI, SWIR1 and SWIR2 to the median composite values for those same three bands.

Wiener, 2002). Mean decrease in node impurity identifies the amount that node variance (an indication of impurity) decreases at each split and averages that decrease by predictor variable.

The ten variables whose removal would contribute to the largest decrease in node impurity were ranked, with the largest decrease given a point value of 10 (for most important), and the 10th largest decrease given a 1. All other variables were given a ranking of 0. For each of the four model classes, these rankings were added for all 500 model iterations. Thus, a variable that was considered to be most important for all 500 iterations would receive a value of 5000, while variables that were not considered to be in the top ten most important for any iterations, would be given a point value of 0. This method reflects a similar rank score technique used in Schroeder et al. (2017). Due to the presence of correlated variables in our suite of predictor variables, these rank interpretations were evaluated primarily by comparing the relative contribution of relevant groups of predictor variables instead of the individual variables themselves.

2.6.2. Inherent dimensionality

The inherent dimensionality of each of the four predictor variable sets was assessed using principal components analysis (PCA). PCA utilizes a linear transformation which rotates the input images so that the original axes are orthogonal to one another with decreasing variance for each successive component (Eklundh and Singh, 1993). This methodology is commonly used for many remote sensing applications, including data reduction and feature extraction (Ren et al., 2014; Eklundh and Singh, 1993). PCA assumes (1) a high signal-to-noise ratio, and, (2) for some analytical applications, that input variables have a Gaussian distribution. Landsat data, the only satellite imagery used in this study, has a high signal-to-noise ratio (Barsi and Markham, 2013), and use of PCA for descriptive, exploratory applications does not require strict adherence to distribution assumptions (Jolliffe et al., 2016).

In PCA, eigen values resulting from the transformation can be used to determine the proportion of total variance explained in each resulting component. Typically a large portion of the variance from the original dataset can be found in the first principal component, with decreasing contributions in each successive component. The final components usually have relatively little variance. Therefore, using only the sequential number of bands which collectively represent a large proportion of the original data (usually 95–99%), data can be reduced. Inherent dimensionality is defined in this study as the number of principal components necessary to explain 99% of the cumulative variance. This method is used by Schlamm et al. (2008), as well as Lungu et al. (2017)). We utilize the inherent dimensionality to quantify and compare the variance explained by different groups of predictor variables as a proxy for the amount of unique information they provide.

2.6.3. Mapping TCC across the landscape

For each of the four model ensembles (per date/region), maps were made using the model whose pseudo R^2 was the closest to the mean

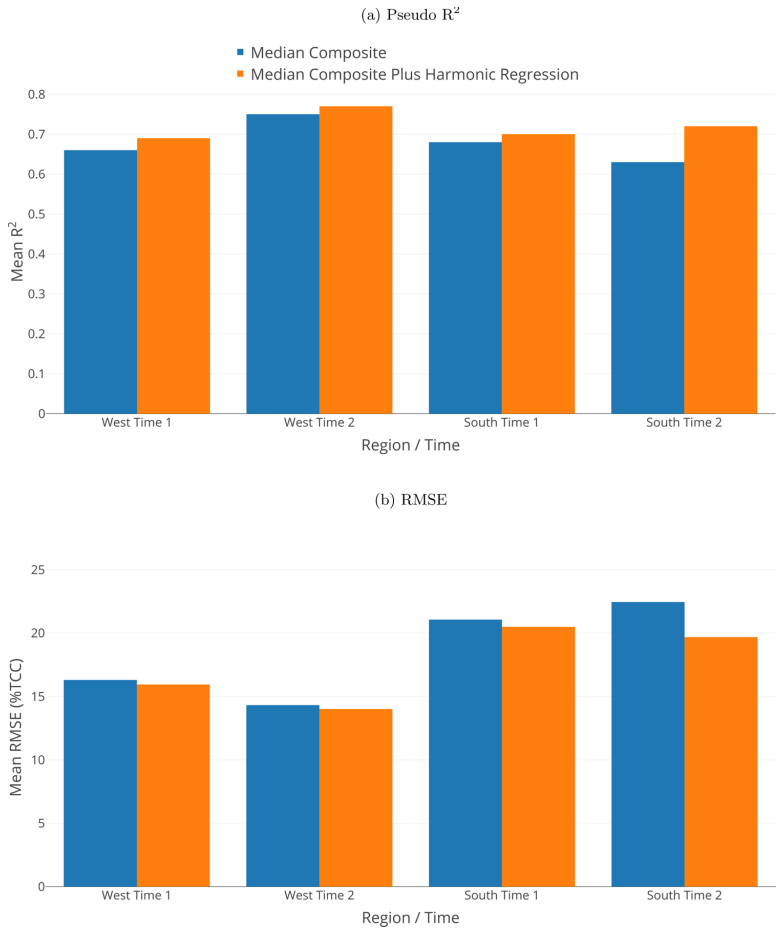

Fig. 7. Comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using the full suite of median composite and terrain variables used in the 2011 TCC product *plus* the harmonic regression coefficients from NDVI, SWIR1, and SWIR2.

pseudo R^2 from 500 iterations. We did not use the model with the highest pseudo R^2 to avoid selecting an over-fit model. This model was applied on a pixel-by-pixel basis using a stack of all model inputs. We then conducted both qualitative visual inspections and quantitative assessments. Simple image differencing was used to highlight spatial differences in models. These difference maps were reclassified to identify areas with greater than |30%| difference in tree canopy cover to identify patterns or land features where the models disagreed.

2.6.4. Variance of predictions across land cover types and TCC values

For each plot location used in the training dataset the 500 out-of-bag predictions were used to calculate the plot variance in predicted TCC. Plot variance for predictions can give a sense for how consistently a value of TCC is predicted at a location. Land cover classes were obtained from NLCD for each plot location from the training dataset. Plot variance metrics were averaged for plots in each NLCD land class. These values were plotted in a bar chart to compare the variance associated with each land cover class and to observe patterns relating these metrics to land cover.

3. Results

Mean pseudo R^2 was stable, across all four models, at 200 iterations and beyond. As such, the mean pseudo R^2 and mean RMSE at 500 iterations was compared among the resulting 500 random forest models. Using the same input bands (NDVI, SWIR1, and SWIR2; objective 1), harmonic regression coefficients slightly outperformed the median composite values (Fig. 6). Minimum and maximum pseudo R^2 values for the 500 random forest models (not shown) range from a low of 0.55 to a high of 0.82 for the harmonic regression model and from a low of 0.44 to a high of 0.80 for the median composite model. A similar slight improvement with harmonic regression was found for RMSE, with a reduced mean RMSE in all instances (Fig. 6).

To address our second research question (comparing the models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using all those variables *plus* the harmonic regression coefficients from NDVI, SWIR1, and SWIR2) we again compared the mean pseudo R^2 and mean RMSE at 500 iterations. The full (hybrid) model slightly outperforms the other models in all instances (Fig. 7). The full (hybrid) model also has the

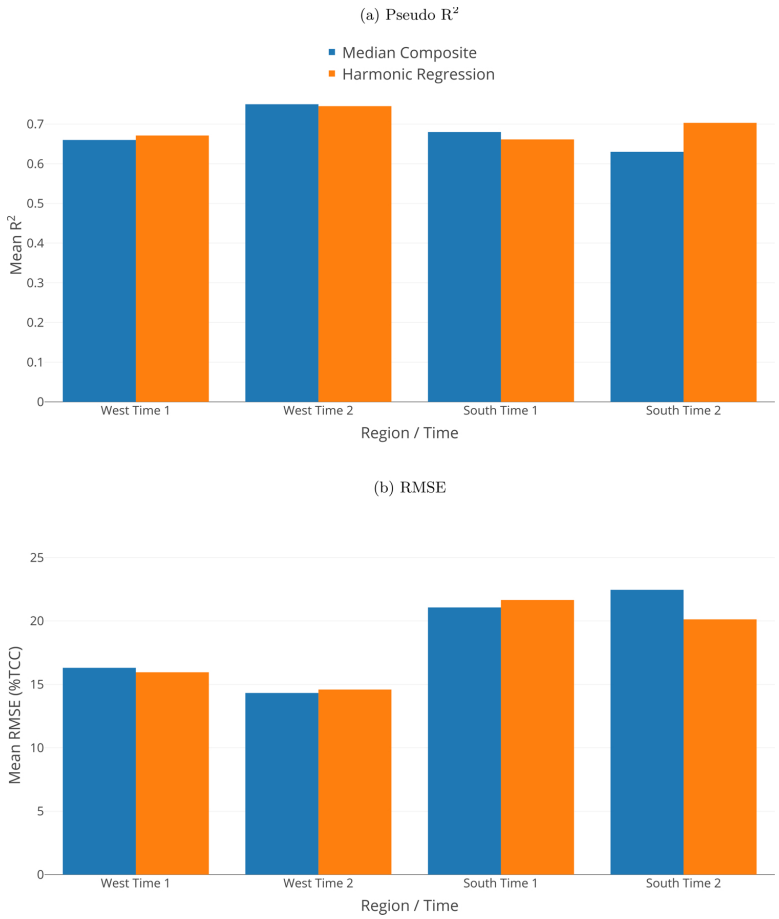

Fig. 8. Comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using the harmonic regression coefficients for NDVI, SWIR1, and SWIR2.

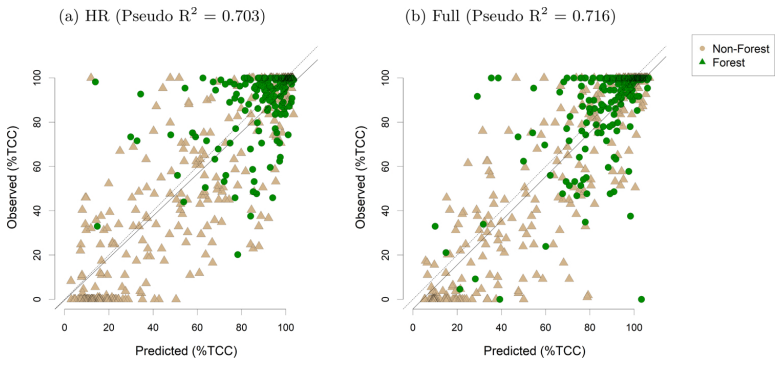

Fig. 9. Observed vs. predicted TCC for mean (with respect to pseudo R^2 out of 500) Time 2 models for the southern study area for the two models using harmonic regression coefficients as part of the predictor variable suite, colored by 2011 NLCD land cover at sample locations. Brown triangles are non-forest; green circles are forest. The dotted line represents the 1:1 relationship between x and y axes ($y = x$) and the solid line represents the linear relationship between observed and predicted TCC. The equation for this linear relationship is $y = 0.9489 * x - 0.2913$ for the HR model and $y = 0.996 * x - 4.301$ for the full model. Variable to axis assignment follows current best practice guidelines (Piñeiro et al., 2008). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

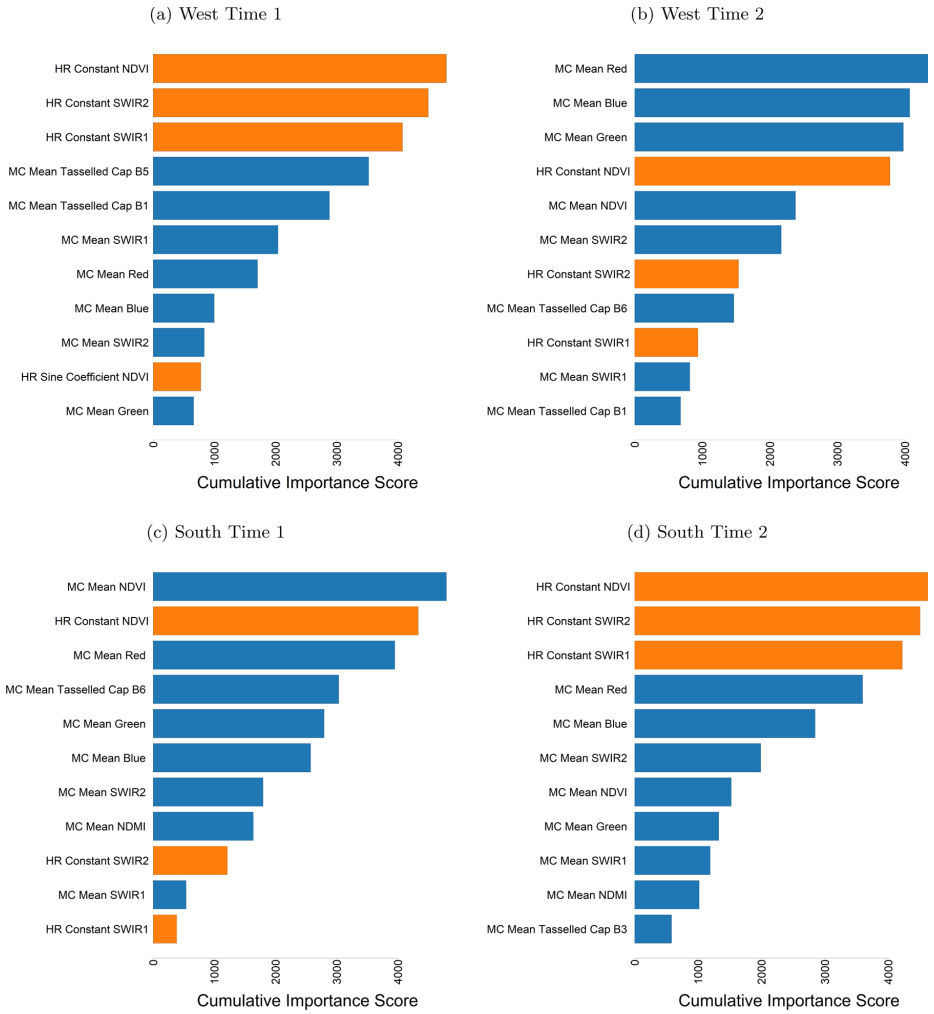

Fig. 10. Top 11 variable importances from full random forests model for each region/time. Median composite means are shown in blue, harmonic regression coefficients in orange. Variable importances were calculated as follows: (1) For each iteration, rank the ten variables whose removal would contribute to the largest increase in node impurity, largest = 10 and 10th (smallest) = 1. (2) Assign all other variables zero. (3) Repeat for 500 iterations (independent model runs), sum to obtain cumulative importance score for each variable. A variable that was considered to be most important for all 500 iterations would receive a value of 5000, while variables that were not in the top ten most important for any iterations, would be given a point value of 0. Abbreviations: HR, harmonic regression; MC, median composite; NDVI, normalized difference vegetation index; SWIR, shortwave infrared. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

highest min/max pseudo R^2, in all instances (not shown), ranging from a low of 0.58 to a high of 0.83.

A comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product (38 variables) to the harmonic regression coefficients from NDVI, SWIR1, and SWIR2 (9 variables) revealed mixed model performance (Fig. 8). The 38 median composite and terrain variables outperformed the 9 harmonic regression coefficients in two regions/times (West Time 2 and South Time 1), with the harmonic coefficients leading to an improved model in the other two (West Time 1 and South Time 2).

Observed vs. predicted scatterplots show similar trends. An example for the HR NDVI and SWIR South Time 2 model with the maximum R^2 (0.78), and differentiated by forest/nonforest (according to NLCD land cover class), is shown as Fig. 9. There is more unexplained variance in the south, which is consistent across all models (not shown). It is also evident that the observed data contains more 0% and 100% values than are predicted by the model. In the observed data in the south, most of the forested land cover has high tree canopy cover, while the lower tree

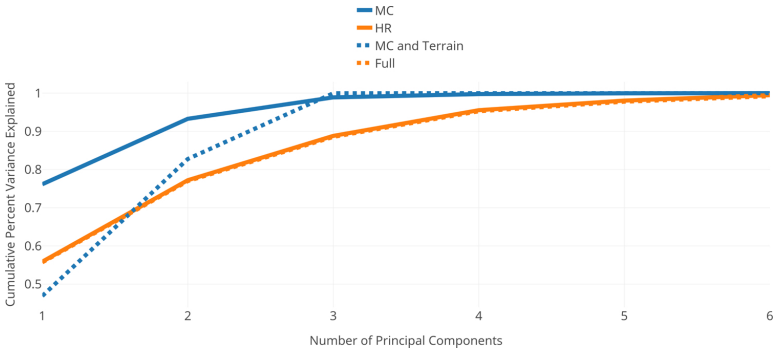

Fig. 11. Eigen analysis of predictor variable sets (South Time 2 shown). The cumulative percent variance explained for a given principal component is calculated as the sum of the eigen values up to and including that principal component divided by the sum of all eigen values for that predictor variable set. Only six principal components are shown for clarity, but the number of principal components for a given predictor variable set is the same as the number of predictor variables in that set. Inherent dimensionality as used here is the number of principal components it takes to explain nearly all the variance, defined as 99%. Note the following: (1) harmonic regression coefficients, alone, or in combination with other predictor variables, have higher inherent dimensionality (and thus information content), and (2) the quantity of available information provided by the harmonic regression coefficients is not substantially increased when looking at the full suite of predictor variables (47 bands), though this is not an indication of the utility of the information provided in predicting TCC.

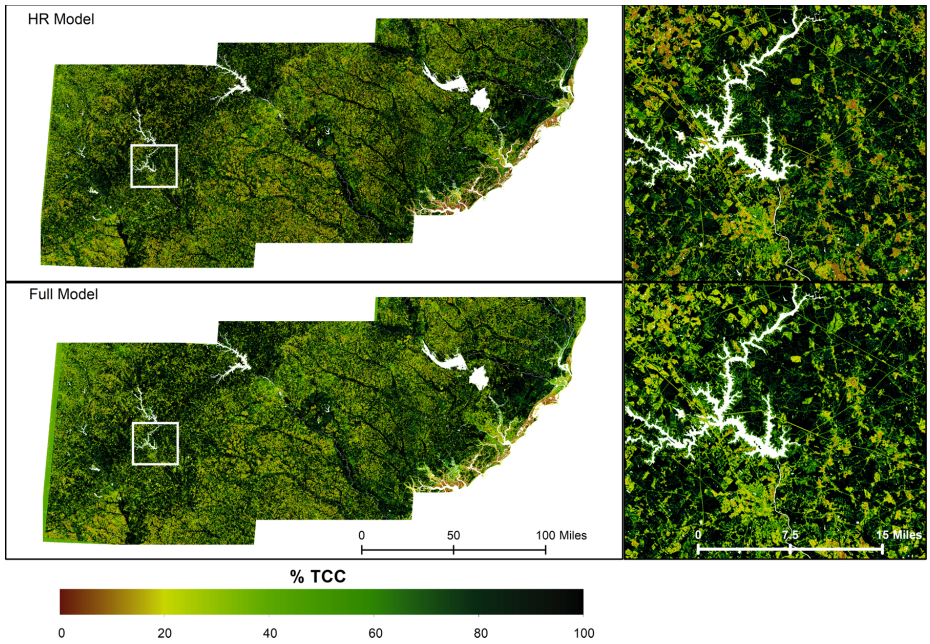

Fig. 12. TCC maps (Time 2) for the southern study area for the two models (*HR* and *Full*) using harmonic regression coefficients as part of the predictor variable suite.

canopy cover values are dominated by nonforest, which includes pasture, grassland, scrub/shrub, etc. There is a more uniform distribution of canopy cover for the samples in the west (excluding zeroes; Fig. 4), which at least partially explains the stronger model performance there.

3.1. Variable importance and correlations

Variable importances for the full (hybrid) model are shown in (Fig. 10). The harmonic regression NDVI constant is always one of the top four most important predictors. In all cases, all three harmonic regression constants and all three median composite mean visible bands are within the top 11 most important variables. The correlations between our top predictor variables and tree canopy cover are shown for the top three variables from each of the models (Table 8). Note that the three predictor variables most correlated with TCC are the harmonic regression constants, and that NDVI is the most important of the bands/ratios. Table 9 shows the correlations between the two NDVI predictor variables and tree canopy cover for each region/time. Note that the harmonic regression constant is always as well or better correlated with tree canopy cover than the median composite value.

3.2. Inherent dimensionality

The principal components analysis of our different sets of input parameters revealed that the models that include harmonic coefficient variables require more principal components to reach 99% cumulative variance than models without these variables (Fig. 11). The full suite of variables (*Full*) had the greatest inherent dimensionality, followed by the harmonic coefficients alone (*HR*), then the median composite and terrain variables (*MC and Terrain*), and finally the median composite

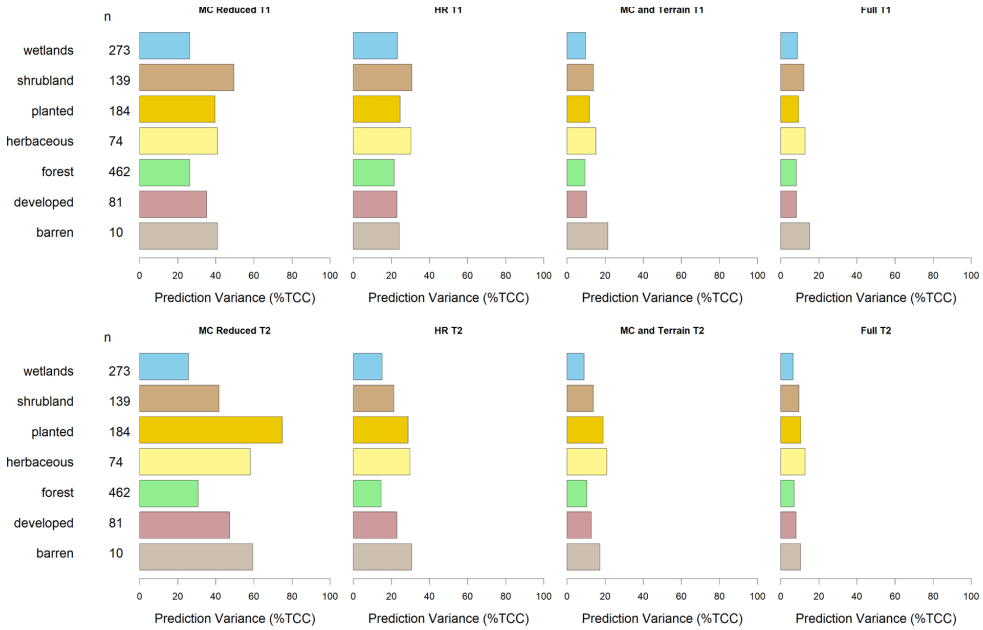

Fig. 13. Mean variance by NLCD land cover class (water excluded) for South, all models.

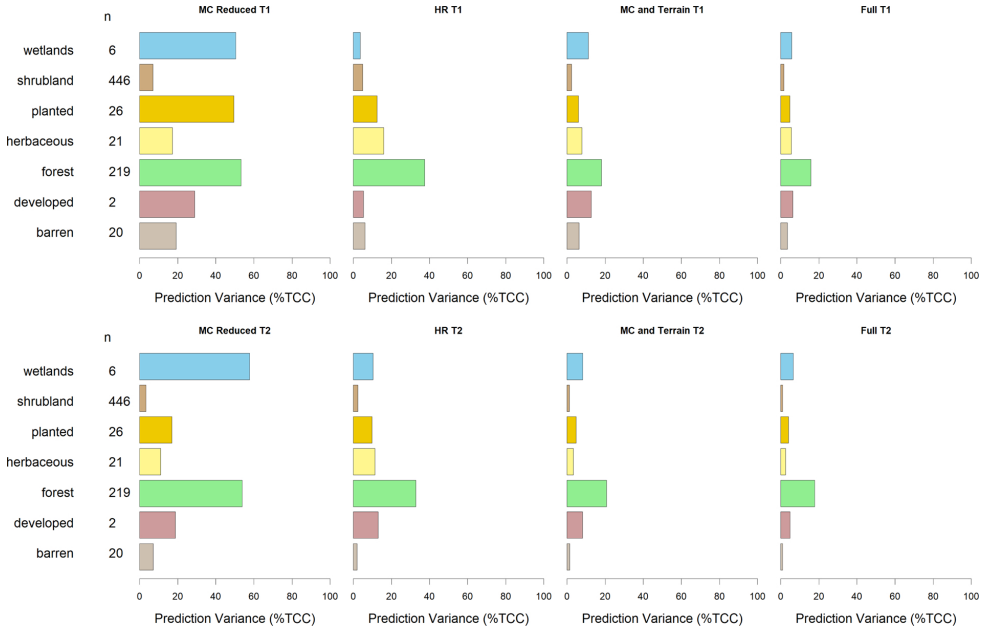

Fig. 14. Mean variance by NLCD land cover class (water excluded) for West, all models.

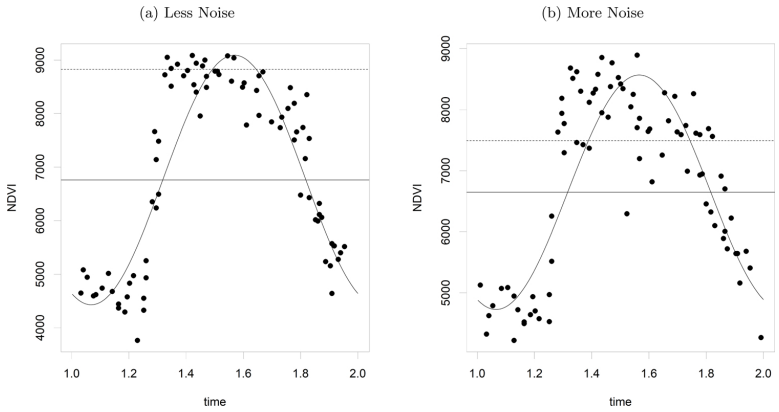

Fig. 15. Observations from two similar South Time 1 plots with fitted harmonic regression line (solid harmonic fit), harmonic regression constant (solid horizontal line), and median composite value (dotted horizontal line) shown. Note that the median composite value is properly located near the peak of the growing season maximum when little noise is present (a, left), but is lower with noisy observations (b, right).

Table 7
Model-specific predictor variables. Abbreviations: MC, median composite; HR, harmonic regression; NDVI, normalized difference vegetation index; NIR, near infrared; SWIR, shortwave infrared; NDMI, normalized difference moisture index; TC, tasseled cap; DEM, digital elevation model.

Model	Number of predictor variables	Variable class	Variables
MC reduced	6	MC[a]	NDVI
			SWIR1
			SWIR2
HR	9	HR[b]	NDVI
			SWIR1
			SWIR2
MC and terrain	38	MC[a]	Blue
			Green
			Red
			NIR
			SWIR1
			SWIR2
			NDVI
			NDMI
			TC (all)
		Terrain[a]	DEM
			Slope
			Aspect
			sin, cos aspect
Full	47	MC[a]	Blue
			Green
			Red
			NIR
			SWIR1
			SWIR2
			NDVI
			NDMI
			TC (all)
		Terrain[a]	DEM
			slope
			aspect
			sin, cos aspect
		HR[b]	NDVI
			SWIR1
			SWIR2

[a] *Focal $\bar{x}$ and *std* in 3 × 3 window.
[b] a_0, a_1, b_1 (Eq. (1)), the constant, cosine amplitude, and sine amplitude of the Fourier fit.

NDVI and SWIR variables (*MC Reduced*). The fact that the harmonic regression coefficients for three bands alone convey more information than the 38 variables in the *MC and Terrain* model is striking. However, increased information in a set of predictor variables does not necessarily mean that the target variable will be better estimated.

3.3. Mapping TCC across the landscape

Maps generated from the harmonic regression and full models are generally consistent across the landscape (Fig. 12). Careful examination of these images revealed some anomalies due to cloud cover or striping from (SLC-off) Landsat 7. This causes a reduction in the number of input observations that inform the original harmonic regression fit. Areas with very few observations due to annually or seasonally persistent cloud cover could result in similar anomalies, though these were not observed.

3.4. Variance of predictions across land cover types and TCC values

Tree canopy cover models show some variability in plot-level variance when aggregated by land cover class (Figs. 13 and 14). Mean variance in predicted value by land cover reveals the greatest inconsistency in prediction across land cover classes in the *MC Reduced* model with the most consistent predictions overall in the *Full* model. In the South, variance is lower in the forest and wetland categories. Higher variance can be observed in the other land cover types, without consistent pattern across model types and time periods. In the West, there is higher variance in forest plots, and in some cases (*MC Reduced* T1, *MC and Terrain* T1, *MC Reduced* T2) in the wetland plots as well. There is also lower variance in shrubland plots, and in some cases barren plots. Across land cover classes, models using harmonic regression coefficients outperform those in which the median composite values are used for the SWIR bands and NDVI (*HR* vs. *MC Reduced*). Adding other bands/indices along with the terrain variables resulted in additional decreases in variance across land cover classes (*MC and Terrain*, *Full*).

4. Discussion

All predictor sets performed reasonably well for a given region. However, in the direct comparison (objective 1; *MC Reduced* vs. *HR*) using the same index (NDVI) and wavelength bands (SWIR1 and SWIR2), the models using harmonic regression coefficients consistently outperformed those using median composite means and standard deviations with respect to both R^2 and RMSE across study areas and time periods (Fig. 7). This is partially explained by the consistently high variable importance metrics from the harmonic regression coefficients, particularly the constants, despite the high correlation among predictors. The improved utility of harmonic regression coefficients with respect to median composite values is also supported by the higher

Table 8
Correlation matrix of seven-most important predictor variables and tree canopy cover from the South, time 2 model. The three most important variables from the full random forests model for each region/time (Fig. 10) are shown; there are only seven since the same predictor variables could be among the three most important across regions/times. Abbreviations: TCC, tree canopy cover; MC, median composite; B, blue; G, green; R, red; a_0, harmonic regression constant; NDVI, normalized difference vegetation index; SWIR, shortwave infrared.

	MC B	MC G	MC R	MC NDVI	a_0 NDVI	a_0 SWIR1	a_0 SWIR2
TCC	−0.64	−0.64	−0.66	0.61	0.74	−0.68	−0.70
MC B		0.96	0.96	−0.69	−0.74	0.67	0.76
MC G			0.96	−0.57	−0.71	0.73	0.79
MC R				−0.68	−0.74	0.73	0.80
MC NDVI					0.78	−0.32	−0.44
a_0 NDVI						−0.66	−0.75
a_0 SWIR1							0.96

Table 9
Correlations between TCC in each region/time and the two NDVI predictor variables. Abbreviations: TCC, tree canopy cover; MC, median composite; a_0, harmonic regression constant; NDVI, normalized difference vegetation index.

	MC NDVI	a_0 NDVI
West Time 1	0.77	0.82
West Time 2	0.83	0.85
South Time 1	0.74	0.74
South Time 2	0.61	0.74

inherent dimensionality of the harmonic regression predictors (Fig. 11) and the high correlations of the harmonic regression constants with TCC (Table 8; note that the constants have the highest correlation coefficients of the seven most important predictor variables).

While there are multiple possible explanations for why the harmonic regression constant outperforms the median composite value for a given band, the two most likely are (1) the decreased sensitivity of the harmonic regression constant to systematic noise (Moody and Johnson, 2001; De Beurs and Henebry, 2010) and (2) the use of another sensor (MODIS, via GLOVIS) to define the peak NDVI interval within which the median value is computed. Noise tends to lower the median composite value (Fig. 15), leading to inconsistencies in the median composite value for the same TCC. With respect to the second possible explanation, the increased temporal resolution of MODIS makes it a sensible choice for choosing the interval within which the median value is computed, but MODIS-derived NDVI, while similar to NDVI computed from the Landsat sensors, is slightly different (Steven et al., 2003).

The coefficients derived from harmonic regression of all-available cloud-free moderate resolution earth resource satellite data have been used in a variety of earth science applications, including image classification (Liang, 2001), stratified estimation of dynamic forest biophysical parameters (Brooks et al., 2016), estimation of static forest biophysical parameters besides TCC (Wilson et al., 2018), and, with this study, TCC. However, despite these successes using harmonic regression, there are alternative approaches to curve fitting that have been developed or tested using the same or similar data sets that might prove to be superior to harmonic regression for this and related applications. These include the double hyperbolic tangent model (Vrieling et al., 2018) applied to Sentinel 2 observations and a Bayesian hierarchical modeling framework (Senf et al., 2017) applied to Landsat observations, among others. While Atkinson et al. (2012) found harmonic regression the best of four tested approaches (Fourier analysis, asymmetric Gaussian model, double logistic model, and the Whittaker filter) for estimating phenological parameters reliably using Medium Resolution Imaging Spectrometer (MERIS) composite data, their study did not use moderate resolution earth resource satellite data nor did it test the more modern approaches just noted. As such, use of alternative curve fitting approaches as a precursor to TCC estimation using moderate resolution earth resource satellite data is a potentially fruitful area of further inquiry.

Mapping yielded visually consistent results across models. However, due to the inclusion of Landsat 7 SLC-off images in our time-series, subtle striping can be observed in models with harmonic regression variables in regions with fewer unmasked pixel observations. This finding is consistent with other studies (Wilson et al., 2018), and may be a reason to exclude Landsat 7 observations when feasible.

Mean variance in predicted values yielded inconsistent results across regions (Figs. 13 and 14). In the South, the two land cover types with more samples (wetlands and forest areas) showed lower variance, as expected since there was more training data available for those types. In the West, shrublands were the most prevalent land cover type and had consistently low variance values as well. However, contrary to these other examples, the land cover type with most consistently high variance, at least in the West, was forest, which is the second most prevalent land cover type in terms of sample points. There are several possible reasons for the high mean prediction variance in NLCD forest in the West, including the following: (1) It is common to confuse shrublands (which, if there are no trees present, should have a TCC value of 0) with forest areas, both spectrally and during photo interpretation. This could lead to spectrally similar training data having drastically different observed TCC values, thus limiting the predictive ability of the model. (2) The majority of the forest types in the west are coniferous which may not provide as much contrast in reflectance between tree and non-tree pixels and may not show as much variation over time resulting in more influence from noise. In all cases (land cover, region, and time), models with large numbers of predictor variables (*MC and Terrain*, 38; *Full*, 47; Table 7) reduced variance. Additionally, this observation may also indicate that the *MC and Terrain* variables, while possessing less information overall, may have a greater explanatory power than the harmonic regression coefficients in predicting TCC.

Given recent advances in high performance computing platforms that host the Landsat archives (i.e., Google Earth Engine, Gorelick et al., 2017; NASA Earth Exchange, Nemani, 2011), the selection of an algorithm that reduces input data and pre-processing streams (i.e., image downloads, cloud masking, and development of predictor variables) is not a major factor of consideration. Instead, the focus can be more on striving for analysis techniques that result in more parsimonious predictive models, with fewer variables but higher explanatory power. This concept represents the biggest advantage of harmonic regression over the median composite approach. The inherent dimensionality shown in Fig. 11 highlights the increased information contained in the models that contain harmonic regression predictors. This makes sense, given the large number of input bands/observation dates over which the Fourier curve is fit (ranging from 60–192 for the 3-year training period). This is illustrated with a simple correlation between TCC and the important predictor variables (Table 8), which indicates that the harmonic regression constants are more strongly correlated to TCC than any median composite mean or standard deviation. Moreover, though not an objective of this study, a comparison of Figs. 6 and 7 reveals

instances where *HR* (9 predictor variables) is equivalent to or outperforms *MC and Terrain* (38 predictor variables).

The variable importance results (Fig. 10 and Table 8) highlight the importance of the visible bands to the prediction of TCC, although the correlation matrix (Table 8) confirms that these bands are highly correlated. Our analysis did not include a visible band in the harmonic regression predictor variables, so direct comparison is not possible. However, the results suggest that harmonic regression coefficients from a visible band (or panchromatic band as shown in Fig. 2) could be valuable to TCC prediction. While this study found the median composite values derived from the tasseled cap bands to be less important than those derived using the visible bands, harmonic regression coefficients derived from tasseled cap Landsat time series stacks (especially, and atypically, TC band 6) could also be useful.

Finally, we would like to reiterate that there are instances where striping from Landsat 7 SLC-off images has produced visually inconsistent predictions from harmonic regression coefficients. We have seen that these inconsistencies relate to differences in the number of observations available to the harmonic regression algorithm. Even absent striping, cloud-prone areas or areas with cloud cover over a consistent portion of the season may not enable sufficient observational representation of annual phenology, thus reducing the value of harmonic regression coefficients fit to a poorly articulated curve. Median composite methodologies could also be susceptible to these issues, especially if observations are lost during the peak of the growing season from striping or clouds.

5. Conclusions

We conclude that harmonic regression coefficients are better predictors of TCC than median composite variables for the same vegetation index and wavelength bands. Correlations support this finding, with improved R^2 and RMSE values for the harmonic regression coefficient models when compared to the median composite reduced models, as well as increased inherent dimensionality in harmonic regression coefficient variables. Further, plot-level prediction variance is lower for the harmonic regression coefficient models than it is for the median composite reduced models for almost every land cover type, regardless of region or time period.

We further conclude that harmonic regression coefficients can be added to a full suite of median composite and terrain variable to improve the quality of TCC models. Variable importance analysis indicates that harmonic regression variables are uniquely instrumental to model success, with the harmonic regression constants appearing consistently in the top 11 predictor variables. Correlations further support this finding, with improved R^2 and RMSE values for the full model when compared to the median composite and terrain model, and overall. Finally, plot-level prediction variance is lowest for the full models than any other model for almost every land cover type, regardless of region or time period. Overall, we have found that harmonic regression coefficients can be used to estimate TCC, both on their own, and when combined with a suite of median composite and terrain variables. The 9 harmonic coefficient variables alone successfully predict TCC without sacrificing predictive power when compared to models drawing from greater numbers of variables (*MC and Terrain*, $n = 38$; *Full*, $n = 47$; Table 7).

Acknowledgements

This study was funded by the 'TCC 2021 Research and Development' grant, the 'NLCD Percent Tree Canopy Cover Layer - Evaluation of Data Sources to Predict Canopy Cover', and the 'NLCD Percent Tree Canopy Cover Layer - Evaluation of the 2011 Product and Alternative Approaches to Detect Change' grant from the USDA Forest Service. It was also supported in part by the Virginia Agricultural Experiment Station and the McIntire-Stennis Program of NIFA, USDA, "Detecting and Forecasting the Consequences of Subtle and Gross Disturbance on Forest Carbon Cycling").

References

Asner, G.P., Lobell, D.B., 1996. A biogeophysical approach for automated SWIR unmixing of soils and vegetation. Remote Sens. Environ. 74, 99–112.
Atkinson, P.M., Jeganathan, C., Dash, J., Atzberger, C., 2012. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. Remote Sens. Environ. 123, 400–417.
Barsi, J.A., Markham, B.L., 2013. Early radiometric performance assessment of the Landsat-8 Operational Land Imager (OLI). Proc. SPIE 8866:8866 – 8866 – 12.
Brandt, M., Hiernaux, P., Tagesson, T., Verger, A., Rasmussen, K., Diouf, A.A., Mbow, C., Mougin, E., Fensholt, R., 2016. Woody plant cover estimation in drylands from earth observation based seasonal metrics. Remote Sens. Environ. 172, 28–38.
Brooks, E.B., 2019. Exponentially Weighted Moving Average Change Detection – Script and Sample Data. Blacksburg: Virginia Tech Department of Forest Resources and Environmental Conservation. Virginia Tech Department of Forest Resources and Environmental Conservation, Blacksburg, doi:10.7294/W4WD3XHK.
Brooks, E.B., Coulston, J.W., Wynne, R.H., Thomas, V.A., 2016. Improving the precision of dynamic forest parameter estimates using Landsat. Remote Sens. Environ. 179, 162–169. https://doi.org/10.1016/j.rse.2016.03.017.
Brooks, E.B., Thomas, V.A., Wynne, R.H., Coulston, J.W., 2012. Fitting the multitemporal curve: a Fourier series approach to the missing data problem in remote sensing analysis. IEEE Trans. Geosci. Remote Sens. 50 (9), 3340–3353. https://doi.org/10.1109/TGRS.2012.2183137.
Coulston, J.W., Blinn, C.E., Thomas, V.A., Wynne, R.H., 2016. Approximating prediction uncertainty for random forest regression models. Photogramm. Eng. Remote Sens. 82 (3). https://doi.org/10.14358/PERS.82.3.189.
Coulston, J.W., Moisen, G.G., Wilson, B.T., Finco, M.V., Cohen, W.B., Brewer, C.K., 2012. Modeling percent tree canopy cover – a pilot study. Photogramm. Eng. Remote Sens. 78ID – 22 (7), 715–727. https://doi.org/10.14358/PERS.78.7.715.
De Beurs, K.M., Henebry, G.M., 2010. Spatio-temporal statistical methods for modelling land surface phenology. Phenological Research: Methods for Environmental and Climate Change Analysis. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3335-2.
DeFries, R.S., Field, C.B., Fung, I., Justice, C.O., Los, S., Matson, P.A., Matthews, E., Mooney, H.A., Potter, C.S., Prentice, K., Sellers, P.J., Townshend, J.R.G., Tucker, C.J., Ustin, S.L., Vitousek, P.M., 1995. Mapping the land surface for global atmosphere-biosphere models: toward continuous distributions of vegetation's functional properties. J. Geophys. Res. 100 (D10), 20. https://doi.org/10.1029/95JD01536.
Eklundh, L., Singh, A., 1993. A comparative analysis of standardised and unstandardised principal components analysis in remote sensing. Int. J. Remote Sens. 14 (7), 1359–1370. https://doi.org/10.1080/01431169308953962.
Gessner, U., Machwitz, M., Conrad, C., Dech, S., 2013. Estimating the fractional cover of growth forms and bare surface in savannas. A multi-resolution approach based on regression tree ensembles. Remote Sens. Environ. 129, 90–102. https://doi.org/10.1016/j.rse.2012.10.026.
Goeking, S.A., Liknes, G.C., Lindblom, E., Chase, J., Jacobs, D.M., Benton, R., 2012. A GIS-Based Tool for Estimating Tree Canopy Cover on Fixed-Radius Plots Using High-Resolution Aerial Imagery. In: Morin, Randall S., Liknes, Greg C. (Eds.), Moving from status to trends: Forest Inventory and Analysis (FIA) symposium 2012; 2012 December 4-6. U.S. Department of Agriculture, Forest Service, Northern Research Station, Baltimore, MD, pp. 237–241 Newtown Square, PA Gen. Tech. Rep. NRS-P-105.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.
J.A. Greenberg. spatial.tools: R Functions for Working with Spatial Data 2018; R package version 1.6.0 https://CRAN.R-project.org/package=spatial.tools.
Hansen, M.C., DeFries, R.S., 2004. Detecting long-term global forest change using continuous fields of tree-cover maps from 8-km Advanced Very High Resolution Radiometer (AVHRR) data for the years 1982–99. Ecosystems 7 (7), 695–716. https://doi.org/10.1007/s10021-004-0243-3.
R.J. Hijmans. Raster: Geographic Data Analysis and Modeling 2018; R package version 2.8-19 https://CRAN.R-project.org/package=raster.
Immerzeel, W., Quiroz, R., De Jong, S., 2005. Understanding precipitation patterns and land use interaction in Tibet using harmonic analysis of SPOT VGT-S10 NDVI time series. Int. J. Remote Sens. 26 (11), 2281–2296. https://doi.org/10.1080/01431160512331326611.
Jennings, S., 1999. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. Forestry 72 (1), 59–74. https://doi.org/10.1093/forestry/72.1.59.
Jolliffe, I.T., Cadima, J., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. 374, 16. https://doi.org/10.1098/rsta.2015.0202.
Jönsson, P., Eklundh, L., 2004. TIMESAT – a program for analyzing time-series of satellite sensor data. Comput. Geosci. 30 (8), 833–845. https://doi.org/10.1016/j.cageo.2004.05.006.
Karlson, M., Ostwald, M., Reese, H., Sanou, J., Tankoano, B., Mattsson, E., 2015. Mapping tree canopy cover and aboveground biomass in Sudano-Sahelian woodlands using landsat 8 and random forest. Remote Sens. 7 (8), 10017–10041. https://doi.org/10.3390/rs70810017.
Liang, S., 2001. Land-cover classification methods for multi-year AVHRR data. Int. J.

Remote Sens. 22 (8), 1479–1493. https://doi.org/10.1080/01431160120833.
Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22. https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf.
Lobell, D.B., Asner, G.P., Law, B.E., Treuhaft, R.N., 2001. Subpixel canopy cover estimation of coniferous forests in Oregon using SWIR imaging spectrometry. J. Geophys. Res. Atmos. 106 (D6), 5151–5160. https://doi.org/10.1029/2000JD900739.
Lungu, C., Ersali, S., Szefler, B., Pîrvan-Moldovan, A., Basak, S., Diudea, M.V., 2017. Dimensionality of big data sets explored by Cluj descriptors. Studia Universitatis Babes-Bolyai, Chemia 62 (3). https://doi.org/10.24193/subbchem.2017.3.16.
Microsoft Corporation, Weston, S., 2017a. doSNOW: Foreach Parallel Adaptor for the 'snow' Package. R package version 1.0.16.. https://CRAN.R-project.org/package=doSNOW.
Microsoft Corporation, Weston, S., 2017b. Foreach: Provides Foreach Looping Construct for R. R package version 1.4.4. https://CRAN.R-project.org/package=foreach.
Moody, A., Johnson, D.M., 2001. Land-surface phenologies from AVHRR using the discrete Fourier transform. Remote Sens. Environ. 75, 305–323. https://doi.org/10.1016/S0034-4257(00)00175-9.
Muoghalu, J.I., Isichei, A.O., 1991. Effect of tree canopy cover on the yield, crude protein and fibre content of forb species in Nigerian Guinea savanna. Vegetatio 95 (2), 167–175. https://doi.org/10.1007/BF00045215.
Nemani, R., 2011. NASA earth exchange: next generation earth science collaborative. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences – ISPRS Archives, Vol. XXXVIII-8/W20. Bhopal. ISPRS.
North, M.P., Kane, J.T., Kane, V.R., Asner, G.P., Berigan, W., Churchill, D.J., Conway, S., Gutiérrez, R.J., Jeronimo, S., Keane, J., Koltunov, A., Mark, T., Moskal, M., Munton, T., Peery, Z., Ramirez, C., Sollmann, R., White, A.M., Whitmore, S., 2017. Cover of tall trees best predicts California spotted owl habitat. Forest Ecol. Manag. 405 (August), 166–178. https://doi.org/10.1016/j.foreco.2017.09.019.
Owens, M.K., Lyons, R.K., Alejandro, C.L., 2006. Rainfall partitioning within semiarid juniper communities: effects of event size and canopy cover. Hydrol. Process. 20 (15), 3179–3189. https://doi.org/10.1002/hyp.6326.
Pandit, R., Polyakov, M., Sadler, R., 2014. Valuing public and private urban tree canopy cover. Aust. J. Agric. Resour. Econ. 58 (3), 453–470. https://doi.org/10.1111/1467-8489.12037.
Peterson, D.W., Reich, P.B., Wrage, K.J., 2007. Plant functional group responses to fire frequency and tree canopy cover gradients in oak savannas and woodlands. J. Veg. Sci. 18, 3–12. https://doi.org/10.1111/j.1654-1103.2007.tb02510.x.
Piñeiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M., 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed? Ecol. Model. 216 (3–4), 316–322. https://doi.org/10.1016/j.ecolmodel.2008.05.006. arXiv: arXiv:1011.1669v3.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ren, J., Zabalza, J., Marshall, S., Zheng, J., 2014. Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. IEEE Signal Process. Mag. 31 (4), 149–154. https://doi.org/10.1109/MSP.2014.2312071.
Rouse, J.W., J. Haas, R., Deering, D., Schell, J., Harlan, J., 1974. Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation. Technical Report. NASA/GSFC Type III Final Report, Greenbelt, MD, USA.
Ruefenacht, B., 2016. Comparison of Three Landsat TM Compositing Methods: A Case Study Using Modeled Tree Canopy Cover 82. Photogramm. Eng. Remote Sens., pp. 199–211. https://doi.org/10.14358/PERS.82.3.199.
Ruefenacht, B., Benton, R., Johnson, V., Biswas, T., Baker, C., Finco, M., Megown, K., Coulston, J., Winterberger, K., Riley, M., 2015. Forest Service Contributions to the National Land Cover Database (NLCD): Tree Canopy Cover Production. In: Stanton, S.M., Christensen, G.A. (Eds.), Pushing boundaries: new directions in inventory techniques and applications: Forest Inventory and Analysis (FIA) symposium 2015. 2015 December 8–10. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, Oregon, pp. 241–243 Gen. Tech. Rep. PNW-GTR-931.
Schlamm, A.A., Messinger, D.W., Basener, W.F., 2008. Geometric estimation of the inherent dimensionality of a single material cluster in multi- and hyperspectral imagery. Proc. SPIE, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIV, 69661G 69661G, 394–403. https://doi.org/10.1117/12.776903.
Schroeder, T.A., Schleeweis, K.G., Moisen, G.G., Toney, C., Cohen, W.B., Freeman, E.A., Yang, Z., Huang, C., 2017. Testing a Landsat-based approach for mapping disturbance causality in U.S. forests. Remote Sens. Environ. 195, 230–243. https://doi.org/10.1016/j.rse.2017.03.033.
Senf, C., Pflugmacher, D., Heurich, M., Krueger, T., 2017. A Bayesian hierarchical model for estimating spatial and temporal variation in vegetation phenology from Landsat time series. Remote Sens. Environ. 194, 155–160. https://doi.org/10.1016/j.rse.2017.03.020.
Shureman, P., 1940. Department of Commerce Coast and Geodetic Survey Manual of Harmonic Analysis, Special Publication No. 98. Technical Report. U.S. Department of Commerce Coast and Geodetic Survey, Washington.
Smith, S.W., 1997. Continuous signal processing. In: Smith, S.W. (Ed.), Digital Signal Processing: A Practical Guide for Engineers and Scientists. Elsevier, pp. 243–260. https://doi.org/10.1016/B978-0-7506-7444-7/50050-9.
Steven, M.D., Malthus, T.J., Baret, F., Xu, H., Chopping, M.J., 2003. Intercalibration of vegetation indices from different sensor systems. Remote Sens. Environ. 88 (4), 412–422. https://doi.org/10.1016/j.rse.2003.08.010.
Toney, C., Shaw, J.D., Nelson, M.D., 2009. A stem-map model for predicting tree canopy cover of forest inventory and analysis (FIA) plots. In: USDA Forest Service Proceedings. January 2009. pp. 1–19.
USDA Farm Service Agency, 2018. NAIP Coverage 2002–2018. US Department of Agriculture Farm Production and Conservation- Business Center, Salt Lake CIty, UT. https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/APFO/status-maps/pdfs/NAIP_Coverage_2018.pdf.
Vrieling, A., Meroni, M., Darvishzadeh, R., Skidmore, A.K., Wang, T., Zurita-Milla, R., Oosterbeek, K., O'Connor, B., Paganini, M., 2018. Vegetation phenology from Sentinel-2 and field cameras for a Dutch barrier island. Remote Sens. Environ. 215 (July 2017), 517–529. https://doi.org/10.1016/j.rse.2018.03.014.
Weisstein, E.W., 2010. Fourier Series. Wolfram Research, Inc. (accessed 4 October 2018). http://mathworld.wolfram.com/FourierSeries.html.
Wilson, B.T., Knight, J.F., McRoberts, R.E., 2018. Harmonic regression of Landsat time series for modeling attributes from national forest inventory data. ISPRS J. Photogramm. Remote Sens. 137, 29–46. https://doi.org/10.1016/j.isprsjprs.2018.01.006.
Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028. eprint:9808047v2 (arXiv:cond-mat).

Chapter 3

Consistency in expert and crowdsourced interpretations of aerial photographs to measure tree canopy cover

Consistency in expert and crowdsourced interpretations of aerial photographs to measure tree canopy cover

Jill M. Derwin[a], Valerie A. Thomas[a,*], Randolph H. Wynne[a], Karen G. Schleeweis[e], John W. Coulston[d], S. Seth Peery[b], Kurt Luther[c], Greg C. Liknes[f], Stacie Bender[g]

[a] *Virginia Tech, Forest Resources and Environmental Conservation, Blacksburg, Virginia 24061 USA*
[b] *Virginia Tech, Enterprise GIS, Blacksburg, Virginia 24060 USA*
[c] *Virginia Tech, Department of History and Department of Computer Science, Arlington, Virginia 22203 USA*
[d] *United States Forest Service, Southern Research Station, Blacksburg, Virginia 24061 USA*
[e] *United States Forest Service, Rocky Mountain Research Station, Ogden, Utah 84401 USA*
[f] *United States Forest Service, Northern Research Station, St Paul, Minnesota 55108 USA*
[g] *United States Forest Service, Geospatial Technology and Applications Center, Salt Lake City, Utah 84119 USA*

Abstract

This study presents a method for the collection of photointerpreted tree canopy cover values from crowdsourcing. Interpretations were collected by experts and crowd workers using an interface developed for Amazon's Mechanical Turk platform. We collected multiple interpretations at sample locations from both crowd and expert interpreters, and sampled these data in a Monte Carlo framework to estimate a classification model predicting the "reliability" of each crowd interpretation using expert interpretations as a benchmark. Using this model, we identified the most important variables in estimating the relationship between a location's characteristics and interpretation behaviors which affect consistency in interpretations between crowd and expert workers. Overall, we show low agreement between crowd and expert groups, as well as between individual experts. We show that the number of plots interpreted was the strongest indicator of the reliability of an individual's interpretations, further evidenced by apparent fatigue effects in crowd interpretations. The second most important variable related to the use of the false color display during interpretation followed by a variable related to the use of the natural color display during interpretation, reflecting the differences in interpretation methodologies used by crowd workers and experts and the impact display has on the interpretation of tree canopy cover. Finally, we discuss recommendations for further study and future implementations of crowdsourced photointerpretation.

*Corresponding author. Tel.: +1 540 231 0958.
Email addresses: jmd59@vt.edu (Jill M. Derwin), thomasv@vt.edu (Valerie A. Thomas), wynne@vt.edu (Randolph H. Wynne), karen.schleeweis@usda.gov (Karen G. Schleeweis), john.coulston@usda.gov (John W. Coulston), sspeery@vt.edu (S. Seth Peery), kluther@vt.edu (Kurt Luther), greg.liknes@usda.gov (Greg C. Liknes), stacie.bender@usda.gov (Stacie Bender)

1. Introduction

In the age of big data, algorithms are growing more sophisticated to accurately and precisely estimate relationships between variables and predict values where known data is unavailable. For environmental applications, these models are essential to monitor trends, understand the current state of our resources, and to more accurately forecast potential threats and opportunities for improved decision-making. However, as algorithms become more complex, they require the collection of more training data, a process which is usually time-consuming and costly. In this study, we test a combination of two methods of data collection which are known to improve upon traditional methods in terms of their ability to gather large volumes of data in less time: photointerpretation and crowdsourcing.

Photointerpretation is "the act of examining the photographic images of objects for the purpose of identifying those objects and deducing their significance" (Colwell et al., 1952, page 375) [1]. With the help of aerial photographs, photointerpretation has been widely used throughout the last century for military and civil engineering applications and in forestry and natural resources. Photointerpretation tasks can include classification, enumeration, measurement, and delineation completed using cues from images' tone, texture, shadow, pattern, association, shape, size, and site characteristics (like topography) [2]. Using photointerpretation, training data can be collected in less time at a lower cost than field data collection, as it involves indirect measurements of imagery rather than field observations which require more travel time and resources.

Similarly, crowdsourcing has risen to prominence as an avenue to reduce timing and cost of data collection by employing a large group of anonymous non-expert participants in data collection from around the world. In addition to many other science applications, crowdsourcing has been used to collect geographic information by way of classification, digitization, and conflation activities [3], and to improve and validate existing remote sensing models. A 2015 paper by Yu et al. demonstrates the power of crowdsourcing for the identification of cloud-covered pixels in Landsat imagery. This study showed that crowdsourcing is an option to improve cloud and cloud-shadow detection. They show that crowdsourcing performed better than automated models, particularly when using a majority voting rule across interpreters, but that crowd interpreters struggled to identify cirrus clouds. This study also showed that quality of interpretations was insensitive to wage, a finding also reflected in other studies [5], [6]. The investigators also found that accuracy deteriorates with time on task [4].

1.1. Background: Measurement of Tree Canopy Cover

Tree canopy cover (TCC), which is defined as the measure of the proportionate area covered by the vertical components of tree crowns [7], is an important measurement in forest management, as it can provide insight into forest health, disease, degradation, growth, and productivity. While it can be measured in the field with various methods, it can also be measured using photointerpretation of aerial imagery.

Tree canopy cover is measured in many ways. In the field, labor-intensive methods such as grid sampling using a Cajanus tube [8] can be used as well as line-intersect sampling [9], [10]. There are also more efficient techniques such as below-canopy digital photography or spherical densiometer measurements (which measure canopy closure, a related, but different metric) and ocular estimates [11]. These methods are considered to be the most robust measures of tree canopy cover, however field-based data collection is a time-consuming endeavor and harmonizing the collection of field observations with satellite imagery at a national scale could be a very costly and challenging endeavor.

Tree/non-tree classification maps using fine scale imagery have also be used to calculate tree cover to compare to coarser resolution models. Hansen et al. (2002) classified five 4 m resolution

IKONOS data and calculated the relative area of forest at the 30-m pixel level to obtain estimates of tree canopy cover. They used those data to train a Landsat-based model of tree canopy cover, which was then aggregated (using simple averaging) and used to validate exiting MODIS products at 250 and 500 m resolution in the Western Province area of Zambia. These products were validated at each level including a field validation of the 30 m product [12]. Ma et al. (2017) used National Agriculture Imagery Program (NAIP) (1 m resolution) and WorldView-2 (0.5 -2 m) imagery to create classifications and obtain relative tree canopy cover at a 30 m scale. They compared these estimates to ground-based measurements using a sighting tube (which were not found to be a suitable comparison to other methods) and LiDAR estimates of tree canopy cover, also aggregated to 30 m. The NAIP and WorldView-2 tree canopy cover models were shown to be comparable to LiDAR estimates, these models overestimated sparse forests and underestimated dense forests [8].

These classification-aggregation models offer promising results, although it should be noted that the use of existing models to validate other models is often discouraged due to the fact that errors in the reference model can be conflated with errors in the model being tested [13]. If the accuracy of the existing map is unknown, the extent to which it could affect the validation will remain a mystery. Likewise, if data derived from modeled TCC values were to be used for model training whatever error the 'training' model possesses sets the bar for the accuracy of the output model.

In contrast to these model-derived outputs, photointerpreted tree canopy cover measurements are measured values, derived using aerial imagery. These interpretations, collected by experts, have been used to train and test nationwide models for the National Land Cover Database (NLCD) since 2011 [14], [15]. The method used for these applications, described in Toney et al. (2012) and Coulston et al. (2012), uses a 105-dot, 90x90 meter triangular grid overlaid on 1 m resolution aerial imagery from the National Agriculture Imagery Program (NAIP) [14], [16]. The interpreter uses a two-sample approach, identifying whether each dot in the grid lands on tree canopy and then aggregating these observations to determine the proportion of all the dots that represent tree canopy, and thus percent tree canopy cover for the stand. In most cases, interpreters had ground-based knowledge of the areas they were interpreting [14]. An ArcGIS plug-in has also been developed to streamline the photointerpretation process for the 2012 forest inventory, as described in Goeking et al. (2012). This paper cites a 109 dots grid overlaid on round plots with a 43.9 m radius to correspond to Forest Inventory and Analysis (FIA) plot dimensions and locations across the nation. Analysts both interpret the plot and provide additional information as well as a rating of confidence in their interpretation [17].

The tree canopy cover photointerpretation method described above has been thoroughly tested. Toney et al. (2012) compares stem-mapped predictions to photointerpreted tree canopy cover and object-based tree canopy cover trained with photointerpretation. Their results shows that measurements from photointerpretation overestimate tree canopy cover compared to ground observations, though some of this may be due to the tendency of the stem-map approach to under-estimate tree canopy cover. This effect is exacerbated in object-based estimates. This study also shows lower variability in photointerpreted measurements and smaller difference between photointerpretation and ground estimates in the west [16].

Jackson et al. (2012) tested the repeatability of interpretations between trained experts at five sites across the US, measuring inter-interpreter agreement based on the proportion of interpreters' measurements that fell within 10% TCC of each other. Although there was a single interpreter whose measurements differed considerably from others at the sites they interpreted in Georgia, with that individual excluded, the over 90% of the other interpreters' measurements were within 10% agreement. In Oregon 70% of the measurements from all interpreters were within 10% agreement. Other sites showed more than 63% of measurements from all interpreters were within 10% agreement, with most interpreters having more than 70% of plots within 10% agreement. The authors

conclude that the potential for repeatability is high with proper training and pre-processing of data [18].

While photointerpretation can be used to measure tree canopy cover more efficiently than the common field methods, and is an appropriate method for both model training and validation, collecting these interpretations at all FIA plot locations across the United States remains a time-consuming and relatively costly proposal. To address this, Tipton et al. (2012) tested several sampling methods to make tree canopy cover photointerpretation more cost-effective. Their experimental design made adjustments to the tree canopy cover sampling intensity across a study area and the dot grid intensity within plots. The authors showed that fluctuations to model fit leveled off at sampling intensities comparable to the existing FIA sampling grid and that there is little return for the time spent to interpret dot grids with more than 40 dots, though the authors acknowledged that the effect of reducing the dot grid on predictive models is unknown [19].

Riemann et al. (2016) compared photointerpreted tree canopy cover measurements to field observations, stem-map modeling, and a canopy dataset that was derived from object-based image analysis using NAIP and LiDAR. They showed a strong overall agreement between field and photointerpreted tree canopy cover, but noted that photointerpretation tends to result in higher values than field observations at lower levels of tree canopy cover. In this study, the object-based image analysis-derived canopy dataset demonstrated higher tree canopy cover values than all other methods and had the highest level of agreement with the photointerpretation method, especially above 20% TCC. A portion of the photointerpreted plots in this study were also re-measured by a second interpreter for quality assurance. This exercise showed a high level of agreement between interpreters, though the authors noted that the photointerpreted plots had high levels of consistency for low tree canopy cover plots with higher variation at higher levels of tree canopy cover [20].

The results of these studies should be taken into consideration; however it should also be noted that the different methods of measuring tree canopy cover utilize different physical perspectives thus providing different information [21], [22]. Congalton and Green (2009) state that "ground verification of aerial estimates of vegetative cover is problematic as the estimates of cover from the ground (i.e. below tree canopies) are fundamentally different from estimates made from above the canopy [13]. Spurr (1960) asserts that forest crown closure is overestimated from aerial imagery and underestimated from the ground. Therefore using ground estimates as reference data for aerial cover estimates can be like comparing apples and oranges" [23]. Tree canopy cover is also scale-dependent [14], [24], so inconsistencies in the scale of data collection could lead to different results.

1.2. Objectives

The primary objective of this study is to determine the how consistent photointerpreted tree canopy cover values can be and what factors influence that consistency. To accomplish this, we measure the differences in interpretations collected by highly trained expert interpreters and between interpretations collected by separate groups of crowd and expert workers. To understand the factors which impact those differences, we identified the plot-specific variables and worker-interpretation-behavior variables that influence the level of "reliability" in crowd interpretations as they compare to experts' interpretations, and investigated the nature of that influence. We have used this information to propose recommendations for the design of future studies in order to improve consistency and agreement in photointerpretation of tree canopy cover by non-expert crowd workers.

2. Data and Methods

2.1. Study Area and Sample Site Selection

This study was conducted at sites across the conterminous United States. Public coordinates from the FIA program were selected as a sampling frame since the continuous hexagonal sampling grid ensures that our plots would not be clustered, and is unlikely to be aligned with regularly spaced landscape features [25]. This reduced the likelihood that classes with fewer samples would be in close proximity, since selecting closely clustered of samples would increase the likelihood that they were very similar and artificially reduce the estimated variability of that class. Alternatively, if we increased our sample size for this study, all 30 meter grid cells that encompassed the standard error layer could have been used without issue since smaller classes could have more samples. Samples were selected across the US, however areas that were predicted to have 0% TCC based on the 2016 USFS Analytical Tree Canopy Cover product were masked from sampling as well as pixels categorized as background, non-tree, water, ice, snow, and agriculture. 0% TCC pixels represented a large portion of the landscape and therefore would have represented a large portion of our sample, however the interface we developed would not have yielded any information about interpretation behavior for plots that were interpreted to have 0% TCC, and, though time required to interpret a 0% TCC plot is negligible, the payment would have been the same, resulting in a relatively costly set of data yielding little information. Therefore we opted to remove those areas.

We used a stratified random sampling protocol with strata delineated using the 1% classes from the standard error layer provided with the USFS Analytical Tree Canopy Cover product [26]. The standard error layer was created by calculating the standard error of the many predictions produced by the random forest model at each pixel location in the tree canopy cover model. Samples were selected using stratified random sampling with optimal allocation which allocates the number of samples in each stratum by a combination of within-stratum variance or standard deviation and stratum area [27]. Optimal allocation is a standard probability sampling design with known inclusion probabilities, where the number of samples (n_h in stratum h is defined by:

$$n_h = n * \frac{a_h * s_h}{\sum (a_h * s_h)} \tag{1}$$

where n is the total sample size, a_h is the area of the strata h, a is the population area, and s_h is the prediction uncertainty within stratum h [27]–[29]. Strata definitions and prediction uncertainty values were obtained from the standard error layer of the 2016 USFS Analytical Tree Canopy Cover product [26]. Strata areas were calculated by multiplying pixel counts for each of the error raster's values by the pixel area (900 m^2).

Because strata are not allocated proportionally to the relative area of the strata in the population, summary data for this study comparing the mean values of different variables across groups were calculated using a weighted average of that variable ($\bar{x}_w$), multiplying strata-level estimates by strata weights (w_h) as seen in Equations 2 and 3.

$$w_h = \frac{a_h}{a}; \sum w_h = 1 \tag{2}$$

where a_h is the area of strata h, and a is the total study area [29].

$$\bar{Y} = \sum (w_h * \bar{Y}_h) \tag{3}$$

where $\bar{Y}_h$ is the mean of a given variable within stratum h and w_h is the stratum weight.

An estimated sample size of 2000 plots was selected and confirmed to be appropriate for this study based on Equation 4.

$$n = \frac{(\sum (N_h * s_h))^2}{\frac{N^2 * E^2}{t^2} + \sum (N_h * s_h^2)} \quad (4)$$

where n is the total sample size, N_h is the number of units in the population in stratum h, s_h is the standard error value of stratum h, N is the population size based on the number of unmasked pixels in the population (comprising 2,989,614,429 30-meter pixels across the conterminous United States), t is *student's t* based on the 95% confidence interval and large population size (1.960), and E is the desired precision of the sample mean to the population mean (which we determined to be 1% TCC) [27]. This results in a value of 1384, was significantly lower than the selected sample size. After fractional plots were rounded during optimal allocation sampling the final sample size for crowd interpretations is 1998.

For the expert sample, a subset of the crowd plots were selected using the same sample allocation method but a smaller sample size. A target sample of 500 plots was used for expert sampling and after fractional plots were rounded during optimal allocation sampling the final number of expert samples is 502.

Because this sampling method could over-emphasize areas with high expected error or under-emphasize areas with low expected error, sample-wide statistics were calculated by first determining the estimate within each strata and then combining those estimates using a weighted average with weighted defined by proportionate strata area so that the strata-level estiamtes were represented proportionally to the population.

2.2. Mechanical Turk HIT Interface Development

Amazon's Mechanical Turk (https://www.mturk.com/) is a network interface for crowdsourcing tasks. This interface enables potential analysts to be trained for and undertake HITs for a small monetary payment. The preliminary design of the Amazon Mechanical Turk interface used for this study has two sectors.

A local interface which hosts spatial data such as plot locations, dot grids, and NAIP imagery (from existing ArcGIS online web services [30], [31]) is located on Virginia Tech's GIS servers (as in Yu et al., 2015) [4], and a Mechanical Turk Shell, which has the capacity for recruitment, worker identification, and payment, interpretation tasks, database output, and interface interaction monitoring. The design harmonizes these two sectors into a single interface using ESRI's ArcGIS API for Java script.

Color-infrared NAIP imagery was obtained using ArcGIS online with dates close to 2011 for time 1 and and 2016 for time 2, corresponding to timing for the 2011 and 2016 USFS Analytical Tree Canopy Cover products. 1 m spatial resolution NAIP imagery is collected by the USDA Farm Service Agency from aerial surveys with partial coverage each year. Since 2009, full national coverage of NAIP imagery is collected every three years. Images are collected on clear days to ensure that images have less than 10 percent cloud cover per quarter quad tile. The images used for this study have three visible bands and one near infrared band, enabling a natural-color or false-color composite.([32])

For this research, we served 1998 HITs, one for each sample plot selected. Each HIT involves the interpretation of a 100-dot grid at each plot location for two time periods. The dot grid consists of evenly-spaced dots aligned with a 30-m Landsat pixel which intersects the public FIA plot location. Upon the completion of each HIT, a crowd worker received a payment of $0.22 and Amazon received a 20% payment of $0.05. We took reasonable measures to ensure that interpretations were completed by different interpreters. There were 9 available assignments for each HIT, meaning that 9 different interpreter accounts accessed and completed the same plot

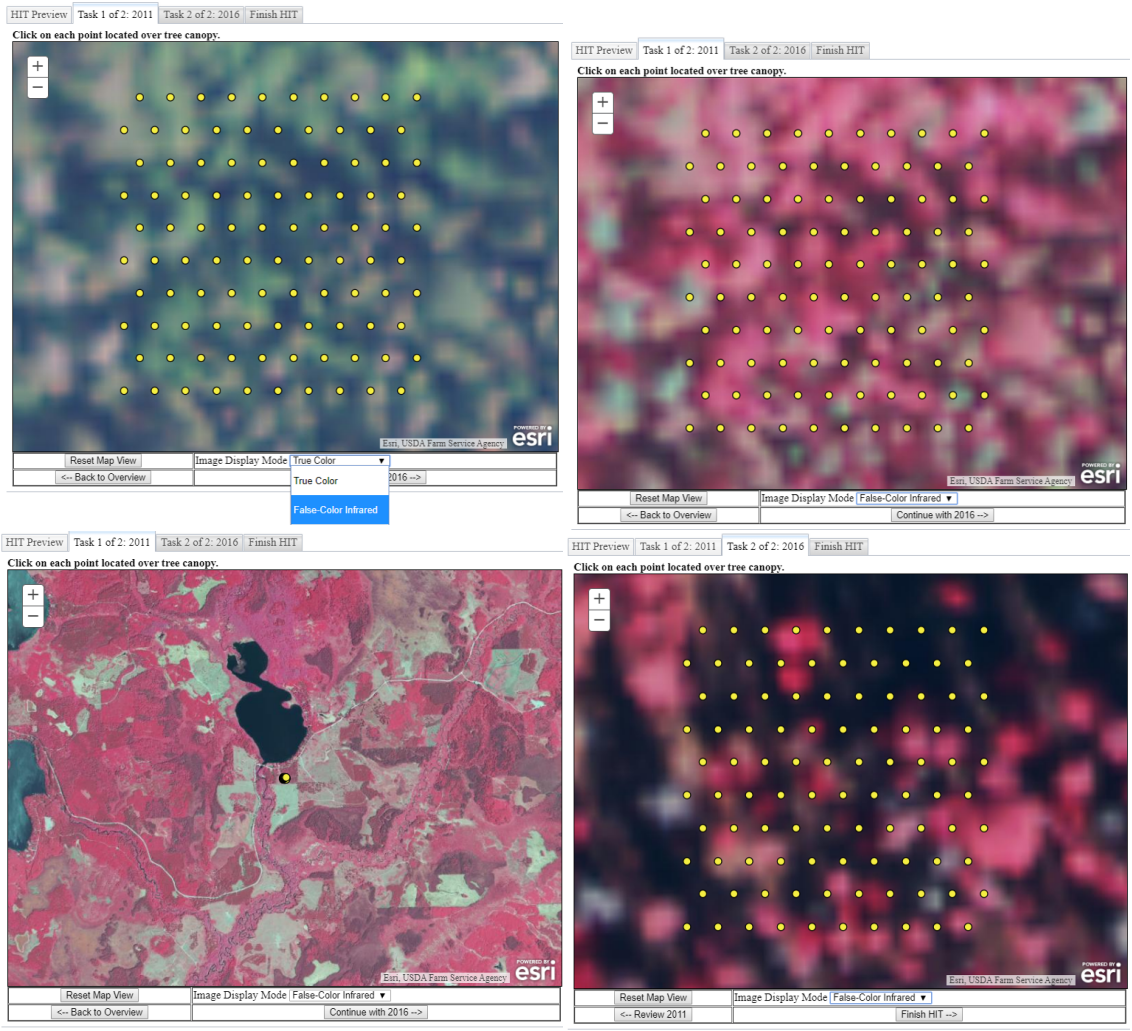

Figure 1: Mechanical Turk interface. (top-left) 2011 image in natural color display, with display dropdown menu expanded. (top-right) 2011 image in false color display. (bottom-left) 2011 image zoomed out (using scale buttons in top left corner of map). (bottom-right) 2016 image in false color (note the tab progression at the top of the window above the map).

interpretation. One interpreter's account cannot access and interpret the same HIT twice, however if an interpreter had multiple accounts they may have been able to work the same hit twice.

During interpretation, the interpreter could change the display between true color and false color, zoom in and out, and reset the zoom to be centered around the dot grid. To interpret TCC on the provided plot, interpreters would click on the dots in the dot grid. The first time the interpreter clicked a dot, the color would turn blue and the designation would change to indicate that the dot feel on tree canopy. If they clicked a dot a second time, this would revert the dot's appearance and change the designation to indicate that the dot did not fall on tree canopy. Once the interpreter completed the 2011 interpretation they would repeat the process, this time using a 2016 NAIP image. They could also return to the 2011 interpretation to review and modify their work. Once both 2011 and 2016 interpretations were completed, the interpreter would certify and submit their HIT (Figure 1).

The back-end of the interface consisted of a database which collected information on the interpretations for each dot from each interpreter. In addition to the interpretation, when an interpreter clicked on a dot, the interface collected the display status (false color/true color) the image scale, and a time-stamp. The data was organized by plot ID and dot ID and additionally included fields with the x and y coordinates of each dot. After each phase of this study, these data were compiled into a JSON file and imported into R where they were read to a table.

Expert interpretations took place using an identical interface located in the Amazon Mechanical Turk Worker Sandbox. This interface operates in much the same way as the outward-facing Mechanical Turk interface, but is used for development and testing before HITs are released for completion. In order to ensure that the designated experts for this project were performing the interpretations allotted, access to these hits was limited to only the accounts owned by those individuals, as confirmed using their email addresses, and the individual experts were provided with a link to their specific HIT collection.

2.3. Training and Advertising

A training document was developed by our team in order to ensure that both expert and crowd interpreters were aware of the objective of the task. This document provided definitions of tree canopy cover, background on how it is collected, and many examples of images that may be difficult to interpret. For example, images depicting defoliated (leafless) branches (which are considered to be tree canopy) and shrubs (which are not considered to be tree canopy) were shown as well as images of dense and sparse forest, agricultural areas, and residential trees. It also showcased the features of the interface that could be helpful for this interpretation, such as false color composite options and zoom. This training document was provided before crowd workers were given their qualification test. It was also available for review at any time from a link included in the interpretation interface.

To qualify for this HIT, crowd workers were expected to review the training protocol and then pass a qualification exam. The exam provided 11 questions showing a single dot on aerial imagery and the worker was expected to accurately determine if the dot landed on a tree canopy. The dot in question was shown in both true and false color, and at smaller and larger scale for context. If the crowd worker passed the exam with an 80%, they were then qualified for the HIT. An informed consent document was linked for review and a question of agreement was included in the exam to ensure that all workers were aware of what data was being collected and any possible risks. A captcha task was also also included in the qualification as the incidents of bots on mechanical turk are rising [33]–[35].

Crowd HITs were released in three phases. After each release, the HITs were advertised on popular sub-reddit channels 'r/HitsWorthTurkingFor', 'r/mturk', and 'r/TurkerNation'. The ad-

vertisement was subsequently re-posted to the 'TurkerNation' Slack page. The first and second phase consisted of plots that corresponded to expert-interpreted plots, while the third phase represented the remaining 1496 plots. The first phase was limited to only 50 plots for the purpose of quality control and to ensure that the interface was working properly. These 50 plots were randomly selected from the set of 502 that the experts were also interpreting. During the first release of this phase we discovered an issue which resulted in all hits being directed to only a single plot. This was corrected before the phase was completed and only the first nine interpretations of this plot from different workers were used for analysis. Phase 1 plots were re-released after the system was corrected, replacing the one plot that had already been interpreted with another randomly selected plot. Phase 3 plots were released using three separate instances on Mechanical Turk to reduce output file sizes and improve data retrieval.

2.4. HIT Release and Quality Control

	Description	n Plots	Initiated	Completed	Approx. Timing
Phase 1	Quality Control	1	24-Oct	29-Oct	6 days
Phase 1b	Quality Control Re-launch	50	19-Nov	23-Nov	5 days
Phase 2	Also Interpreted by Experts	451	18-Dec	21-Dec	4 days
Phase 3	Additional Plots	1498	2-Mar	5-Mar	4 days

Table 1: Schedule and characteristics of each HIT release including phase, description, number of plots, dates, and timing.

All crowd interpretations for this study were collected between October 24, 2019 and March 5, 2020. Data collection took between 4 and 6 days. Technical issues with the initial phase 1 request resulted in individual workers interpreting the same plot location many times. As an exploratory quality control measure, these interpretations were plotted compiled and mapped to show how many of their own dot interpretations agreed. This information was qualitatively assessed to be satisfactory before the HITs were re-released. The first nine (chronologically) interpretations from different workers who responded to this initial release were retained and used, but the other interpretations were removed from subsequent analysis. Once phase 1 was completed, plot interpretations were mapped out over NAIP and qualitatively spot checked to ensure that workers were interpreting the imagery correctly and that the interpretations were properly aligned with NAIP in case of projection issues. Interpretations that resulted in 0% TCC were also identified and compared to ensure that workers were not submitting empty HITs to increase their wage. After phase 2, histograms were produced to confirm there was not an over-abundance of 0% TCC interpretations compared to the expected distribution. Direct quantitative comparisons were not performed for quality control as one of the objectives of this study was to test the ability of crowd workers to collect consistent interpretations with experts, and adjusting the interface based on the results of these comparisons could bias that outcome.

Expert interpreters began collecting their interpretations well in advance of the crowd interpreters for this study. This allowed them to provide feedback about the interface and make improvements.

2.5. Data Analysis

This study generated up to nine crowd interpretations at 1998 plot locations and up to three expert interpretations at 477 plot locations across the United States. No clear benchmark for 'truth' exists among these photointerpretations of TCC, and so we rely on analysis of agreement

between crowd and expert groups. Witihin-group agreement was measured using the standard deviation across interpretations at each plot. The reliability of crowd worker interpretations was measured using the proportion of plots where crowd workers and experts agreed within 10% TCC. This proportion was calculated using the weighted average of the proportion of a crowd worker's interpretations that are within 10% of expert interpretations (p_{ah}) (Equations 2,3, and 5).

$$p_{ah} = \frac{n_{ah}}{n_h} \tag{5}$$

where n_{ah} is the number of plots in agreement in stratum h and n_h is the sample size within stratum h.

The proportion of plots with agreement within 10% TCC was also calculated between individual experts to better understand their level of agreement. Expert interpretations were also compared with one another and to crowd interpretations using a weighted Pearson's correlation coefficient and a weighted Kolmogrov-Smirnov test. Weighted Pearson's correlation coefficients (r_{xy}) (Equation 6) between groups x and y were calculated by:

$$r_{XY} = \frac{\sum (w_h * (X_h - \bar{X}) * (Y_h - \bar{Y}))}{\sqrt{\sum (w_h * (X_h - \bar{X})^2) * \sum (w_h * (Y_h - \bar{Y})^2)}} \tag{6}$$

where w_h is the stratum weight of stratum h, X_h and Y_h are the values of from groups X and Y within stratum h, and $\bar{X}$ and $\bar{Y}$ the weighted means of X and Y across all strata [36], [37].

The weighted Kolmogorov-Smirnov test uses a weighted cumulative distribution function ($\hat{F}_W(x)$)

$$\hat{F}(X) = \sum_{i=1}^{n} w_i * I(X_i < X) \tag{7}$$

where X_i is the i^{th} sample in X, $I(X_i< x)$ is an indicator function identifying elements X_i of X that are less than the other elements of X, and w_i is the weight calculated by equation 8.

$$w_i = \frac{w_h}{n_h} \tag{8}$$

where w_h is the stratum weight and n_h is the number of samples in stratum h.

These weighted cumulative distribution functions are combined in the weighted Kolmogrov-Smirnov test as follows using Equation 9.

$$max_X(\hat{F}(X) - \hat{G}(X)) = max_X(\sum_{i=1}^{n} w_i * I(X_i < X) - \sum_{i=1}^{m} v_i * I(Y_i < X)) \tag{9}$$

where $\hat{F}(X)$ is the cumulative distribution function used to pull n samples from X with weights w_i and $\hat{G}(X)$ is the cumulative distribution function used to pull m samples from Y with weights v_i [38].

The p-value for the weighted Kolmogrov-Smirnov test is

$$Prob(D > d) = 1 - 2 * \sum_{i=1}^{\infty} (-1)^{(i-1)} * e^{(-2i^2 z^2)} \tag{10}$$

[38] where z is defined as:

$$z = d\sqrt{\frac{n_{w1} * n_{w2}}{n_w}} \tag{11}$$

[38] where values of n_w are weighted sample sizes, defined by:

$$n_w = \frac{\Sigma(w_i)^2}{\Sigma(w_i^2)} \tag{12}$$

where w_i are weights.[38]

The weighted Kolmogrov-Smirnov test was implemented in R using functions from the Fast Doubly Robust Estimation, R package (fastDR) [39] and the weighted p-value was obtained using the stats package.

Plot interpretations were also aggregated to find mean and median interpretation values at each plot. These aggregated interpretations were compared to expert interpretations using the weighted mean, weighted Pearson's correlation coefficient (r), weighted Kolmogrov-Smirnov index (d), and weighted Analysis of Variance (ANOVA).

2.6. Reliability Model

Dependent Variable	Explanation
Reliability Class	Class "Unreliable" or "Reliable" based on agreement frequency with experts

Table 2: Dependent variable explanation for reliability model.

Parameter	Setting	Explanation
type	'classification'	Classification model (not regression)
importance	TRUE	Return predictor importance assessment
ntree	500*	Grow 500 classification tree models
mtry	floor(sqrt(ncol(x)))=4*	Number of randomly selected variables to include in each model
replace	TRUE*	Select in-bag samples with replacement
sampsize	nrow(x)*	Number of in-bag samples to select
nodesize	1*	Minimum size of terminal nodes
maxnodes	NULL*	Maximum number of terminal nodes trees in the forest can have (NULL limits this only by nodesize)

Table 3: Parameter settings for the randomForest classification model in R. * indicates a default setting.

TCC Variables	Explanation
Interpreted TCC	Worker-interpreted % Tree Canopy Cover using dot grid.
Standard Deviation of Crowd Interpretations	Standard deviation of crowd interpretations at a given plot location
2016 US Forest Service TCC Analytical Model Standard Error [26]	Standard Error Estimate by Pixel for the 2016 US Forest Service Analytical TCC Model
2016 US Forest Service TCC Analytical Model [26]	Modeled % Tree Canopy Cover value from the 2016 FS Analytical Tree Canopy Cover Model
Interpretation Variables	**Explanation**
Plot Interpretation Time	Number of seconds (to the nearest 60 seconds) it took the worker to complete the interpretation
Number of Plots Interpreted	Number of interpretations completed before a given interpretation, an indication of worker fatigue
Number of HITs Available	Number of remaining HITs available to a worker at the time of a given interpretation, an indication of potential earnings
Release Phase	HIT release phase (Phases 1, 2, or 3)
Dot-level Interpretation Variables	**Explanation**
% Dots in False Color	Proportion of all dots in interpretation clicked with the false color display activated
% Dots in Natural Color	Proportion of all dots in interpretation clicked with the natural color display activated
% Dots Zoomed In	Proportion of dots in interpretation clicked while zoomed in from the starting scale
% Dots Zoomed Out Beyond Image Resolution	Proportion of dots in interpretation clicked while zoomed out beyond the interpretation scale
% Dots Zoomed Out	Proportion of dots in interpretation clicked while zoomed out from the starting scale
% Dots at Starting Zoom	Proportion of dots in interpretation clicked while still the starting scale
Ecological Plot Variables	**Explanation**
2016 NLCD Land Cover [40]	2016 National Land Cover Database land cover class
Level-2 Ecoregion	Level-2 Eco-region

Table 4: Predictor variable explanation for reliability model.

Crowd workers' interpretations were compared to expert interpretations to obtain ratings of crowd worker reliability. Because there were multiple crowd and expert interpretations, a data frame was created with records for all possible comparisons of crowd and expert workers' interpretations at each plot. The percentage of these comparisons that were within 10% TCC was obtained for each crowd interpreter using the weighted mean agreement as described above (Equations 2, 3, and 5).

Benchmarks for crowd worker reliability were set by the upper and lower bounds of agreement between the three experts. The minimum proportion of all 2011 and 2016 observations that fell within 10% TCC between any two experts (0.20) was used to define the upper bound of "unreliable" crowd workers, and the maximum proportion of all 2011 and 2016 observations that fell within 10% TCC between any two experts (0.50) was used to define the lower bound of "reliable" crowd workers. Crowd workers whose agreement frequency fell between 0.20 and 0.50 and those individuals who only interpreted HITs in phase 3 (which were not also interpreted by experts) were considered to be 'Other/Unknown' and were left out of the training sample.

A subset of variables relating to plot characteristics and interpretation behavior were collected for each interpretation (Table 4). Training samples were then selected using a Monte Carlo framework and used to estimate a series of 5000 random forest classification models predicting reliability class (Tables 3 and 2)[41]. Because of the disparity in the number of "reliable" interpretations compared with the more prevalent "unreliable" class, bootstrapped sampling was used to randomly choose a balanced sample from the training data for each of the 5000 models. The "reliable" interpretations were randomly selected with replacement to identify an in-bag sample and an out-of-bag testing sample. The resulting in-bag dataset was then sampled with replacement again to amplify the number of interpretations in the sample until the "reliable" class accounted for 50% of interpretations used to train our model. The "unreliable" data were also sampled using this method to select an "unreliable" sample also accounting for 50% of interpretations used to train our model. The resulting training dataset had two classes with 2637 interpretations in each. "Unreliable" interpretations that were not selected for the training data set were held out for the out-of-bag testing dataset. The class samples were combined to create a training data set and a testing dataset, and a random forest classification model was estimated to predict interpreter reliability. This process was repeated 5000 times, creating training and testing data for each model.

For interpretation purposes, predictor variables were organized in categories, including 'interpretation variables', related to the sequence and timing of interpretation, 'dot-level interpretation variables', related to user interaction metrics recorded at each click, 'TCC variables', related to tree canopy cover at each plot, and 'ecological plot variables', describing the eco-region and land cover at the plot location. Records with any null values were removed to comply with the random forest modeling approach used.

While data was collected for both 2011 and 2016, initial testing on both years of data showed strong correlations between results, so while the reliability analysis presented in this study relates only to 2016 data, similar trends were observed for 2011. The benchmarks described above, however are defined using both 2011 and 2016 data.

Variable importance metrics and in-bag and out-of-bag predictions were obtained for each of the 5000 models. Out-of-bag predictions were used to estimate overall model accuracy for each model iteration and predictions on all data (in-bag and out-of-bag) were used to calculate a percent reliability metric for each interpretation. This metric reflects the percent of models which predicted that interpretation to be "reliable." To test variable importance, the mean increase in error resulting from each predictor's removal from the model was collected and ranked for each iteration. The top 10 most important variables were given a score between 10 (for the most important variable) and 1 (for the 10^{th} most important variable) and those scores were added by variable across iterations

to get an overall variable importance score [42].

Different subsets of crowd interpretations (all crowd, 2011, 2016, reliable, and unreliable) were compared to expert interpretations to test if they came from statistically different distributions. These statistics were approximated using a combination of all available interpretations at each plot location. To ensure representation of all interpretations we randomly sampled a single interpretation from (up to 9) available crowd interpretations and a single interpretations from the (up to 3) available expert interpretations at each of the 477 plots where both crowd and expert interpreters worked and repeated this process 5000 times. For each of the 5000 resulting data sets, subsets of data were obtained by querying by the interpretation year and, for 2016 data, by interpretation reliability class. Expert interpretations were compared to crowd interpretations for each subgroup using weighted Pearson's correlation coefficient (r), weighted Kolmogrov-Smirnov index (d), and weighted analysis of variance (ANOVA) using a weighted linear model with a response variable of TCC and a predictor variable of expert/crowd grouping.

3. Results

3.1. Interface Technical Issues and Feedback

We received feedback on reddit and via email detailing several aspects of our interface. The primary technical issue that users experienced with our HITs was that maps did not load reliably, yielding a dot grid over an empty screen. We expect that this could be related to limited bandwidth and internet speeds in some areas, which could prevent the high resolution imagery from rendering.

Crowd workers also expected a higher payment for their work. The total amount spent per assignment was $0.27, determined based on our existing budget, required sample size, and previous experience working with Amazon's Mechanical Turk before our interface was developed. Amazon has a 20% fee, so, for each completed assignment workers were paid $0.22. We had planned for 10 assignments per HIT instead of 9, however, Amazon's fee increases from 20% of the payment to 40% when 10 assignments or more are issued per HIT, so we reduced the number of assignments to 9. We estimated the time required to complete a complex HIT (an interpretation of both 2011 and 2016 imagery with difficult to interpret, densely forested coverage) to be approximately two minutes, with many HITs taking considerably less time. In fact, experts averaged 4.6 minutes per HIT overall (with a median value of 3 minutes). Crowd workers, on the other hand, spent 2.02 minutes to complete a HIT on average, with a median value of 2 minutes per HIT.

Assignments Completed	n Plots
7	13
8	266
9	1719

Table 5: Number of plots by number of assignments completed by crowd workers.

Assignments Completed	n Plots
1	65
2	223
3	189

Table 6: Number of plots by number of assignments completed by expert workers.

Some assignments were left incomplete for expert HITs resulting in 7 and 8 tree canopy cover interpretations per plot instead of the 9 interpretations expected. Expert workers also left some

assignments incomplete, resulting in 1 and 2 interpretations per plot instead of the 3 interpretations expected. While less than 1% of crowd plots were missing two interpretations and a significant majority of crowd plots were complete, about 13% of expert plots were missing two interpretations and the majority of expert plots were missing one interpretation. (Table 5 and Table 6). There was also some difficulty compiling and downloading the large volume of data from phase 2, so phase 3 interpretations were split into three separate requests.

Expert interpretations were collected through Amazon Mechanical Turk's worker sandbox. This interface is meant for testing and development, and is not explicitly designed to collect data in the manner we used it. As such, there were several issues where data was lost. In particular, some expert interpretations collected while advancing through assignments using an auto-accept feature which was being beta-tested in the worker sandbox were not recorded in the database. Additionally, data stored in the worker sandbox is deleted after 120 days, which resulted in lost data.

3.2. All Interpretations

Statistic	Value (n)	Number of Workers Who Submitted At Least n Assignments	Number of Workers Who Submitted Only 1 Assignment
Minimum	1	170	49
First Quartile	1	170	
Median	6	86	
Mean	110	23	
Third Quartile	33	43	
Maximum	1927	1	

Table 7: Sumary statistics for the number of plots interpreted by individual crowd workers.

Worker Type	Year	Phases	n Workers	n Plots	n Interpretations	Weighted Mean % TCC	Weighted Mean Disagreement (% TCC)	Weighted Confidence Interval Width
Expert	2011	1 & 2	3	502	1,078	63.6	16.8	35.6
Expert	2016	1 & 2	3	502	1,078	63.6	16.4	34.6
Crowd	2011	1 & 2	108	502	4,432	41.7	23.3	23.5
Crowd	2016	1 & 2	108	502	4,432	43.6	21.9	22.7
Crowd	2011	1, 2, & 3	170	1998	17,690	38.8	21.2	25.8
Crowd	2016	1, 2, & 3	170	1998	17,690	39.6	20.5	24.2

Table 8: Summary values for interpretations as collected. Disagreement is the standard deviation of all interpretations of each plot by worker type The weighted mean confidence interval width refer to the confidence interval around the mean for all group-specific interpretations at each plot.

Crowd interpreters returned 17,690 assignments during this study, each representing a TCC interpretation for both years, while expert interpreters collected a total of 1,078 observations for 2011 and 2016. Interpretations in phases 1 and 2 were interpreted by both crowd and expert workers. These account for 4,432 crowd interpretations total (at 502 out of the total 1998 plot locations), while 13,258 crowd interpretations occurred at plots that were not also interpreted by expert workers. The 170 crowd workers who participated in this study completed between 1 and

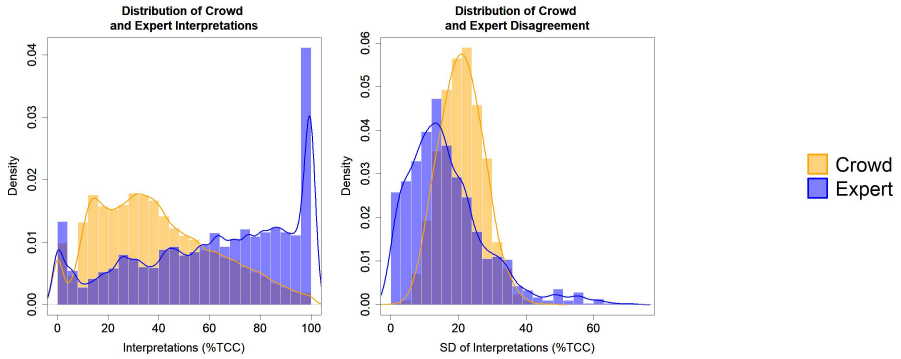

Figure 2: Distribution of expert and crowd interpretation values (left) and disagreement, calculated using the standard deviation across interpreters' measurements at each plot (right) for 2016. Data from 2011 show similar trends and distributions.

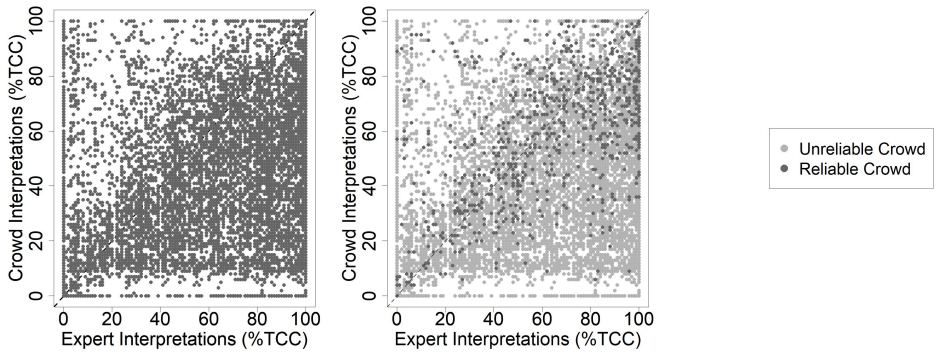

Figure 3: (left) All combinations of expert vs. crowd interpretations from 2016 and (right) all combinations of expert vs. crowd interpretations from 2016 split into "unreliable" and "reliable" classifications based on model predicitons.

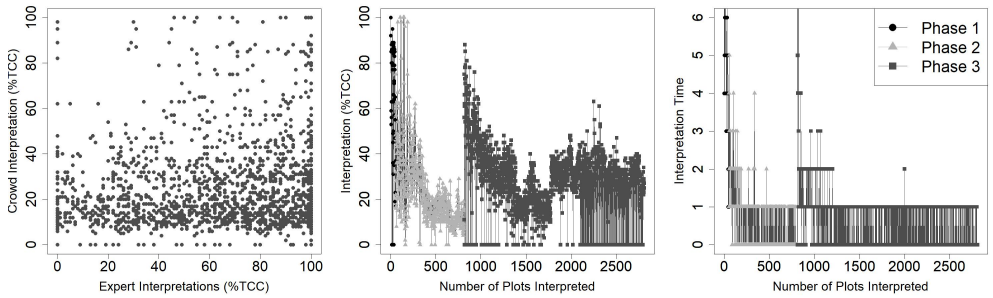

Figure 4: Plots from an individual crowd worker's interpretations showing the experts' interpretations vs. the worker's interpretations (left), the sequence of the worker's interpretations vs. the interpretation's value (center) and the sequence of the worker's interpretations vs. the time it took to complete the interpretation (right).

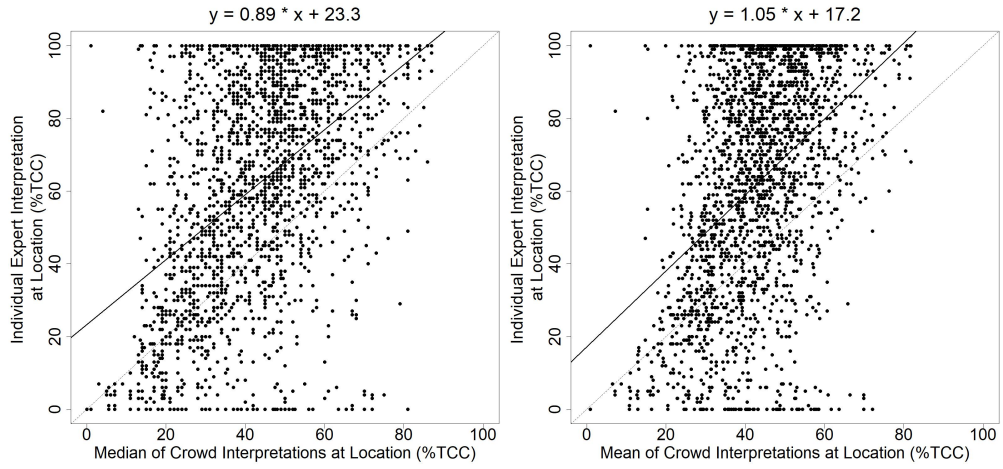

Figure 5: Plots comparing the median (left) and mean (right) of all assignments submitted for each HIT to the individual expert interpretations evaluated at the same plot locations. The bold line is the fitted trend line, and the dashed line is the 1:1 relationship.

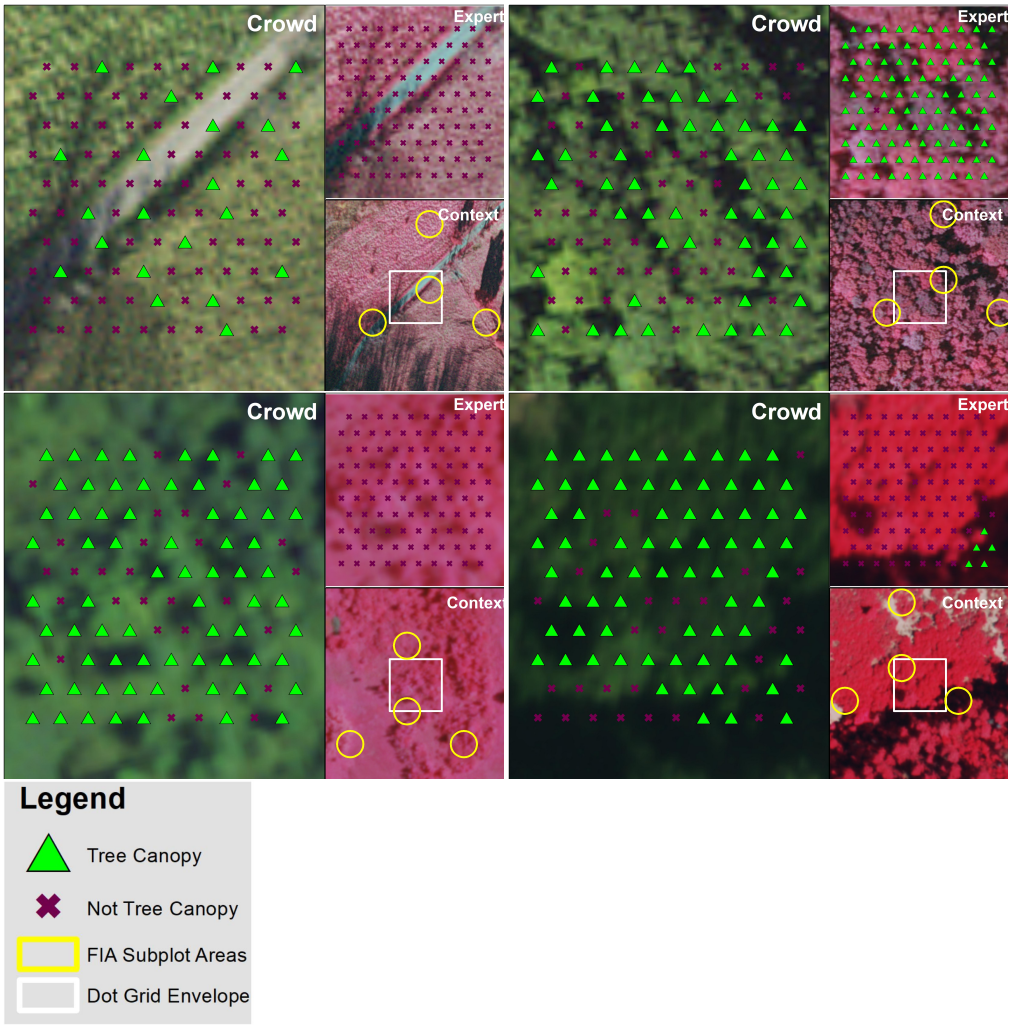

Figure 6: Each image shows a panel with a crowd interpretation (and the underlying NAIP imagery in the display used during the majority of that interpretation) at an approximate scale of 1:700, an expert interpretation (and the underlying NAIP imagery in the display used during the majority of that interpretation) at an approximate scale of 1:1000, and a context image, zoomed to an approximate scale of 1:3250. Base layers from ESRI and the USDA Farm Service Agency.

1,927 assignments individually, each involving the interpretation of a single plot with imagery from two dates, 2011 and 2016. The mean number of assignments completed was 11, and the median number of assignments was 6. The majority of workers submitted at least 6 assignments, but many workers submitted fewer, and 49 workers submitted only 1 assignment (Table 7).

Mean tree canopy cover values from crowd interpretations are substantially lower than those from expert interpretations (Table 8). When phase 3 is accounted for, the weighted mean of the crowd interpretations is even lower for both years, however there were no expert interpretations at phase 3 plots for comparison. The disparity in crowd and expert interpretations is apparent in the histograms from 2016 which show that crowd interpreters interpreted lower tree canopy cover values more often than experts (Figure 2). By contrast, expert histograms show that experts interpreted very high tree canopy cover values more often than crowd workers (Figure 2). Data from 2011 show similar trends and distributions.

Disagreement was measured using the standard deviation of interpretations for each plot by group. Disagreement between crowd workers is higher overall than it is for experts in 2011 and 2016 for phase 1 and 2 data only. This trend is maintained when all three phases of crowd data are accounted for (Table 8). Histograms also show that disagreement values tend to be higher for crowd interpreters than for experts in 2016 (similar results were observed for 2011), but experts' disagreement levels have more plots with high disagreement.This could be due to the smaller number of interpreters (Figure 2).

The width of the confidence interval around the mean for crowd interpretations is consistently smaller than for expert interpretations, though the confidence intervals for crowd workers were wider when phase 3 plots were included (Table 8).

Diagnostic plots for individual crowd workers and maps of plot interpretations show several patterns in crowd interpreter behavior. For example, a scatter plot between an individual crowd worker's interpretations (later deemed "unreliable") and expert interpretations show a concentration of crowd interpretation values between 5 and 25% TCC. Comparing interpretation values to the number of plots interpreted show a sharp decline in TCC values as the worker interpreted more plots, with recovery at the beginning of each HIT release phase which were separated by up to 72 days. Additionally, interpretation times decline as the worker interpreted more plots, also recovering at the beginning of each interpretation phase (Figure 4).

Scatter plots comparing expert interpretations to mean and median values of assignments submitted by crowds for each HIT show undersampling of TCC among crowd workers compared to experts. The intercept of the trendlines of each plot support this as well, indicating that, in aggregate, the value of crowd interpretations were lower than expert's interpretations. The scatter plot comparing expert interpretations to the median values across crowd interpretations has a wider range of values, extending to higher TCC values, and a lower slope in its trendline than the plot comparing expert interpretations to the mean values across crowd interpretations. This indicates that a few very low TCC values from individual assignments may have impacted the mean of crowd interpretations for each HIT, causing even lower aggregate values in this case (Figure 5).

Maps of selected crowd interpretations show the following cases:

- Crowd workers clicked a series of disconnected dots throughout the plot resulting in an overestimate of the TCC value compared to experts
- A plot with heavy texture and shadow where a crowd worker underestimated TCC compared to experts
- A plot showing features which a crowd worker interpreted as trees, and which the expert interpreted as non-tree features resulting in a higher crowd interpretation value compared to

the expert

- A plot with heavy shadow and nearby trees where a crowd worker overestimated TCC compared to experts

(See Figure 6.)

Finally, scatter plots pairing all expert interpretations against all crowd interpretations show that crowd workers routinely underestimate TCC compared to experts, and that there is little correlation between the groups' interpretations (Figure 3).

3.3. Expert Interpreters and Inter-Expert Agreement

Expert	n Plots
Expert 1	385
Expert 2	247
Expert 3	446

Table 9: Number of interpretations by expert.

Comparison	n Interpretations	Weighted Mean Proportion within 10% (2011 and 2016)	Weighted d (2011)	Weighted d (2016)	Weighted r (2011)	Weighted r (2016)
Exp 1 v Exp 2	388	0.53	0.31	0.28	0.64	0.74
Exp 2 v Exp 3	484	0.32	0.32	0.32	0.77	0.82
Exp 3 v Exp 1	708	0.20	0.45	0.47	0.73	0.67
All Experts	1580	0.31				

Table 10: Expert-to-expert comparison showing number of plots interpreted between experts, proportion of interpretations within 10%, weighted Kolmogrov-Smirnov index (d), and weighted Pearson's correlation coefficient (r). An asterisk next to the value of d (*) indicates that the two sets of data come from the same distribution as demonstrated by a p-value of greater than or equal to 0.05.

The expert interpreters collected between 247 and 446 plot interpretations for all plots from 2011 and 2016 (Table 9). Experts were compared individually to determine if their interpretations came from similar distributions. The number of single-date interpretations paired between experts ranged from 388 (between experts 1 and 2) and 708 (between experts 1 and 3). For all experts, 31% of interpretation were within 10% TCC of the other experts, with the lowest agreement frequency between experts between experts 1 and 3 and the highest frequency of agreement between experts 1 and 2. These agreement proportions account for interpretations from both 2011 and 2016. The distributions were found to be different between all pairs of experts based on weighted Kolmogrov Smirnov tests showing p vales of less than 0.05 (Table 10).

Finally, weighted Pearson's correlation coefficients (r) show correlations between experts (comparing data from individual years) that range between 0.64 and 0.82. Interpretations collected by experts 2 and 3 had the strongest correlations for both years. For 2016 imagery, interpretations collected by experts 1 and 2 and experts 2 and 3 were more strongly correlated than those collected using 2011 imagery. The weakest correlation in 2011 was between experts 1 and 2, while the weakest correlation in 2016 was between experts 1 and 3.

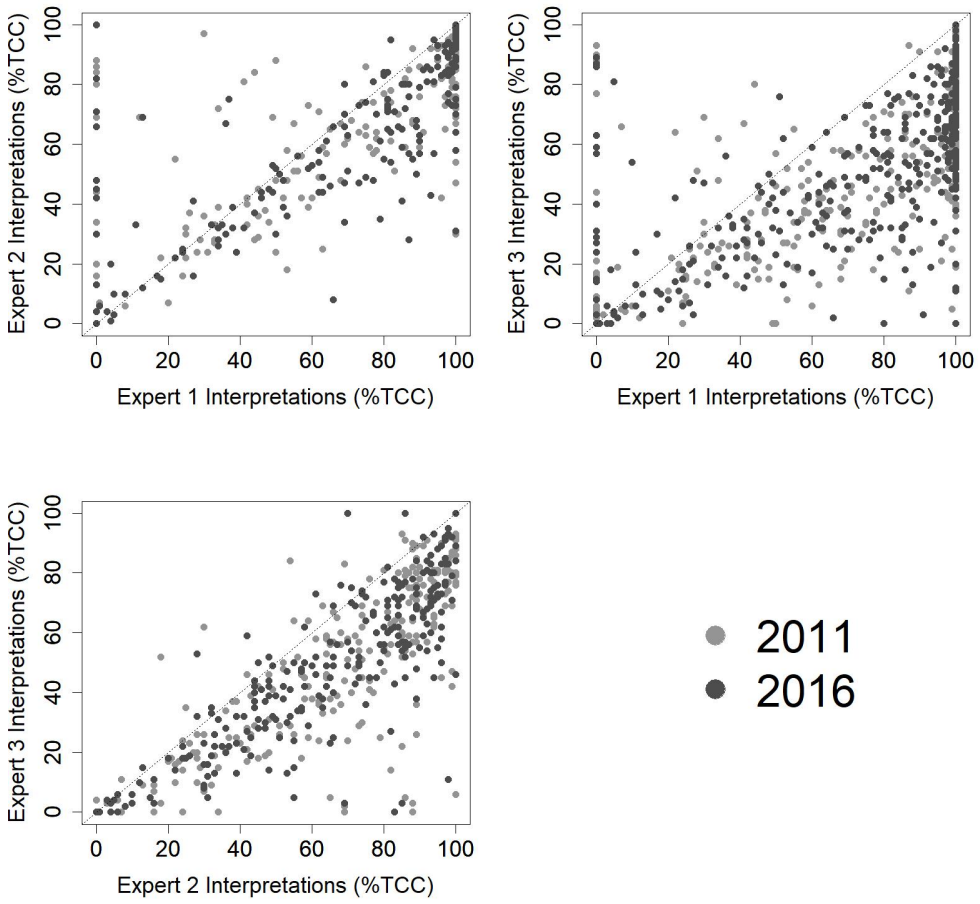

Figure 7: Scatter plot between individual experts for 2011 and 2016.

Scatterplots of TCC interpretations between expert workers show some correlation with the interpretations of expert 1 and expert 2 being closest. Expert 1 had higher TCC interpretations compared to other experts, and expert 2 had higher TCC interpretations than expert 3. Plots show little difference in agreement between interpretations of 2011 imagery and 2016 imagery (Figure 7).

Plots that compare expert 1's interpretations to experts 2 and 3 show disagreement between what constitutes tree canopy (producing plots interpreted as 0% TCC while the other interpreter produced a plot with greater than 0% TCC) and canopy vs. ground (producing plots interpreted as 100% TCC while the other interpreter produced a plot with less than 100% TCC) (Figure 7).

3.4. Reliability

3.4.1. Reliability Model

Statistic	Proportion of Plots within 10%	d	n crowd interpreters with p >= 0.05
Minimum	0.00	0.00	44 out of 108
First Quartile	0.22	0.01	
Median	0.28	0.33	
Mean	0.31	0.38	
Third Quartile	0.40	0.79	
Maximum	1.00	1.00	

Table 11: Agreement between crowd and expert interpreters as demonstrated by the summary metrics for the crowd workers' proportion of plots within 10% TCC of experts summary metrics for the Kolmogrov-Smirnov test results between each crowd worker's interpretations and expert interpretations, as well as the number of workers whose interpretations are likely to have come from the same distribution as the expert interpretations.

The average proportion of plots from a single crowd worker that were within 10% agreement with experts was 0.31, with a minimum agreement proportion of 0 and a maximum agreement proportion of 1. Kolmogrov-Smirnov distance values between individual crowd workers and all experts had a mean of 0.38, and 44 out of 108 workers had p-values greater than 0.05 indicating that the data came from the same distribution. 21 of those 44 workers had only provided one interpretation and only 11 of those workers had at least five interpretations in 2016 (Table 11).

As described above, model training data was obtained by classifying all of a crowd workers' interpretations as "reliable" and "unreliable" based on the proportion of their plot interpretations which agreed with expert workers within 10% TCC. The benchmarks for this classification was based on the minimum and maximum agreement between experts, thus "reliable" crowd workers were selected by having a proportional agreement greater than or equal to 0.53, while "unreliable" crowd workers were selected by having a proportional agreement less than 0.20. This process resulted in a training sample with 95 "reliable" interpretations at 93 plots from 10 workers and 5168 "unreliable" interpretations at 1952 plots from 23 workers. It should be noted that while these benchmarks were obtained using data from both 2011 and 2016, the reliability model was only estimated and analyzed for the 2016 data.

The predictor variables across all training and testing data had a maximum absolute correlation of 0.71 between % of dots at starting zoom and % of dots in false color. The second highest correlation was between % of dots at starting zoom and Interpreted TCC. Interpreted TCC had a correlation greater than 0.5 % of dots in natural color display, as well as with % of dots in false color display. % of dots in natural color display had a correlation greater than 0.5 with % of dots zoomed out. All of these correlations were positive, indicating direct relationships between variables

Figure 8: Correlation matrix showing a subset of variables which have a correlation greater than |0.5| with at least one other variable in the subset.

(Figure 8) Since no variable correlations were greater than or equal to 0.8, the most common cutoff for determining multi-collinearity, no variables were removed from our models [43] as cited by [44].

Comparison	Reliable (Observed)	Unreliable (Observed)	Predicted Totals
Reliable (Predicted)	25.4	7.5	32.9
Unreliable (Predicted)	9.4	3098.4	3107.7
Observed Totals vs. Expert	34.8	3105.8	99.5%

Table 12: Average number of interpretations predicted and/or observed to be "reliable" and "unreliable." Average Overall % accuracy across all 5000 models is in the bottom right cell.

The Monte Carlo modeling framework produced 5000 models, each with their own class predictions ("reliable" or "unreliable") for all plots, out-of-bag error estimates, and variable importance rankings. The out-of-bag accuracy of all classification models are normally distributed with a mean of 99.5%, a minimum model accuracy of 99.0%, and a maximum model accuracy of 99.8%. The aggregate confusion matrix for these models shows that, on average, 73% of interpretations that were observed to be "reliable" were also predicted to be "reliable", and >99% of plots that were observed to be "unreliable" were also predicted to be "unreliable" (Table 12).

Reliability scores by interpretation (obtained from percent of models which predicted an interpretation to be "reliable") have a binomial distribution with a very large peak representing interpretations that were never predicted to be "reliable" and a much smaller peak representing interpretations that were almost always predicted to be "reliable". Only 75 interpretations have a perfect reliability score of 100%. The mean reliability score is 2.62%.

The final reliability classification separates all interpretations into "reliable" and "unreliable" classes using these scores. If more than 50% of models classified an interpretation as "reliable", the final classification was "reliable". If less than or equal to 50% of the models classified an interpretations "reliable", the final classification was "unreliable". The final class of "reliable" interpretations have a higher level of agreement with expert's interpretations than those interpretations classified as "unreliable". Where the full set of crowd data generally shows crowd workers under-estimating TCC values compared to experts, a larger proportion of the interpretations deemed "reliable" are close to the 1:1 line or above the 1:1 line (Figure 3).

Variable importance scores, based on the rank score method described in Schroeder, et al. (2017) demonstrate that *the number of plots interpreted* is the most significant variable in determining the reliability of a plot interpretation [42]. The second most important variable overall was *the % of dots interpreted with a false color display*, which was ranked most important for 90 models, second most important in 4599 models, and lower rankings for 311 models, with an overall score of 44733. The *% of dots interpreted with a natural color display* and *the interpreted TCC value* are closely ranked, with overall scores of 36531 and 35915, respectively. Categorically, two out of four of the interpretation-scale variables (*Number of plots interpreted* and *Number of HIT's available*) have importance scores ranked in the top five, as well as two out of four of the dot-level interpretation variables (*% of dots interpreted using false color display* and *% of dots interpreted using natural color display*). One of the TCC variables was also ranked in the top five (*interpreted TCC*). Neither of the two ecological plot variables have importance scores ranked in the top five, however *level-2 ecoregion* was the lowest ranked variable in the top half (Figure 9).

Pairs of crowd and expert interpreters were randomly selected using the bagging approach. Selected interpretations were evaluated for correlation, and to test whether data came from the same distribution. These metrics were calculated using all available plots for the 2011, 2016, and full set of data, to identify different patterns across the dates interpreted. There were 477 plots that both

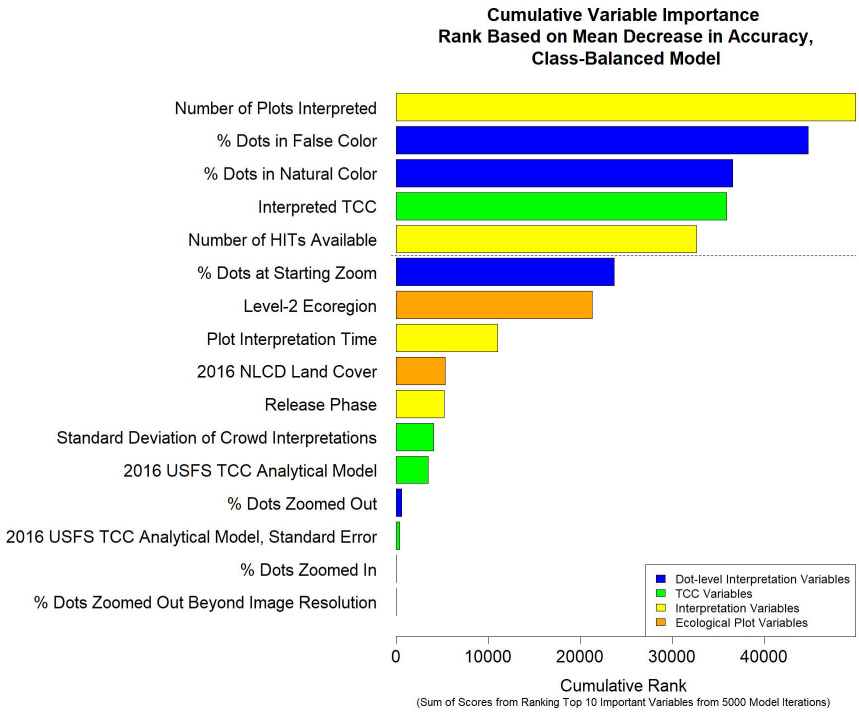

Figure 9: Variable importance scores, calculated using a rank-score method which assigns points to the top ten most important variables in each of the 5000 models and adds them together (2016 only). Dashed line indicates the top five most important variables across all models. These scores were derived from the mean decrease in accuracy metric supplied by the randomForest function in R.

Year	n Interpretations	Weighted r	Weighted d
2011	477	0.19	0.30
2016	477	0.23	0.31
All	954	0.21	0.30
Reliable (2016)	41.2	0.53	0.23*
Unreliable (2016)	434.9	0.21	0.32

Table 13: Weighted Pearson's correlation coeffiecients (r) and weighted Kolmogrov-Smirnov indices (d) between all crowd and experts for 2011 and 2016 and all dates, and between "reliable" crowd and experts and "unreliable" crowd and experts for 2016. An asterisk next to the value of d (*) indicates that the two sets of data come from the same distribution as demonstrated by a p-value of greater than or equal to 0.05. Records with NA values were removed from calculations. All comparisons are calculated using bootstrapped interpretations (one at each plot) so *n Interpretations* is also equivalent to the number of plots represented.

Year	n Interpretations	df	Sum Sq	Mean Sq	F Value
2011	477	1	45.2	45.2	112.3
2016	477	1	40.4	40.4	109.6
All	954	1	85.3	85.3	221.0
Reliable (2016)	41.2	1	0.34	0.34	1.12*
Unreliable (2016)	434.9	1	42.7	42.7	116.6

Table 14: Weighted ANOVA between all crowd and experts for 2011 and 2016 and all dates, and between "reliable" crowd and experts and "unreliable" crowd and experts for 2016. An asterisk next to the value of F (*) indicates that the two sets of data come from the same distribution as demonstrated by a p-value of greater than or equal to 0.05. All comparisons are calculated using bootstrapped interpretations (one at each plot) so *n Interpretations* is also equivalent to the number of plots represented.

crowd and expert workers interpreted for 2011 and 2016 (adding to 954 single-date interpretations). "Reliable" interpretations accounted for a smaller proportion of the interpretations in all of the 5000 bootstrapped samples, accounting for about 8.6% of each sample on average.

When comparing all crowd and expert interpreters, 2016 data had a slightly higher correlation compared to 2011, but data from individual years and both years combined showed a low correlation between crowd and expert interpreters. Kolmogrov-Smirnov test and ANOVA tests for 2011, 2016, and data from both years combined show that the crowd and expert interpretations are significantly different, as evidenced by p-values less than 0.05. The mean squared difference is the same for all comparisons because the degrees of freedom is 1 (since there are only two categories, 'crowd' and 'expert' present in our linear model). For "unreliable" 2016 interpretations, many of the statistics follow the same patterns as the full set of 2016 interpretations, with even lower correlations between crowd and expert interpretations and p-values of less than 0.05, indicating the groups are significantly different. However, "reliable" 2016 data show a stronger (though unremarkable) correlation between crowd and expert interpretations, and the Kolmogrov-Smirnov test and ANOVA tests show p-values of greater than 0.05, indicating that "reliable" crowd interpretations and expert interpretations are not significantly different.

3.5. Important Variables

3.5.1. Number of Plots Interpreted

Number of plots interpreted is ranked the most important variable overall in determining crowd interpretation reliability in photointerpretation of TCC compared to experts. It was ranked most

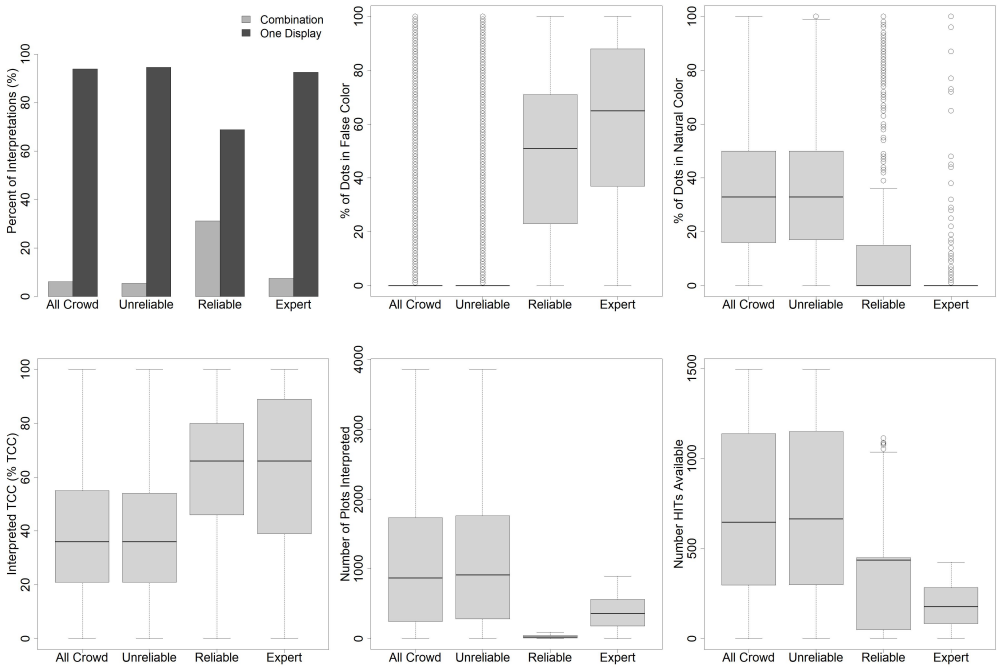

Figure 10: Summary plots for important variables, including a bar plot showing percent of interpretations which use a single display vs. a combination of displays (top left), and box plots showing percent of dots in false color display for each interpretation (top center), percent of dots in natural color display for each interpretation (top right), interpreted TCC (bottom left), number of plots interpreted (bottom center),and number of HITs available at the time of interpretation (bottom right) by interpretation class (all crowd, expert, "unreliable" crowd, and "reliable" crowd).

Variables	All Crowd Interpretations	Unreliable Crowd Interpretations	Reliable Crowd Interpretations	Expert Interpretations
Number of Plots Interpreted	1095.3	1124.8	30.8	378.9
% Dots in False Color	5.8	4.7	46.5	62.0
% Dots in Natural Color	34.8	35.3	17.0	1.4
Interpreted TCC	39.6	39.0	64.3	62.9
Number of HITs Available	707.7	715.6	398.9	186.9

Table 15: Weighted mean values of the top five most important variables by interpreter class for 2016 only.

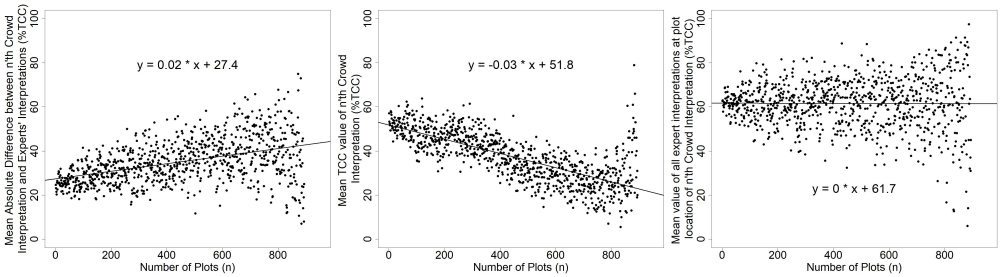

Figure 11: Scatter plots demonstrating an increase in error and decrease in TCC value with additional plots as interpreted by crowd workers. (left) Absolute error of the n'th crowd interpretation, averaged across all workers who interpreted n plots vs. the number of plots interpreted (n). (center) Mean TCC value of the n'th crowd interpretation, averaged across all workers who interpreted n plots vs. the number of plots interpreted (n). Please note that HITs were assigned randomly and not in sequence, therefore the n'th plot for one worker is likely to fall in a different location than the n'th plot for another worker. Therefore, this data represents the average value of crowd interpretations across multiple locations. (right) Mean TCC value of the expert interpretations at the plots locations corresponding to the n'th crowd interpretation, averaged across all available expert interpretations at plots that were interpreted as any crowd worker's n'th interpretation vs. the number of plots interpreted (n). Dots displayed in all three plots represent interpretations of 2016 imagery at plots that were interpreted by experts and crowd workers.

important variable for 4910 out of 5000 models, and second for the remaining 90 models. The number of plots interpreted is higher among all crowd interpretations and "unreliable" crowd interpretations, and even lower among "reliable" crowd interpretations than among experts who were only expected to complete 502 of the 1998 total plots theoretically available to crowd workers. This indicates that "reliable" interpretations were more likely to occur early on in a worker's interpretation sequence (Table 15 and Figure 10).

When the number of plots interpreted (n) is plotted against the mean absolute difference between the n'th crowd interpretation and all available expert interpretations, the trend line has a positive slope of 0.02. Based on this linear model, the predicted error of the first interpretation all workers interpreted is 27.4%, while the predicted error of the 400th plot interpreted is 35.4%, and the predicted error of the 894th plot (the maximum value of number of plots interpreted by a crowd worker across plots that were also interpreted by experts) interpreted is 45.3% (Figure 11).

3.5.2. % of Dots in False Color Display and % of Dots in Natural Color Display

Two image display options were available to interpreters, natural color, which is a composite of visible bands (R = red band, G = green band, B = blue band), and false color, which uses the near infrared band along with the red and green visible bands (R = Near Infrared Band, G = Red Band, B = Green Band). The variables "% of Dots in False Color" and "% of Dots in Natural Color" refer to the percent of all 100 dots in a each single-date plot interpretation that were clicked in each display. For expert interpretations, the majority of plots were interpreted using the false color display, while crowd workers usually interpreted images in a natural color display, which was the default. Like expert interpretations, "reliable" crowd interpretations utilized the false color display more than "unreliable" crowd interpretations, but "reliable" crowd workers were more likely to use a combination of displays, changing display in the course of 31.1% of their interpretations compared to all other groups who changed display settings in the middle of an interpretation less than 10% of the time (Table 15 and Figure 10).

3.5.3. Interpreted TCC

Interpreted TCC represents the resulting value of the interpreted plots. The plot interpretation value for all crowd workers was significantly lower than experts with an average of 39.6% TCC compared to 62.9% TCC for experts at the same plots. "Reliable" crowd interpretations have a higher TCC value on average than expert interpretations, with a mean of 64.3% TCC, and interpreted TCC values from experts have a wider distribution than "reliable" interpretations. "Unreliable" crowd interpretations have the lowest value, with an average of 39.0% TCC. (Table 15 and Figure 10).

3.5.4. Number of HITs Available

The number of HITs available represents the actual number of available hits for each worker at the time of interpretation. "Unreliable" crowd interpretations were performed when there was a larger number of HITs available than "reliable" interpretations. For 'unreliable' and all crowd interpretations, box plots show a wider, more balanced distribution of values than they do for "reliable" crowd interpretations, which has a skewed distribution. The majority of "reliable" crowd interpretations were completed while there were fewer than the mean number of HITs available for that group. Expert interpretations show a narrow, balanced distribution, around a fairly low mean number of HITs available, reflecting the smaller number of expert interpreters who accepted fewer assignments overall with no competition for assignments (Table 15 and Figure 10).

4. Discussion

4.1. The Wisdom of Crowds

Voting mechanisms are common in the analysis of crowdsourced observations. A wide range of studies demonstrate that, in aggregate, group estimates are better than individuals, from Francis Galton's 1907 example of a weight-judging competition where a group of farmers' median estimate was 1% of an ox's actual weight on average (but with an individual probable error of 3.1%) [45]. Other, more contemporary studies have also shown similar results with crowdsourced data. For example, Foody, et al. (2018) used data from Geo-Wiki.org to show that consensus approaches such as simple voting or weighted voting can improve land cover classifications over those generated from individual observations. The researchers explored several methods including basic majority voting as well as methods to remove low-accuracy contributions or emphasize high-accuracy contributions. They found for their study, that some strategies for the removal of inaccurate data improved results more than weighting for the emphasis of high-accuracy interpretations and simple voting, but that

simple voting provided higher accuracy (76.6%) than the most accurate individual worker (73.9% accuracy) [46]. Hsueh, et al. (2009) also show that multiple noisy crowd-sourced observations can be used to generate better observations to train supervised models used to identify sentiment in political blog posts. Their study compared aggregate data from workers recruited from Mechanical Turk to individual responses from experts in a lab setting. While the annotations from individual workers were noisy, in aggregate, crowd workers provided usable training data [47].

Crowd wisdom can be impeded by various factors, such as estimation bias or social influence [48]. Salk et al. (2016) found that, while a volunteer's cropland classifications were in high agreement (94.4% overall) with the majority rule aggregation of those classifications, neither individual or aggregated classifications were closely related to experts'. In this study 22% of images were mis-categorized using a majority voting technique, with an increase in mis-classification for more difficult images [49].

Figures 2, 3, and 5 demonstrate that crowd interpretations were systematically lower than expert interpretations, both individually and in aggregate using the mean and median of crowd interpretations submitted from each HIT. Figure 5 also shows that, at high expert-interpreted TCC values, mean crowd interpretations tended to be lower than median crowd interpretations. This could be further evidence a pattern observed in figure 4 and elsewhere in the data collected, where many of the HITs requested received one or more very low value interpretations as compared to the other assignments submitted by crowd workers and compared to the interpretations submitted by experts. These very low value interpretations caused the mean of crowd TCC interpertations to be generally lower than the median values in plots with high expert-interpreted tree canopy cover.

These results are unexpected in the context of the wisdom of crowds, however it should be noted that both the structure of our HIT request and the incentive structure of crowdsourcing in general would lead to a dataset that was biased toward lower crowd TCC values. Our HIT only required workers to click on dots they identified as landing on tree canopies, and workers were paid for each assignment submitted with no regard for the time invested. Crowdsourcing is commonly associated with task-defined payments rather than time-defined payments, so it follows that spending less time on task by submitting systematically lower-value interpretations would yield a higher wage. Additionally, assignments on Mechanical Turk are usually distributed on a first-come, first-serve basis and the number of assignments available for a given HIT are finite, thus working too slowly could reduce the number of HITs a worker has access to and, in turn, their potential earnings. Finally, the competition for available assignments could also lead a worker to submit assignments for many hours, possibly working through fatigue to ensure they had access to as many HITs as possible. Some of these time-money motivations are also mentioned in the literature related to Mechanical Turk and various methods have been explored to improve quality and reduce the influence of workers who prioritize speed over quality [50], [51].

The visible patterns in Figures 5 and 4 may indicate different methods for reducing time-on-task such as submitting a number of fully-interpreted assignments followed by a few empty uninterpreted assignments versus clicking only a portion of the dots that landed on trees in each assignment. Patterns characterized by progressively lower TCC interpretations or intermittent 0% TCC observations at a higher frequency than would be expected based on the distribution of the data were observed in several individual crowd workers' interpretations, though they could also be explained by technical issues visualizing the assignment or inexperience interpreting this type of imagery. These patterns may also indicate broader differences in the interpretation approach taken between crowd and expert workers or possibly the affects of fatigue on crowd workers.

4.2. Interpretation Consistency and Reliability

This study found that crowd interpretations were generally unreliable with respect to expert interpretations. Visual inspection of various interpretations revealed that crowd workers had some difficulty distinguishing trees from vegetated background, and identifying canopy gaps from shadows in dense canopy areas. Additionally, crowd workers had difficulty resolving separated tree crowns when they were situated among heavily vegetated shrublands.They also appeared to miss contextual cues which could be determined by zooming out from a plot to see other trees for comparison. This resulted in over-sampling croplands and shrubslands as tree canopy cover. By contrast, crowd interpreters were successful identifying trees with separated canopies on bare, un-vegetated soil.

It is also clear that some workers clicked randomly and erratically, not following any pattern related to the vegetation on the landscape. For some individuals, scatter plots of their interpretations compared to expert interpretations showed most of their interpretations fell between 10 and 30% TCC with no association with expert interpretations. Some scatter plots for individual workers show that 0% TCC interpretations were prevalent even when expert workers interpreted high values of TCC, possibly an indication they they occasionally submitted empty interpretations to increase their payout. Quality of crowd work has declined on Mechanical Turk in the last few years, from the rise of bots [33], [34]. Starting in 2018, research states that bots have become more prevalent (though our captcha task and the complexity of our HIT makes bots unlikely respondents for our study), as well as 'troll' workers who provide insincere responses, and workers masking their location information to participate in geographically restricted studies (this study was not geographically restricted). For example, Ahler, Roush, and Sood (2021) cited an increase in untrustworthy responses from 6-13% in 2015 to 25% in 2018. These combined interpretations characterized by misrepresented location (i.e., using servers to mask IP address), and insincere responses (which comprised 5-7% of responses in 2018) [35]. Another study by Chmielewski and Kucker (2019) performed a 4-wave survey data collection over the summer of 2018 and used response-validity indicators to ensure the quality of responses. They showed that in wave 1, only 10.4% of respondents were flagged. This increased to 13.8% in wave 2. In wave 3, the first of the 'post-concern' waves, they reported 62.0% of respondents were flagged. In wave 4, this was slightly reduced to 38.2% [52].

Some cases also show a sharp decline in tree canopy cover value and interpretation time with interpretation sequence, perhaps an indication of interpretation fatigue (Figures 4 and 6).

Consistency between groups of crowd interpreters and expert interpreters was also a challenge. The inter-comparison of expert data shows that each worker's interpretations follow a different distribution and the experts with the highest level of agreement interpreted 53% of plots to be within 10% TCC of one another. While the maximum weighted correlation is 0.82, the minimum correlation is only 0.64, suggesting that experts had difficulty with many plot interpretations and lacked consistency in some instances (Table 10). Some reasons for this inconsistency can be seen in the scatter plots comparing their interpretations. Plots that compare expert 1's interpretations to experts 2 and 3 show that expert 1 often considered a plot to have 0% TCC while other experts considered that plot to have greater than 0% TCC. There are also instances where expert 1 interpreted 100% TCC at a plot, but other experts interpreted less than 100% TCC at that plot. This is likely due to a disagreement between experts relating to what features constitute trees and the distinction between canopy gaps that reveal sub-canopy tree cover or ground (Figure 7).

While more interpretations among experts at each plot could improve our understanding of their consistency, several methodological decisions may also improve interpretation consistency among interpretations in future studies. First, our experts were not trained using a consistent protocol as seen in Jackson et al. (2012), where between 63-90% of measurements from expert interpreters were within 10% agreement with other experts [18]. Additional differences exist between these studies. For example, our experts were not necessarily familiar with the locations they were interpreting,

and interpreted plots across a broader region. Second, our workers did not interpret plots that were modeled to have 0% TCC based on the 2016 USFS Analytical Tree Canopy Cover product product. Finally, tree canopy cover is scale-dependent [14], [24]. The ramifications for this work being that, though the definitions of tree and the photointerpretations methods that produced this dataset are similar to those for the original training data for the Forest Service TCC model, the interpretations are performed at different scales, yielding different results. Interpretations from this study were performed using a 100-dot grid that covered a 30x30 meter plot, while the TCC models training data used a 105-dot grid over a 90x90 meter plot [14], while the resolution of the base imagery (NAIP) remained the same. This change could greatly impact all interpreters' ability to resolve features across the landscape, leading to reduced precision in interpretations and larger differences.

We also looked at the standard deviation in crowd and expert interpretations at each plot to understand the level of disagreement. Weighted mean standard deviation values were 16.8% in 2011 and 16.4% in 2016 for experts, indicating that the consistency of expert tree canopy cover interpretations is relatively low. The standard deviation of crowd workers is higher, as expected, ranging between 20.5-23.3% for different subset of data and years. Confidence intervals are much wider for experts than for crowd workers, likely due to the smaller number of interpretations at each plot (up to 3 interpretations for experts compared to 7-9 for crowd workers). The weighted mean confidence interval width of 35.6 for 2011 and 34.6 for 2016 for experts and 22.7-25.8 for crowd workers across years and subsets of data indicates that this study lacked consensus among experts and among crowd workers. Between groups of workers there is very little agreement. The weighted correlation between crowd and expert interpretations is 0.21 overall. Even "reliable" crowd interpretations only have a weighted correlation of 0.53 with experts.

Additionally, and perhaps more notably, our expert interpreters had varying levels of training in photointerpretation, and very different experiences with the subject of this study, tree canopy cover. Crowd workers were provided a document to reference and guide their training and were required to pass a multiple-choice qualification. A consistent and rigorous training protocol for expert and crowd interpreters could improve results going forward.

4.3. Interpreter Fatigue

The most important variable for determining plot reliability was the number of plots interpreted. This is no surprise given the discussion of worker 'fatigue' in the literature related to crowdsourcing and photointerpretation ([53], [54] and [55] as referenced by [53]). The longer a worker stays on task, the more the quality of their work tends to decline. Figure 11 demonstrates that the error of an interpretation increases as a worker interprets more plots, thus the longer a worker stays on task, the more the value of their interpretations tends to diverge from experts'. The fitted model in that figure suggests that the error increases by 0.02 for every plot interpreted. If this relationship holds in phase 3, then the expected absolute difference between crowd and expert interpretations for the 1927th plot interpreted (the maximum number of plots interpreted by any crowd worker) would be 65.9%. The spread of the scatter plot widens as the number of plots interpreted increases. This is likely due to the fact that substantially fewer workers interpreted a larger number of plots. For example, nine workers interpreted the maximum number of plots that were also interpreted by experts (894 plots) while 49 workers interpreted 1 plot.

4.4. Display

The second most important variable in our reliability model was the % of dots that were clicked using the false color composite and the third most important variables in our reliability model was the % of dogs that were clicked using the natural color composite. These variables represents

a clear difference between crowd and expert interpreters. For experts the majority of plots were interpreted using the false color display, while crowd workers usually interpreted images in a natural color display.

Since "reliable" interpretations and expert interpretations were typically performed using the false color display it is possible that the visible differences between the natural color imagery and the false color imagery led to the different interpretations of TCC at the same plots. In general, crowd workers under-sampled areas where experts indicated there should be high tree canopy cover. There are examples of this sort of behavior in the literature, where inexperienced workers view shadows on other canopies as gaps to the ground (Avery, 1969, page 30) [21]. The false color display, which uses the near infrared band to distinguish strong vegetation signatures, even in some cases when they are obstructed by shadow could have resulted in higher TCC values at plots with a lot of gaps and shadows. We see examples of this in this study as well (See Figure 6).

While the false color display was provided to workers, and information regarding the usefulness of false color imagery for TCC interpretation was described in training, workers rarely used the false color composite imagery. The natural color display was the default setting when workers started a HIT, as we expected workers to be better acquainted with this view. In order to change the display, the map would need to be reloaded which would take several seconds depending on web traffic and an individual's internet connection, so perhaps those workers who were motivated to work quickly chose to use the default display setting. Additionally, Campbell and Wynne (2011) suggest that the use of non-visible bands can be a source of confusion to non-expert interpreters [2]. Thus it's possible that workers who provided "reliable" interpretations were more experienced interpreting imagery and were therefore more comfortable with the false color imagery, or that those who provided "unreliable" interpretations were more comfortable with the natural color imagery. Since false color and natural color imagery highlight different landscape characteristics, more consistent interpretations may have been obtained if all interpreters used the same display.

4.5. Potential Earnings

The fifth-most important variable, the number of HITs available, was included in our assessment to get an understanding of how financial motivators were affecting workers' attention to their interpretations. For workers on mechanical turk, the number of HITs available could represent the potential earnings from working a request (which can be estimated by multiplying the payment by the number of HITs available).

Access to these potential payments depends on a worker's ability to complete assignments quickly since working a HIT on Mechanical Turk is a no-sum game. There are a finite number of workers who can participate in each HIT (determined by the number of assignments available) and once a worker completes to the last assignment available for a HIT, then no other worker gets the opportunity to work that HIT. "Reliable" interpretations were often completed toward the end of a request when there were relatively few HITs available, and less competition to complete those HITs.

High potential earnings for a HIT may create a motivation to move through HITs faster in order to get to the potential pay out before another worker. If a worker interprets HITs in a request slowly, many of the available HITs could close due to all assignments being completed while that worker is completing a single assignment themselves, and their potential payment declines sharply. In our study, plot interpretation time was not as significant a determinant to the reliability of an interpretation as the number of HITs available. This could indicate that the number of HITs available is an indication of the type of motivation a worker has for participating in the HIT, but that the actual interpretation time is unaffected. Rogstadius, et al. (2011) demonstrates that intrinsic motivators can improve the quality of a crowdsourcing task and that extrinsic motivators

(like financial incentives) could actually create a 'crowding out' effect where workers have less interest in tasks they are paid to perform [56].

A complicating factor in drawing conclusions about plot interpretation time is that our interface measured time to the nearest minute, and since these interpretations took crowd workers only 2.02 minutes on average to complete, there may not have been enough granularity in the measurement of time to show a relationship to reliability.

4.6. Interpreted TCC

Interpreted TCC was the fourth most important variable. Crowd workers consistently undersampled TCC in their plot interpretations compared to experts, but "reliable" crowd interpretations tended to have higher TCC values, with an average value that exceeded the expert average.

TCC values could be associated with worker fatigue. Our interface only required workers to click on a dot to indicate that it landed on a tree canopy. Dots that landed on non-tree features could be left alone, so as a worker grew tired they may not have been as diligent about checking and clicking every dot that fell on a tree.

Lower TCC values could also be associated with display settings, as false color display could potentially reveal vegetation within gaps or shadows that would be visible in natural color imagery (as discussed above). This variable was also somewhat correlated with the display variables, with a correlation of 0.56 with % of dots interpreted in false color display and 0.59 with % of dots interpreted in natural color display, possibly suggesting that the importance of these three variables is underestimated as they serve similar roles in partitioning the data (Figure 8).

Finally, a workers potential earnings could also motivate them to work more quickly. Higher TCC interpretations cost the worker more time and therefore represent lower earnings.

4.7. Recommended Improvements

Congalton and Green (2009) also suggest that measures can be taken during the planning phase of a study to improve the consistency between field observations and photointerpretation for image classification and accuracy assessment, such as consistency in training and data collection, a rule-based process of interpretation, and ground validation at a subset of plots [13].

An interpreter could also have various personal qualities that could improve photointerpretation performance. Van Coillie et al. (2014) cites a long history of psychological studies related to remote sensing image interpretation. The earliest examples relate to 'vigilance' or fatigue effects observed in WWII operators watching radar screens [53]. In these studies, fatigue effects result in a drop in performance over time [55] as referenced by [53]). The attitudes of an operator (pessimism or optimism) has also been shown to affect their response to fatigue mitigation training [54]. Research was also reviewed that demonstrated the importance of task-specific training to create higher levels of agreement between interpreters [57], and that training for systematic methods of interpretation as opposed to random search caused higher performance [58]. Van Collie et al. (2014) measured these effects using photointerpretation by enlisting volunteers to digitize lamp posts, water bodies, road networks, olive trees, olive parcels, and vine rows from remote sensing imagery. They tested various psychological variables as well as external variables and their association with interpreter performance. The results demonstrate that desire to obtain good results, confidence in interpretations, faith in the performance of the other volunteers, experience, the amount of time the interpreter was willing to spend, and the number of visual working memory representations the interpreter could retain are indicators of interpreter performance. These human factors were responsible for 26% of inter-individual differences in performance. They also cited that external factors that influenced performance included busy or noisy work environment (which had a negative impact on work performance), and they observed a clear fatigue effect [53].

There are possible ways to improve this study for future work. The biggest improvement would be systematic quality control throughout the study. We performed thorough qualitative analysis on the first phase of interpretations, but this study would have benefited from additional assessments throughout. These could take the form of review by expert individuals, comparison to expert interpretations, or occasional repeated plot interpretations by the same workers to evaluate their consistency with themselves. Requiring workers to review and certify their work using a visual inspection could also be helpful. Yu, et al (2014), suggest a combination of "social" and "learning strategies" to increase intrinsic motivation to incentivize higher quality work and improve participant engagement [59]. This means establishing a clear training goal or skill-development outcome for crowd participants to work toward and also enabling a collaborative or social aspect to participation, such as a chat window or a way of comparing interpretations with other workers. Their findings show that learning strategies improved data quality while social strategies improved participant engagement.

Requiring workers to click every dot in a plot would require every worker to evaluate each dot, perhaps using an interface that moved from dot to dot at a finer scale. This recommendation is an example of a 'defensive design' element, ensuring that 'cheating' at the task is not easier than doing it well [60]. This could also ensure that workers spent more time on task, reducing the incentive to speed through tasks and make more money. The increased attention required per assignment could also reduce the likelihood of interpretations that don't follow the prompt, such as those where dots are randomly clicked without following the pattern in the image or empty interpretations submitted for images that are densely forested.

Payment should reflect time on task and a study should be conducted to assess the required time using a group of non-expert interpreters. We underestimated the time it would take a non-expert worker to complete an interpretation in our study significantly,predicting that most interpretations could be completed in under 2 minutes. This resulted in an over-estimation of the average hourly wage of our workers. This time was estimated using a few complex examples from expert interpreters, with the assumption that interpretation time would be significantly reduced in areas with sparse vegetation. Feedback from workers as well as the interpretation times reported in this study communicated that this payment was too low for a task of this complexity and that we underestimated the time required to make detailed interpretations. While other studies [4], [5] suggest that wage did not have a strong influence on interpretation quality, efforts should be made to estimate a fair wage based on time required to complete "reliable" crowd interpretations if similar studies are to be attempted in the future. Mason and Watts (2010) study the affect of payment on crowd performance. They observed an "anchoring" effect where workers who were paid more perceived the value of their work to be greater and were therefore not more motivated than workers who were paid less for the same task. They showed that a quota system resulted in better work for less pay, which would be the best option if it could be operationalized [5]. Additionally, based on worker's comments, with a higher wage we might expect to see better retention of workers who were willing and motivated to spend time and perform high quality interpretations with a higher wage. In future endeavors it would be wise to perform a separate study using non-expert workers to estimate the average time required for interpretation and in turn, payment.

Limiting the number of hits a worker could complete before a new qualification or by releasing hits in smaller groups could also overcome the motivation to speed through tasks since the number of HITs available would be smaller. This could also enable quality control on subsets of the data without long delays, reducing repercussions to the requester from negative online reviews about payment time. A smaller number of HITs requested per release would also result in more manageable file sizes for compiling and downloading data for analysis.

This study showed that most work was completed by crowd workers using the interface defaults.

Most crowd workers used the natural color display and interpreted images at the default scale. Changing these settings required the map to reload and render new information, which would lead to a longer interpretation time for workers. We recommend that a decision be made on best practices for interpretation, including both interpretation scale and display. These settings should be used in the interface as the default settings. This may increase the number of crowd interpretations that used the same settings as experts and, if so, could reduce the possibility that difference in interpretation between groups arose from the use of a different display.

Further contextual information may be helpful in improving these interpretations. For example, land cover information within the plot could be provided using the FIA data (which represents a different, confidential location nearby) or LiDAR point clouds could be supplied where available to provide additional information about vegetation height and structure. LiDAR could also be used to supplement interpretations or to provide quality control by differentiating canopy measurements from the ground and obtain estimates of stand-level % tree canopy cover. [8], [61], [62]

Not all of the interpretations in this study were unreliable. It is clear that many workers on Mechanical Turk participate with the intention of providing good work and getting paid for it. One way to improve work performed on Mechanical Turk could be to incentivize good work. Workers found to be more reliable could be offered increased qualifications and higher value tasks in the future. Bonuses for work quality are also common on Mechanical Turk.

Additionally, adding intrinsic motivators to the HIT design could improve worker retention and the quality of output work. In a review of approaches to quality control in crowdsourcing systems, Allahbakhsh et al. (2013) suggest that a combination of intrinsic and extrinsic incentives can increase the participation of honest workers in a crowdsourcing task. They further suggest that intrinsic motivators may have a more significant positive influence on data quality than extrinsic motivators [60]. Rotman et al. (2012) discusses the motivations of experts and non-experts in voluntary citizen science collaborations. They found that volunteers' interest was initially driven by egoism, basing the choice to participate on commonality with past experience, existing interests or hobbies, or to build career-related skills [63]. Goodchild (2007) also cites these type of motivators, such as self-promotion, personal use, and personal satisfaction as reasons volunteers contribute geographic information to various online mapping interfaces [64]. While the decision to contribute to a crowdsourcing request with Mechanical Turk is more often predicated on financial incentives [65], providing certificates for professional accolades, or advertising to specific communities could increase the number of participants driven by these intrinsic ego-driven motivators. Additionally, adding elements to "gamify" the existing crowdsourcing system could provide additional intrinsic motivation.

5. Conclusion

This study was aimed at identifying the factors that could predict whether an interpretation could be classified as "reliable" or "unreliable" and how those variables were related to reliability in order to make recommendations for future studies, both with regards to interface design and quality control.

We conclude that, while crowd workers and experts are both capable of measuring tree canopy cover using photointerpretation, agreement among experts is limited, and agreement among experts and non-experts is confounded by financial motivations, fatigue, and differences in the utilization of the user interface. In particular, our data shows trends that indicate a priority for timely interpretations over high quality interpretations, which is a supported by the incentive structure of Mechanical Turk. The reliability of crowd interpretations also appears to be affected by the potential earnings a worker perceives, with "reliable" interpretations occurring more often when

the number of HITs available (and therefore the potential payout) was lower. Fatigue affects were apparent as consistency between crowd and expert interpretations declined with the number of plots interpreted for crowd workers. Finally, crowd workers prioritized the default settings of our interface. Incidentally, these settings included a natural color composite which may be more intuitive for a non-expert worker to interpret, but would likely have less available information compared to the false color composite, which was less commonly used. Differences in interpretations arise from both the display settings used and each worker's level of experience interpreting color-infrared imagery and resolving differences between background and foreground in highly vegetated landscapes. These factors may have contributed to differences in quality and consistency among crowd workers compared to experts.

We make several recommendations in the event that this work were to be operationalized, utilizing crowd interpretations to obtain measures of tree canopy cover from photointerpretation. First, we recommend the implementation of increased quality control measures to identify and reduce interpretations that follow the prompt, such as those where dots are randomly clicked without following the pattern in the image or empty interpretations submitted for images that have greater than 0% TCC. We further suggest that workers be required to evaluate and click every dot in the grid, reducing the likelihood of incomplete interpretations. We also suggest a thorough analysis of required interpretation time for non-expert workers to ensure payment is fair and reasonable. Limiting the number of HITs per worker and reducing the number of HITs in each request could mitigate some of the fatigue affects we observed and reduce the perception of competition and the potential earnings from a HIT. To make interpretations more consistent, the user interface should be set up so that the ideal interpretation environment is the default setting since most crowd workers did not change these settings. Additionally, more contextual information about each plot could be provided (in the form of LiDAR data or land cover classifications) to improve a user's understanding of the landscape they're interpreting. Finally, incentives to reward high quality interpretation or quality contingent pricing [66] could yield better results.

While the our models show promise in being able to distinguish between "reliable" and "unreliable" interpretations, these insights are aimed at improving the interface to yield better measurements in future studies. Using models such as this one to filter interpretations or select preferred data would risk eliminating sincere interpretations or including erroneous data and would violate the assumption that the data were part of an equal probability sample. The result would therefore be inappropriate for use in model validation or model training. What's more, expert interpreters did not agree perfectly with one another during this exercise, highlighting the subjective nature and innate challenges involved in the photointerpretation of tree canopy cover and the risk of using an imprecise benchmark to select crowd interpretations.

6. Acknowledgements

This work was supported by the USDA National Institute of Food and Agriculture McIntire Stennis project entitled "Precision Forestry for Southern Pine Carbon Monitoring" and the overall research project was funded through by the United States Forest Service (USFS) under the grant "TCC 2021 Research and Development, 18-JV-11242305-035". Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the National Institute of Food and Agriculture (NIFA) or the United States Department of Agriculture (USDA). We would like to thank Susmita Sen and Muna Khatiwada for their contributions to the data collection for this study and we would like to acknowledge the contributions of the crowd workers who participated in our HIT as well as those who provided meaningful feedback.

References

[1] R. N. Colwell, K. E. Bradshaw, H. T. U. Smith, R Thoren, and C. A. J. von Drabbe, "Report of the commission vii (photographic interpretation) to the international society of photogrammetry," *Photogrammetric Engineering*, vol. 18, 375—451, Jun. 1952.

[2] J. B. Campbell and R. H. Wynne, "Image interpretation," in *Introduction to Remote Sensing*, 5th ed. New York, NY: Guilford Press, 2011, ch. 5, pp. 130–153.

[3] J. a. Porto de Albuquerque, B. Herfort, and M. Eckle, "The tasks of the crowd: A typology of tasks in geographic information crowdsourcing and a case study in humanitarian mapping," *Remote Sensing*, vol. 8, no. 859, 1–22, 2016. DOI: 10.3390/rs8100859. [Online]. Available: http://dx.doi.org/10.3390/rs8100859.

[4] L. Yu, S. B. Ball, C. E. Blinn, K. Moeltner, S. Peery, V. A. Thomas, and R. H. Wynne, "Cloudsourcing: Using an online labor force to detect clouds and cloud shadows in landsat images," *Remote Sensing*, vol. 7, no. 3, 2334–2351, 2015, ISSN: 2072-4292. DOI: 10.3390/rs70302334. [Online]. Available: http://dx.doi.org/10.3390/rs70302334.

[5] A. W. Mason and J. D. Watts, "Financial incentives and the "performance of crowds"," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108, 2010.

[6] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011, PMID: 26162106. DOI: 10.1177/1745691610393980. [Online]. Available: https://doi.org/10.1177/1745691610393980.

[7] S. Jennings, N. Brown, and D Sheil, "Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures," *Forestry: An International Journal of Forest Research*, vol. 72, no. 1, pp. 59–74, Mar. 1999, ISSN: 0015-752X. DOI: 10.1093/forestry/72.1.59. eprint: https://academic.oup.com/forestry/article-pdf/72/1/59/6746101/720059.pdf. [Online]. Available: https://doi.org/10.1093/forestry/72.1.59.

[8] Q. Ma, Y. Su, and Q. Guo, "Comparison of canopy cover estimations from airborne lidar, aerial imagery, and satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 9, pp. 4225–4236, 2017.

[9] C. Toney, J. D. Shaw, and M. D. Nelson, "A stem-map model for predicting tree canopy cover of forest inventory and analysis (fia) plots.," in *Forest Inventory and Analysis (FIA) Symposium 2008*, ser. e-Gen. Tech. Rep. RMRS-P-56CD, Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, 2009, 1—19. [Online]. Available: https://www.fs.usda.gov/treesearch/pubs/33381.

[10] P.-A. Esseen, K. U. Jansson, and M. Nilsson, "Forest edge quantification by line intersect sampling in aerial photographs," *Forest Ecology and Management*, vol. 230, no. 1, pp. 32 –42, 2006, ISSN: 0378-1127. DOI: https://doi.org/10.1016/j.foreco.2006.04.012.

[11] R. G. Congalton and G. S. Biging, "A pilot study evaluating ground reference data collection efforts for use in forest inventory," *Photogrammetric Engineering & Remote Sensing*, vol. 58, no. 12, pp. 1669 –1671, 1992.

[12] M. Hansen, R. DeFries, J. Townshend, L Marufu, and R Sohlberg, "Development of a modis tree cover validation data set for western province, zambia," *Remote Sensing of Environment*, vol. 83, no. 1, pp. 320 –335, 2002, The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring, ISSN: 0034-4257. DOI: https://doi.org/10.1016/S0034-4257(02)00080-9. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425702000809.

[13] R. G. Congalton and K. Green., *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. 2nd ed. Boca Raton, FL: CRC Press, 2009, ISBN: 978-1-4200-5512-2.

[14] J. W. Coulston, G. G. Moisen, B. T. Wilson, M. V. Finco, W. B. Cohen, and C. K. Brewer, "Modeling percent tree canopy cover: A pilot study," *Photogrammetric Engineering & Remote Sensing*, vol. 78, no. 7, pp. 715–727, Apr. 2012, ISSN: 2220-9964. [Online]. Available: https://www.fs.usda.gov/treesearch/pubs/40860.

[15] J. M. Derwin, V. A. Thomas, R. H. Wynne, J. W. Coulston, G. C. Liknes, S. Bender, C. E. Blinn, E. B. Brooks, B. Ruefenacht, R. Benton, M. V. Finco, and K. Megown, "Estimating tree canopy cover using harmonic regression coefficients derived from multitemporal landsat data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 86, p. 101 985, 2020, ISSN: 0303-2434. DOI: https://doi.org/10.1016/j.jag.2019.101985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243419302119.

[16] C. Toney, G. Liknes, A. Lister, and D. Meneguzzo, "Assessing alternative measures of tree canopy cover: Photo-interpreted naip and ground-based estimates.," in *Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists.*, ser. e-Gen. Tech. Rep. SRS-157, Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, 2012, 209—215. [Online]. Available: https://www.fs.usda.gov/treesearch/pubs/41009.

[17] S. Goeking, G. Liknes, E. Lindblom, J. Chase, D. Jacobs, and R. Benton, "A gis-based tool for estimating tree canopy cover on fixed-radius plots using high-resolution aerial imagery.," in *Moving from Status to Trends: Forest Inventory and Analysis (FIA) Symposium 2012*, ser. Gen. Tech. Rep. NRS-P-105, Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station, 2012, 237—241. [Online]. Available: https://www.fs.usda.gov/treesearch/pubs/42752.

[18] T. A. Jackson, G. G. Moisen, P. L. Patterson, and J Tipton., "Repeatability in photo-interpretation of tree canopy cover and its effect on predictive mapping.," in *Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists.*, ser. e-Gen. Tech. Rep. SRS-157, Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, 2012, 189—192. [Online]. Available: https://www.fs.usda.gov/treesearch/pubs/41006.

[19] J. Tipton, G. Moisen, P. Patterson, T. A. Jackson, and J. Coulston, "Sampling intensity and normalizations: Exploring cost-driving factors in nationwide mapping of tree canopy cover.," in *Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists.*, ser. e-Gen. Tech. Rep. SRS-157, Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, 2012, 201—208. [Online]. Available: https://www.srs.fs.usda.gov/pubs/41008.

[20] R. Riemann, G. Liknes, J. O'Neil-Dunne, C. Toney, and T. Lister, "Comparative assessment of methods for estimating tree canopy cover across a rural-to-urban gradient in the mid-atlantic region of the usa.," *Environ Monit Assess*, vol. 188, no. 297, pp. 1 –17, 2016. DOI: https://doi.org/10.1007/s10661-016-5281-8.

[21] T. E. Avery., *Forester's Guide to Aerial Photo Interpretation.* Agriculture Handbook No. 308. Washington, D.C.: U.S. Department of Agriculture Forest Service., 1969.

[22] L D. Schepaschenko, M. L. See, J. F. Bastin, D. Mollicone, N. E. Tsendbazar, L. Bastin, I. McCallum, J. C. L. Bayas, A. Baklanov, C. Perger, M. Dürauer, and S. Fritz., "Recent advances in forest observation with visual interpretation of very high-resolution imagery.," *Surveys in Geophysics*, vol. 40, pp. 839–862, 2019.

[23] S. Spurr, *Photogrammetry and photo-interpretation. With a section on applications to forestry.* New York, USA: The Ronald Press Company, 1960.

[24] J. W. Coulston, S. N. Oswalt, A. B. Carraway, and W. B. Smith, "Assessing forestland area based on canopy cover in a semi-arid region: a case study," *Forestry: An International Journal of Forest Research*, vol. 83, no. 2, pp. 143–151, Jan. 2010, ISSN: 0015-752X. DOI: 10.1093/forestry/cpp039. eprint: https://academic.oup.com/forestry/article-pdf/83/2/143/1394613/cpp039.pdf. [Online]. Available: https://doi.org/10.1093/forestry/cpp039.

[25] G. J. Brand, M. D. Nelson, D. G. Wendt, and K. K. Nimerfro., "The hexagonal/panel system for selecting fia plots under an annual inventory.," R. E. McRoberts, G. A. Reams, and P. C. V. Deusen, Eds., Gen. Tech. Rep. NC-213, St. Paul, MN: USA: Department of Agriculture, Forest Service, North Central Research Station., 2000, pp. 8–13.

[26] United States Department of Agriculture, Forest Service, *2016 usfs "analytical" tcc*, 2019. [Online]. Available: https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/.

[27] T. E. Avery and H. E. Burkhart, *Forest Measurements*, 4th ed. McGraw-Hill, 1994, ISBN: 9780070025561.

[28] F. Freese, *Elementary forest sampling.* Ser. Agriculture Handbook. Washington, DC: U.S. Department of Agriculture, Forest Service, 1962, vol. 232. [Online]. Available: http://www.fs.fed.us/fmsc/ftp/measure/cruising/other/docs/AgHbk232.pdf.

[29] W. G. Cochran, *Sampling Techniques*, 3rd ed. John Wiley & Sons, Inc., 1977, ISBN: 9780471162407.

[30] Esri and USDA Farm Service Agency, *Usa naip imagery: Natural color*, Revised in 2020, 2014. [Online]. Available: https://naip.arcgis.com/arcgis/rest/services/NAIP/ImageServer.

[31] Esri and USDA Farm Service Agency, *USA NAIP Imagery: Color Infrared*, Revised in 2020, 2014. [Online]. Available: https://naip.arcgis.com/arcgis/rest/services/NAIP/ImageServer.

[32] USDA Farm Service Agency, *NAIP Coverage 2002-2018*, 2018. [Online]. Available: https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/APFO/status-maps/pdfs/naipcov{_}2018.pdf (visited on 10/30/2018).

[33] C. Stokel-Walker, "Bots on amazon's mechanical turk are ruining psychology studies," *New Scientist*, 1—4, Aug. 2018, Accessed on 8/11/2018.

[34] A. J. Moss and L. Litman, *After the bot scare: Understanding what's been happening with data collection on mturk and how to stop it [blog post]*. Sep. 2018. [Online]. Available: https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/.

[35] D. J. Ahler, C. E. Roush, and G. Sood, "The micro-task market for lemons: Data quality on amazon's mechanical turk (working paper)," *Political Science Research and Methods*, pp. 1–39, 2021. [Online]. Available: http://www.gsood.com/research/papers/turk.pdf.

[36] F. Pozzi, T. D. Matteo, and T. Aste, "Expoential smoothing weighted correlations," *Eur. Phys. J. B*, vol. 85, no. 175, pp. 1–21, 2012. DOI: 10.1140/epjb/c2012-20697-x.

[37] Tim, *Such thing as a weighted correlation?* Cross Validated, Accessed: 2021-03-24. eprint: https://stats.stackexchange.com/q/222107. [Online]. Available: https://stats.stackexchange.com/q/222107.

[38] G. Ridgeway, *Email communication describing the appropriate equations corresponding to the weighted ks-test function.* Personal Communication, Apr. 2021.

[39] G. Ridgeway., *Fast doubly robust estimation*, version 0.9-3, 2021. [Online]. Available: https://rdrr.io/github/gregridgeway/fastDR/.

[40] J. Dewitz, *National Land Cover Database (NLCD) 2016 Products: U.S. Geological Survey data release*, 2019. [Online]. Available: https://doi.org/10.5066/P96HHBIE.

[41] L. Breiman, A. Cutler, A. Liaw, and M. Wiener., *Breiman and cutler's random forests for classification and regression*, version 4.6-14, 2018. [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/.

[42] T. A. Schroeder, K. G. Schleeweis, G. G. Moisen, C. Toney, W. B. Cohen, E. A. Freeman, Z. Yang, and C. Huang, "Testing a landsat-based approach for mapping disturbance causality in u.s. forests," *Remote Sensing of Environment*, vol. 195, pp. 230–243, 2017. DOI: 10.1016/j.rse.2017.03.033.

[43] W. Berry and S. Feldman, *Multiple Regression in Practice (Quantitative Applications in the Social Sciences)*. Thousand Oaks, CA: SAGE Publications, 1985.

[44] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar, "Multicollinearity in regression analyses conducted in epidemiologic studies," *Epidemiology (Sunnyvale)*, vol. 6, no. 2, pp. 1–20, 2016. DOI: 10.4172/2161-1165.1000227.

[45] F. Galton, "Vox populi (the wisdom of crowds)," *Nature*, vol. 75, no. 1949, pp. 450–451, 1907.

[46] G. Foody, L. See, S. Fritz, I. P. Moorthy, C. Perger, C. Schill, and D. Boyd, "Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data," *ISPRS International Journal of Geo-Information*, vol. 7, no. 3, p. 80, Feb. 2018, ISSN: 2220-9964. DOI: 10.3390/ijgi7030080. [Online]. Available: http://dx.doi.org/10.3390/ijgi7030080.

[47] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: A study of annotation selection criteria," Association for Computational Linguistics; Boulder, Colorado, 2009, pp. 27–35.

[48] A. B. Kao, A. M. Berdahl, A. T. Hartnett, M. J. Lutz, J. B. Bak-Coleman, C. C. Ioannou, X. Giam, and I. D. Couzin, "Counteracting estimation bias and social influence to improve the wisdom of crowds.," *J. R. Soc. Interface*, vol. 15, no. 20180130, pp. 1–9, 2018.

[49] C. F. Salk, T. Sturn, L. See, and S. Fritz, "Limitations of majority agreement in crowdsourced image interpretation," *Transactions in GIS*, vol. 21, no. 2, pp. 207–223, 2016. DOI: 10.1111/tgis.12194.

[50] C. Eickhoff and A. P. de Vries, "Increasing cheat robustness of crowdsourcing tasks.," *Inf Retrieval*, vol. 16, pp. 121–137, 2013.

[51] C. Eickhoff, C. G. Harris, and A. P. de Vries, "Quality through flow and immersion: Gamifying crowdsourced relevance assessments.," Portland, Oregon, 2012, pp. 871–879.

[52] M. Chmielewski and S. C. Kucker, "An mturk crisis? shifts in data quality and the impact on study results," *Social Psychological and Personality Science*, vol. 11, no. 4, pp. 464–473, 2019. DOI: 10.1177/1948550619875149.

[53] F. Van Coillie, S. Gardin, F. Anseel, W. Duyck, L. P. C. Verbeke, and R. R. D. Wulf, "Variability of operator performance in remote-sensing image interpretation: The importance of human and external factors," *International Journal of Remote Sensing*, vol. 35, no. 2, pp. 754–778, 2014. DOI: 10.1080/01431161.2013.873152. [Online]. Available: https://doi.org/10.1080/01431161.2013.873152.

[54] J. L. Szalma, P. A. Hancock, J. S. Warm, W. N. Dember, and K. S. Parsons, "Training for vigilance: Using predictive power to evaluate feedback effectiveness," *Human Factors: The Journal of Human Factors and Ergonomics Society*, vol. 48, pp. 682 –692, 2006.

[55] R Parasuraman, "Vigilance, monitoring, and search," in *Handbook of Human Perception and Performance, Volume 2: Cognitive Processes and Performance*, J. R. Boff, L. Kaufmann, and J. P. Thomas, Eds., New York: Wiley, 1986, 1–49.

[56] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic, "An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets," 1, vol. 5, 2011, pp. 321–328. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14105.

[57] J. S. Cooper, S. K. Mukherji, A. Y. Toledano, C. Beldon, I. M. Schmalfuss, R. Amdur, S. Sailer, L. A. Loevner, P. Kousouboris, K. K. Ang, J. Cormack, and J. Sicks, "An evaluation of the variability of tumor-shape definition derived by experienced observers from ct images of supraglottic carcinomas (acrin protocol 6658)," *International Journal of Radiation Oncology*Biology*Physics*, vol. 67, no. 4, pp. 972–975, 2007, ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2006.10.029.

[58] M.-J. J. Wang, S.-C. Lin, and C. G. Drury, "Training for strategy in visual search," *International Journal of Industrial Ergonomics*, vol. 20, pp. 101–108, 1997.

[59] L. Yu, P. André, A. Kittur, and R. Kraut., "A comparison of social, learning, and financial strategies on crowd engagement and output quality.," Baltimore, MD, USA, 2014, pp. 967–977. DOI: 10.1145/2531602.2531729.

[60] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Computing*, vol. 17, no. 2, pp. 76–81, 2013. DOI: 10.1109/MIC.2013.20.

[61] D Gatziolis., "Comparison of lidar- and photointerpretation-based estimates of canopy cover.," in *Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists.*, ser. e-Gen. Tech. Rep. SRS-157, Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, 2012, 231—235. [Online]. Available: https://www.fs.usda.gov/treesearch/pubs/41012.

[62] A. M. Smith, M. J. Falkowski, A. T. Hudak, J. S. Evans, A. P. Robinson, and C. M. Steele, "A cross-comparison of field, spectral, and lidar estimates of forest canopy cover," *Canadian Journal of Remote Sensing*, vol. 35, no. 5, pp. 447–459, 2009. [Online]. Available: https://doi.org/10.5589/m09-038.

[63] D. Rotman, J. Preece, J. Hammock, K. Procita, D. Hansen, C. Parr, D. Lewis, and D. Jacobs, "Dynamic changes in motivation in collaborative citizen-science projects," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, ser. CSCW '12, New York, NY, USA: Association for Computing Machinery, 2012, 217–226, ISBN: 9781450310864. DOI: 10.1145/2145204.2145238. [Online]. Available: https://doi.org/10.1145/2145204.2145238.

[64] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography.," *GeoJournal*, vol. 69, pp. 211–221, 2007. [Online]. Available: https://doi.org/10.1007/s10708-007-9111-y.

[65] C. Heipke, "Crowdsourcing geospatial data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 550 –557, 2010, ISPRS Centenary Celebration Issue, ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2010.06.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271610000602.

[66] J. Wang, P. G. Ipeirotis, and F. Provost, "A framework for quality assurance in crowdsourcing," no. CBA-13-06, 2013.

pi.bib

Chapter 4

Sampling methods for the accuracy assessment of continuous variables derived from remotely sensed data

Sampling methods for validating continuous variables estimated using remotely sensed data.

Jill M. Derwin[a], John W. Coulston[b], Valerie A. Thomas[a,*], Randolph H. Wynne[a]

[a] *Virginia Tech, Forest Resources and Environmental Conservation, Blacksburg, Virginia 24061 USA*
[b] *United States Forest Service, Southern Research Station, Blacksburg, Virginia 24061 USA*

Abstract

Data collection for the accuracy assessment of remote sensing models can be a costly endeavor. The careful selection of samples through stratifiedified sampling can improve the confidence in estimates of model accuracy compared to simple random sampling while using the same sample size. However, while best practices for sample selection and accuracy assessment for categorical data products have been established, there is little guidance in the literature for evaluating continuous variables. We demonstrate the utility of stratified sampling with optimal allocation for reducing the variance in estimates of model accuracy and compare it to simple random sampling, as well as stratified random sampling with equal allocation and with proportional allocation. Our method allocates samples based on prediction uncertainty, which can be derived from commonly used ensemble modeling approaches like random forest regression. Our analysis of sampling protocols includes three cases utilizing simulated data: one with normally distributed reference data, one with non-parametric reference data, and one with reference data derived from the expected relationship with an existing model. For each case, we draw 250 sets of samples with each sampling protocol and compare the distributions of sample estimates of the mean reference Y value and RMSE between protocols using measures of accuracy and precision. For the stratified random sampling protocols, when estimating RMSE the population was stratified using a pixel-wise estimate of prediction uncertainty and when estimating the mean of the reference Y we use $\widehat{Y}$ values. We also test the effect of using smaller sample sizes on the effectiveness of the various sampling protocols. We show that stratified random sampling with optimal allocation provides the most precise estimate of both the mean of the reference Y and the RMSE of the population. We also demonstrate that all sampling methods demonstrate reasonably accurate estimates on average. Additionally we show that, as sample sizes are increased with each sampling method, the precision generally increases, eventually reaching a level of convergence where gains in estimate precision from adding additional samples would be marginal. We also demonstrate a method to report the accuracy of continuous models for domains that are either regionally constrained or numerically defined based on other variables or the modeled quantity itself.

*Corresponding author. Tel.: +1 540 231 0958.
Email addresses: jmd59@vt.edu (Jill M. Derwin), jcoulston@fs.fed.us (John W. Coulston), thomasv@vt.edu (Valerie A. Thomas), wynne@vt.edu (Randolph H. Wynne)

1. Introduction

The field of Remote Sensing has gained a foothold within the scientific community, with satellite imagery feeding models at local and global scales and spanning decades into the past and future. Computational advances and data availability have allowed for the use of machine learning algorithms to produce global maps of various biophysical variables such as terrestrial carbon and forest extent. However, as these models become more complex, the need for large volumes of consistent reference data has increased [1].

Accuracy assessment is an essential part of the modeling process for any remote sensing study, both in order to provide statistical evidence for the robustness of the methods and models used therein [2]. Despite the importance of accuracy assessment, studies reporting the accuracy of remote sensing models often neglect to provide the full details of the sampling protocol used to collect reference data [3], [4]. Furthermore, there are very few guidelines or reference materials designed specifically for the sampling of continuous remote sensing products for the purpose of accuracy assessment [5].

Additionally, the validation of remote sensing products provides unique challenges in that the population for remote sensing models tend to be large, wall-to-wall datasets [6]. For example, a 30-meter resolution model spanning the conterminous United States has a sampling frame of almost 9 billion pixels. Typically, sample size is defined based on statistical power. If the sample size is too small then the accuracy cannot be estimated to a meaningful degree of precision, resulting in measures of accuracy that can be difficult to interpret and lack a robust statistical basis [7]. However, the number of samples measured can also be limited by the project's budget. Each additional sample adds to the cost of data collection, and, especially for field surveys, this margin can be considerable [8]. Stehman and Foody (2019) suggest that an adequate sample size has been achieved if the standard error and width of the confidence intervals obtained for the accuracy determined by the sample are small enough to meet the needs of the general user [9].

While the number of samples has the most influence on the confidence interval around the sample estimate, stratified sampling designs can also improve the precision of estimates. The main goal of stratification is to reduce the variance in attributes of interest by separating the population into homogeneous groups [10]. Cochran (1977) suggests that a correctly implemented stratified sampling protocol almost always results in a smaller variance for the estimated mean compared to simple random sampling approaches [11]. In particular, stratified sampling with optimal allocation determines the number of plots in each stratum in order to minimize the standard error [12], [13].

The ideal variable for the stratification is the variable being surveyed. When estimating the accuracy of continuous maps, measures of agreement and correlation are typically used, such as root mean square error (RMSE), the coefficient of determination (r^2), and adjusted r^2 which measure covariance between model predictions and reference data [3], [14], [15]. However, these measures cannot be quantified a priori. Cochran (1977) suggests that a high relative precision through stratification can still be achieved if characteristics of the population vary widely in area, if the variables to be measured are closely related to the areas of these characteristics, and if there is a good measure of the area for setting up the strata [11].

In this case, prediction uncertainty may be a useful variable to base a stratification on in the case of ensemble models such as random forest. Coulston et al (2016) demonstrate a method to estimate prediction uncertainty at each pixel by calculating the variance among the predictions from each of the individual regression trees in a random forest model [16]. This measure is critical to convey the utility of the related map product to the user in a spatially explicit manner.

In most cases, accuracy for continuous maps is reported across the full population extent, but it can also be reported by category or map subset. For example, Hansen, et al (2002) reported

the accuracy of their tree cover data set over the Western Province of Zambia using accuracy in five strata, each with 25 randomly selected reference sites [17]. There is also the opportunity to demonstrate map accuracy based on spatially explicit classes, such as regions or land cover type.

The objective of this study is to compare sample allocation methods for the purpose of assessing the accuracy of continuous models. We focus specifically on the method of stratified random sampling with optimal allocation using measures of prediction uncertainty derived from random forest regression models. Our analysis of sampling protocols includes three cases utilizing simulated data. For one of those cases, we also test the effect of using smaller sample sizes on the effectiveness of the various sampling protocols. Finally, we select a single set of sample data for this case, and demonstrate a method to report the accuracy of continuous models for domains that are either regionally constrained or numerically defined based on other variables or the modeled quantity itself.

2. Data and methods

This study demonstrates the use of optimal allocation sampling for the estimation of error metrics and population mean using three cases. The first two cases use data that are fully simulated, and the third case uses an existing model.

2.1. Simulation of predictors and response variable

For cases 1 and 2, raster images were simulated using the RandomFields library for R [18], [19] in order to generate a population of X and Y variables to use in drawing samples with various sampling methods. We constructed six populations representing X variables using a combination of Gaussian random fields with different levels of spatial correlation following, Coulston et al. 2016 [16] (See Figures 1 and 2). We created 30 1416 x 1770-pixel tiles, resulting in a population of 75,189,600 pixels. The data were organized in tiles in order to reduce memory utilization during computation, however each of the tiles have the same properties. We chose to create a fairly large population in order to represent an overall sampling intensity across the population that was reasonably comparable to a real-world situation.

We determined a relationship between these X variables to simulate the distribution of the response variable, Y_{dist}. For the first case, we assumed a normally distributed Y_{dist} following Equation 1 and Figure 4, while the second case assumes a non-parametric Y_{dist} with a left-skewed distribution bounded between 0 and 100. This relationship was engineered to mimic the distribution of the 2016 USFS Tree Canopy Cover (TCC) Analytical product [20] across the United States and follows Equation 2 and Figure 4.

$$Y_{dist} = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + N[0,1] \tag{1}$$

$$Y_{dist} = \frac{(X_1 + X_2 + X_3)^2 - (X_4 + X_5 + X_6)^2}{(X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2} + \frac{(X_4 + X_5 + X_6)}{10} + N[0,1] \tag{2}$$

We introduced a small amount of error to each Y_{dist} in order to represent random noise following a normal distribution with an average error of 0 and a standard deviation of 1. In order to enable direct comparisons between estimates from the different models, these equations were scaled from 0 to 100. (See Equation 3)

$$Y = 100 * \frac{Y_{dist} + |Y_{min}|}{max(Y_{dist} + |Y_{min}|)} \tag{3}$$

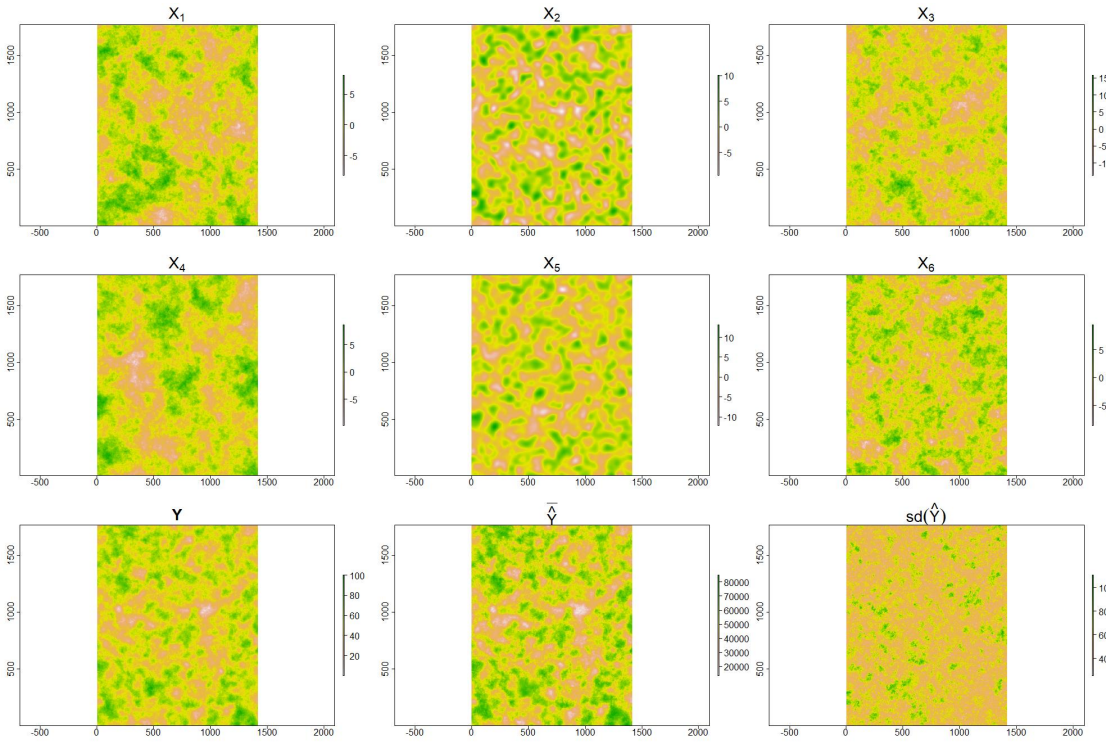

Figure 1: Raster maps of predictor variables (X_1-X_6, top two rows) response variable (Y, bottom-left) as well as the mean of random forest regression tree predictions ($\overline{\hat{Y}}$, bottom-center) and the standard deviation of predictions ($sd(\hat{Y})$), bottom-right) for the normally distributed case. Only one 1416 x 1770 pixel tile (out of 30) is shown, however all tiles are similar.

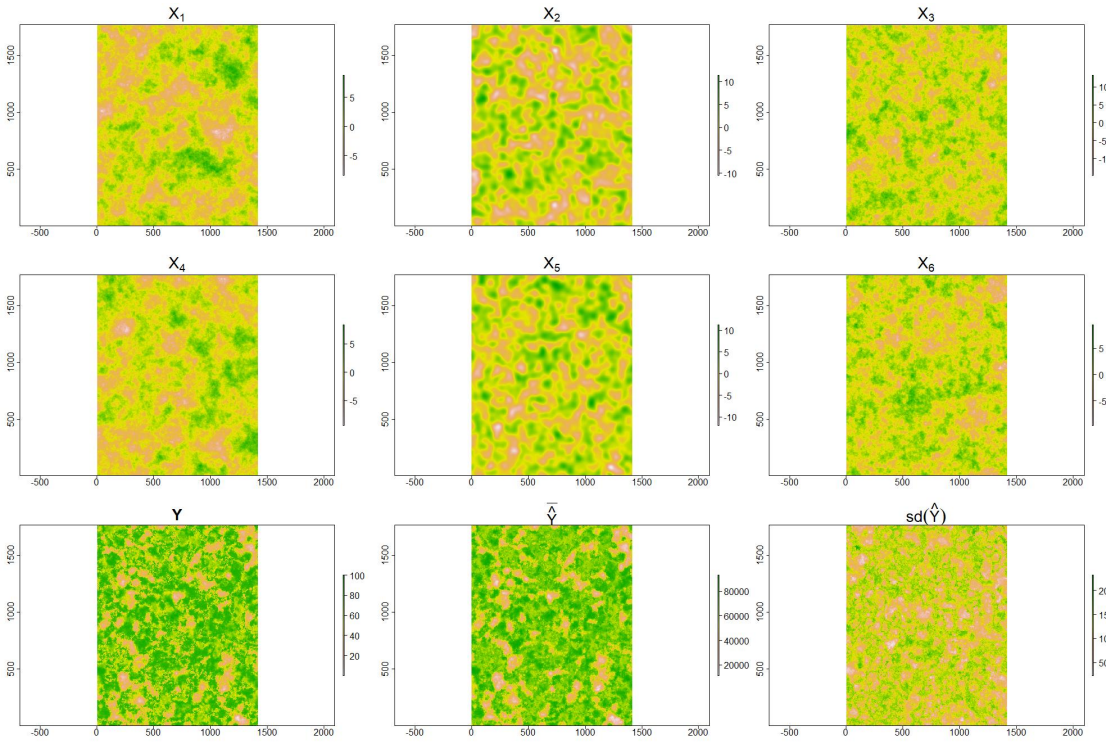

Figure 2: Raster maps of predictor variables (X_1-X_6, top two rows) response variable (Y, bottom-left) as well as the mean of random forest regression tree predictions ($\overline{\hat{Y}}$, bottom-center) and the standard deviation of predictions ($sd(\hat{Y})$), bottom-right) for the non-parametric case. Only one 1416 x 1770 pixel tile (out of 30) is shown, however all tiles are similar.

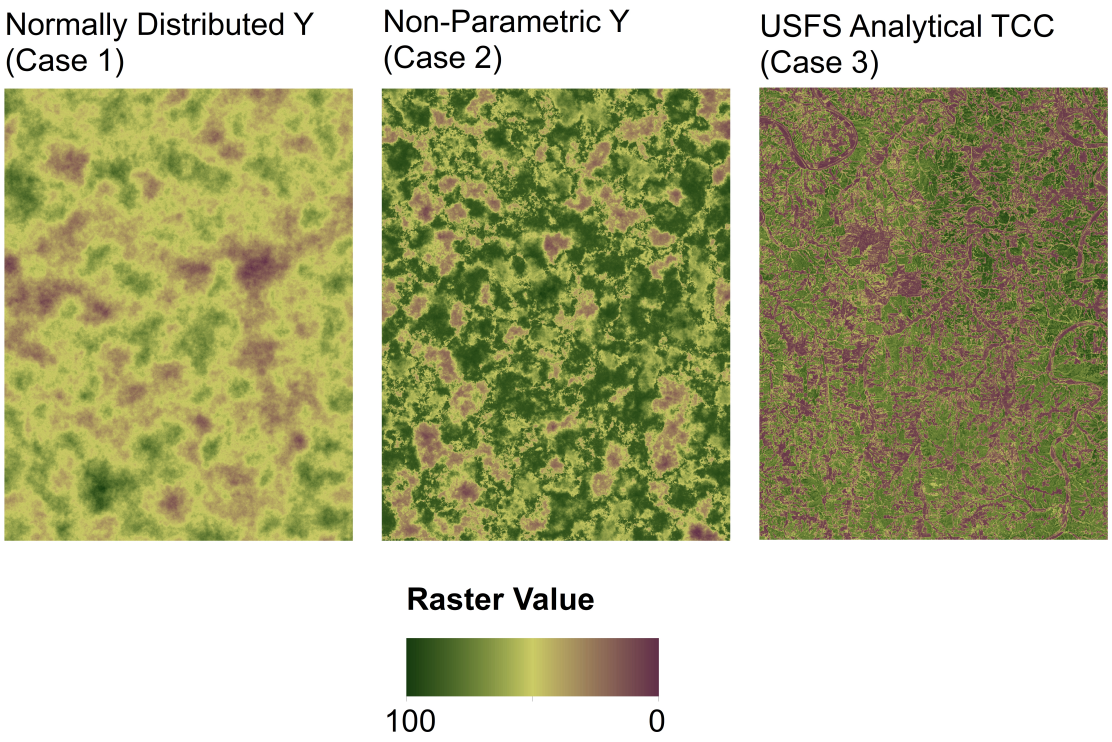

Figure 3: Raster maps of normally distributed Y for case 1 (left), non-parametric Y for case 2 (center), and the USFS Analytical TCC product for case 3 (right). 1416 x 1770 pixel tiles are shown for each case.

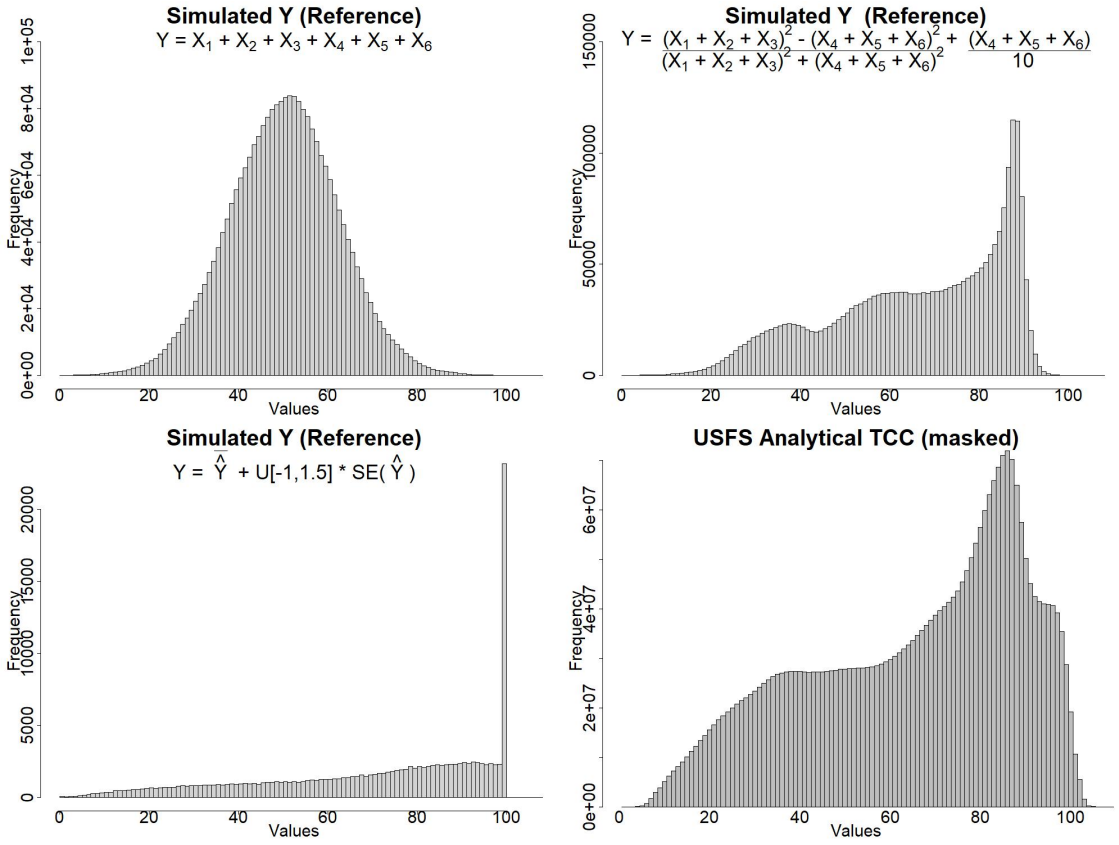

Figure 4: Distribution of Y values for the normally distributed case (case 1) (top-left), the non-parametric case (case 2) (top-right), the USFS Analaytical TCC map product extracted to 142,246 unmasked FIA public coordinate plots (case 3) (bottom-left), and the USFS Analaytical TCC map product (the basis for the distribution of case 2) (bottom-right) . For cases 1 and 2, only one 1416 x 1770 pixel tile (out of 30) is represented, however all tiles are similar.

where Y_{min} and Y_{max} are the minimum and maximum of the distributions generated using Equations 1 and 2 before scaling.

Instead of simulating X and Y variables, for case 3 we used the 2016 USFS Analytical TCC product as well as the Standard Error layer [20] provided as our modeled values and prediction uncertainty values, respectively. (See Figure 5.) Pixels categorized as background, non-tree, water, ice, snow, and agriculture were masked from analysis. In this case, the "true" Y value for the TCC model was unavailable, and so it was simulated using the expected relationship between "truth" and the model and prediction uncertainty measures. (See Equation 4 and Figure 4) These Y values were bounded between 0 and 100 with negative values being set to 0 and values over 100 being set to 100.

$$Y_{dist}[0, 100] = \overline{\hat{Y}} + U[-1, 1.5] * SE(\hat{Y}) \tag{4}$$

2.2. Model prediction

The training datasets for cases 1 and 2 were selected using systematic sampling with a fishnet grid across the simulated tiles, with 720 sample pixels selected from each of the 30 tiles, totalling 21600 samples. For both of these cases, a random forest model relating the X and Y variables was estimated from these samples from the randomForest package in R [19], [21]. Model predictions from each regression tree were obtained for each pixel in the population. Correlations between predictor and response variables are less than 0.8 between all pairs of variables, as shown in Figure 6 for cases 1 and 2. The mean and standard deviation of these predictions was obtained for each pixel. The standard deviation of predictions was used as a proxy for prediction uncertainty.

As previously described, the 2016 USFS Analytical TCC product and Standard Error layer [20] were used for case 3. The TCC product represents the mean of predictions from each regression tree of a random forest model trained with data derived from remote sensing imagery [24], and the Standard Error layer is obtained at each pixel from the standard error of the predictions. The standard error layer represents prediction uncertainty for this model. The standard error is the standard deviation of predictions divided by the square root of the number of predictions. Since the models in cases 1 and 2 each produce 500 predictions at each pixel, the standard error would be the same as the standard deviation (which was used in those cases to approximate prediction uncertainty) scaled by a constant (22). These model outputs were extracted to 364,242 non-intensified Forest Inventory and Analysis (FIA) public coordinate plot locations for sampling and estimation. 47 of those plots fell on null-value pixels in the TCC model, and 221,949 plots landed on pixels classified as background, non-tree, water, ice, snow, or agriculture in the 2016 NLCD land cover product [25], leaving 142,246 plots for sampling.

2.3. Sampling

We test the following four sampling protocols: simple random sampling, stratified random sampling with equal allocation, stratified random sampling with proportional allocation, and stratified random sampling with optimal allocation.

In order to choose a sample size that was appropriate for each case and sampling protocol, we used the following equation for simple random sampling without replacement.

$$n = \frac{1}{(\frac{E}{t*s})^2 + \frac{1}{N}} \tag{5}$$

where n is the total sample size N is the population size based on number of unmasked pixels in the population, t is *student's t* based on the 95% confidence interval and large population size (1.960), E

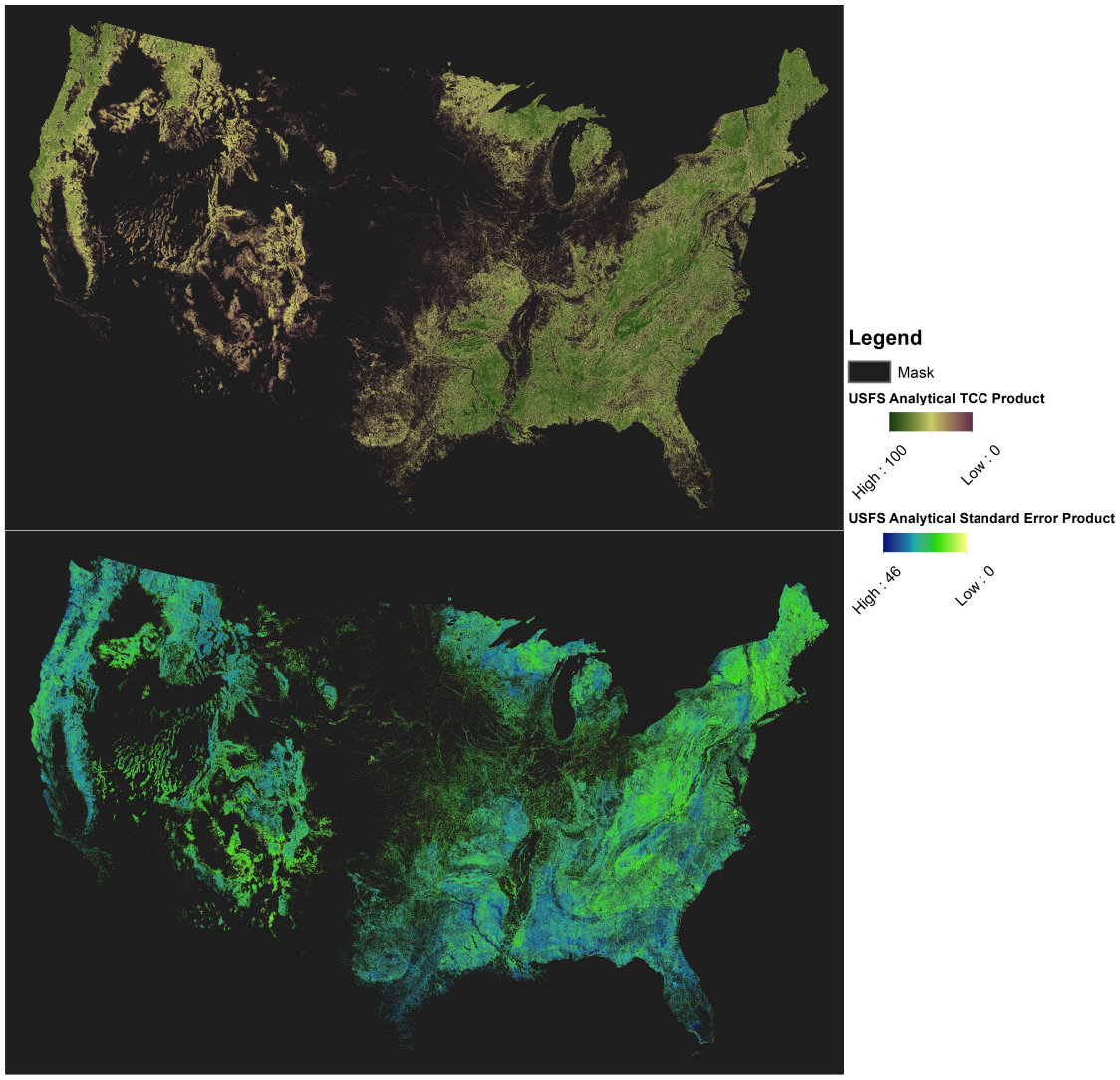

Figure 5: 2016 USFS Analytical TCC layer (top) and Standard Error layer (bottom) [20].

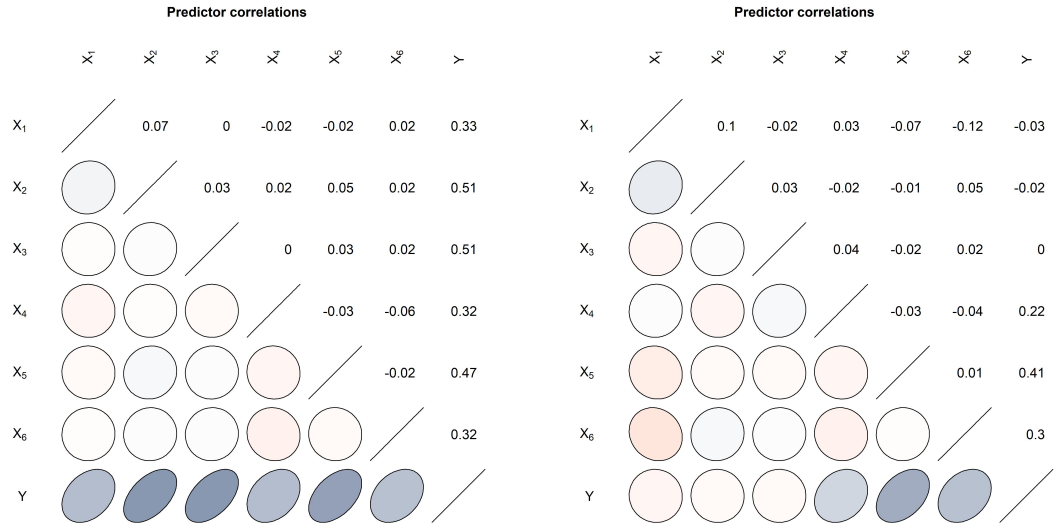

Figure 6: Correlation matrix for X and Y variables for the normally distributed case (case 1) (left) and the non-parametric case (case 2) (right) showing a lack of multicollinearity (based on a threshold of 0.8 as established by [22] as cited by [23]). Only one 1416 x 1770 pixel tile (out of 30) is represented for both cases, however all tiles are similar.

is the desired precision of the sample mean to the population mean (which we determined to be 1% TCC), and s is the population standard deviation of the variable being estimated. Because these cases are supposed to demonstrate a real-world sampling activity where measures of the standard deviation of the reference data would be unavailable, the model outputs were used. The variable with the highest standard deviation was the prediction uncertainty variable for all cases. Using this equation, we determined that a sample size of 409 would be adequate for case 1, 869 would be adequate for case 2, and 2095 would be adequate for case 3. Thus, a sample size of 2000 was used for all three cases. For case 3, samples of size 1000 and 500 were also drawn to test the differences between sampling protocols where larger samples aren't possible.

2.3.1. Simple random sampling

Simple random sampling, which randomly select samples from the sampling frame has equal inclusion probabilities across all samples in the sampling frame. This method may not capture the full variance across the population, since there is nothing to prevent spatial or informational clusters of samples [2], [26]. Alternatively, samples that are too far away may lead to gaps in coverage, missing rare classes or over-emphasizing prevalent classes [26].

2.3.2. Stratified random sampling with equal allocation

Under stratified random sampling the population is divided into h strata. A sample of n_h is drawn at random from the N_h population elements in the stratum. In this case the population is made up of all pixels in the map simulated for a given case. Stratified random sampling uses defined strata and defines sampling intensities for each stratum using an allocation method. Within those strata, the designated number of plots is randomly selected.

Under stratified random sampling, the inclusion probability for each stratum is

$$\pi_h = \frac{n_h}{N_h} \tag{6}$$

and the strata weights are defined as:

$$w_{\mathrm{h}} = \frac{a_h}{\sum a_h}; \sum w_h = 1 \tag{7}$$

where a_h is the area of strata h.

For stratified random sampling protocols, samples are not allocated proportionally to the relative area of the strata in the population, resulting in unequal inclusion probabilities in different strata . These unequal inclusion probabilities can be remedied by re-balancing the resulting accuracy metrics using a weighted average to calculate $\overline{Y}$ defined by:

$$\overline{Y} = \sum (\overline{Y}_h * w_h) \tag{8}$$

where w_h is the weight of stratum h and $\overline{Y}_h$ is the mean of a given variable within stratum h, defined by:

$$\overline{Y}_h = \frac{\sum Y_{ih}}{n_h} \tag{9}$$

where Y_{ih} is the i^{th} observation in h.

The estimated variance of the mean is:

$$v(\overline{Y}) = \frac{\sum (w_h^2 * s_h^2)}{n_h} \tag{10}$$

where s_h is the within-stratum sampling variance calculated by:

$$s_h = \frac{\sum (Y_{ih}^2 - \overline{Y}_h^2)}{(n_h - 1)} \tag{11}$$

The standard error of the mean is defined as:

$$SE(\overline{Y}) = \sqrt{v(\overline{Y}} \tag{12}$$

$$\overline{Y}_h = \frac{\sum Y_{ih}}{n_h} \tag{13}$$

where Y_{ih} is the i^{th} observation in h. Stratified random sampling with equal allocation uses an equal number of samples in each stratum. For equal allocation, strata-level sample sizes (n_h) were determined using Equation 14.

$$n_h = \frac{n}{|h|} \tag{14}$$

where $|h|$ is the number of strata. [12], [13]

2.3.3. Stratified random sampling with proportional allocation

Proportional allocation allocates samples proportionally based on stratum area. For proportional allocation, strata-level sample sizes (n_h) were determined using Equation 15.

$$n_h = n * \frac{a_h}{\sum a_h} \tag{15}$$

where a_h is the area of the strata h. [12], [13].

2.3.4. Stratified random sampling with optimal allocation

Optimal allocation weights the number of samples in each stratum by a combination of within-stratum variance or standard deviation and stratum area [13]. Therefore, a sample in a more variable stratum may be more likely to be included than a sample in a less-variable stratum. According to Burkhart and Avery (1994), this is the preferred method of sample allocation when the area of each strata and the standard deviation of the variable of interest within that stratum can be determined reliably [13].

For optimal allocation, strata-level sample sizes (n_h) were determined using Equation 16.

$$n_h = n * \frac{a_h * s_h}{\sum (a_h * s_h)} \tag{16}$$

where n is the total sample size, a_h is the area of the strata h, and s_h is the mean of the prediction uncertainty values within stratum h [12], [13]. For our purposes, we used the mean of the pixel-level standard deviation of the random forest regression tree predictions within stratum h for cases 1 and 2 and the mean of the values from the standard error layer from the USFS Analytical TCC Product within stratum h for case 3.

2.4. Statistical analysis

For all cases, the population mean (μ) and the Root Mean Squared Error (RMSE) were calculated for the population. (See Equations 17 and 19.)

$$\mu = \frac{\sum Y_i}{N} \tag{17}$$

where Y_i is the value of the i^{th} element of Y and N is the number of pixels in the population (75,189,600).

$$SSR = \sum (Y_i - \overline{\hat{Y}_i})^2 \tag{18}$$

where $\overline{\hat{Y}_i}$ is the mean of the regression-tree predictions for the i^{th} element of Y.

$$RMSE = \sqrt{\frac{SSR}{N}} \tag{19}$$

To estimate the sample mean, Equations 9 and 13 were used. To calculate the sample *RMSE* for samples collected using stratified random sampling protocols, the sample *SSR* was first calculated. Because *SSR* is a cumulative sum, the weighted average of the strata-level sample Mean of Squared Residuals (MSR_h) estimates were calculated and then multiplied by the sample size (n). (See Equations 20 and 21.)

$$MSR_h = \frac{SSR_h}{n_h} \tag{20}$$

where SSR_h is the SSR for stratum h and n_h is the sample size within stratum h.

$$SSR_{sample} = n * \sum (w_h * MSR_h) \tag{21}$$

where n is the total sample size. This value was then used to calcualte sample *RMSE* using equation 19.

To capture the variability of the estimates provided by each sampling protocol, 250 iterative samples were drawn following with the same sample size and sampling intensities for each protocol

and sample mean and sample RMSE were calculated for each iteration. The mean, standard deviation, difference from population mean, and width of the 95% confidence intervals were calculated for these sets of estimates.

2.5. Domain estimation

Domain estimation allows for validation across smaller subsets the population that may or may not be aligned with sampling strata. Indicator variables are used to obtain estimates of the mean within the domain in order to utilize the full strata sample. Within the domain of interest, the indicator variable's value is 1, while outside the domain of interest, the indicator variable's value is 0.

We demonstrate domain estimation techniques using both 10% TCC strata and land cover classes extracted from the 2016 NLCD Land Cover product. Samples used in the calculation of domain estimates for this study come from a single selection of 2000 samples chosen using stratified random sampling with optimal allocation using model outputs from the 2016 USFS Analytical TCC product and simulated reference data used in case 3.

Within-domain means are calculated, first by determining the stratum means within the domain ($\overline{Y}_{dh}$) using Equation 22.

$$\overline{Y}_{dh} = \frac{\sum (Y_{ih} * I_{ihd})}{n_h} \tag{22}$$

where $\bar{x}_h$ is the mean of a given variable within stratum h. Where Y_{ih} is the i^{th} observation in the set of samples in stratum h, I_{ihd} is the indicator value of the i^{th} observation in the set of samples in stratum h defined by domain d, and n_h is the number of samples in stratum h.

and then by determining the domain mean using Equation 23.

$$\overline{Y}_{d} = \sum (w_h * \overline{Y}_{dh}) \tag{23}$$

where w_h is the stratum weight of stratum h.

Strata sampling variance for a given domain can be approximated to understand the uncertainty surrounding the estimated mean. This is calculated by first calculating within-strata sampling variance for the domain using Equation 24.

$$s^2_{hd} = \frac{\sum (Y^2_{ih} * I_{ihd} - \overline{Y}^2_h)}{n_h - 1} \tag{24}$$

where ${Y_{ih}}^2$ is the i^{th} observation in the set of samples in stratum h squared, ${\overline{Y_{\mathrm{h}}}}^2$ is the mean of the sample estimate within stratum h squared, I_{ihd} is the indicator value of the i^{th} observation in the set of samples in stratum h and domain d, and n_h is the number of samples in stratum h.

The variance of the sample estimate within the domain is then calculated by:

$$var(\overline{Y_d}) = \frac{\sum (s^2_{hd} * w^2_h)}{n_{hd}} \tag{25}$$

where ${w_h}^2$ is the square of the stratum weight of stratum h, and n_{hd} is the number of samples in stratum h and domain d.

We also produce a confusion matrix showing the number of samples within domains using modeled and reference values.

Case	Variance of Reference Y	Mean of Reference Y	Variance of Squared Residuals	RMSE
Case 1: Normal Distribution	159.234	50.294	25.365	4.146
Case 2: Non-Parametric Distribution	389.964	65.428	125.979	9.123
Case 3: Extracted 2016 USFS Analytical TCC	661.918	71.204	297.198	14.655

Table 1: Population variance and mean of the reference Y and population variance of squared residuals and RMSE for all cases.

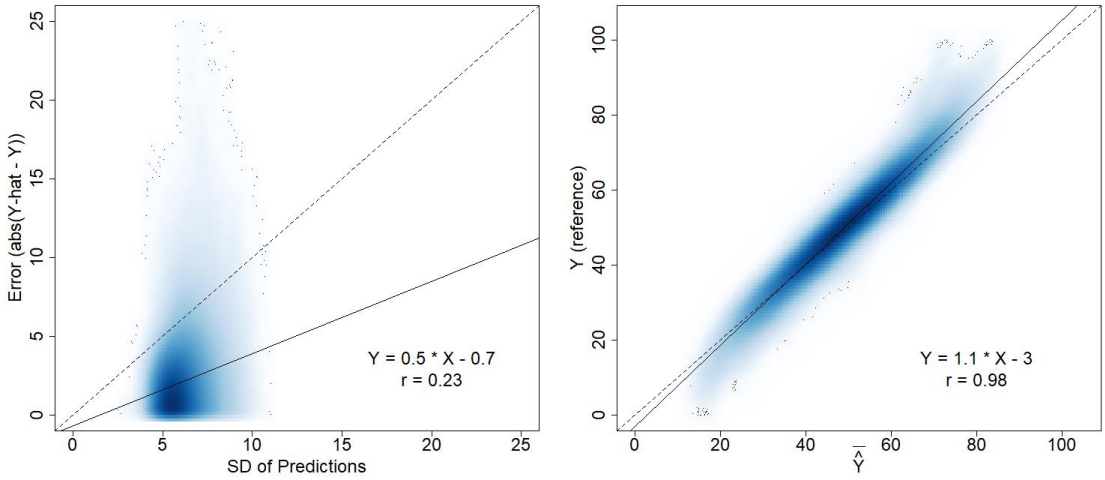

Figure 7: Scatter plots showing the correlation between the variable to be estimated and the stratification variables used. (left) Error vs the standard deviation across random forest regression tree predictions. (right) Reference Y vs the mean of regression tree predictions. Only one 1416 x 1770 pixel tile (out of 30) is shown, however all tiles are similar.

3. Results

3.1. Sampling protocol

3.1.1. Case 1: Normally distributed Y

Case 1, which has a normally distributed response variable (Y), has the lowest variance and mean in both of the variables that were estimated using sampling. (See Table 1.) For this case, the absolute error between predictions and reference values of Y is poorly correlated to the standard deviation of predictions compared to other cases. The range of standard deviation of predictions falls between 2.4 and 11.3, while the absolute error ranges from 0 to 28. There is a much stronger correlation between the reference Y and the mean of model predictions compared to other cases, which ranges from 13 and 85. (See Figure 7.)

For the 250 sets of samples drawn using each sampling protocol, the mean of the RMSE estimates are within 0.01 of eachother, and the average absolute difference between these estimates and the population RMSE is less than 0.006. The standard deviation of estimates varied from 0.016 to 0.096, and the width of the confidence intervals ranged from 0.016 to 0.024. The widest confidence interval and largest standard deviation of estimates was found using stratified random sampling

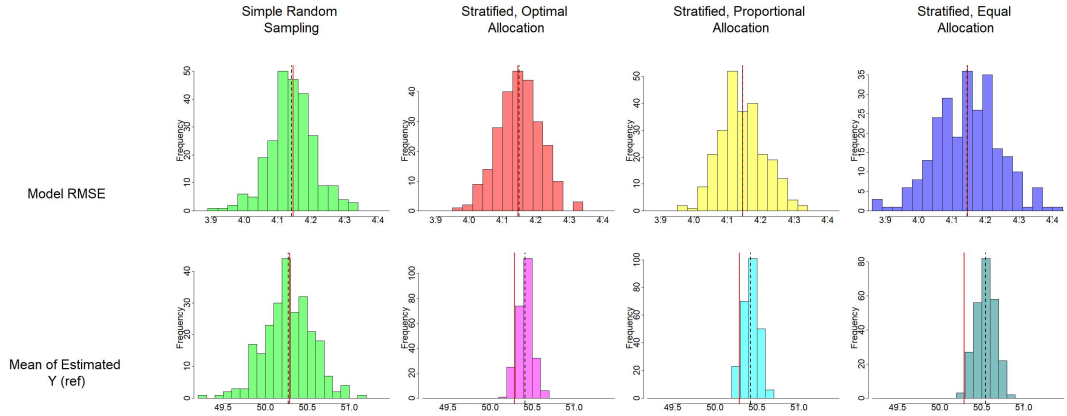

Figure 8: Histograms showing the distribution of ensemble estimates of the RMSE (top row) and mean Y (bottom row) across the population, calculated from 250 sets of samples using each sampling protocol. For stratified sampling protocols, the estimates displayed in the top row used the standard deviation of $\hat{Y}$ for stratification and the estimates displayed in the bottom row used $\overline{\hat{Y}}$. The black dashed line represents the estimated mean, and the solid red line represents the true population value.

with equal allocation. The narrowest confidence interval and standard deviation of estimates was found using stratified random sampling with optimal allocation. (See Table 3.)

The mean of the average reference Y estimates falls between 50.300 and 50.551, and the average absolute difference between these estimates and the population mean (50.29) is ranges between 0.006 and 0.257 across sampling protocols. The standard deviation of estimates varied from 0.093 to 0.286, and the width of the confidence intervals ranged from 0.023 to 0.071. The widest confidence interval and largest standard deviation of estimates was found using simple random sampling. The narrowest confidence interval and standard deviation of estimates were tied between stratified random sampling with optimal allocation and proportional allocation. (See Table 2.)

Hisotgrams show that sample estimates of RMSE range from 3.8 and 4.5 for all sampling protocols. Proportional and optimal allocation protocols show a narrower distribution with more sample estimates falling closer to the mean. The mean of the sample estimates of RMSE is close to the population RMSE for all sampling protocols. The sample estimates of the average reference Y range from 49.2 to 51.4 for the simple random sampling protocol. The stratified random sampling protocols approximate the average reference Y value with a smaller range of values, concentrated around the mean of the sample estimates. The mean of the sample estimates falls closest to the average of the reference Y for the population for the simple random sampling protocol, followed by stratified random sampling with optimal allocation, and then stratified random sampling with proportional allocation. The mean of the sample estimates drawn using stratified random sampling with Equal allocation is farthest from the average reference Y value for the population. (See Figure 8.)

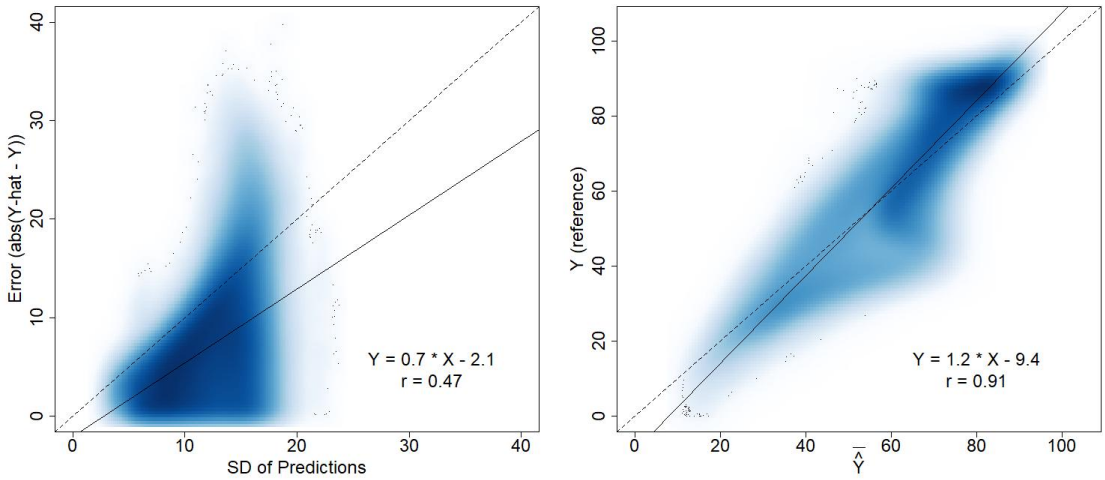

Figure 9: Scatter plots showing the correlation between the variable to be estimated and the stratification variables used. (left) Error vs the standard deviation across random forest regression tree predictions. (right) Reference Y vs the mean of regression tree predictions. Only one 1416 x 1770 pixel tile (out of 30) is shown, however all tiles are similar.

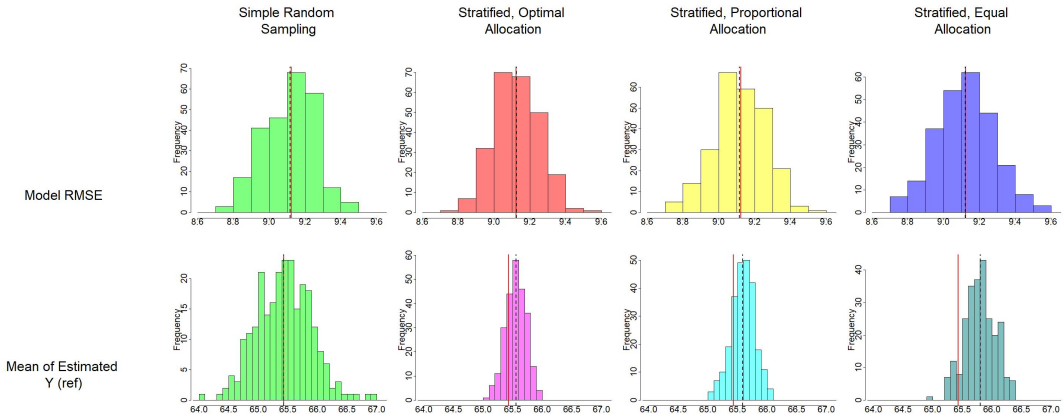

Figure 10: Histograms showing the distribution of ensemble estimates of the RMSE (top row) and mean Y (bottom row) across the population, calculated from 250 sets of samples using each sampling protocol. For stratified sampling protocols, the estimates displayed in the top row used the standard deviation of $\hat{Y}$ for stratification and the estimates displayed in the bottom row used $\overline{\hat{Y}}$. The black dashed line represents the estimated mean, and the solid red line represents the true population value.

Sampling Protocol	Average of Estimates	Standard Deviation of Estimates	Width of the 95% Confidence Interval (CI) of Estimates	Bias: Mean Estimate vs Population Mean (50.29)
Simple Random	50.275	0.290	0.072	-0.018
Optimal Allocation by $\overline{\hat{Y}}$	50.416	0.088	0.022	0.123
Proportional Allocation by $\overline{\hat{Y}}$	50.427	0.089	0.022	0.134
Equal Allocation by $\overline{\hat{Y}}$	50.545	0.116	0.029	0.252

Table 2: Average of estimated population mean, standard deviation of population mean estimates, width of the 95% confidence interval of population mean estimate, and mean difference from actual population mean for 250 iterative sample estimates. Stratification is based on the mean of model predictions at each pixel ($\overline{\hat{Y}}$).

Sampling Protocol	Average of Estimates	Standard Deviation of Estimates	Width of the 95% Confidence Interval (CI) of Estimates	Bias: Mean Estimate vs Population RMSE (4.15)
Simple Random	4.141	0.069	0.017	-0.005
Optimal Allocation by u	4.149	0.016	0.016	0.004
Proportional Allocation by u	4.145	0.067	0.017	0.000
Equal Allocation by u	4.144	0.096	0.024	-0.002

Table 3: Average of estimated RMSE, standard deviation of RMSE estimates, width of the 95% confidence interval of RMSE estimate, and mean difference from actual population RMSE for 250 iterative sample estimates. Stratification is based on the standard deviation of model predictions at each pixel (a proxy for prediction uncertainty, u).

295 *3.1.2. Case 2: Non-parametric Y*

296 Case 2, whose response variable (Y) has a non-parametric distribution, has a variance and
297 mean that fall between the other two cases in both of the variables that were estimated using
298 sampling. (See Table 1.) For this case, the absolute error between predictions and reference values
299 of Y is poorly correlated to the standard deviation of predictions. The range of standard deviation
300 of predictions falls between 2.1 and 23.8, while the absolute error ranges from 0 to 40. There is a
301 much stronger correlation between the reference Y and the mean of model predictions which range
302 between 11.1 and 93.3. (See Figure 7.)

303 For the 250 sets of samples drawn using each sampling protocol, the mean of the RMSE estimates
304 falls between 4.138 and 4.151, and the average absolute difference between these estimates and the
305 population RMSE (4.15) is less than 0.01 for all sampling protocols. The standard deviation of
306 estimates varied from 0.064 to 0.091, and the width of the confidence intervals ranged from 0.016 to
307 0.023. The widest confidence interval and largest standard deviation of estimates was found using
308 stratified random sampling with equal allocation. The narrowest confidence interval and standard
309 deviation of estimates was found using stratified random sampling with optimal allocation. (See
310 Table 3.)

311 The mean of the average reference Y estimates falls between 50.300 and 50.551, and the average
312 absolute difference between these estimates and the population RMSE (50.29) is ranges between
313 0.006 and 0.257 across sampling protocols. The standard deviation of estimates varied from 0.093 to
314 0.286, and the width of the confidence intervals ranged from 0.023 to 0.071. The widest confidence
315 interval and largest standard deviation of estimates was found using simple random sampling.
316 The narrowest confidence interval and standard deviation of estimates were tied between stratified

Sampling Protocol	Average of Estimates	Standard Deviation of Estimates	Width of the 95% Confidence Interval (CI) of Estimates	Bias: Mean Estimate vs Population Mean (65.43)
Simple Random	65.434	0.462	0.115	0.006
Optimal Allocation by $\overline{\hat{Y}}$	65.555	0.174	0.043	0.127
Proportional Allocation by $\overline{\hat{Y}}$	65.590	0.199	0.050	0.162
Equal Allocation by $\overline{\hat{Y}}$	65.801	0.260	0.065	0.373

Table 4: Average of estimated population mean, standard deviation of population mean estimates, width of the 95% confidence interval of population mean estimate, and mean difference from actual population mean for 250 iterative sample estimates. Stratification is based on the mean of model predictions at each pixel ($\overline{\hat{Y}}$).

Sampling Protocol	Average of Estimates	Standard Deviation of Estimates	Width of the 95% Confidence Interval (CI) of Estimates	Bias: Mean Estimate vs Population RMSE (9.12)
Simple Random	9.117	0.144	0.036	-0.006
Optimal Allocation by u	9.125	0.128	0.032	0.002
Proportional Allocation by u	9.115	0.143	0.036	-0.009
Equal Allocation by u	9.122	0.163	0.041	-0.002

Table 5: Average of estimated RMSE, standard deviation of RMSE estimates, width of the 95% confidence interval of RMSE estimate, and mean difference from actual population RMSE for 250 iterative sample estimates. Stratification is based on the standard deviation of model predictions at each pixel (a proxy for prediction uncertainty, u).

317 random sampling with optimal allocation and proportional allocation. (See Table 2.)

318 Hisotgrams show that sample estimates of RMSE range from 8.6 and 9.7 for all sampling
319 protocols. Proportional and optimal allocation protocols show a narrower distribution with more
320 sample estimates falling closer to the mean. The mean of the sample estimates of RMSE is close to
321 the population RMSE for all sampling protocols. The sample estimates of the average reference Y
322 range from 64.0 to 67.1 for the simple random sampling protocol. The stratified random sampling
323 protocols approximate the average reference Y value with a smaller range of values, concentrated
324 around the mean of the sample estimates. The mean of the sample estimates falls closest to the
325 average of the reference Y for the population for the simple random sampling protocol, followed
326 by stratified random sampling with optimal allocation, and then stratified random sampling with
327 proportional allocation. The mean of the sample estimates drawn using stratified random sampling
328 with Equal allocation is farthest from the average reference Y value for the population. (See Figure
329 8.)

330 *3.1.3. Case 3: 2016 USFS Analytical TCC Product extracted to FIA plots*

331 Case 3, which used the existing 2016 USFS Analytical TCC product and standard error layer to
332 approximate the reference Y, has the highest variance and mean in both of the variables that were
333 estimated using sampling. (See Table 1.) For this case, the absolute error between predictions and
334 reference values of Y is more correlated to the standard deviation of predictions than in the other
335 cases. The range of standard deviation of predictions falls between 0 and 45, while the absolute
336 error ranges from 0 to 55. The scatter plot between the reference Y and the mean of model
337 predictions shows a concentration of points at high Y reference and $\overline{\hat{Y}}$. The correlation between

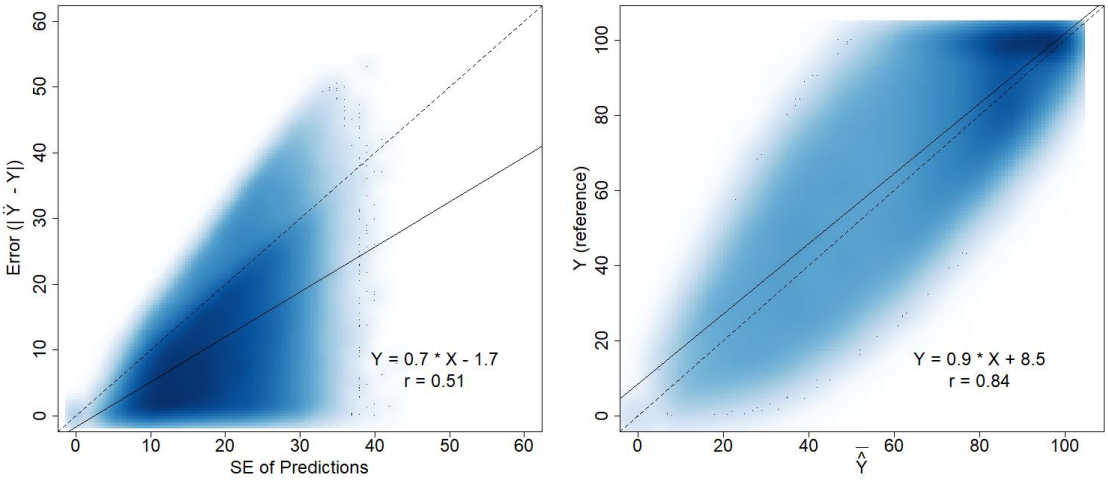

Figure 11: Scatter plots showing the correlation between the variable to be estimated and the stratification variables used. (left) Error vs the standard error layer value from the 2016 Analytical TCC product. (right) Reference Y vs the 2016 Analytical TCC product.

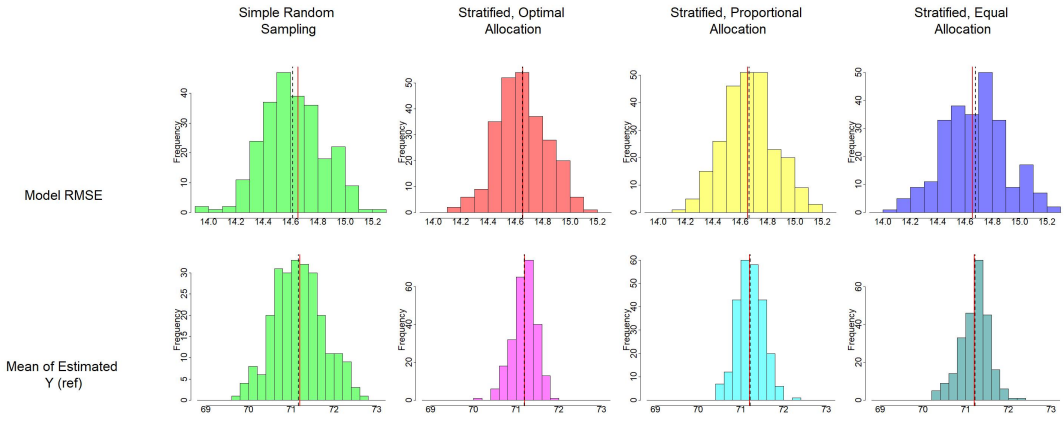

Figure 12: Histograms showing the distribution of ensemble estimates of the RMSE (top row) and mean Y (bottom row) across the population, calculated from 250 sets of samples using each sampling protocol. For stratified sampling protocols, the estimates displayed in the top row used the standard deviation of $\hat{Y}$ for stratification and the estimates displayed in the bottom row used $\overline{\hat{Y}}$. The black dashed line represents the estimated mean, and the solid red line represents the true population value.

Sampling Protocol	Average of Estimates	Standard Deviation of Estimates	Width of the 95% Confidence Interval (CI) of Estimates	Bias: Mean Estimate vs Population Mean (71.20)
Simple Random	71.166	0.576	0.143	-0.038
Optimal Allocation by $\overline{\hat{Y}}$	71.191	0.278	0.069	-0.013
Proportional Allocation by $\overline{\hat{Y}}$	71.218	0.310	0.077	0.014
Equal Allocation by $\overline{\hat{Y}}$	71.215	0.337	0.084	0.010

Table 6: Average of estimated population mean, standard deviation of population mean estimates, width of the 95% confidence interval of population mean estimate, and mean difference from actual population mean for 250 iterative sample estimates. Stratification is based on the mean of model predictions at each pixel ($\overline{\hat{Y}}$).

Sampling Protocol	Average of Estimates	Standard Deviation of Estimates	Width of the 95% Confidence Interval (CI) of Estimates	Bias: Mean Estimate vs Population RMSE (14.66)
Simple Random	14.618	0.223	0.056	-0.037
Optimal Allocation by u	14.652	0.187	0.047	-0.003
Proportional Allocation by u	14.667	0.189	0.047	0.011
Equal Allocation by u	14.676	0.230	0.057	0.021

Table 7: Average of estimated RMSE, standard deviation of RMSE estimates, width of the 95% confidence interval of RMSE estimate, and mean difference from actual population RMSE for 250 iterative sample estimates. Stratification is based on the standard deviation of model predictions at each pixel (a proxy for prediction uncertainty, u).

these two variables is lower than the other two cases. (See Figure 11.)

For the 250 sets of samples drawn using each sampling protocol for this case, the average absolute difference between these estimates and the population RMSE (14.66) falls between 0.003 and 0.038 for all sampling protocols. The standard deviation of estimates varied from 0.187 to 0.230, and the width of the confidence intervals ranged from 0.047 to 0.057. The confidence intervals are wider for this case than for the other two cases which used fully simulated data. The widest confidence interval and largest standard deviation of estimates was found using stratified random sampling with equal allocation. The narrowest standard deviation of estimates was found using stratified random sampling with optimal allocation, and the narrowest confidence interval of estimates was found using stratified random sampling with either optimal allocation or proportional allocation. (See Table 7.)

The average absolute difference between the estimated RMSE and the population RMSE (71.20) ranges between 0.010 and 0.038 across sampling protocols. The standard deviation of estimates varied from 0.278 to 0.576, and the width of the confidence intervals ranged from 0.069 to 0.143. The widest confidence interval and largest standard deviation of estimates was found using simple random sampling. The narrowest confidence interval and standard deviation of estimates were found for stratified random sampling with optimal allocation. (See Table 6.)

Histograms show that sample estimates of RMSE range from 13.9 and 15.3 for all sampling protocols. Proportional and optimal allocation protocols show a narrower distribution with more sample estimates falling closer to the mean. The mean of the sample estimates of RMSE is close to the population RMSE for all sampling protocols. The sample estimates of the average reference Y range from 68 to 74 for the simple random sampling protocol. By comparison, the three stratified

random sampling protocols approximate the average reference Y value with a smaller range of values, concentrated around the mean of the sample estimates. (See Figure 12.)

3.2. Number of samples

Sampling Protocol	Sample Size	Width of the 95% Confidence Interval (CI) of Estimates of the Mean	Width of the 95% Confidence Interval (CI) of Estimates of RMSE
Simple Random	500	0.296	0.110
	1000	0.212	0.076
	1500	0.173	0.064
	2000	0.143	0.056
Optimal Allocation	500	0.144	0.088
	1000	0.102	0.059
	1500	0.089	0.052
	2000	0.069	0.047
Proportional Allocation	500	0.158	0.098
	1000	0.106	0.078
	1500	0.087	0.052
	2000	0.077	0.047
Equal Allocation	500	0.157	0.101
	1000	0.119	0.082
	1500	0.093	0.065
	2000	0.084	0.057

Table 8: Width of the 95% confidence interval of estimates of the mean and estimates of RMSE for each sampling protocol at various sample sizes using case 3 data. For estimates of the mean, stratification is based on TCC values from the USFS Analytical TCC product from 2016, and for RMSE estimates, stratification is based on the standard error values from the USFS Analytical Standard Error product from 2016 (a proxy for prediction uncertainty, u).

For all sampling methods, the width of the confidence interval across estimates of both the average reference Y value and the RMSE grow as the number of samples gets smaller. When estimating the average reference Y value, the width of the confidence intervals for simple random sampling are considerably higher for all other sampling protocols for all sample sizes. That difference is larger for the smallest sample size. When estimating RMSE, the width of the confidence interval for all sampling protocols are more similar. The sampling protocol that produced the most precise estimates is stratified random sampling with optimal allocation, with the exception of the estimation of the average Y reference value using 1500 samples, where proportional allocation has a more narrow confidence interval. When estimating both the average Y reference value and the RMSE, the width of the confidence interval for the optimal allocation sampling protocol declines more steeply between 500 and 1000 samples than it does with additional samples, appearing to stabilize. A similar trend is visible for the proportional allocation sampling protocol stabilizing after 1000 samples when estimating the average reference Y value and after 1500 samples when estimating the RMSE. Likewise, the equal allocation sampling protocol appears to stabilize after 1500 samples when estimating the average reference Y value. This stabilization is not as apparrent for estimates from samples collected using simple random sampling based on the sample sizes tested in this study, however the gains in precision do get smaller with added samples (See Tables 8 and Figure 13).

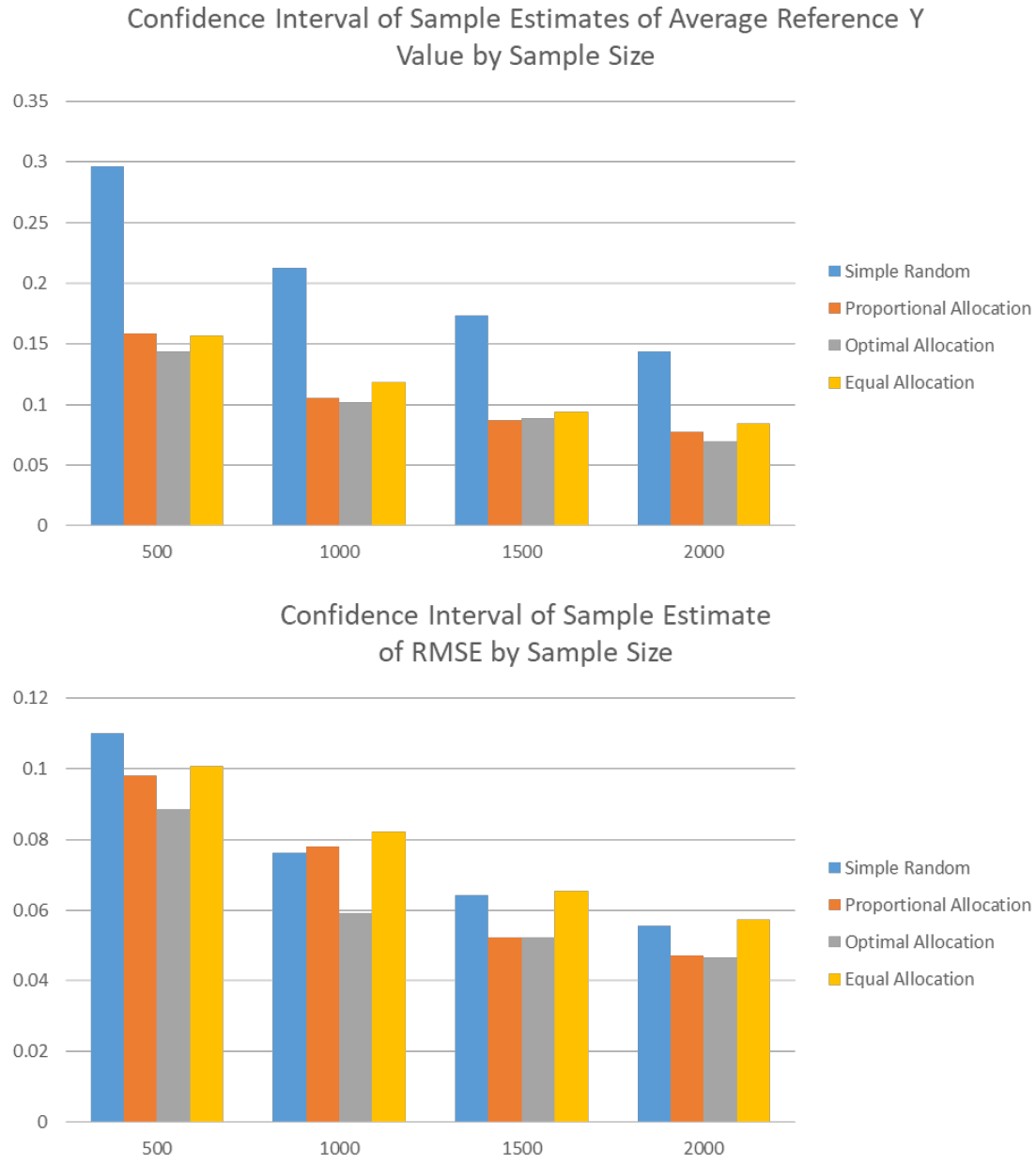

Figure 13: Width of the confidence intervals of samples estimates of the average reference Y value (top) and the average RMSE value (bottom) by sampling protocol and sample size.

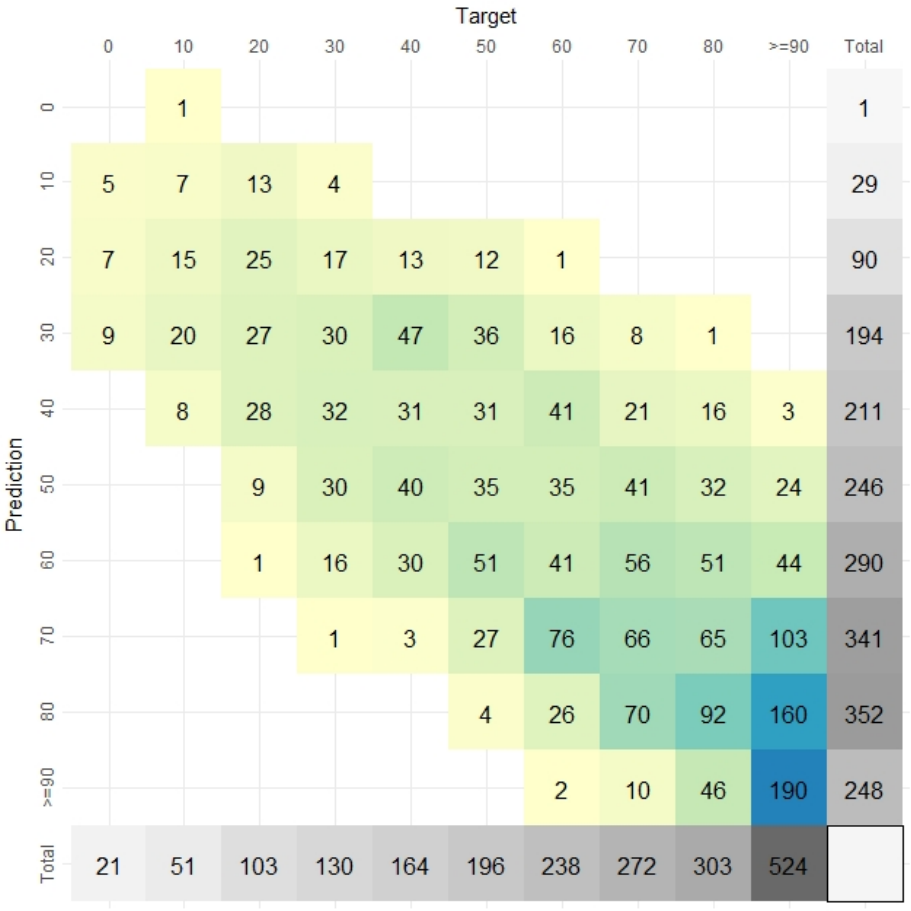

Figure 14: Confusion matrix showing the number of samples in target and prediction classes.

Modeled TCC Class (Domain)	RMSE Value Within Domain	Variance of Squared Error Values within Domain	Standard Error of Squared Error Values within Domain	n Samples within Domain
[0,10)	0.152	0.001	0.023	1
[10,20)	1.525	0.389	0.624	29
[20,30)	3.215	2.509	1.584	90
[30,40)	5.065	6.664	2.581	194
[40,50)	5.646	9.034	3.006	211
[50,60)	6.166	9.999	3.162	246
[60,70)	6.133	8.738	2.956	290
[70,80)	5.839	5.904	2.430	341
[80,90)	5.046	2.791	1.671	352
>= 90	2.676	0.484	0.696	248

Table 9: Within-domain estimates of RMSE, variance in pixel-wise errors, standard error of pixel-wise errors, and sample size for one set of samples. The samples are drawn with stratified random sampling using optimal allocation. The strata are defined by the prediction uncertainty metric (the standard error values extracted from the USFS Analytical TCC Standard Error product for 2016) and the domains are defined by 10% intervals of the simulated TCC reference data.

3.3. Domain estimation

Domain-level error reporting for a single sample using simulated response data, shows that, if our simulated reference Y accurately portrays the true value of TCC on the ground at the FIA plot locations selected, the RMSE of the TCC model is expected to be highest between 50 and 60% TCC and lowest under 20% TCC. Variance and standard error values of simple error estimated from the sample is expected to be highest between 50 and 60% TCC. Using land cover to define domains, RMSE would be expected to be highest in "Deciduous Forest" and "Evergreen Forest" categories, and lowest over "Open Water", "Developed, Medium Intensity", and "Barren Land (Rock/Sand/Clay)". Additionally, variance and standard error of simple error estimated from the sample is highest in the "Evergreen Forest" class and lowest in the "Open Water" and "Barren Land (Rock/Sand/Clay)" classes (See Tables 10 and 9).

The confusion matrix constructed for the 10-class TCC domains demonstrate that the majority of samples fall in high TCC classes, as defined by both $\overline{\hat{Y}}$ and the simulated reference Y. The majority of modeled values landed in the correct category based on the reference data, or in the category immediately above or below the reference category (See Figure 14).

4. Discussion

4.1. Sampling protocol

For all three cases, when estimating either the mean reference Y or the RMSE, the optimal allocation is the most precise sampling protocol, also tied with proportional allocation when estimating the mean reference Y in case 1 and when estimating RMSE in case 3.

When estimating the mean reference Y, samples collected using simple random sampling were the least precise, and when estimating the RMSE, samples collected using stratified random sampling with equal allocation were the least precise. Since the objective of stratification is to reduce the variance in the target variable, t. (See Table 1.)

Most of the sample estimates of the population mean have a wider confidence interval and larger standard deviation than the estimates of RMSE for the same case and sampling protocol.

Land Cover Class (Domain)	RMSE Value Within Domain	Variance of Squared Error Values within Domain	Standard Error of Squared Error Values within Domain	n Samples within Domain
Open Water	0.108	0.000	0.012	1
Developed, Open Space	3.229	2.923	1.710	75
Developed, Low Intensity	1.713	0.643	0.802	25
Developed, Medium Intensity	0.426	0.033	0.182	1
Barren Land (Rock/Sand/Clay)	0.444	0.020	0.141	2
Deciduous Forest	7.163	10.177	3.190	510
Evergreen Forest	7.838	12.008	3.465	607
Mixed Forest	4.108	3.598	1.897	189
Shrub/Scrub	4.748	5.700	2.388	180
Grassland/Herbaceous	3.222	3.064	1.750	60
Pasture/Hay	3.103	2.632	1.622	77
Cultivated Crops	1.082	0.210	0.459	10
Woody Wetlands	4.815	4.956	2.226	254
Emergent Herbaceous Wetlands	1.505	0.994	0.997	11

Table 10: Within-domain estimates of RMSE, variance in pixel-wise errors, standard error of pixel-wise errors, and sample size for one set of samples. The samples are drawn with stratified random sampling using optimal allocation. The strata are defined by the prediction uncertainty metric (the standard error values extracted from the USFS Analytical TCC Standard Error product for 2016) and the domains are defined by the NLCD land cover product for 2016.

There are also larger gains in precision from stratification when estimating the mean reference Y Additionally, the estimates of each respective sampling protocol decreased in precision with each case. The width of the 95% confidence intervals increased between cases 1 and 2 and again between cases 2 and 3. For each of the cases, the sample sizes are the same. Whats more, the percent of the population sampled is actually higher for case 3 than cases 1 or 2, yet the estimates are less precise. This can be explained by the overall variance in the estimated quantity across the population. The variance in the mean reference value (Y) is higher than the variance in the squared error of the population in all cases. The variance in both quantities also increases between cases 1 and 2 and cases 2 and 3. Thus a larger sample is required to capture the variance in case 3 with the same level of confidence as was achieved in case 1. However, the precision of estimates is still higher for all cases and sampling protocols than the precision expected using Equation 5 based on a 95% confidence interval, 2000 samples, and the case-specific standard deviation and population size.

The bias between sample estimates and population values is higher when estimating the population mean than when estimating the population RMSE by case and sampling protocol. In cases 1 and 2, simple random sampling produced the most accurate estimates of the mean reference Y on average, and in cases 2 and 3 stratified random sampling with optimal allocation produced the most accurate estimates of RMSE on average, tied with equal allocation in case 2. All stratified sampling protocols overestimated the mean reference Y for cases 1 and 2, albeit by a very small amount since the estimates produced by all sampling protocols have a bias of less than 1.0. (See Tables 2, 3, 4, 5, 6, and 7.)

4.2. Sample size

The width of the 95% confidence interval of the estimates of the mean reference Y from 1000 samples selected using any stratified random sampling protocol was smaller than the width of the

confidence interval of the estimates from 2000 samples collected using simple random sampling. Further, the confidence interval of the estimates from only 500 samples selected using optimal allocation sampling was also very close to the estimates from 2000 samples selected using simple random sampling. This finding suggests that if a smaller sample is required, even for very small samples, stratified random sampling is preferred to simple random sampling.

When estimating RMSE, all sampling protocols are closer in precision when using the same sample sizes than they are when estimating the mean reference Y. Estimates of RMSE from samples drawn using proportional allocation and optimal allocation sampling protocols with 1500 samples exceed the precision of simple random sampling with 2000 samples. Thus, gains in precision of estimates from optimal allocation or proportional allocation sampling, could still justify a slightly smaller sample size than simple random sampling.

The results of this analysis show that, for all sampling protocols there is a point of convergence where additional samples only provide marginal gains in the confidence of the estimate. Exploratory analysis using progressive sampling as demonstrated in Blatchford et al (2021) may help to determine the ideal sample size when there are limitations to budget or other barriers to sampling [27].

We also show, that for estimates of the mean reference Y using 1000 or more samples, optimal allocation and proportional allocation, perform better than equal allocation in terms of estimate precision. This provides a corollary to findings in other studies, such as Zhu, et al (2016) which showed that training data extracted using an optimal strategy drastically improved the overall accuracy of classification model, followed by proportional allocation which performed better than equal allocation so long as samples sizes are large enough that rare classes are adequately represented [28]. In our study, we also see that, when only 500 samples are used, estimates of reference Y with stratified sampling equal allocation are slightly higher in precision than those that use proportion allocation, likely due to the inability to capture rare classes with proportional allocation sampling.

4.3. Domain estimation

The domain estimation exercise shown above indicates that RMSE in the 2016 USFS Analytical TCC product [20] is expected to be highest between 50 and 60% TCC and lowest under 20% TCC. The variance and standard error in squared errors across the domain is also highest between 50 and 60% TCC. This could reflect challenges during the collection of training data, particularly, difficulties in interpreting mixed landscapes due to shadow and mis-classification of features like shrubs, as well as improved interpretation over bare land (See Table 9). These observations are reflected when error is reported across land cover domains, where the highest errors are expected in deciduous and evergreen forest classes, and the lowest errors are expected in open water, developed, medium intensity, and barren land classes. Higher variance in errors in deciduous and evergreen forest classes is likely due to the same challenges in interpretation (See Table 10).

The confusion matrix constructed demonstrate that the majority of samples fall in high TCC classes, as defined by both $\overline{\hat{Y}}$ and the simulated reference Y. This is likely due to the mask applied to the product before sampling, which removed non-forest pixels from analysis. Based on this confusion matrix, the precision of the TCC model is expected to fall within +/- 15% TCC (See Figure 14)..

5. Conclusion

We demonstrate the feasibility of using the suite of predictions from ensemble modeling approaches like random forest models as a proxy for within-cell variance in optimal allocation sampling. We use these values to determine the distribution of samples across strata in order to

estimate both the mean reference value in the population, and the RMSE across the population. Additionally, when estimating RMSE, we used the values of prediction uncertainty to delineate our stratification, and we use the mean of predictions to delineate the stratification when estimating the population mean. We show that stratified random sampling can improve the precision of sample estimates if used in place of simple random sampling, possibly allowing for smaller sample sizes, especially when estimating quantities with high variance across the population. We observe that differences in the precision of estimates can be attributed to both the variance in the estimated quantity across the population as well as the sample size, and we see a non-linear convergence of precision with added samples across both estimated quantities and several sampling protocols indicating that a progressive sampling approach could be useful. We recommend caution as we show that the equal allocation sampling protocol can out-perform proportional allocation sampling when small sample sizes are used likely due to the inability for proportional allocation to capture rare classes. We further observe that optimal allocation provides the highest-precision estimates of all of the sampling methods, likely due to its ability to increase the sampling intensity over strata with high prediction uncertainty. This observation holds across simulated examples with normally distributed and non-parametric data, in the estimation of both the population mean and the population RMSE.

6. Acknowledgements

This work was supported by the USDA National Institute of Food and Agriculture McIntire Stennis project entitled “Precision Forestry for Southern Pine Carbon Monitoring” and the overall research project was funded through by the United States Forest Service (USFS) under the grant “TCC 2021 Research and Development, 18-JV-11242305-035”. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the National Institute of Food and Agriculture (NIFA) or the United States Department of Agriculture (USDA).

References

[1] C. Justice, A. Belward, J. Morisette, P. Lewis, J. Privette, and F. Baret, "Developments in the 'validation' of satellite sensor products for the study of the land surface," *International Journal of Remote Sensing*, vol. 21, no. 17, pp. 3383–3390, 2000. DOI: 10.1080/014311600750020000. eprint: https://doi.org/10.1080/014311600750020000. [Online]. Available: https://doi.org/10.1080/014311600750020000.

[2] R. G. Congalton and K. Green., *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices.* 2nd ed. Boca Raton, FL: CRC Press, 2009, ISBN: 978-1-4200-5512-2.

[3] C. M. Roelfsema and S. R. Phinn, "Validation," in *Coral Reef Remote Sensing*, J. A. Goodman, S. J. Purkis, and S. R. Phinn, Eds. New York, NY: Springer, 2013, ch. 14, pp. 375–401, ISBN: 9789048192915.

[4] S. Ye, R. G. P. Jr., and R. Rakshit, "A review of accuracy assessment for object-based image analysis from per-pixel to per-polygon approaches.," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 141, pp. 137–147, 2018. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2018.04.002.

[5] S. V. Stehman and G. M. Foody, "Accuracy assessment," in *The SAGE Handbook of Remote Sensing*. Asheville, NC: SAGE Publications Inc., 2020, 11—26. [Online]. Available: https://dx.doi.org/10.4135/9780857021052.

[6] S. Brogaard and R. Olafsdottir, "Ground-truths or ground-lies? : Environmental sampling for remote sensing application exemplified by vegetation cover data," L. U. Department of Physical Geography, Ed., vol. 1, 1997, pp. 121–137.

[7] G. M. Foody, "Sample size determination for image classification accuracy assessment and comparison," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5273–5291, 2009. DOI: 10.1080/01431160903130937. eprint: https://doi.org/10.1080/01431160903130937. [Online]. Available: https://doi.org/10.1080/01431160903130937.

[8] D. M. Hawkins and J. Kraker, "Deterministic fallacies and model validation," *Journal of Chemometrics*, vol. 24, no. 3-4, pp. 188–193, 2010. DOI: 10.1002/cem.1311. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.1311. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.1311.

[9] S. V. Stehman and G. M. Foody, "Key issues in rigorous accuracy assessment of land cover products," *Remote Sensing of Environment*, vol. 231, pp. 1–23, 2019. [Online]. Available: https://doi.org/10.1016/j.rse.2019.05.018.

[10] W. A. Bechtold and P. L. Patterson, "The enhanced forest inventory and analysis program–national sampling design and estimation procedures," in, ser. Gen. Tech. Rep. SRS-80, Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, 2005. [Online]. Available: https://doi.org/10.2737/SRS-GTR-80.

[11] W. G. Cochran, *Sampling Techniques*, 3rd ed. John Wiley & Sons, Inc., 1977, ISBN: 9780471162407.

[12] F. Freese, *Elementary forest sampling*, ser. Agriculture Handbook. Washington, DC: U.S. Department of Agriculture, Forest Service, 1962, vol. 232. [Online]. Available: http://www.fs.fed.us/fmsc/ftp/measure/cruising/other/docs/AgHbk232.pdf.

[13] T. E. Avery and H. E. Burkhart, *Forest Measurements*, 4th ed. McGraw-Hill, 1994, ISBN: 9780070025561.

[14] A. K. Waljee, P. D. Higgins, and A. G. Singal, "A primer on predictive models," *Clinical and translational gastroenterology*, vol. 5, no. 1, pp. 1–4, 2013. [Online]. Available: https://doi.org/10.1038/ctg.2013.19.

[15] L. Ji and K. Gallo, "An agreement coefficient for image comparison," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 7, pp. 823–833, 2006.

[16] J. W. Coulston, C. E. Blinn, V. A. Thomas, and R. H. Wynne, "Approximating prediction uncertainty for random forest regression models," *Photogrammetric Engineering and Remote Sensing*, vol. 82, no. 3, 2016, ISSN: 00991112. DOI: 10.14358/PERS.82.3.189.

[17] M. Hansen, R. DeFries, J. Townshend, L Marufu, and R Sohlberg, "Development of a modis tree cover validation data set for western province, zambia," *Remote Sensing of Environment*, vol. 83, no. 1, pp. 320 –335, 2002, The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring, ISSN: 0034-4257. DOI: https://doi.org/10.1016/S0034-4257(02)00080-9. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425702000809.

[18] M. Schlather, A. Malinowski, P. J. Menck, M. Oesting, and K. Strokorb, "Analysis, simulation and prediction of multivariate random fields with package RandomFields," *Journal of Statistical Software*, vol. 63, no. 8, pp. 1–25, 2015. [Online]. Available: http://www.jstatsoft.org/v63/i08/.

[19] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/.

[20] United States Department of Agriculture, Forest Service, *2016 usfs "analytical" tcc*, 2019. [Online]. Available: https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/.

[21] L. Breiman, A. Cutler, A. Liaw, and M. Wiener., *Breiman and cutler's random forests for classification and regression*, version 4.6-14, 2018. [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/.

[22] W. Berry and S. Feldman, *Multiple Regression in Practice (Quantitative Applications in the Social Sciences)*. Thousand Oaks, CA: SAGE Publications, 1985.

[23] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar, "Multicollinearity in regression analyses conducted in epidemiologic studies," *Epidemiology (Sunnyvale)*, vol. 6, no. 2, pp. 1–20, 2016. DOI: 10.4172/2161-1165.1000227.

[24] J. W. Coulston, G. G. Moisen, B. T. Wilson, M. V. Finco, W. B. Cohen, and C. K. Brewer, "Modeling percent tree canopy cover–a pilot study," *Photogrammetric Engineering and Remote Sensing*, vol. 78ID - 22, no. 7, pp. 715–727, 2012, ISSN: 00991112. DOI: 10.14358/PERS.78.7.715. [Online]. Available: internal-pdf://76.174.207.146/2012-Coulston-PhotogrammetricEngineeringand.pdf{\%}0Ahttp://dx.doi.org/10.14358/PERS.78.7.715.

[25] J. Dewitz, *National Land Cover Database (NLCD) 2016 Products: U.S. Geological Survey data release*, 2019. [Online]. Available: https://doi.org/10.5066/P96HHBIE.

[26] P. M. ATKINSON, "Optimal ground-based sampling for remote sensing investigations: Estimating the regional meant," *International Journal of Remote Sensing*, vol. 12, no. 3, pp. 559–567, 1991. DOI: 10.1080/01431169108929672. eprint: https://doi.org/10.1080/01431169108929672. [Online]. Available: https://doi.org/10.1080/01431169108929672.

[27] M. L. Blatchford, C. M. Mannaerts, and Y. Zeng, "Determining representative sample size for validation of continuous, large continental remote sensing data," *International Journal of Applied Earth Observations and Geoinformation*, vol. 94, pp. 1–11, 2021. [Online]. Available: https://doi.org/10.1016/j.jag.2020.102235.

[28] Z. Zhu, A. L. Gallant, C. E. Woodcock, B. Pengra, P. Olofsson, T. R. Loveland, S. Jin, D. Dahal, L. Yang, and R. F. Auch, "Optimizing selection of training and auxiliary data for operational land cover classification for the lcmap initiative," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 206 –221, 2016, ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2016.11.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271616302829.

sampling.bib

Chapter 5

Conclusions

5.1 Summary and conclusion

Tree canopy cover is a critically important measure in the management of forests. As such, there are many studies which endeavor to estimate models of tree canopy cover across the globe, often using data-hungry machine-learning algorithms. This research focused on the modeling of tree canopy cover using methods made possible due to recent improvements in computing, data availability, and innovations in data collection.

The opening of the Landsat archive has enabled the utilization of a nearly complete record of imagery collected since the 1970's. We've used this data to create harmonic regression variables which were used in the first chapter of this dissertation to estimate models of tree canopy cover. In the study sites selected, we found that these variables improved existing models and that the harmonic regression variables alone contained more information than a suite of other median composite variables, suggesting that these data may be useful in other applications as well.

In the second chapter, we explored the use of crowdsoucing with Amazon's Mechanical Turk to produce tree canopy cover interpretations from high resolution aerial imagery. This study compared expert and crowd interpretations and modeled crowd interpretation reliability in order to identify factors which resulted in crowd interpretations that were more similar to experts'. We found that crowd workers were highly susceptible to fatigue, as the number of plots interpreted was the strongest indicator of interpretation reliability. We also found that most crowd workers used the natural color display, while most experts used the false color display. The crowd interpretations completed using mostly the false color display proved to be more similar to expert interpretations. Finally, when crowd workers that were motivated to complete interpretations quickly, either due to competition for access to HITs or high potential earnings, the reliability of their interpretations declined. Since the payment structure of Amazon's Mechanical Turk is based on the number of tasks completed instead of the quality of the work produced, this was a notable finding.

Finally, we explored the use of stratified random sampling with optimal allocation to improve the precision of estimates of the reference mean and model RMSE. We repeatedly allocated samples using measures of prediction uncertainty derived from the predictions of random forest regression models to represent the variability of the target measurement within

each stratum. We also collected repeated samples using simple random sampling, stratified random sampling with proportional allocation, and stratified random sampling with equal allocation and compared the variability of estimates from each set of samples using each sampling protocol. We showed that stratified random sampling with optimal allocation can produce more precise estimates, even with smaller sample sizes in some cases.

The results presented here highlight the challenges and opportunities in the spatial modeling and evaluation of tree canopy cover datasets. These insights could support future studies using ensemble models or machine learning algorithms where large quantities of reference data are required for both training and validation.

5.2 Future work

Going forward, the results of this work could be applied to modeling change in tree canopy cover. In our first study, the two time intervals for the reference imagery used were too close to result in a significant amount of training sites that showed a detectable change.

In the second study, we provide recommendations to be used in similar studies in the future. In particular, it would be of interest to see if additional data sources, such as LiDAR point clouds could be used to improve tree canopy cover interpretations.

Finally, for our third study, it would be useful to demonstrate the utility of stratified random sampling with optimal allocation using reference data that was collected on the ground rather than simulated data. It would be of interest to test this method using various sample sizes to better understand whether smaller sample sizes would produce an adequate estimate of the reference mean and RMSE.