Making national forest inventory data relevant for local forest management

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

Barry Tyler Wilson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Joseph F. Knight, Advisor

July 2018

Acknowledgements

This dissertation would not have been possible without the guidance and assistance of so many. I would like to thank my committee members Joe Knight, Ron McRoberts, Glen Meeden, and Chris Edgar for the countless hours of advice, intellectual debate, and encouragement they have provided me over the past several years. I am grateful for the support of Dennis May, program manager of Forest Inventory and Analysis, and Tom Schmidt, former assistant director of the United States Forest Service Northern Research Station, who gave me the opportunity to pursue this research. I would like to acknowledge Marv Bauer and Tom Burk for their pivotal roles in setting me on this path many years ago, when I first arrived as a graduate student at the University of Minnesota. I am also grateful for numerous colleagues in the Forest Service who willingly provided a sounding board for so many of the ideas contained here. Finally and most importantly, I would like to thank my wife and children for their inspiration, patience, and tireless encouragement. I truly could not have done this without you.

Abstract

The national forest inventory conducted by the United States Forest Service Forest Inventory and Analysis (FIA) program provides information for strategic level decisions regarding national and regional management of forest ecosystem goods and services. However, the sampling intensity typically limits the application of traditional direct estimators to areas the size of a large county, if not larger. This dissertation describes methods for combining FIA data with auxiliary information to enhance its relevance for local forest management.

Background information is provided on the way population estimates are currently produced, and how precision can be improved via satellite imagery. A study is described that uses features extracted from dense time series of Landsat imagery with a model-assisted direct estimator. The study examined the relative predictive power of land cover models incorporating extracted spectro-temporal features versus composite imagery alone. Non-parametric models were fitted for multiple attributes measured on FIA plots using all archived Landsat scenes for Minnesota from 2009-2013. The estimated coefficients developed by harmonic regression of the time series imagery were shown to be moderately to highly correlated with tree-level and land cover attributes. When comparing results for spectro-temporal features to monthly image composites, regression models had greater explained variance and classification models had greater overall and individual class accuracies.

Finally, a study is presented that tested the performance of a proposed variant of the *k*-nearest neighbors algorithm for areas too small to use a direct estimator. Spectro-temporal features were extracted for one ecological unit in Minnesota. A simulated population of tree canopy cover was sampled at FIA plot locations. The proposed algorithm was used to fit a non-parametric model to predict tree canopy cover that incorporates the spectro-temporal features. The model was used to construct predictive intervals for spatial domains over a range of domain sizes, and the resultant tests showed the coverage probability approached the theoretical value for areas as small as 1200 hectares. The study suggests that, given good auxiliary data and models, the scale of valid inference using FIA data can approach what is needed for local decision makers.

Table of Contents

List of Tables

List of Figures

Introduction

The United States Forest Service (USFS) Forest Inventory and Analysis (FIA) program conducts a continuous annual national forest inventory (NFI) of the forests of the United States. The FIA program collects information on a quasi-systematic sample of permanent plots established at a base sampling intensity of approximately one plot per 2,400 hectares (Reams et al., 2005). Field crews collect data for an extensive suite of traditional forest mensuration variables on all measured plots and, in a subsequent phase, also collect data for multiple forest health variables on a subset of these plots. For many of these variables, data collection is only possible in the field since they cannot be measured accurately, or at all, by remote means such as through airborne or spaceborne sensors. As of FIA's 2016 fiscal year, the annual program costs were almost $76 million USD for the base system of over 323,000 plots covering the United States, excluding interior Alaska, with approximately 12% of these plots (~39,000) measured annually either by field visits or aerial photo interpretation (Vogt & Smith, 2017). Of the approximately $61 million USD in direct expenses incurred by the program, just over 50% were associated with those components closely related to the plot measurement effort, and do not include costs associated with data management, nor the analyses based on these data. Using these values, the average annual plot measurement cost works out to be more than $800 USD per plot, including all sampling phases of the inventory, and average costs being substantially greater for the ~14,000 base forested plots measured in the field.

Many land management entities in the United States require information for forests that cover areas containing few FIA plots at the base sampling intensity (1 plot per 2,400 ha). One such entity might be a ranger district within the USFS National Forest System, responsible for managing forest fuels to mitigate wildfire risk in a 100,000 ha district (~40 FIA plots), 10,000 ha watershed (~4 plots), or 100 ha stand (no plots). Another might be a 2,500 ha State Wildlife Management Area (~1 plot), requiring information on forest structure and composition to manage for habitat of a critical wildlife species. Even discounting the average FIA per-plot cost because of savings achieved by restricting both the sample and subsequent field travel to a much smaller area, conducting a comparable annual forest inventory at the sampling intensity needed to support local forest management decisions could be prohibitively expensive. An attractive option for such land managers might be to use the existing FIA sample, borrowing strength from plots in similar neighboring areas and in conjunction with relatively inexpensive remote sensing imagery, in an attempt to fulfill their information needs.

This dissertation explores the issue of small area estimation (SAE), in the context of NFI in general but with a special focus on FIA, where the small area sample size is too small to make use of the traditional direct estimators. Through the use of SAE techniques, with auxiliary remote sensing imagery, the spatial scale of application of NFI data can be brought to a level approaching what is needed to support local forest management decisions. The dissertation is presented as a series of three interrelated chapters, though each was prepared as a manuscript that could stand on its own. Chapter 1 provides an organizing framework and review of SAE techniques in general, as well as a summary of numerous studies employing those techniques with NFI, or similar sample survey data, and remote sensing imagery. Chapter 2 demonstrates, via a case study in Minnesota, the utility of harmonic regression for feature extraction from dense time series of Landsat imagery for the purposes of modeling several continuous and categorical variables from FIA data. Chapter 3 proposes a variant of the nonparametric *k*-nearest neighbors algorithm, called Bamboo *k*NN, that simultaneously optimizes the model, performs feature selection, and estimates prediction bias. The chapter includes an application study of Bamboo *k*NN, using one ecological region in Minnesota and a sample drawn from a simulated tree canopy cover dataset at FIA plot locations with the features extracted in Chapter 2 as auxiliary variables. Chapter 2 was published in the *ISPRS Journal of Photogrammetry and Remote Sensing*. Chapters 1 and 3 are being prepared for submission to peer-reviewed journals.

Chapter 1: A review of small area estimation techniques using national forest inventory and remotely sensed auxiliary data

Summary

National forest inventory (NFI) data provide the primary source of information for strategic analysis of the status and trends of a nation's forest resources. For large domains that contain many sample units, a design-based mode of inference is an appropriate choice, since estimators under this approach are generally assumed to be unbiased. However, for small areas the sample size is often too small to make reliable estimates of population parameters using direct estimators, which rely strictly upon the sample units drawn from the domain. When auxiliary variables are available for all population units, a model-based mode of inference can be used to provide more precise estimates, yet the estimators are potentially biased. This review documents the current state of small area estimation (SAE) approaches for use with NFI data. It provides an organizing framework for categorizing approaches to SAE, with emphasis given to estimators that use auxiliary variables that are known for all population units, and can be collected efficiently over a large spatial extent, such as the case of satellite remote sensing imagery.

Introduction

A multi-resource national forest inventory (NFI) provides the scientific foundation for strategic (i.e. national to regional in scale) analysis of the status and, in the case of a continuous NFI, trends of a nation's forest resources as they are managed for several, often conflicting benefits like timber, bioenergy, wildlife habitat, recreation, and clean water. Data from an NFI can also be used to satisfy a nation's international reporting requirements, such as those of the United Nations (UN) Framework Convention on Climate Change for greenhouse gas emissions from the forestry sector and the UN Food and Agriculture Organization Global Forest Resources

Assessments for general status and trends of the world’s forests. However, because of the expense associated with establishing forest inventories, and particularly with collecting the data, there is growing interest in using NFI data for local (i.e. smaller than regional) analyses that involve estimating population parameter values of interest for ever smaller areas, often pushing the limits of the NFI sample design.

The objective of this review is to document the current state of small area estimation (SAE) using national forest inventory and remotely sensed auxiliary data, as well as to provide an organizing framework for categorizing approaches to SAE. Emphasis is given to estimators that use auxiliary variables that are known for all population units, particularly where the auxiliary variables are continuous rather than categorical, available in a digital raster format for ease of processing, and can be collected efficiently over a large spatial extent, such as the case of satellite remote sensing imagery (McRoberts, 2011).

In the context of this manuscript, a domain is defined as a subset of a population for which some measure, such as the mean or total, is sought. SAE is then defined as the situation in finite population sampling and estimation where the sample size is too small relative to the variability in the domain of interest for the use of direct estimators, which rely solely upon the sample values of the units drawn from the domain for statistical inference. SAE techniques borrow strength, or information, from all sample units, including those outside of the domain, to make inferences about the domain population, typically resulting in less uncertainty than can be achieved using a direct estimator. For future reference, the terms *domain* and *small area* are used interchangeably (Rao, 2003).

The Oxford English Dictionary defines inference as “the drawing of a conclusion from known or assumed facts or statements.” In the context of survey sampling, statistical inferences are made about population parameters based upon a sample, generally accompanied by some measure of uncertainty, and expressed in probabilistic terms (Dawid, 2006). Gregoire (1998) and Little (2004) describe two primary modes of inference used in the analysis of finite population sampling data: design-based and model-based approaches. The two approaches make different assumptions about the nature of random variation in a sample.

In the design-based approach, the values associated with population units are assumed to be fixed, apart from measurement error, while the values indicating whether or not a population unit is included in the sample, or indicator values, are considered random variables, and each sample unit's inclusion probability is known from the sample design. In this case, the uncertainty in any estimate of a finite population parameter stems from the distribution of the indicator values. Design-based estimators of these population parameters are inherently unbiased, or nearly so, and result in deviations from the true value that are assumed to be approximately normally distributed for large samples. However, for any particular sample, the deviation between the estimate and the true value may be substantial.

Under the model-based mode of inference, the values associated with population units are represented as random variables that are realizations of an underlying stochastic (i.e. non-deterministic) superpopulation model. Two types of models are used: frequentist and Bayesian. In the case of frequentist superpopulation models, the population values are assumed to be generated from a random sample from a larger "superpopulation" with a probability distribution based on fixed parameters that are estimated from the sampled values. Inferences are based on the joint distribution of the population and indicator values. Bayesian superpopulation models require the specification of a prior distribution for the population parameter values. Inferences are based on the posterior predictive distribution of the non-sampled values of the population given the sampled values.

Direct estimators

Direct estimators use only the sample values from the small area of interest, along with weights for sample units that stem from the sample design, and in some cases auxiliary information about the population of interest. All of the direct estimators described in this section use design-based inference. Horvitz and Thompson (1952) developed an estimator that provides a general framework for direct estimation under multiple sample designs, whether or not auxiliary variables are available. The Horvitz-Thompson (HT) estimator for the population total Y is,

$$\hat{Y}_{ht} = \sum_{i=1}^{N} I_i \, d_i y_i,$$

where, for the *i*th unit in a population of size *N*, I_i is a random variable that indicates whether or not the unit is in the sample, d_i is the unit's design weight, and y_i is the observation of the variable of interest for the unit. The design weight of a unit is the inverse of its probability of inclusion in the sample, π_i, or $d_i = \pi_i^{-1}$. The inclusion probabilities are determined by the sample design, which defines whether or not the sample units are to be drawn, for example, from a simple random sample (SRS), systematic sample, or cluster sample. Yates and Grundy (1953) developed an estimator of the variance of the HT estimator,

$$\widehat{Var}(\hat{Y}_{ht}) = \sum_{i=1}^{n-1} \sum_{j=1+1}^{n} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} + \frac{y_j}{\pi_j}\right)^2,$$

where π_{ij} is the joint inclusion probability of units *i* and *j*. Under SRS with a sample of size *n*, these estimators are much simplified, becoming

$$\hat{Y}_{ht} = N\bar{y},$$

and

$$\widehat{Var}(\hat{Y}_{ht}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)S^2, \text{ where } S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2,$$

respectively. Because of the fundamental assumption in SAE of an insufficient sample for the calculation of a reliable direct estimate, the results from studies utilizing these estimators will not be described, except for the case of model-assisted estimation.

Auxiliary variables

If there are auxiliary variables available for all units in the population, or known population totals for these variables, then stratified or post-stratified estimators can be used in an effort to reduce the variance of the estimator. For example, satellite imagery can be used to classify all population units (pixels) into mutually exclusive groups (strata) of units with similar spectral characteristics that are assumed to represent similar biophysical characteristics related to sample observations. Once the population has been stratified, the proportion of units assigned to each stratum can be calculated, as well as the within-stratum means and variances of the sample unit observations.

The estimators based on Cochran (1977) are presented. The stratified estimator of the total for a population of size *N* and sample of size *n* is,

$$\hat{Y}_{str} = N\sum_{j=1}^{J} w_j\, \bar{Y}_j,$$

where $\bar{Y}_j$ is the mean of the j^{th} stratum ($\sum_{i=1}^{n_j} y_i/n_j$) and w_j is proportion of population units assigned to the j^{th} stratum (N_j/N). The stratified estimator of the variance of the total is,

$$\widehat{Var}(\hat{Y}_{str}) = N^2\sum_{j=1}^{J} w_j^2\hat{\sigma}_j^2/\,n_j,$$

where $\hat{\sigma}_j^2$ is the within-stratum variance of the *j^th^* stratum,

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\bar{Y}_j - y_i)^2.$$

The post-stratified estimator of variance is similar; however, it includes an additional term accounting for the fact that sample sizes within strata are determined after the sample has been collected and are also random variables.

While both of these approaches can lead to greater precision than the direct estimator under SRS, particularly with large samples, both exacerbate the problem of insufficient sample sizes for SAE since sample units are distributed across multiple strata. In the case of post-stratification, there may even be strata that contain no sample units. Within the context of NFI and satellite imagery, McRoberts and Tomppo (2007) provide a succinct review of the progression from double sampling for stratification techniques, originally used with aerial photography, to more recent applications using post-stratified estimation.

Model-assisted estimation

While one approach to estimation is to use the stratified estimator when auxiliary variables are available, an alternative is to use the model-assisted estimator. This is also known as the calibration estimator, or the general/generalized regression (GREG) estimator (Deville & Särndal, 1992). The GREG estimator is a generalization of a class of estimators, such as the ratio and regression estimators, that use values of one or more auxiliary variables for all population units with an assisting model to calibrate the direct estimator. It still uses the design weights and is therefore fundamentally design-based. It also does not borrow strength from sample units outside of the small area of interest. As described in Rao (2011), suppose that the superpopulation model that describes the relationship between unit-level obervations of the variable of interest and the auxiliary variables is,

$$y_i = x_i'\beta + \varepsilon_i,$$

where β are the model parameters and ε_i is the model error. The errors are assumed to be uncorrelated with mean of zero and variance proportional to a known constant q_i. The design-weighted estimator of β given sample s is then,

$$\hat{\beta} = \sum_{i \in s}(d_i x_i x_i'/q_i)^{-1}(d_i x_i y_i/q_i),$$

where d_i is the design weight of the i^{th} unit. The GREG estimator of the population total Y is given by,

$$\hat{Y}_{greg} = \hat{Y} + \hat{\beta}'(X - \hat{X}),$$

where X are the known population totals of the auxiliary variables and $\hat{Y}$ and $\hat{X}$ are the corresponding estimated values for the variable of interest and auxiliary variables using the sampled units and their design weights. The working model used with the GREG estimator does not need to be a linear model, and could instead be nonlinear or nonparametric.

Opsomer et al. (2007) compared design-based estimators for a study area in northeastern Utah, USA using NFI response variables (forest/non-forest and total wood volume) and auxiliary variables from a digitial elevation model, classified Landsat imagery, and a vegetation index derived from an Advanced Very High Resolution Radiometer biweekly composite. The comparison included the direct estimator under SRS, a post-stratified estimator using the classes from a classification of Landsat imagery as strata, and two GREG estimators using a generalized additive model as the working population model. For both variables of interest, the results indicated that the SRS estimate of the population total had the largest variance, followed by the post-stratified estimate, with the two GREG estimators having the smallest, and approximately equal, variances.

Næsset et al. (2011) used a model-assisted approach to estimation of total above-ground forest biomass for a study area in southeastern Norway, as well as for individual tracts (subdivisions of municipalities) within the larger study area. Sample unit observations of above-ground biomass were collected on a stratified sample of forest inventory plots. Auxiliary variables were derived from airborne LiDAR and spaceborne InSAR data. The results showed that estimates of standard errors using the model-assisted approach with LiDAR data were approximately 45% of the associated direct estimate, both for individual tracts and the entire study area. Even the model-assisted approach using the coarser InSAR data produced standard errors that were somewhat smaller than those produced using the direct estimators. The authors also identified two other benefits of the approach: 1) it assures that estimates for sub-populations always sum up to the estimate for the population as a whole, and 2) it produces a map of the attribute of interest.

Indirect estimators

Unlike direct estimators, indirect estimators increase the effective sample size for a small area of interest, and thus decrease the variance of the estimate, by borrowing strength from sample units outside of the domain (Rao, 2003). Indirect estimators can use either design-based or model-based inference.

Implicit superpopulation models

Estimators that use implicit superpopulation models, describing the relationship between the variables of interest and one or more auxiliary variables, also make use of a design-based mode of inference and usually have design variances (i.e. variances that account for the probability of selection) that are smaller than those of direct estimators. However, they are potentially design-biased, meaning that these estimators use the sample design weights, yet the expected value of the estimator differs from the true value. If the implicit superpopulation model is approximately correct, then the design bias should be small and lead to a mean squared error (MSE) that is smaller than that of a direct estimator. This is also true for the explicit superpopulation models and is a hallmark of all indirect estimators.

Synthetic estimators

Synthetic estimators (Steinberg, 1979) are used to produce estimates for a small area of interest under the assumption that it has similar characteristics to the larger sampled population as a whole, for which there is an adequate sample size to specify a superpopulation model between the variable of interest and any auxiliary variables. Synthetic estimators are similar in form to the GREG estimator. However, they do not include the terms that attempt to correct for any bias in the model-based estimator. So, while synthetic estimators do not rely upon a direct estimate and can provide estimates with small variances even for small areas that contain few or no sample units, they can potentially be biased, particularly when small areas are not homogenous across the population. One commonly used synthetic estimator is the synthetic regression estimator,

$$\hat{Y}_{syn} = \hat{\beta}'X,$$

where $\hat{\beta}$ is as defined for the GREG estimator and X is the known population total of the auxiliary variables. While it is relatively straightforward to estimate the variance of a synthetic estimator for a small area, the difficulty is in providing a reliable estimate of the associated bias, which is essential given that the MSE of an estimator is equal to the variance plus the square of the bias.

Katila (2006) compared the performance of three synthetic estimators with a direct estimator, as well as the *k*-nearest neighbors method, for a study area in southeastern Finland. The study examined estimates of forest area and tree volume (total and by species) for small areas of approximately 10,000 ha and 100 ha in size. Observations of variables of interest were collected on NFI plots and auxiliary variables were derived from Landsat ETM+ imagery. A global (i.e. applicable to the entire study area) synthetic ratio estimator, similar to the one described above but using a simple ratio rather than a regression model, was developed. The study also examined two modified synthetic ratio estimators, one that restricts the ratio model to nearby plots and another that uses post-stratified ratios. The results showed that the synthetic estimators tended to underestimate forest area for the larger units and that the global synthetic ratio estimates of tree volume differed significantly from the direct estimates for many of the areas evaluated. The modified synthetic ratio estimators performed somewhat better, in terms of both bias (compared to design-unbiased direct estimators) and precision, though not as well as the k-nearest neighbors method especially for the smallest units.

Composite estimators

Composite, or shrinkage, estimators hedge against the potential design bias (i.e. bias in an estimator that uses the design weights of the sample units) and small variance of a purely synthetic estimator by using a weighted average of a synthetic estimator and a design-unbiased direct estimator that may have large variance because of the small sample size. The James-Stein approach (James & Stein, 1961) is a particularly popular composite estimator that uses common averaging weights for all small areas in the population. A composite estimator for a small area may be written as,

$$\hat{Y}_{comp} = \gamma\hat{Y}_{ht} + (1-\gamma)\hat{Y}_{syn},$$

where γ is a weighting factor between 0 and 1 that controls the relative shrinkage toward the direct estimate depending upon the sample size within the small area. Assuming that small areas can be enumerated a priori, individual weights can be specified for each small area. The optimal values for γ minimize the MSE of $\hat{Y}_{comp}$ and can be estimated as,

$$\hat{\gamma} = \frac{\widehat{MSE}(\hat{Y}_{syn})}{\widehat{MSE}(\hat{Y}_{syn}) + \widehat{Var}(\hat{Y}_{ht})}.$$

The James-Stein estimate of γ is based on the minimization of the MSE of $\hat{Y}_{comp}$ assuming a common value for all small areas.

Nonparametric models

Nonparametric models do not assume a specific functional form, nor do they assume normality in the distribution of deviations of observations from their means. Examples include kernel regression, multivariate adaptive regression splines, classification and regression trees, tree-based ensemble methods like Random Forests and boosted trees, as well as some generalized additive models. One of the more common nonparametric approaches used in forestry applications is *k*-nearest neighbors (*k*NN) estimation. *k*NN is defined as a class of techniques that includes a broad range of nearest neighbor imputation approaches that differ only in specific modeling choices, such as the number of nearest neighbors to use for imputation and the distance metric used to determine proximity. It is a form of kernel regression where the kernel density is fixed rather than the kernel width (Wand & Jones, 1995). One of the principal benefits of *k*NN is that it is both a nonparametric and multivariate method, providing predicted values for all population units of all variables observed for sample units. Population parameters can then be estimated from these predicted values, accounting for prediction uncertainty and covariance among predicted values.

Because *k*NN is a nonparametric approach to making unit-level predictions. This means that estimates of small area population parameters will be based on a collection of unit-level predictions. As with all model-based approaches, the corresponding small area estimates of variance must account for covariance among unit-level predictions, each of which is a random variable. Numerous studies have reported examinations of *k*NN methods for unit-level prediction and mapping of NFI variables. Eskelson et al. (2009) provide an excellent review of the development of nearest neighbor imputation approaches to prediction of missing values using NFI data with auxiliary variables. Some key early studies include the work of Tomppo (1990), Tokola et al. (1996), and Katila and Tomppo (2001) in Finland, and Moeur and Stage (1995), Franco-Lopez et al. (2001), Ohmann and Gregory (2002), and McRoberts et al. (2002) in the United States. However, most of these studies, as well as many others that have followed, have focused on the effects of an assortment of modeling choices on prediction errors for individual population units, such as predictor variables, number of neighbors, weighting functions, distance metrics, etc. Relatively few have addressed *k*NN methods in the context of SAE and attempted to quantify the uncertainty associated with estimates based on the summarization of pixel values within a small area of interest. McRoberts (2012) provides a thorough review of the *k*NN estimator for SAE with NFI data and auxiliary variables, along with an analysis of the impacts of several modeling choices, such as distance metrics, numbers of neighbors, and weighting of neighbors.

McRoberts et al. (2007) developed a model-based approach to SAE using the *k*NN technique. In the terminology for *k*NN used by the authors, the auxiliary variables constitute the feature space. The set of population units for which observations of both response and auxiliary variables are available (i.e. the sampled units) is defined as the reference set. The set of population units for which estimates of the response variables are required is defined as the target set. An observation of a response variable for the i^{th} unit of the population can be expressed as,

$$y_i = \mu_i + \varepsilon_i,$$

where ε_i is the random deviation of the observation, y_i, from its mean, μ_i. In the case of *k*NN, the estimate of μ_i is,

$$\hat{\mu}_i = \tilde{y}_i = (\sum_{j=1}^{k} w_{ij})^{-1} \sum_{j=1}^{k} w_{ij}\, y_{ij},$$

where y_{ij} is the observation associated with the *j*th nearest neighbor in the reference set to the *i*th unit of the target set, $w_{ij} = d_{ij}^{-t}$, d_{ij} is the distance in feature space between the *i*th target set unit and the *j*th nearest reference set unit with respect to the distance metric, and typically $0 \leq t \leq 2$. In the absence of spatial autocorrelation among observations in the reference set, an estimate of the variance of $\hat{\mu}_i$ is,

$$\hat{\sigma}_i^2 = k^{-1} \sum_{j=1}^{k} (y_{ij} - \hat{\mu}_i)^2.$$

The population mean of a small area, μ, is then estimated as,

$$\hat{\mu}_{knn} = N^{-1} \sum_{i=1}^{N} \hat{\mu}_i = N^{-1} \sum_{i=1}^{N} \tilde{y}_i,$$

where *N* is the small area population size. An estimate of the variance of $\hat{\mu}_{knn}$ is,

$$\widehat{Var}(\hat{\mu}_{knn}) = \widehat{Var}(N^{-1} \sum_{i=1}^{N} \hat{\mu}_i) = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{Cov}(\hat{\mu}_i, \hat{\mu}_{\mathrm{j}}).$$

In the absence of spatial autocorrelation, an estimate of $Cov(\hat{\mu}_i, \hat{\mu}_{\mathrm{j}})$ is,

$$\widehat{Cov}(\hat{\mu}_i, \hat{\mu}_{\mathrm{j}}) = \hat{\sigma}_\iota \hat{\sigma}_j m_{ij} k^{-2},$$

where m_{ij} is the number of common nearest neighbors used to calculate $\hat{\mu}_i$ and $\hat{\mu}_{\mathrm{j}}$.

The study used NFI data from the FIA program in the United States, using a spatially balanced random sample, along with auxiliary data on visible and near-infrared spectral reflectance from Landsat 5 TM and Landsat 7 ETM+ imagery, for one scene in northeastern Minnesota, USA. Estimates of proportion forest area, volume, basal area, and stem density and their associated variances were made for 15 circular small areas of interest, each with a radius of 10 kilometers and containing approximately 20-25 sample plots, using both design-based and model-based approaches, where k=5 and neighbors were given equal weights. Variograms indicated that the effective range of spatial autocorrelation was much smaller than the average distance between sample plots and could therefore be ignored.

The results did not suggest any apparent bias in the *k*NN estimator for any of the response variables. Small area model estimates for all response variables were generally within two standard errors of the corresponding direct estimates. Model-based estimates of variance averaged almost an order of magnitude smaller than the design-based estimates. However, no formal measure of bias was estimated for small areas, which precluded making estimates of mean square error. Also, calculation of the estimates of covariance among the unit-level predictions proved to be quite computationally intensive for the small areas used in the study, although the authors did demonstrate an approach to working around this problem that reduced computation time by more than a factor of 50.

Magnussen et al. (2009) developed a similar model-based estimator for *k*NN approaches to SAE. Their estimator differs from the one developed earlier by including models describing the relationship between the auxiliary and response variables thereby defining a new feature space, factoring in unequal weighting of the k reference values based on feature space distance, and providing an estimate of bias for unit-level predictions. The results of their study, based again on FIA survey data and Landsat ETM+ imagery from Minnesota, confirmed that the model-based MSE of an estimate for a small area can be several times larger than the naïve MSE that fails to account for the covariance among unit-level predictions. Their variance estimator showed good agreement with that of McRoberts et al. (2007) for large values of *k*, but poorer agreement for small values of *k*. They concluded that the agreement was best when the true value of the response variable was located inside the convex hull spanned by the values of the *k*-nearest neighbors, which is more likely to be the case with a larger value of *k*. They also cautioned that the need to fit additional models required a larger reference set and recommended a lower limit of 300 reference units.

Baffetta et al. (2009) used a design-based, model-assisted approach to variance estimation with the *k*NN technique. Their empirical difference estimator is shown to be approximately unbiased and appropriate for use with data from a probability sample, such as is typically the case with an NFI. It uses the sample design weights and is closely related to the calibration estimator discussed. However, it is a direct estimator that requires an adequate number of sample units from the area of interest, a condition that is assumed not to be met in the case of SAE. Also, complex sample designs may preclude the use of the empirical difference estimator because of difficulties in quantifying inclusion probabilities of sample units.

Magnussen et al. (2010) proposed a resampling approach to variance estimation of small area population totals of NFI variables via *k*NN called the modified balanced repeated replication (BRR) estimator. Resampling methods are nonparametric approaches that involve repeatedly sampling from the sample in order to make inferences about population parameters. The modified BRR estimator accounts for covariance among predicted small area population unit values and is calculated from a small number (approximately 100) of balanced half-samples. In the proposed approach, the classic BRR estimator is modified by imputing response variables from the included half-sample units to the complementary half-sample units prior to imputation to the non-sampled units. The estimator can be used with NFI data collected under SRS and cluster sampling designs. The results suggested comparable performance to the empirical difference estimator for large areas and improved performance for small areas. However, the modified BRR estimator proved to be computationally intensive and was limited to sample sizes smaller than 1,984 units, although the authors suggested some faster shortcuts giving approximate results.

Breidenbach, Nothdurft, and Kändler (2010) compared three methods for computing distance metrics to each other for use with the *k*NN technique, when utilizing the model-based parametric variance estimator of McRoberts et al. (2007), as well as to the design-based direct estimator. The study included forest inventory and airborne laser scanner (ALS) data for a 50-km^2 study area of state and municipal forests near Freiburg, Germany. The study area was tesselated into a fine mesh of hexagonal units of 452 m^2, approximately the size of a sample plot, and height and density metrics extracted from the ALS data were summarized for each hexagonal unit and used as auxiliary variables with the *k*NN approach to estimation.

The canonical correlation analysis (CCA) distance metric used in the most similar neighbor (MSN) imputation approach (Moeur & Stage, 1995; LeMay & Temesgen, 2005), as well as the Random Forests (RF) (Breiman, 2001), and Random Forests based on conditional inference trees (CF) (Strobl et al., 2008) distance metrics were calculated by fitting the auxilary variables to the response variable, in this case total standing timber volume. Under each of the three methods, both forward selection and backward elimination procedures were used to determine the best subset of predictor variables to include in the corresponding model. Estimates of total volume, along with the individual volumes for the primary tree species in the study area, were computed for each hexagonal unit via *k*NN with $k = 8$, using each of the three distance metrics.

The results indicated that the *k*NN estimator using the RF distance metric was consistently more biased for unit-level predictions of total timber volume than using either the CCA or CF distance metric. However, for all three methods the root mean squared deviation was substantially smaller than the corresponding standard deviation in timber volume observed on the inventory plot. The three methods performed almost equally well for SAE at the stand level, with no indication of bias in the estimator for stand mean values. The parametric variance estimates tended to be smaller than the corresponding direct estimates for all stands containing at least two inventory plots. A comparison of species-specific means was difficult to conduct because of the large variances associated with the design-based estimates. The authors concluded that the MSN method was best suited to their application because of 1) the apparent lack of bias in the estimator for unit-level predictions, 2) slightly smaller variances for large stands relative to the CF method, and 3) fast analytical solutions to the determination of nearest neighbors.

McRoberts et al. (2011) compared the parametric variance estimator using the *k*NN technique discussed earlier with resampling methods, in particular the jackknife and bootstrap estimators. Using the jackknife approach for a sample of size n, the j^{th} jackknife sample is defined to be the original sample with the j^{th} unit removed. The j^{th} estimate of the population parameter, $\hat{\mu}_j$, is obtained from the j^{th} jackknife sample. The jackknife estimate of the population parameter is,

$$\hat{\mu}_{jack} = n^{-1} \sum_{j=1}^{n} \hat{\mu}_j.$$

The jackknife estimate of bias is,

$$\widehat{Bias}\ (\hat{\mu}_{jack}) = (n-1)\ (\hat{\mu}_{jack} - \hat{\mu}),$$

where $\hat{\mu}$ is the estimate obtained using the complete sample. The jackknife estimate of variance is,

$$\widehat{Var}\ (\hat{\mu}_{jack}) = (n-1)\ n^{-1} \sum_{j=1}^{n} (\hat{\mu}_{jack} - \hat{\mu}_j)^2.$$

Similarly, for a sample of size *n*, the *b*th bootstrap sample is defined to be a random sample with replacement of size *n* from the original sample. The *b*th estimate of the population parameter, $\hat{\mu}_b$, is obtained from the *b*th bootstrap sample. The bootstrap population estimate is,

$$\hat{\mu}_{boot} = {n_{boot}}^{-1} \sum_{b=1}^{n_{boot}} \hat{\mu}_b,$$

where n_{boot} is the number of bootstrap samples. The bootstrap estimate of bias is,

$$\widehat{Bias}\ (\hat{\mu}_{boot}) = \hat{\mu}_{boot} - \hat{\mu},$$

where $\hat{\mu}$ is the estimate obtained using the complete sample. The bootstrap estimate of variance is,

$$\widehat{Var}\ (\hat{\mu}_{boot}) = (n_{boot} - 1)^{-1} \sum_{b=1}^{n_{boot}} (\hat{\mu}_{boot} - \hat{\mu}_b)^2.$$

For *k*NN applications the bootstrap variance estimator is preferred to the jackknife estimator, due to computational intensity for large samples (i.e. n_{boot} is generally smaller than *n*) and to a violation of the jackknife estimator’s assumption that the statistic of interest is smoothly varying.

In addition to using the earlier site in Minnesota, the study area was expanded to include Landsat ETM+ scenes and NFI data for North Karelia, Finland and Molise, Italy. The Finnish NFI data were collected using a systematic cluster sample of variable radius plots, while the

Italian NFI data came from a systematic random sample of fixed radius plots. Small areas of interest, approximately 8 km by 8 km in size, were defined and tree volumes per unit area were estimated for each small area using a *k*NN approach similar to that described in the earlier study. Estimates of means and variances for each small area were calculated using the parametric and bootstrap estimators, as well as the jackknife estimator for the Molise and Minnesota study areas. For the bootstrap estimator, 1000 bootstrap resamples were used.

The results suggested close agreement among all three estimators of means and variances, though care must be taken with the bootstrap approach to appropriately mimic the nature of the sample design. The authors concluded that the results provide strong evidence of validity in the assumptions underlying the parametric estimator. They also recommended the use of the bootstrap estimator for *k*NN approaches where $k = 1$, since use of the parametric estimator is infeasible in such cases, as well as for small values of *k* where the parametric estimator would not provide credible confidence intervals.

Explicit superpopulation models

In the case of explicit superpopulation models, inferences about estimated population parameters are based on the underlying model. This includes spatial models that assume correlation among small area effects. Explicit superpopulation models account for the variability in the relationship between auxiliary and response variables among small areas and can be specified as either unit-level or area-level models, depending on whether the auxiliary data are available for the individual population units or only at the aggregate level for each small area. Such models permit the estimation of area-specific MSE values, unlike in the case of purely synthetic estimators where the estimated MSE is averaged over all small areas. However, small areas must be defined a priori.

Frequentist models

General linear mixed models include both fixed effects of auxiliary variables across small areas as well as random effects associated with each small area. Under these models, Empirical Best Linear Unbiased Predictor (EBLUP) estimators can be used to simultaneously estimate the parameters associated with the fixed and random effects (Henderson, 1975). The name EBLUP arises from the fact that these estimators are based on linear functions of the data, have an expected value that is equal to the true value of the parameter being estimated, result in the minimum MSE for the class of all linear unbiased estimators, and are empirical in the sense that they use variances estimated from the data rather than based on theoretical values. Assuming that the small areas can be enumerated a priori, a general linear mixed model, or mixed effects model, can be expressed as,

$$y = X\beta + Zv + \varepsilon,$$

where y is the vector of observations, X is the design matrix for the fixed effects (those associated with auxiliary variables), Z is the design matrix for the random effects (those associated with the individual small areas), β are the model coefficients, and v and ε are vectors of random effects and random model errors respectively that are assumed to be independently and normally distributed with means of zero, each with some unknown covariance. In the case of SAE, the interest is estimation of a population mean μ, which is described by the model,

$$\mu = l'\beta + m',$$

where l and m are vectors of constants derived from the design matrices for the fixed and random effects respectively. The EBLUP estimator of μ is,

$$\hat{\mu}_{eblup} = l'\hat{\beta} + m'\hat{v},$$

where $\hat{\beta}$ is the generalized least squares estimator of β and $\hat{v}$ is the BLUP estimator of the vector of random effects.

Goerndt et al. (2011) compared EBLUP estimators used with general linear mixed models to several other estimators for a study area in northwestern Oregon, USA. Variables of interest from variable-radius forest inventory plots included tree density, quadratic mean diameter, basal area, height, and volume. Auxiliary variables were derived from airborne LiDAR data. An area-level mixed effects model was assumed for the EBLUP estimator. Stand-level estimates of the variables of interest were calculated using direct, synthetic, composite, EBLUP, and nonparametric imputation estimators. Because of the assumption of an inadequate sample for direct estimation under SAE, the study also examined the relative precision and bias of the estimators under a range of simulated sampling intensities. The results indicated that the direct estimator had a smaller relative root MSE (RRMSE) and estimated relative bias (RB) than the others for almost every variable of interest given the greatest sampling intensity. However, for smaller sampling intensities, the EBLUP estimator had the smallest RRMSE and one of smallest RB for all response variables, followed closely by the composite estimator based on a stand-level multiple linear regression model. The synthetic estimator generally had the largest RRMSE and RB when estimating small area population parameters for all response variables. The authors cautioned that the performance of any estimator that relies upon regression will depend largely upon the strength of the relationship between the variable of interest and the auxiliary variables.

Breidenbach and Astrup (2012) conducted a similar study in southeastern Norway using above-ground forest biomass data from NFI plots and a photogrammetric canopy height model developed from digital aerial photography. They compared the performance of direct, synthetic, GREG, and EBLUP estimators for multiple municipalities in Vestfold County, which had from 1 to 35 NFI plots. No estimate of MSE was calculated for the synthetic estimator because the bias could not be adequately evaluated. The authors found that both the EBLUP and GREG estimators resulted in a smaller MSE than the direct estimator for all municipalities. Furthermore, the EBLUP estimator usually had a smaller MSE than the GREG estimator, and was precise even for municipalities having few or even just one sample unit.

Bayesian models

All the estimators considered to this point, even the model-based ones, are derived from a frequentist interpretation of probability, in contrast to the Bayesian interpretation. For a readily accessible introduction to Bayesian statistics, see Bolstad (2007). Under the frequentist interpretation, probability refers to the long-term frequency of an observation (the evidence) given repeated outcomes from an experiment, or repeated samples from a population (the hypothesis). The underlying parameters that describe this repeated process are assumed to be unknown but fixed. Under the Bayesian interpretation, probability refers to the plausibility of a set of underlying parameters describing the random process (the hypothesis) given the fixed set of observations available (the evidence). In this view, the parameters are not fixed, but come from a distribution of possible values. This can be expressed in terms of Bayes theorem,

$$\Pr(H|E) = \Pr(E|H)\Pr(H)/P(E),$$

where Pr(H|E) is the conditional probability of the hypothesis given the evidence (the posterior probability), Pr(E|H) is the conditional probability of the evidence given the hypothesis (the likelihood), Pr(H) is the unconditional probability of the hypothesis (the prior probability), and Pr(E) is the unconditional probability of the evidence (a normalizing constant). When applied to model-based approaches to SAE, a Bayesian interpretation of the superpopulation models typically leads to two general parametric approaches: Empirical Bayes and Hierarchical Bayes (Ghosh & Rao, 1994).

Empirical Bayes (EB)

In the case of linear superpopulation models, EB is equivalent to the EBLUP approach via general linear mixed models. This is not a fully Bayesian approach because it does not incorporate a prior probability distribution of model parameters, but instead estimates values for model parameters based on the sample units using maximum likelihood, which is fundamentally frequentist. EB approaches can be considered an approximation to a fully Bayesian approach to SAE (Datta et al., 1999).

Hierarchical Bayes (HB)

In the HB approach, a prior probability distribution of model parameters is specified. This, in conjunction with the values of the sample units, induces a posterior probability distribution of the small area parameter of interest via Bayes theorem. In practice, closed-form expressions of the posterior probability distribution are typically not available. In much the same way that bootstrapping has enabled the development of estimators for complex frequentist model-based approaches, Markov Chain Monte Carlo (MCMC) methods have paved the way toward development of fully Bayesian models. With recent advances in computer hardware and parallel-processing, MCMC methods, such as Gibbs sampling and the Metropolis-Hastings algorithm, have ushered in a range of Bayesian approaches to prediction and SAE. For a thorough introduction to HB and advanced computational methods, see Gelman et al. (2013).

Banerjee and Finley (2007) developed an HB approach to unit-level predictions that includes a spatial component that accounts for spatial autocorrelation among observations, both at the plot and subplot scale. The study used gross live tree biomass from NFI plots and subplots as the variable of interest and auxiliary variables calculated from multi-date Landsat ETM+ imagery for a heavily forested region in north-central Minnesota, USA. The authors compared one aspatial model to five spatial models, each of which assumed slightly different spatial processes. The results, based on multiple MCMC chains and noninformative priors (i.e. no prior information about the distribution of population parameters was assumed), suggested a clear advantage to the spatial models, particularly the multi-resolution models with nested spatial processes.

Finley et al. (2009) built upon the earlier study by developing a spatially-varying multinomial logistic regression model, again under an HB approach, for unit-level predictions of forest type group for a study area encompassing all forested land across the entire state of Michigan, USA. Again, the variable of interest was observed on NFI plots and the auxiliary variables were calculated from climatic and topographic raster data. A spatially-varying predictive process was modeled on a regular mesh of roughly 200 points across the study area. A comparison was made among four different predictors: a nonparametric approach based on geographic proximity, a *k*NN approach, a hierarchical model with spatially-varying intercepts, and a hierarchical model with all spatially-varying coefficients. The results, also based on multiple MCMC chains and noninformative priors for the HB models, suggested the *k*NN approach and HB model with spatially-varying intercepts performed comparably well, and that the HB model with all spatially-varying coefficients performed best.

Finley et al. (2011) demonstrated similar results for Michigan, USA, for predicting unit-level forest biomass from NFI plots using auxiliary variables derived from Landsat ETM+ imagery, using an HB approach and spatially-varying coefficients model. Finley et al. (2013) conducted a comparable analysis for the Pensobscot Experimental Forest in Maine, USA, for multiple variables of interest extracted from forest inventory plots and auxiliary variables calculated from LVIS airborne LiDAR data and AVIRIS spaceborne hyperspectral imagery, again using a spatially-varying coefficients model. While all these studies feature an HB approach to unit-level prediction and mapping, none specifically addresses SAE. However, it should be apparent that the approach could be readily adapted to SAE by specifying small areas a priori.

Nonparametric Bayes

Parametric Bayesian methods have limited value for large-scale, multipurpose surveys because of the difficulties in validating the parametric assumptions (Rao, 2011). A nonparametric Bayesian approach provides an alternative, but it requires the specification of a likelihood function based on the values of the variables observed for the sample units and a prior distribution of population values. Lazar et al. (2008) provide a nonparametric Bayesian framework for SAE with auxiliary variables based on the Polya posterior, explained in the next paragraph.

This posterior is derived from Polya sampling, and can be described using a scenario with two urns. Urn #1 contains the sample units drawn from the population. Urn #2 contains the non-sampled population units. In Polya sampling, one unit is drawn at random from each urn. The unit drawn from urn #2 is assigned the label of the unit drawn from urn #1, and then both units are returned to urn #1. This process is repeated until urn #2 is empty and urn #1 contains all population units. The end result is that urn #1 contains a simulated copy of the population, conditional upon the sample units drawn from the population, from which population parameters of interest can be calculated. If the procedure is repeated many times, a posterior distribution of population parameters is generated and can be used for making statistical inferences.

Lazar et al. (2008) suggest that the Polya posterior is appropriate when one might use SRS under a design-based approach to inference and yields results with good frequentist properties. The Polya posterior makes use of a likelihood function based on the sample units that is noninformative because all unobserved values of the nonsampled population units have the same likelihood function. Estimation using the Polya posterior is done using resampling methods similar to bootstrapping, with the distinction that bootstrapping replicates samples, whereas Polya sampling replicates populations.

Discussion

This literature review has provided an overview of a variety of estimators available when using NFI data with auxiliary variables for SAE. All of the more familiar direct estimators covered, such as the SRS, stratified, and GREG estimators, are unbiased, or nearly so, and rely upon design-based inference. However, while many have been shown to be more precise than the SRS estimator, all direct estimators use only the NFI sample plots from within a small area of interest, thereby limiting precision gains.

Indirect estimators, whether they rely upon design-based or model-based inference, achieve better precision by using the auxiliary information to effectively increase the sample size within a small area of interest. Unlike direct estimators, none of the indirect estimators reviewed can be assumed to be unbiased, though some have been shown to be nearly so under some conditions.

Nonparametric model-based approaches are particularly attractive because they do not assume a particular functional form for the relationship between the response and auxiliary variables, nor do they require probability samples or assumed normality in the distribution of residual errors. Furthermore, the nonparametric *k*NN estimator accommodates a multivariate response, simultaneously producing estimates for all NFI variables of interest.

There are two fundamental interpretations of probability that can be used for making statistical inferences via superpopulation models: frequentist and Bayesian. Frequentist formulations of the nonparametric *k*NN estimator have been developed for SAE, though there could be considerable computational cost associated with calculating covariance among unit-level predictions. Recent advances in computer processing and algorithm development have paved the way for more widespread application of numerical methods for estimation and inference, such as bootstrapping for the frequentist and Gibbs sampling for the Bayesian, when a closed-form exact solution of the posterior probability cannot be calculated .

Parametric Bayesian approaches to SAE have been developed, using HB with Gibbs sampling. However, there may be difficulties in validating their parametric assumptions, particularly in the case of multi-resource NFI data. Nonparametric Bayesian methods manage to circumvent this issue. While such approaches to SAE have been developed, such as through Polya sampling, they have not yet been applied to NFI data.

Conclusions

There are a variety of approaches to SAE that can be used by integrating NFI data with auxiliary variables derived from remotely sensed imagery. The choice ultimately comes down to the preferred mode of statistical inference and assumptions about the nature of random variation observed in the population. Under the design-based mode of inference, this random variation arises from the distribution of indicator values that determine whether or not a population unit is part of the sample. Under the model-based mode of inference, the observed response values associated with population units are assumed to have been generated by some underlying stochastic superpopulation model, i.e. there is random deviation about the model mean.

The primary advantage of the design-based approach is that the estimators are unbiased, or nearly so, with the principal disadvantage being that estimates are based only upon the sample units within the domain, limiting precision for small areas. Estimators used under the model-based approach borrow strength from sample units outside of the domain, resulting in a tradeoff between the benefit of increased precision and the cost of potential bias. Bayesian models also generally have the further advantage of better performance for small samples.

The nonparametric *k*NN estimator has been used extensively in studies with multi-resource NFI data, likely because of its simplicity and multivariate nature (i.e. simultaneously providing estimates for all response variables). The literature reviewed here suggests that there is utility to a model-based formulation of the *k*NN estimator for SAE. Such an estimator makes minimal assumptions about the relationship between the auxiliary variables and the response variables, or about the distribution of the random variation around the modeled expected value. However, as with all model-based approaches, prediction bias must be properly accounted for to make valid inferences.

Chapter 2: Harmonic regression of Landsat time series for modeling attributes from national forest inventory data

Summary

Imagery from the Landsat Program has been used frequently as a source of auxiliary data for modeling land cover, as well as a variety of attributes associated with tree cover. With ready access to all scenes in the archive since 2008 due to the USGS Landsat Data Policy, new approaches to deriving such auxiliary data from dense Landsat time series are required. Several methods have previously been developed for use with fine temporal resolution imagery (e.g. AVHRR and MODIS), including image compositing and harmonic regression using Fourier series. The article presents a study, using Minnesota, USA during the years 2009-2013 as the study area and timeframe. The study examined the relative predictive power of land cover models, in particular those related to tree cover, using predictor variables based solely on composite imagery versus those using estimated harmonic regression coefficients. The study used two common non-parametric modeling approaches (i.e. *k*-nearest neighbors and Random Forests) for fitting classification and regression models of multiple attributes measured on United States Forest Service Forest Inventory and Analysis plots using all available Landsat imagery for the study area and timeframe. The estimated Fourier coefficients developed by harmonic regression of 'Tasseled Cap Transformation' time series data were shown to be correlated with land cover, including tree cover. Regression models using estimated Fourier coefficients as predictor variables showed a two- to three-fold increase in explained variance for a small set of continuous response variables, relative to comparable models using monthly image composites. Similarly, the overall accuracies of classification models using the estimated Fourier coefficients were

approximately 10 to 20 percentage points greater than the models using the image composites, with corresponding individual class accuracies between six and 45 percentage points higher.

Introduction

The use of remotely sensed data for inventory and mapping of agricultural and natural resources has a long history. Such data have typically been collected on airborne or spaceborne platforms by either passive sensors that use solar radiation as a source of electromagnetic energy, or active sensors that provide their own radiation source. Both passive and active sensors have demonstrated utility as sources of predictor variables that can be used to model a variety of natural resource phenomena.

In the context of using these data in combination with national forest inventory (NFI) data for the purposes of monitoring forest resources over long periods of time and large geographic areas, the Landsat platform has garnered particular interest, as noted by the review of McRoberts, Cohen, Næsset, Stehman, and Tomppo, (2010). This is due to several factors, which will be discussed in greater detail below. First, the Landsat Program provides the longest-running continuous collection of Earth imagery of any satellite program, with sensors designed to maintain consistency in resolution (i.e. spatial, spectral, radiometric, and temporal) across missions. Second, the multispectral characteristics of the Landsat sensors enable the derivation of metrics with biophysical meaning. Third, the spatial resolution of the latter generation of sensors provides a better match to the size of the typical NFI plot than that of satellite sensors with coarser spatial resolution. Finally, with the adoption of the open access data policy for the Landsat Program in 2008, the entire archive of Landsat imagery is freely available for use.

Landsat as a source of auxiliary data

Length of the Landsat record

From the Multispectral Scanner (MSS) sensor used onboard the first five Landsat satellites to the Thematic Mapper (TM) sensor of Landsat-4 and Landsat-5, the Landsat program has striven for consistency in the data record while simultaneously incorporating improved technological capabilities in terms of spatial and spectral resolution. This consistency extends to the sensors used onboard the Landsat satellites in orbit today, as well as those planned for future missions.

One of the two Landsat sensors currently in operation is the Landsat-7 Enhanced Thematic Mapper Plus (ETM+). In May of 2003, the Landsat-7 scan line corrector (SLC), a small rotating mirror that compensates for the forward motion of the spacecraft, failed. This resulted in the loss of data in wedge-shaped areas on either side of the image, with more missing data further away from nadir. These SLC-off gaps amount to a loss of approximately 22% of the data for any given scene (Ju & Roy, 2008). These gaps have limited the utility of ETM+ imagery, although there have been several approaches proposed to dealing with them, including compositing of several images (Roy et al., 2010), interpolation using SLC-on imagery (Maxwell et al., 2007; Chen et al., 2011), data fusion with MODIS imagery (Roy et al., 2008), and geostatistical methods (Zhang et al., 2007; Pringle et al., 2009).

Development of Landsat spectral indices

Multiple spectral indices have been developed to establish the relationship between the spectral and radiometric response measured by remote sensors and the presence of various land covers, especially vegetation. Bannari et al. (1995) reviewed more than forty vegetation indices that have been developed for sensors ranging from ground-based to spaceborne systems. Most such indices are based on the fact that vegetation has wavelength-dependent absorption, transmission, and reflection properties, in particular the differential response in the red and near-infrared (NIR) wavelengths, and have been shown to provide a better indication of the amount of vegetative land cover than any single band alone (Curran, 1980).

Kauth and Thomas (1976) developed a linear transformation of all four of the original MSS bands, named the tasseled cap transformation (TCT), to produce indices related to not only growing vegetation ("green stuff"), but also soils ("brightness"), senescent vegetation ("yellow stuff"), and shadows ("non-such") to better differentiate various crops, as well as phases of crop development over time. Crist and Cicone (1984a, 1984b) extended TCT for use with the six reflective TM bands, simultaneously dropping some features ("yellow stuff" and "non-such") while defining new ones (e.g. "wetness"). Huang et al. (2002) and Baig et al. (2014) developed comparable transformations for ETM+ and the Operational Land Imager (OLI) respectively.

Numerous studies have shown TCT components derived from Landsat imagery to be useful for mapping forest characteristics. These studies, using the components of "brightness", "greenness", and/or "wetness", demonstrate their utility across a range of forest mapping applications, including land cover (Byrne et al., 1980; Yuan et al., 2005), forest types (Dymond et al., 2002), succession (Helmer et al., 2000), stand-replacing disturbance (Cohen et al., 1998; Jin & Sader, 2005; Healey et al., 2005), pest damage (Skakun et al., 2003), growing stock volume (Zheng et al., 2014), canopy cover and biomass (Karlson et al., 2015), and recovery from disturbance (Pickell at el., 2016).

Size and configuration of the NFI plot footprint

The current annual NFI conducted by United States Forest Service (USFS) Forest Inventory and Analysis (FIA) program exhibits many of the characteristics observed for NFI globally that are pertinent to its use with Landsat imagery. A comprehensive description of the FIA program is provided in Bechtold and Patterson (2005). FIA inventory plots are established according to a well-defined sample design, with a sampling frame that is used to generate a spatially-balanced sample of plots. Each plot is a cluster of sub-plots. Cluster plots are often used for NFI because the determination of the relative locations of the sub-plots is typically more accurate, their layout is generally faster due to the smaller distances measured and traveled (possibly over rugged terrain), and they capture more variability in the population than one larger plot (Kangas & Maltamo, 2006).

Since the nationwide start of the annual NFI program in the US in 1998, FIA plots comprise four circular sub-plots, each with a radius of 7.3152 meters. Circular plots are often used in forest inventory because they are easy to establish for small radii and are prone to less error in plot area, for the same reasons given previously for using cluster plots (Kangas & Maltamo, 2006). Circular plots also minimize edge effects, since they have the smallest possible perimeter for a given area by the isoperimetric inequality. The sub-plots are arranged with the center of one sub-plot defining the center of the cluster plot. The centers of each of the other three sub-plots are equally spaced about plot center, oriented so one sub-plot is due north of plot center, and each is 36.576 meters distant from plot center.

Although other NFI programs have different sample designs, sampling frames, and plot configurations, the information presented for the FIA program remains instructive for comparisons to Landsat and other satellite imagery. Each FIA sub-plot constitutes an area of about 168 m^2, approximately 19% of the area of a TM, ETM+, or OLI pixel for the reflective bands, but only about 0.27% the area of a 250-meter MODIS pixel. The smallest circle that circumscribes all four sub-plots has an area of about 6,052 m^2, or just less than seven 30-meter pixels. The four sub-plots cover approximately 11% of this area, and about 1% of the area of a MODIS pixel.

Open data access policy for the Landsat archive

Ease of access to data from the Landsat Program has been variable over time. During the commercial operations period, there were high financial costs and restrictive copyright rules in place that limited sharing of access to imagery. With the assumption of mission operations by the United States Geological Survey (USGS), purchased imagery could be shared more freely. Wulder et al. (2012) suggest that the USGS Landsat Data Policy that took effect in 2008 has allowed the scientific community to finally realize the full value of the Landsat Program, as indicated by the dramatic rise in the use of its data globally. This data policy provides unrestricted access to the entire USGS National Satellite Land Remote Sensing Data Archive (NSLRSDA), with selected products made available for retrieval over the Internet at no financial cost to users (Woodcock et al, 2008).

Approaches to analyzing dense time series of satellite imagery

Free and open access to the Landsat Archive (i.e. NSLRSDA) permits new approaches to image analysis that were not previously available to users of Landsat imagery. With the high costs of Landsat scenes under earlier data policies, users typically selected only those scenes of interest that were predominantly free of clouds, exhibiting a “scene-centric” focus. Under the new data policy, as well as due to continuing advancements in data storage and computing power, users have begun developing methods that utilize all scenes over a period of interest to improve the information collected for pixels of interest, shifting to a “pixel-centric” focus.

There are several approaches to processing and analyzing dense time series of satellite imagery. Most were developed for use with other satellite platforms, particularly those having finer temporal resolution than Landsat. One of the first was published by Goward et al. (1985) for analyzing daily observations of NDVI across North America using the Advanced Very High Resolution Radiometer (AVHRR) onboard the NOAA-7 satellite. The daily data were grouped into three-week bins, and the maximum value for each pixel from each bin was used as the bin's composite value, thereby reducing the effects of cloud contamination. These maximum value composites (MVC) were computed for a 30-week growing season and used to estimate the variability and area under the seasonal Normalized Difference Vegetation Index (NDVI) profile for each pixel. Both metrics were shown to be correlated with seasonal patterns in natural and cultivated vegetation, as well as net primary productivity.

Reed et al. (1994) took a similar approach for a study using AVHRR data to map land cover types for the conterminous US. Four growing seasons of biweekly MVC NDVI were used to develop a set of 12 seasonal NDVI metrics. These included not only NDVI range and area under the curve, as used in the study by Goward et al. (1985), but also time and value at onset and end of greenness and derived rates of green-up and senescence along with a few other metrics. The authors noted some limitations of using biweekly MVC images, such as the compositing period being too long to determine some phenological events and residual cloud contamination. Nevertheless, the results indicated strong agreement with expected characteristics for various land cover types, including assorted agricultural crops, grasslands, shrublands, and forests. DeFries et al. (1995) reported similar results for a study using AVHRR MVC data for global land cover mapping.

Sellers et al. (1994, 1996) were among the first to use a mathematical model to describe time series of AVHRR NDVI data in an effort to correct for residual cloud contamination in MVC images. The complete methodology they proposed included a series of steps, with the first being adjustment of the MVC values by means of a Fourier series. A Fourier series can be used to approximate a periodic function, such as the seasonality of NDVI, with a closer approximation given by using more harmonics in the series. This application of Fourier series is also known as harmonic analysis or harmonic regression. Other harmonic regression methods have also been developed to produce cloud-free AVHRR images (Roerink & Menenti, 2000).

Moody and Johnson (2001) conducted a study to map land cover for a small area in southern California, USA, using AVHRR NDVI monthly MVC images. The coefficients for a Fourier series with two harmonics were estimated for each pixel in the study area. These estimated coefficients were used as feature variables with an unsupervised classification scheme to produce a map of six basic vegetation formations. Comparison with field-based validation data yielded an overall accuracy of 68%. The authors found that the estimated mean NDVI provided discrimination along a continuum from grassland to closed canopy forest, amplitude separated evergreen from deciduous vegetation, and phase distinguished grasslands from shrublands/woodlands/forests and irrigated croplands.

Alternative methods for describing AVHRR time series data have been proposed, using models such as the asymmetric Gaussian (Jönsson & Eklundh, 2002) and the Savitzky-Golay filter (Chen et al, 2004). Hermance (2007) and Bradley et al. (2007) suggested enhancements to the original harmonic regression methods intended to address the issue of higher-order harmonics generating spurious oscillations identified in the two aforementioned studies. Others studies have employed similar approaches with MODIS time series data, using harmonic regression (Potgeiter et al., 2007; Geerken, 2009; Wilson et al., 2012), piecewise logistic models (Zhang et al., 2003), the wavelet transform (Sakamoto et al. 2005), cubic splines (Scharlemann et al., 2008), and the autoregressive integrated moving average (Bayr et al., 2016). Hird and McDermid (2009) conducted a thorough comparison of several of these techniques for reducing noise in NVDI time series of MODIS imagery and concluded that the double logistic and asymmetric Gaussian approaches were generally superior to the others in the study. However, they also cautioned that their conclusions were conditional on the nature of the noise in the imagery.

Using Landsat time series data for vegetation modeling and mapping

Many of the studies reviewed in the previous section primarily focused on methods to remove noise from satellite image time series in coarse spatial resolution satellite imagery. There are comparable studies in the literature that use Landsat imagery. For example, Brooks et al. (2012) demonstrated the effectiveness of using harmonic regression to model NDVI time series derived from TM and ETM+ imagery. Other studies previously reviewed focused instead on the use of these seasonal characteristics to produce models and maps of vegetation. A few such applications using AVHRR and MODIS imagery include mapping net primary productivity (Goward et al., 1985), land cover (Reed et al., 1994; DeFries at al., 1995; Moody & Johnson, 2001), and tree species relative abundance (Wilson et al., 2012).

In terms of mapping vegetation with finer spatial resolution imagery, Badhwar et al. (1982) developed one of the earliest such applications, using temporal profiles of TCT greenness derived from MSS imagery. A nonlinear model was used to fit TCT greenness time series via an iterative Marquardt technique. The two estimated model parameters for each pixel were then used to classify 40 small cropland test sites into “corn”, “soybean”, and “other” crops. The authors found that the estimated model parameters were closely related to the target of interest, and suggested that the inclusion of other TCT metrics could improve the results.

After the implementation of the 2008 open data policy for the NSLRSDA, Zhu et al. (2012) demonstrated that harmonic regression of dense TM and ETM+ time series could be used to map forest disturbance. After masking out clouds and cloud shadows, a Fourier series was fit on a per-pixel basis to the time series of surface reflectance for each reflective band in a 3,600 km^2 study area on the border between Georgia and South Carolina, USA. A measure of forest disturbance was predicted by comparing predicted image values, based on the harmonic regression model parameters estimated for each pixel, with the observed image values. The accuracy assessment of the resultant maps of forest disturbance, based on manual interpretation of the Landsat imagery combined with finer resolution imagery, demonstrated user's and producer's accuracies of greater than 95% for classification into "forest disturbance" and "stable forest" classes.

Hansen et al. (2013) developed a method for mapping forest change at a global scale that leveraged not only the opening of the Landsat Archive and some of the analysis techniques reviewed here, but also the advent of cloud-based computing platforms. This massive study, spanning all global land except for the Arctic and Antarctic regions, used more than 650,000 growing season ETM+ scenes collected between 2000 and 2012. The Landsat data were processed using Google Earth Engine (GEE), a platform for global scientific analysis and visualization of geospatial datasets (Gorelick et al., 2017). The raw ETM+ data were pre-processed by conversion to top-of-atmosphere (TOA) reflectance, screening for clouds, shadows, and water, and normalization using MODIS imagery. Three groups of seasonal metrics were derived from each band of the pre-processed data. These derived metrics were used as predictor variables in a model to predict forest cover, loss, and gain over the time period of the study.

Summary of the literature review

The preceding literature review covered multiple justifications for using Landsat time series imagery with NFI, and in particular FIA data for modeling and mapping land cover or use and especially characteristics of forest resources. These include the long history of the Landsat program and its temporal and geospatial overlap with the annual FIA program, the configuration of the FIA plot and sub-plots and their relative size compared to a Landsat pixel, and the institution of the 2008 open data policy for the Landsat Archive and concomitant development of cloud-based computing platforms like GEE that can readily access and process large volumes of data. It also touched upon the development of spectral indices for monitoring vegetation using satellite imagery as well as a variety of methods for analyzing time series of satellite imagery, from image compositing to the use of mathematical models for predicting seasonality in spectral reflectance associated with vegetative land cover. Finally, it provided examples of how these techniques were used to model vegetation, or changes in vegetation over time.

The literature review provided some examples of studies that have thoroughly examined individual elements of the problem at hand, such as the Hird and McDermid (2009) comparison of noise-reduction techniques in NDVI time series. However, there has not been one that explicitly examines the utility of using harmonic regression of dense time series of TCT metrics derived from TM and ETM+ data to extract auxiliary data for the modeling of forest attributes from NFI data. Therefore, a study is proposed here that will test the primary hypothesis that models using estimated harmonic regression coefficients will produce more accurate predictions than those based on composite images of dense Landsat time series alone. Furthermore, there is a secondary hypothesis associated with the proposed approach for modeling and mapping forest attributes using time series of Landsat imagery, namely that harmonic regression provides a means for overcoming missing data due to either weather-related phenomena like clouds and snow cover, or to instrument related gaps from the SLC failure.

Materials and methods

Study area and data

The study area and timeframe were selected to coincide with an area and timeframe covered by annual forest inventory data collection. The NFI conducted by FIA is based on the annual collection of data from a permanent plot sample of the population. The period over which all plots are measured ranges from either five or seven years in the eastern US to 10 years in the western US, meaning that 10-20% of the plots are measured each year. The sampling intensity of the FIA plot network is approximately one plot per 2,400 hectares. Some state partners of the FIA program contribute additional funds to increase the sampling intensity. The state of Minnesota is one such partner, and has augmented FIA funding to support an increase in sampling intensity to one plot per 1,200 hectares and to retain a measurement cycle of five years.

Because of the combination of a relatively large number of FIA plots due to its land area and increased sampling intensity, as well as the authors' familiarity with the state's forested ecosystems based on previous studies, Minnesota was chosen as the study area. In order to match the time period needed to collect a complete cycle of data for the FIA survey, all TM and ETM+ scenes for the timeframe of 2009-2013 were used. Furthermore, this five-year period provides a larger sample of sensor observations (across years) for each pixel from which to develop a harmonic regression model and increases the likelihood of filling any voids in the seasonal record due to clouds, snow, or SLC-off artifacts.

In the public FIA database (FIADB), an EVALID uniquely identifies a collection of plots used for producing estimates for a specific group of population attributes. All plots included in the FIA sample for the state of Minnesota during the study timeframe 2009-2013 for estimating change in condition area and tree volume were used in the analysis, corresponding to 'EVALID=271303'. Out of a total of 17,500 plots in this collection, a subset of 17,343 was located inside the study area defined using the census boundary for the state of Minnesota stored in the Google cloud computing platform. For each of these plots, a set of 10 forest inventory response variables were extracted from FIADB for the central sub-plot in the cluster, in order to more closely match the spatial extent sampled by a single Landsat TM or ETM+ pixel. Categorical (2) and continuous (8) variables were chosen to examine the utility of the predictor variables under both classification and regression models.

The categorical variables used were the class assignment of the condition, corresponding to an area with similar land use and cover characteristics, at the center of the central sub-plot under two classification schemes. The first scheme assigned the condition to either the 'non-forest' or 'forest' class. The second scheme assigned the condition to one of five non-forest and three forest classes: 'water', 'cropland', 'grassland', 'settlement', 'wetland', 'coniferous forest', 'broadleaf forest', or 'non-stocked forest'. These schemes were chosen to be representative of a range of land cover and use classification levels.

The continuous variables used were the per hectare values of the number, basal area, foliage biomass, and total aboveground biomass of all live trees at least 2.54 cm diameter at breast height on the central sub-plot. The same set of attributes was also calculated for only live broadleaf trees. These variables were selected to be representative of a range of forest inventory attributes, based on counts (number), areas (basal area), and volumes (biomass), and by their nature would be expected to have varying correlations with the multispectral signal detected by the Landsat sensors. Furthermore, the variables associated with live broadleaf trees were chosen to evaluate the utility of dense time series of Landsat imagery for differentiation between deciduous and evergreen vegetation.

Image processing

Given the large volume of imagery required for the study, the GEE platform was used for processing the TM and ETM+ data. GEE provides ready access to registered users to data in the Landsat Archive, as well as a browser-based programming interface to process the imagery and store the results using Google's cloud computing infrastructure.

The number of remote sensing observations collected for a pixel depends on its location within the Worldwide Reference System (WRS) used to catalog Landsat data. There are four distinct zones of overlap in WRS-2, which is used to catalog both TM and ETM+ data, determined by the amount of overlap among adjacent scenes. There are overlap zones 1) with both sidelap and endlap, 2) with sidelap, 3) with endlap, and 4) with neither sidelap nor endlap. Endlap is the area of overlap between adjacent scenes along the path of the satellite. Sidelap is the area of overlap between adjacent scenes on neighboring paths. It should be noted that areas of endlap contain duplicate observations, since a single observation for a given pixel appears in adjacent scenes along the path of the satellite and were collected on the same day. Areas of sidelap do not contain duplicate observations, since the observations for a given pixel are collected on different days.

Landsat Level-1T (L1T) TOA reflectance data for the six reflective bands of TM and ETM+, along with the collection timestamp, were extracted from the Landsat Archive for the study area and timeframe. A total of 4,425 images covering 25 WRS-2 Path/Row scene centers were used in the study, with 1,575 from Landsat-5 and 2,850 from Landsat-7. Although the temporal resolution for approximately the first three years of the study was eight days, it dropped to 16 days for the remainder due to the end of Landsat-5 operational image collection in November 2011. Therefore, a compositing period of about 52 days (i.e. 1/7 of a year) was chosen to ensure that there were at least three unique observations of each pixel for each compositing period in the study timeframe. This resulted in 35 composite images for the study timeframe.

The 52-day composite images were created by first masking out the observations most likely contaminated by clouds or snow, as well as any data flagged as missing during L1T processing, such as data located in SLC-off gaps. A built-in GEE cloud-scoring algorithm, based on the tendency of clouds to have larger reflectance values in blue, visible, and NIR bands but smaller values in thermal infrared bands, was used to compute cloud metrics. The Normalized Difference Snow Index (NDSI) was used to compute snow metrics, and is based on the normalized difference of the green and first shortwave infrared (SWIR) bands (Hall et al. 1995). Only pixels with both a cloud score less than 75 and NDSI less than 0.5 were used to create the composite images.

Maximum value composites have often been used with NDVI data, because NDVI values have been shown to be smaller in the presence of clouds and snow than they would be otherwise. Median value composites were used in this study, since the effect associated with the presence of clouds and snow could be either positive or negative depending on the spectral band. Furthermore, the median is a more accurate measure of central tendency than the mean when dealing with data that are skewed, as is the case for TOA values from any pixels having residual cloud or snow contamination not filtered out by the masking procedure. Each 52-day TOA composite image was generated by computing every pixel’s median value for each band, including the image timestamp extracted from the scene metadata, from the set of observations passing through the cloud and snow filters.

The TCT coefficients developed by Huang et al. (2002) for ETM+ data were used to compute brightness, greenness, and wetness metrics for each composite image. Although composite images might include data from either TM or ETM+, the TCT coefficients developed by Crist and Cicone (1984a, 1984b) for TM data were not used. This choice was made for two reasons. First, it simplified the compositing procedure. Second, it was not readily apparent that the additional set of TCT coefficients would accentuate or mitigate any discrepancies between the data collected by TM and ETM+, since the two sensors have the same spatial and spectral characteristics in the reflective bands and the appropriate sensor metadata had already been used to compute TOA reflectance values.

Harmonic regression models

Ordinary least squares (OLS) regression was used to fit separate Fourier series to each composite time series of the three TCT metrics for each pixel in the study area. A form of the Fourier series based on the one presented in Sellers et al. (1996) was used in the analysis. Each time series of data was approximated as a trigonometric polynomial,

$$\hat{Y}_t = a_0 + \sum_{j=1}^{m} a_j \cos(j2\pi t/n) + b_j \sin(j2\pi t/n)\,,$$

where *t* is the composite timestamp value, *n* is the length of the cycle, and *m* is the order of the polynomial and equal to the number of harmonics in the approximation. Landsat image timestamps are stored as milliseconds since the start of the epoch (January 1, 1970) and were converted to fractional ephemeris days. A value of 365.2421891 ephemeris days per tropical year was used as the length of the annual cycle.

Regression models with one to four harmonics (i.e. 1st order to 4th order Fourier series) were tested to determine the model form that provided the best fit to the data without introducing the spurious oscillations noted by Hermance (2007) and Bradley et al. (2007). Geerken (2009) likewise suggested using between three and five harmonics for time series imagery, noting that between 81 and 99% of variance in a MODIS reference dataset was explained by a 3rd order Fourier series. These harmonics correspond to cycles of approximately 12, six, four, and three months respectively. The estimated coefficients from these Fourier series were stored as three-, five-, seven-, and nine-band images respectively, depending on the number of harmonics used in the regression model, resulting in nine, 15, 21, or 27 coefficients in total. The root mean squares of the residual errors (RMSE) for each order of Fourier series were also stored as three-band images. These RMSE images were used only for evaluation of the harmonic regression models, and not as auxiliary data for the modeling of forest attributes from NFI data.

Models of forest attributes

Because both classification and regression were required due to the nature of the response variables selected for the study, the non-parametric methods of Random Forests (RF) and *k*-nearest neighbors (*k*NN) were used to construct individual models relating the dense time series of Landsat imagery to the set of 10 forest inventory variables. Both methods have been used extensively in modeling and mapping applications that integrate satellite imagery and NFI data. McRoberts, Tomppo, and Næsset (2010) provide an excellent review of parametric and non-parametric modeling approaches that have been used for combining these data sources, including examples of both *k*NN and RF applications.

A brief description of the *k*NN (Fix & Hodges, 1951) and RF (Breiman, 2001) methods will be presented here, using similar terminology to that presented in McRoberts, Tomppo, and Næsset (2010). The set of population units for which both the predictor and response variables have been observed is designated the reference set. The set of population units for which predictions of response variables are desired is designated the target set. The space defined by the predictor variables is designated the feature space. A *k*NN prediction for a unit in the target set is made by calculating the mean (for continuous variables) or mode (for categorical variables) of the observed response variable for the *k* units in the reference set that are nearest to the unit in the target set in the feature space with respect to a distance metric. The *k*NN approach is usually optimized by the modeler, via the choice of *k*, the distance metric, the set of weights used to calculate the mean or mode, and the set of predictor variables and the corresponding weights used to construct the feature space (McRoberts, 2009a).

The RF method is an ensemble approach that requires the construction of a random set of decision trees (i.e. a forest of trees), each of which recursively partitions the reference set using threshold values along individual dimensions of the feature space onto leaves containing population units with similar observed response variables. The stochastic processes by which the trees are constructed include bootstrapping of reference units and random selection of predictor variables in the feature space. An RF prediction for a unit in the target set is made by first assigning the unit to a leaf in each tree in the forest according to the decision rules of the given tree. For each tree in the forest, the target unit is then assigned the mean or mode of the observed response variable for all reference units on the leaf to which it was assigned. Finally, the target unit is assigned the mean or mode of these predictions across all trees in the forest. Unlike *k*NN, RF typically does not require optimization by the modeler.

Two distinct sets of predictor variables derived from the dense time series of Landsat imagery were used to construct the feature space for the *k*NN and RF models. The first feature space was constructed from the seven mean monthly composites, approximately corresponding to the growing season in Minnesota of April to October, of the 3 TCT metrics of brightness, greenness, and wetness. These mean monthly composites were derived by first creating the 60 monthly median value composites of TCT metrics covering the study timeframe, extracting the subset of 35 monthly growing season composites, then calculating the average TCT metric values by month across the five-year timeframe of the study. The second feature space was constructed from the estimated coefficients for each time series of 52-day composites of the 3 TCT metrics using 3rd order Fourier series, resulting in seven coefficients per series and matching the 21 dimensions of the composite feature space.

The *k*NN models were fitted using the ‘rflann’ package in R, which provides an interface to the Fast Library for Approximate Nearest Neighbors (Muja and Lowe, 2009). For each distinct feature space, the ‘Neighbour’ function was called with the default parameters of a kd-tree search with minimal checks on the precision of the result. This function was used to construct a list of the 200 nearest neighbors of each of the 17,343 plots in the study, treating these population units as both reference and target set, conditional on the given feature space. Because a reference unit will always be among its own nearest neighbors, a leave-one-out cross-validation approach was taken to optimize the choice of the value of *k*. Leave-one-out cross validation is the particular instance of the more familiar *K*-fold cross validation when *K* is equal to the number of observations. In the case of a *k*NN model, each observation in turn is withheld and used as the target unit, while the remaining observations are used as the reference set. The same effect can be achieved by including all observations in both the reference and target sets in a single *k*NN model, then excluding the first nearest neighbor when finding the set of *k*-nearest neighbors for each target unit.

One of the most appealing aspects of the *k*NN approach is that the set of *k*-nearest neighbors identified in the given feature space can be used to make multivariate predictions, under the assumption that predictor variables that are correlated with one response variable will also be correlated with the others. The current study also assumes that relationship and uses the same feature space for all *k*NN models. However, to allow for fairer comparison with RF model results, the optimal value of k was allowed to vary among kNN models to maximize the correlation between the predictor variables and each individual response variable. Excluding the reference unit itself, for regression models the coefficient of determination (R^2) was calculated for each model with k ranging from one to 199, with the predicted value for each unit being the unweighted mean of the observed continuous response variable for the k nearest neighbors. Similarly for classification models, excluding the reference unit itself, the overall accuracy (i.e. the percentage of the sample units that were correctly classified according to the plot data) was calculated for each model with k ranging from one to 199, with the predicted value for each unit being the mode of the observed categorical response variable for the k nearest neighbors. The value of k that maximized R^2 or the overall classification accuracy determined the optimal model given the feature space. No other optimization of the kNN models was attempted, because the purpose of the study was a comparison of the two feature spaces used for modeling, not a comparison of the two modeling approaches used. Individual class accuracies were also assessed using producer's accuracy (i.e. the percentage of sample units that were correctly classified for a given class in the plot data) and user's accuracy (i.e. the percentage of sample units that were correctly classified for a given class in the pixel data).

The RF models were fitted using the 'randomForest' package in R (Liaw and Wiener, 2002). For each combination of feature space and response variable, the 'randomForest' function was called with the default parameters of 500 trees in the forest, at least five units assigned per leaf in the tree, and the number of predictor variables p tested at each split in the tree as p/3 for regression and $\sqrt{p}$ for classification models. For convenience, the RF models used the 'out-of-bag' sample units for model validation rather than the leave-one-out cross validation employed for the kNN models. The R^2 value was calculated using the ensemble predictions for each regression model, as was the overall accuracy for each classification model.

Results

Image composites and harmonic regression

For the timeframe used in the study area, pixels in overlap zone 1 (both sidelap and endlap) of WRS-2 appeared in about 325-380 TM or ETM+ scenes, those in overlap zones 2 or 3 (either sidelap or endlap) in 215-265 scenes, and those in overlap zone 4 (no overlap) had 100-135 scenes, shown as an RGB image of the blue (R), NIR (G), and second SWIR (B) bands in Figure 2.1. The pattern of high, medium, and low pixel values for scene totals depicted in the figure corresponds closely with the scene boundaries of WRS-2, indicating zones of overlap. These totals represent the largest possible pool of candidate observations that could be used to construct the composite images of TCT metrics. Observations either flagged as missing or scored as contaminated by snow or clouds were removed from this pool to produce a filtered pool of relatively clean observations.

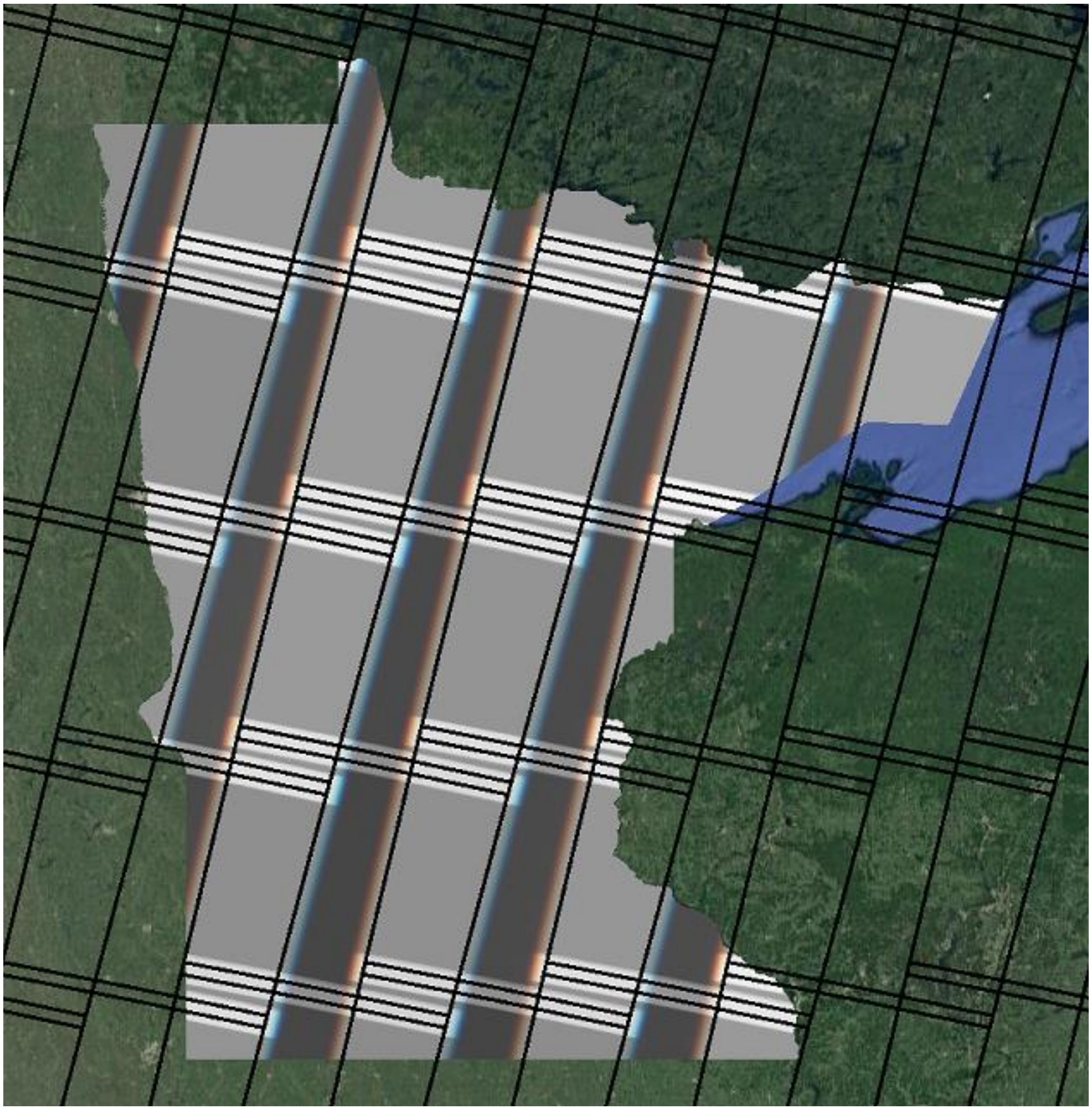

Figure 2.1. RGB image of the total number of Landsat TM and ETM+ observations for each pixel in bands 1 (R), 4 (G), and 7 (B) for study timeframe, 2009-2013. Higher pixel values appear brighter. The black lines represent WRS-2 scene boundaries. Overlap zones are: 1) with both sidelap and endlap (white), 2) with sidelap (medium gray), 3) with endlap (intersection of dark gray and white), and 4) with neither sidelap nor endlap (dark gray).

Figure 2.2 depicts the ratio of the filtered pool size to the total pool size for each reflective band as an RGB image, again showing the blue (R), NIR (G), and second SWIR (B) bands. Water bodies are clearly seen to have the smallest ratio values, meaning that most observations were removed by the snow or cloud filters. Another notable feature is the set of colored lines running parallel to the path of the satellite along the boundaries of overlap zone 4 in a northeast-to-southwest orientation. These are caused by a differential across bands in data flagged as missing, and are also present yet less prominent in Figure 2.1. Also visible is the pattern of higher-value white pixels among the intermediate-value gray pixels. Visual comparison of this pattern to finer resolution imagery suggests a strong correlation with tree cover, with pixels covered by tree canopy having higher pixel values. A final feature, barely visible at the scale of Figure 2.2, is a pattern of alternating brighter and darker bands running perpendicular to the path of the satellite. This will be shown more clearly in the discussion in Figure 2.9.

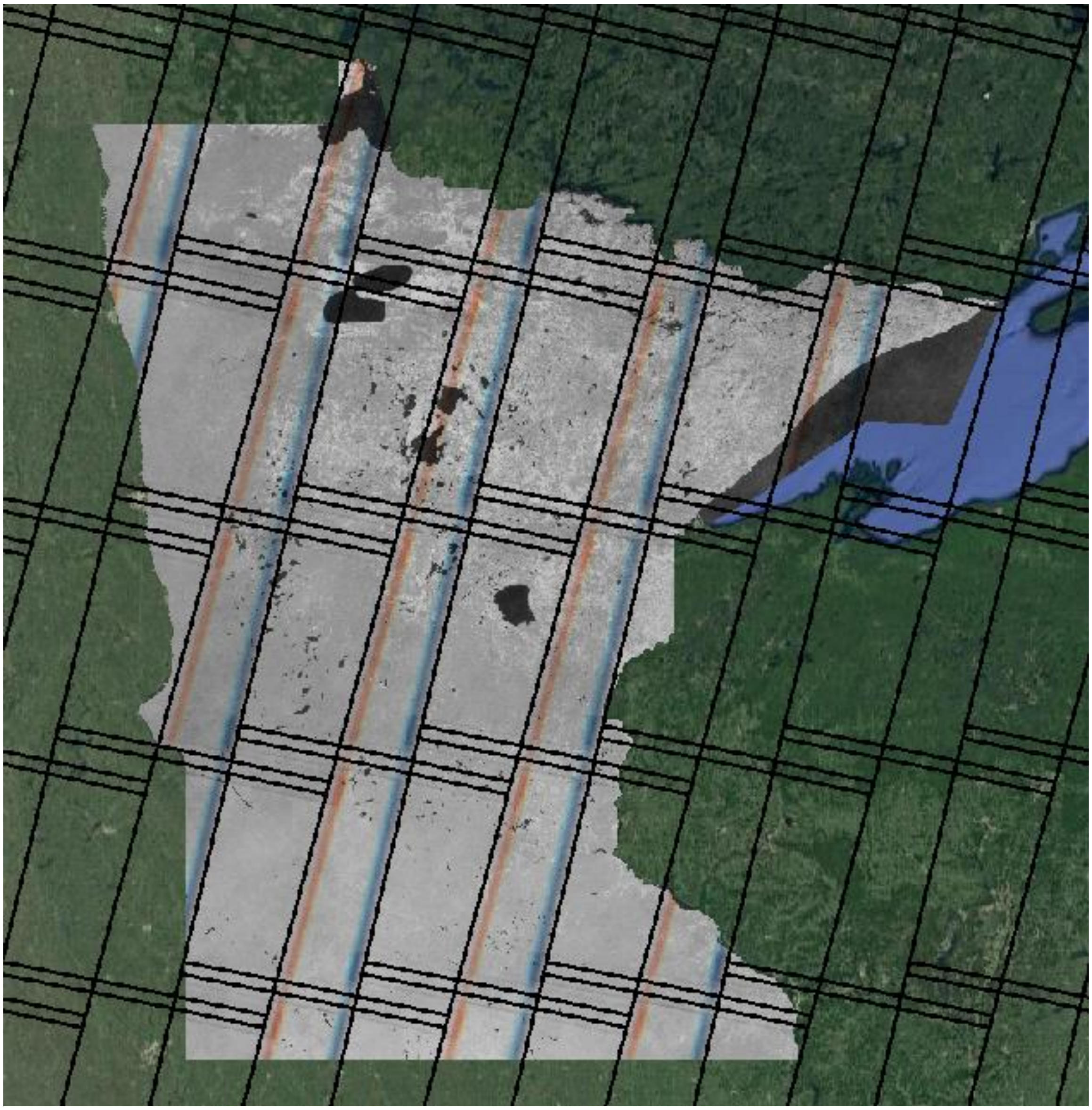

Figure 2.2. RGB image of the ratio of the number of Landsat TM and ETM+ observations per pixel remaining after filtering out cloud and snow to the totals shown in Figure 2.1 for bands 1 (R), 4 (G), and 7 (B). Higher pixel values appear brighter, ranging from 0% remaining after filtering (black) to 100% remaining (white). The black lines represent WRS-2 scene boundaries.

The filtered pool of observations for each band was used to construct 35 median value six-band composite images used for harmonic regression. Figure 2.3 shows the OLS RMSE values as RGB images for TCT brightness (R), greenness (G), and wetness (B) for 1st order to 4th order Fourier series, with the values in each image stretched to a common range (i.e. the band ranges of the 1st order series) for visual comparison. All images show artifacts that correspond with the boundary between overlap zone 4 of WRS-2 and neighboring zones.

The corresponding mean, maximum, and minimum values for the study area are shown as bar charts in Figure 2.4 to highlight the marginal improvement with increasing series order. GEE stores image assets at multiple spatial resolutions to facilitate efficient processing and display of geospatial data. Starting with the native spatial resolution, each subsequent scale (zoom level) differs by a factor of two. This means that pixel values at each coarser scale are assigned the mean of the four corresponding pixel values at the prior finer scale, e.g., 30-m spatial resolution pixels are aggregated to 60-m, 120-m, 240-m, 480-m, 940-m, and 1920-m spatial resolutions. The values in Figure 2.4 were calculated using a pixel resolution of 1,920 meters for efficiency of display and show that mean RMSE for all TCT metrics decreased as the order of the series increased.

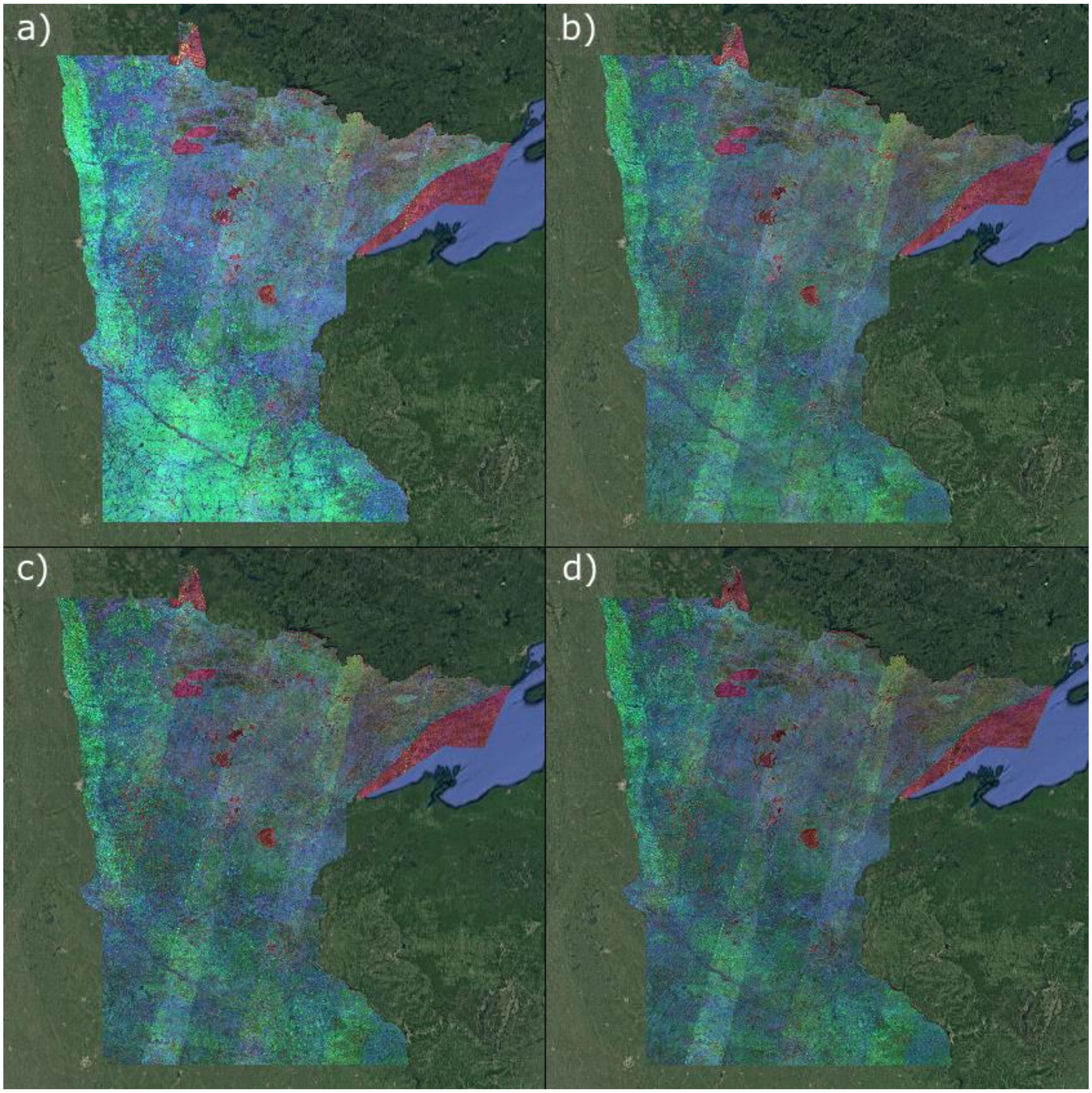

Figure 2.3. Per-pixel RMSE values for 1st to 4th order Fourier series (a-d), displayed as RGB images of TCT brightness (R), greenness (G), and wetness (B). Larger RMSE values appear as brighter.

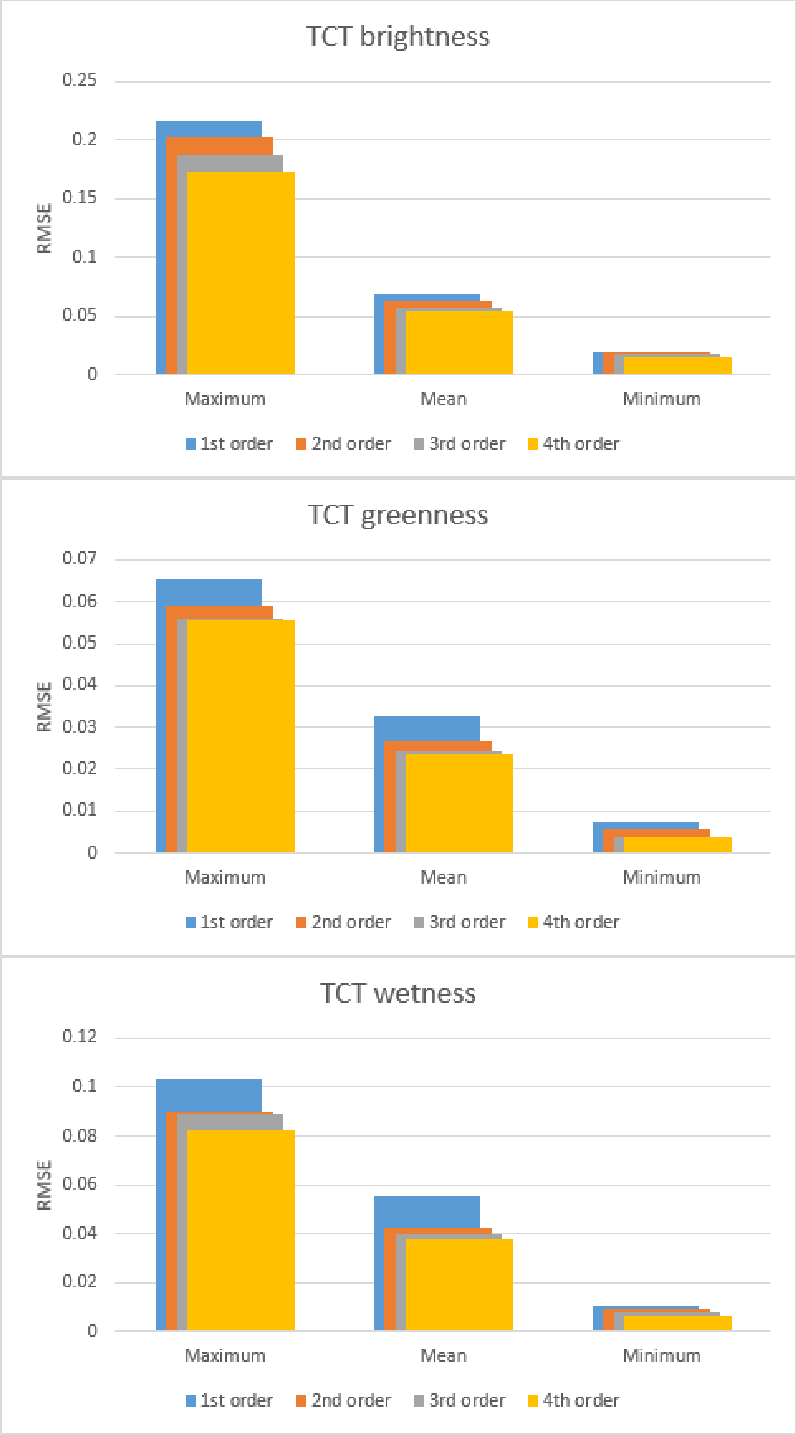

Figure 2.4. Bar charts of maximum, mean, and minimum RMSE for TCT metrics, summarized for the study area from the pixel values in Figure 2.3.

Figure 2.5 shows the estimated OLS regression coefficients for the constant term, corresponding to the mean value, of the model for TCT brightness, greenness, and wetness as RGB images for 1st order to 4th order Fourier series, with the values in each image stretched to a common range for visual comparison. As was the case with Figure 2.3, the images clearly show artifacts that correspond with the boundary between overlap zone 4 of WRS-2 and neighboring zones. Similar patterns appear in the images of the 1st order harmonic terms, which are present in all models, and are shown in Figures 2.6 and 2.7. Comparable images for the 2nd order harmonic terms, which appear only in the 2nd through 4th order series, are shown in Figure 2.8. Images for higher-order harmonic terms were created, but are not shown here.

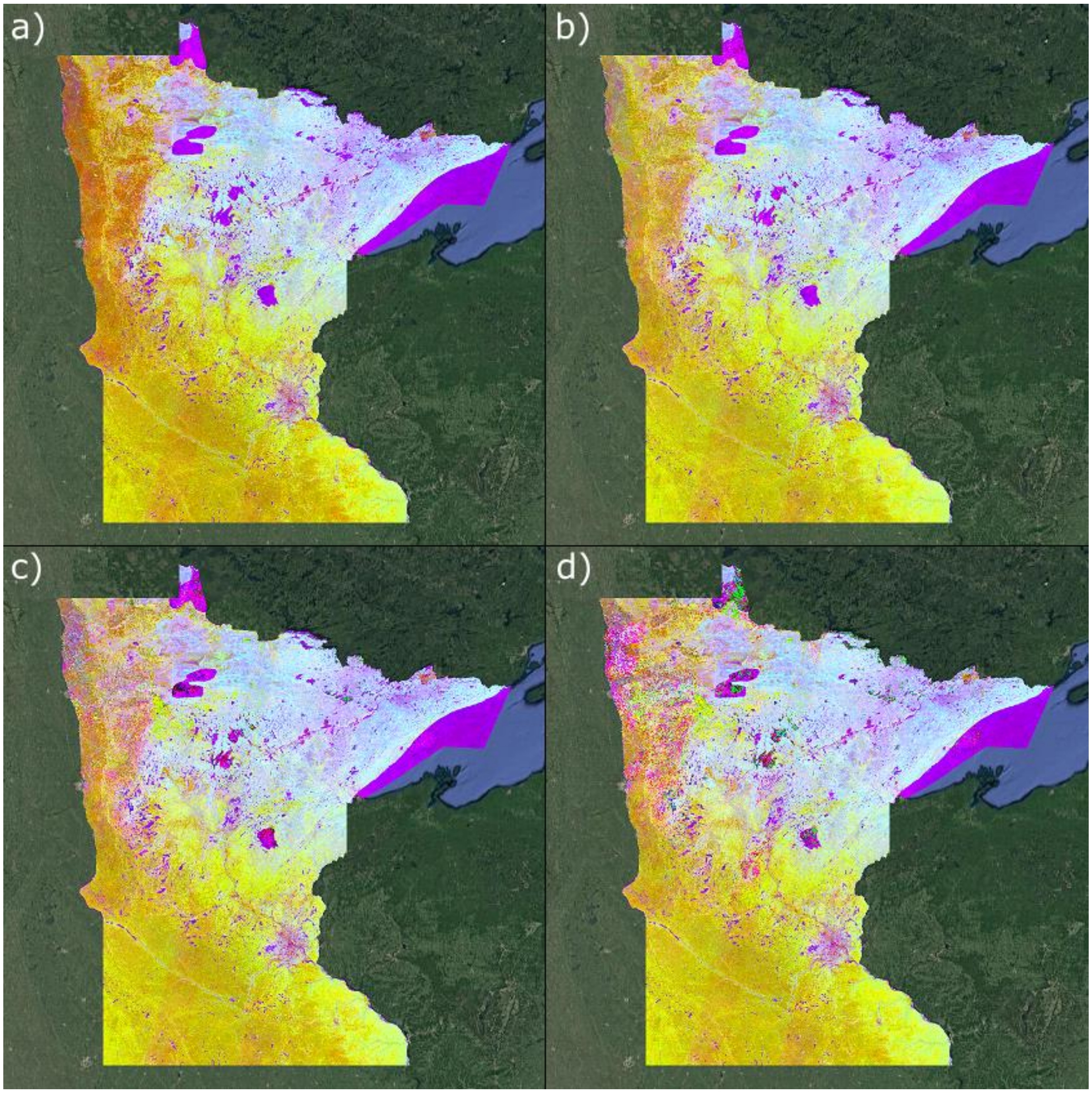

Figure 2.5. RGB images of mean TCT brightness (R), greenness (G), and wetness (B), estimated from 1st to 4th order Fourier series (a-d).

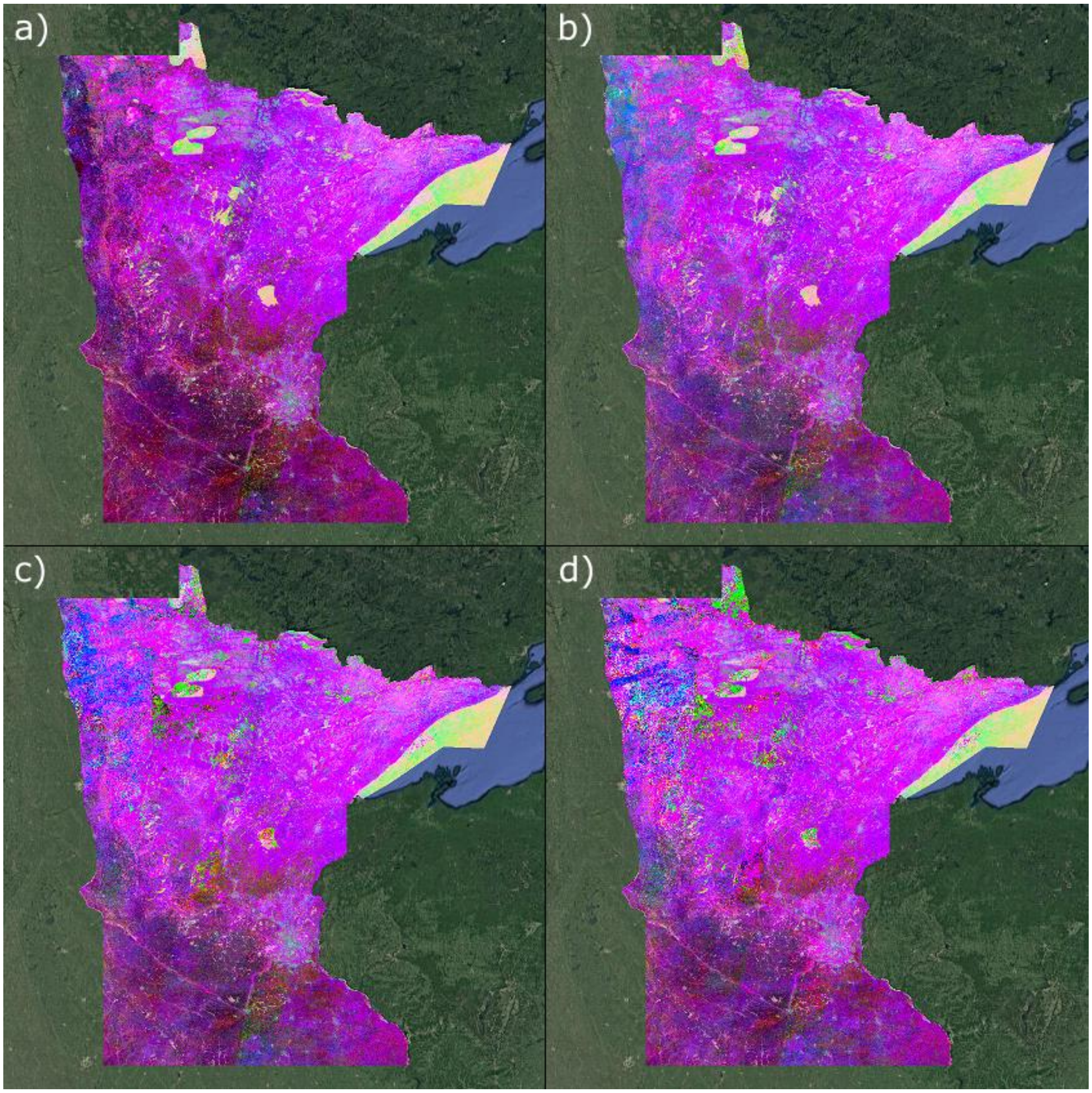

Figure 2.6. RGB images of the estimated coefficient of the cosine term of the 1st harmonic for TCT brightness (R), greenness (G), and wetness (B), estimated from 1st to 4th order Fourier series (a-d).

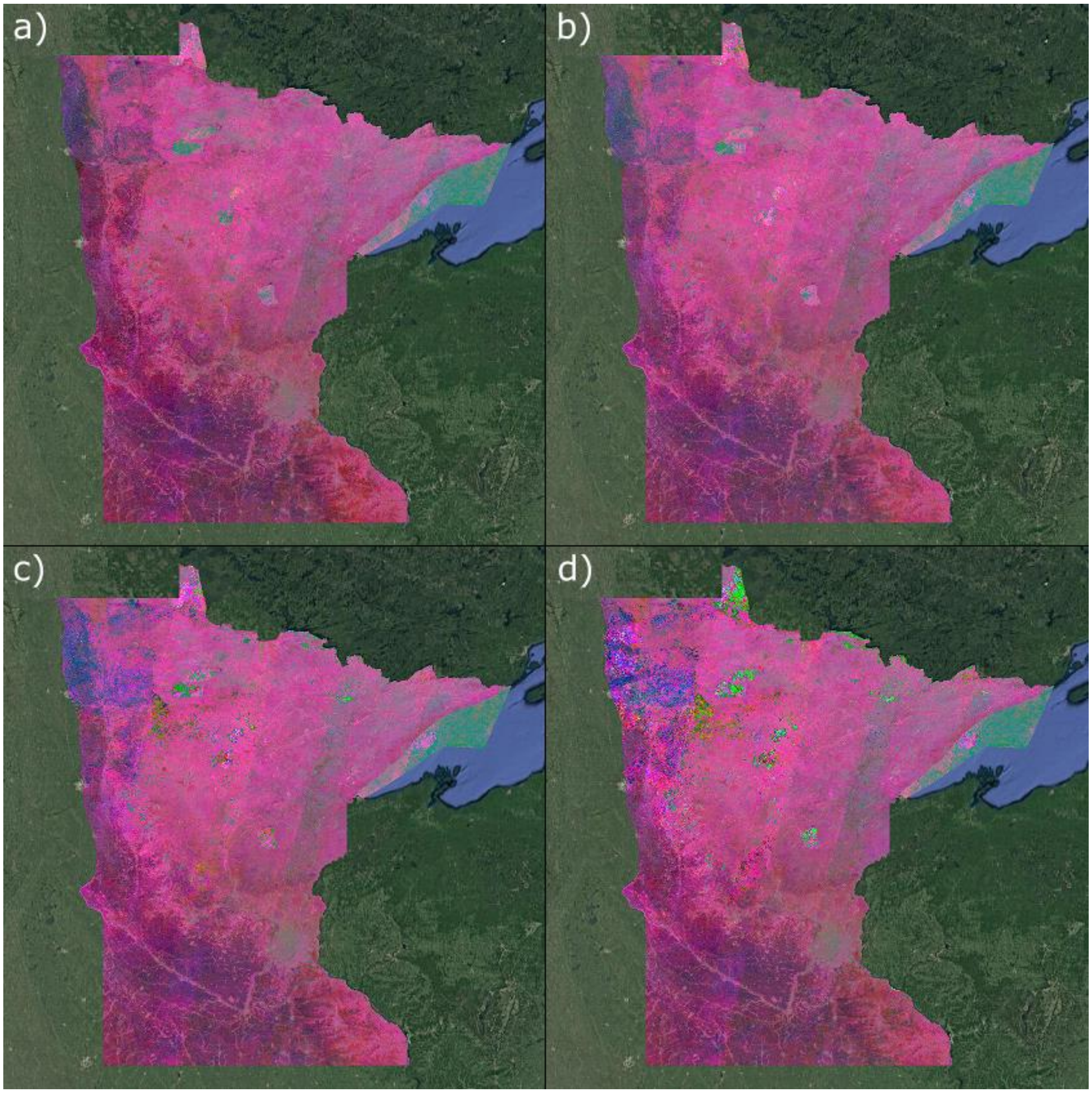

Figure 2.7. RGB images of the estimated coefficient of the sine term of the 1st harmonic for TCT brightness (R), greenness (G), and wetness (B), estimated from 1st to 4th order Fourier series (a-d).

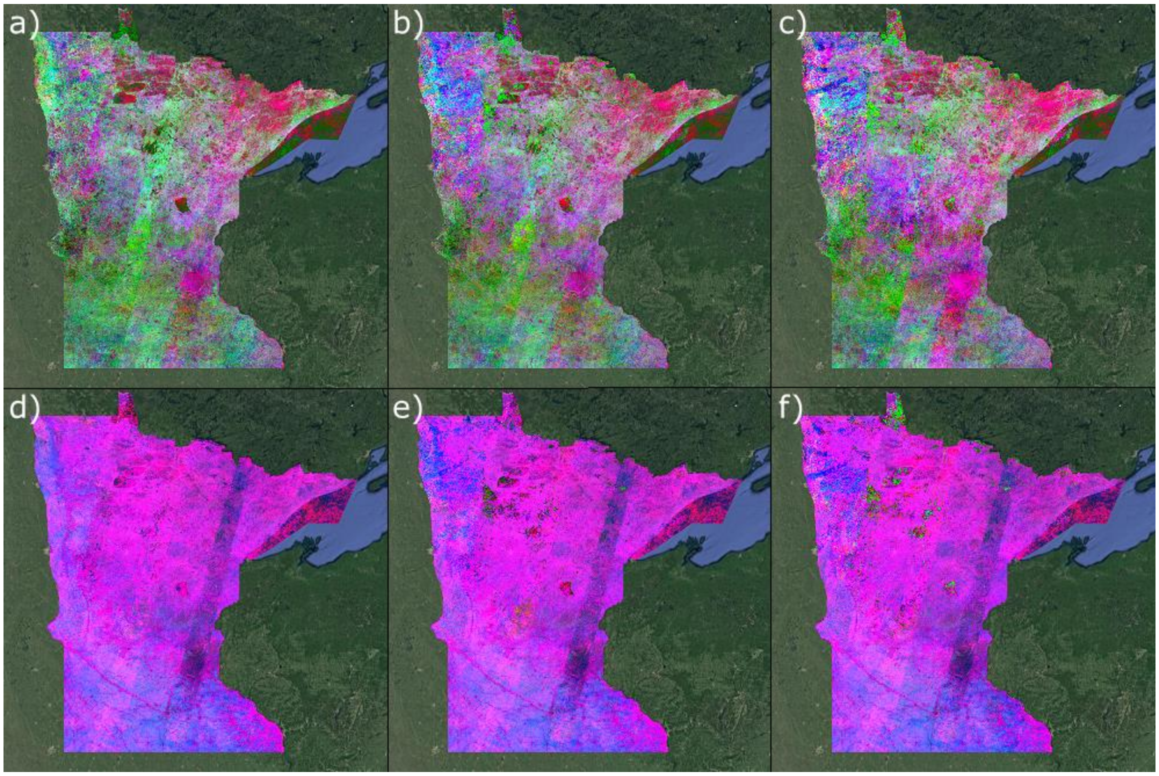

Figure 2.8. RGB images of the estimated coefficients of the cosine (a-c) and sine (d-f) terms of the 2nd harmonic for TCT brightness (R), greenness (G), and wetness (B), estimated from 2nd to 4th order Fourier series.

Models of forest attributes using k-nearest neighbors and Random Forests

The regression results for the *k*NN models are shown in Table 2.1. For each of the 8 continuous response variables, conditional on each of the feature spaces (i.e. composite vs. Fourier), R^2 in the table was estimated using the optimal value of *k* based on leave-one-out cross-validation. The values range from approximately 0.11 to 0.24 for the models using the composite feature space (designated 'composite models') and 0.31 to 0.55 for those using the Fourier feature space (designated 'Fourier models'). The optimal value of *k* varied with both response variable and feature space used in the model, with it being considerably larger for the composite models (93-200) than for the Fourier models (22-69).

***k*NN**	$\bar{Y}$	**RMSE comp.**	**RMSE Fourier**	R^2 **comp.**	R^2 **Fourier**	R^2 **ratio**	**Opt. *k* comp.**	**Opt. *k* Fourier**
trees (#/ha)	62.64	144.2	127.1	0.1658	0.3517	2.122	116	37
trees_bl (#/ha)	40.76	119.6	105.2	0.1103	0.3129	2.837	200	69
barea (m^2/ha)	1.427	2.580	2.004	0.2287	0.5364	2.345	134	29
barea_bl (m^2/ha)	0.9316	2.123	1.681	0.1618	0.4759	2.941	192	47
biofol (kg/ha)	253.5	492.2	377.8	0.2381	0.5517	2.317	93	22
biofol_bl (kg/ha)	119.8	282.2	226.9	0.1537	0.4545	2.957	134	64
bioall (kg/ha)	4824	10070	8294	0.1739	0.4410	2.535	134	64
bioall_bl (kg/ha)	3537	9134	7556	0.1339	0.4090	3.054	143	64

Table 2.1. Results of *k*NN regression models. Includes models of the number (trees), basal area (barea), foliage biomass (biofol), and total aboveground biomass (bioall) of all live trees on the central sub-plot, as well as comparable values for only live broadleaf trees (_bl). Results include the mean of the observations ($\bar{Y}$) and root mean square error (RMSE) of the predictions, as well as the coefficient of determination (R^2) for models using the composite and Fourier feature spaces, their ratio (Fourier/composite), and the optimal value of *k* based on leave-one-out cross-validation.

For each response variable, the ratio of the coefficients of determination using the two feature spaces is also shown, giving a measure of the gain in variance explained by the models using the estimated harmonic regression coefficients, indicating approximately two to three times greater explanatory power. The corresponding regression results for the RF models are shown in

Table 2.2, exhibiting similar ranges for R^2 values and their ratios indicating approximately 2.5 to 3.5 times greater explanatory power of the Fourier models.

RF	$\bar{Y}$	**RMSE comp.**	**RMSE Fourier**	R^2 **comp.**	R^2 **Fourier**	R^2 **ratio**
trees (#/ha)	62.64	145.8	124.9	0.1544	0.3742	2.424
trees_bl (#/ha)	40.76	121.0	104.0	0.09884	0.3281	3.319
barea (m^2/ha)	1.427	2.603	1.958	0.2181	0.5561	2.549
barea_bl (m^2/ha)	0.9316	2.149	1.655	0.1476	0.4906	3.324
biofol (kg/ha)	253.5	496.4	367.6	0.2284	0.5750	2.517
biofol_bl (kg/ha)	119.8	285.9	223.7	0.1378	0.4686	3.399
bioall (kg/ha)	4824	10190	8178	0.1606	0.4556	2.837
bioall_bl (kg/ha)	3537	9255	7454	0.1193	0.4234	3.548

Table 2.2. Results of RF regression models. Includes models of the number (trees), basal area (barea), foliage biomass (biofol), and total aboveground biomass (bioall) of all live trees on the central sub-plot, as well as comparable values for only live broadleaf trees (_bl). Results include the mean of the observations ($\bar{Y}$) and root mean square error (RMSE) of the predictions, as well as the coefficient of determination (R^2) for models using the composite and Fourier feature spaces and their ratio (Fourier/composite).

The classification results for the two-class *k*NN models are shown in the next set of tables. Table 2.3 shows the results for the composite model, with an overall accuracy of almost 82%, individual class accuracies of approximately 69-88%, and an optimal *k* of 33. Table 2.4 shows results for the Fourier model, with an overall accuracy of almost 93%, individual class accuracies of approximately 87-96%, and an optimal *k* of 12. The corresponding results for the two-class RF models are shown in Tables 2.5 and 2.6, indicating similar accuracies to the *k*NN models.

***k*NN composite**
(optimal *k*=33) **Predicted**

		Non-forest	Forest	Total	Producer's accuracy
Observed	**Non-forest**	10521	1492	12013	87.58%
	Forest	1637	3693	5330	69.29%
	Total	12158	5185	17343	
	User's accuracy	86.54%	71.22%		81.96%

Table 2.3. Results of two-class *k*NN classification model using composite feature space. Results include the optimal value of *k* based on leave-one-out cross-validation, number of reference units assigned to each class, individual class user's (commission error) and producer's (omission error) accuracies, as well as overall accuracy.

***k*NN Fourier**
(optimal *k*=12) **Predicted**

		Non-forest	Forest	Total	Producer's accuracy
Observed	**Non-forest**	11267	746	12013	93.79%
	Forest	500	4830	5330	90.62%
	Total	11767	5576	17343	
	User's accuracy	95.75%	86.62%		92.82%

Table 2.4. Results of two-class *k*NN classification model using Fourier feature space. Results include the optimal value of *k* based on leave-one-out cross-validation, number of reference units assigned to each class, individual class user's (commission error) and producer's (omission error) accuracies, as well as overall accuracy.

RF composite		Predicted			
		Non-forest	**Forest**	**Total**	**Producer’s accuracy**
Observed	**Non-forest**	10648	1365	12013	88.64%
	Forest	1759	3571	5330	67.00%
	Total	12407	4936	17343	
	User’s accuracy	85.82%	72.35%		81.99%

Table 2.5. Results of two-class RF classification model using composite feature space. Results include the number of reference units assigned to each class, individual class user’s (commission error) and producer’s (omission error) accuracies, as well as overall accuracy.

RF Fourier		Predicted			
		Non-forest	**Forest**	**Total**	**Producer’s accuracy**
Observed	**Non-forest**	11437	576	12013	95.21%
	Forest	609	4721	5330	88.57%
	Total	12046	5297	17343	
	User’s accuracy	94.94%	89.13%		93.17%

Table 2.6. Results of two-class RF classification model using Fourier feature space. Results include the number of reference units assigned to each class, individual class user’s (commission error) and producer’s (omission error) accuracies, as well as overall accuracy.

The classification results for the 8-class *k*NN models are shown in Tables 2.7 and 2.8, associated with the composite and Fourier models respectively. The composite model had an overall accuracy of almost 62%, individual class accuracies of approximately 41-74% (excluding the rarely sampled ‘grassland’ and ‘non-stocked forest’ classes), and an optimal *k* of eight. The Fourier model had an overall accuracy of approximately 81%, individual class accuracies of approximately 51-92% (again excluding the two rare classes), and an optimal *k* of nine. The corresponding results for the eight-class RF models are shown in Tables 2.9 and 2.10, indicating an improvement in overall and individual class accuracies similar to that shown in the *k*NN results for models using the estimated Fourier coefficients rather than the mean monthly composites.

***k*NN composite (optimal *k*=8)**		**Predicted**									
		Water	**Crop.**	**Grass.**	**Settle.**	**Wet.**	**Conif.**	**Broad.**	**Non.**	**Total**	**Prod. Acc.**
Observed	**Water**	347	275	0	18	48	74	239	0	1001	34.67%
	Crop.	121	7209	0	149	190	115	661	0	8445	85.36%
	Grass.	0	19	0	1	3	0	5	0	28	0.00%
	Settle.	27	568	1	213	31	52	177	0	1069	19.93%
	Wet.	58	668	0	31	190	107	416	0	1470	12.93%
	Conif.	58	197	0	19	78	631	536	0	1519	41.54%
	Broad.	135	825	0	87	178	381	2150	1	3757	57.23%
	Non.	4	12	0	2	9	4	23	0	54	0.00%
	Total	750	9773	1	520	727	1364	4207	1	17343	
	User. Acc.	46.27%	73.76%	0.00%	40.96%	26.13%	46.26%	51.11%	0.00%		61.93%

Table 2.7. Results of eight-class *k*NN classification model using composite feature space. Results include the optimal value of *k* based on leave-one-out cross-validation, number of reference units assigned to each class, individual class user's (commission error) and producer's (omission error) accuracies, as well as overall accuracy.

***k*NN Fourier**
(optimal *k*=9)

Predicted

		Water	**Crop.**	**Grass.**	**Settle.**	**Wet.**	**Conif.**	**Broad.**	**Non.**	**Total**	**Prod. Acc.**
Observed	**Water**	853	14	0	13	50	35	36	0	1001	85.21%
	Crop.	7	7986	1	119	171	4	157	0	8445	94.56%
	Grass.	0	19	0	1	5	0	3	0	28	0.00%
	Settle.	15	340	0	430	76	35	173	0	1069	40.02%
	Wet.	41	405	0	49	549	91	335	0	1470	37.35%
	Conif.	6	15	0	12	68	1079	338	1	1519	71.03%
	Broad.	8	157	0	39	144	220	3189	0	3757	84.88%
	Non.	0	6	0	2	6	8	32	0	54	0.00%
	Total	930	8942	1	665	1069	1472	4263	1	17343	
	User. Acc.	91.72%	89.31%	0.00%	64.66%	51.36%	73.30%	74.81%	0.00%		81.22%

Table 2.8. Results of eight-class *k*NN classification model using Fourier feature space. Results include the optimal value of *k* based on leave-one-out cross-validation, number of reference units assigned to each class, individual class user’s (commission error) and producer’s (omission error) accuracies, as well as overall accuracy.

RF composite		**Predicted**									
		Water	**Crop.**	**Grass.**	**Settle.**	**Wet.**	**Conif.**	**Broad.**	**Non.**	**Total**	**Prod. Acc.**
Observed	**Water**	409	262	0	9	20	66	235	0	1001	40.86%
	Crop.	137	7473	0	57	67	73	638	0	8445	88.49%
	Grass.	0	23	0	0	0	0	5	0	28	0.00%
	Settle.	39	606	0	152	11	44	207	0	1069	14.22%
	Wet.	58	752	0	16	111	106	427	0	1470	7.55%
	Conif.	51	193	0	5	45	663	562	0	1519	43.65%
	Broad.	138	822	0	27	81	317	2372	0	3757	63.14%
	Non.	2	16	0	1	4	7	24	0	54	0.00%
	Total	834	10157	0	267	339	1276	4470	0	17343	
	User. Acc.	49.04%	73.57%	NA	56.93%	32.74%	51.96%	53.06%	NA		64.46%

Table 2.9. Results of eight-class RF classification model using composite feature space. Results include the number of reference units assigned to each class, individual class user’s (commission error) and producer’s (omission error) accuracies, as well as overall accuracy

RF Fourier **Predicted**

		Water	Crop.	Grass.	Settle.	Wet.	Conif.	Broad.	Non.	Total	Prod. Acc.
Observed	**Water**	884	18	0	3	45	16	35	0	1001	88.31%
	Crop.	7	8180	0	48	67	0	143	0	8445	96.86%
	Grass.	0	20	0	1	3	0	4	0	28	0.00%
	Settle.	24	405	0	403	55	37	145	0	1069	37.70%
	Wet.	47	500	0	18	520	71	314	0	1470	35.37%
	Conif.	7	25	0	12	56	1121	298	0	1519	73.80%
	Broad.	7	199	0	22	106	213	3210	0	3757	85.44%
	Non.	0	10	0	1	7	5	31	0	54	0.00%
	Total	976	9357	0	508	859	1463	4180	0	17343	
	User. Acc.	90.57%	87.42%	NA	79.33%	60.54%	76.62%	76.79%	NA		82.56%

Table 2.10. Results of eight-class RF classification model using Fourier feature space. Results include the number of reference units assigned to each class, individual class user's (commission error) and producer's (omission error) accuracies, as well as overall accuracy.

Discussion

Image composites and harmonic regression

A few features shown in the first two figures merit some explanation. The visible mismatch between WRS-2 overlap zones and the regions of high, medium, and low pixel values shown in Figure 2.1 is apparently due to the actual scene dimensions being slightly larger than the dimensions of the scene boundaries used in the WRS-2 index map. This is simply noted and not considered problematic. However, the differential in the amount of data flagged as missing across bands near the zone 4 boundaries, visible in both Figures 2.1 and 2.2, would have the effect of a differential reduction across bands in the pool of candidate observations for construction of composite images. This could result in spurious image artifacts, since each band composite image is calculated from a slightly different pool of observations.

Regarding the small values for water bodies in Figure 2.2, Hall et al. (1995) stated that water bodies may have NDSI values in the range of snow, but lower NIR reflectance. Thus, most water pixels would also be filtered out by the snow filter. However, since the focus of this study is on NFI applications, differentiating water from snow was of secondary importance. Nevertheless, it is noted that the ratio image could be used to mask out water bodies from the study area if desired.

For the large pixel values in Figure 2.2, it is obvious that fewer observations are filtered out for those pixels relative to others. One plausible explanation for the correlation between these pixels and those with tree canopy cover is that snow falling on trees, with or without foliage, does not remain in situ as long as snow falling on the land surface. This result supports the findings of Stueve et al. (2011), who demonstrated the use of snow-covered Landsat time series imagery to improve the mapping of forest disturbances. However, the focus of this study is on the development of methods that can be applied across the US using NFI data, even where snow cover may be rare or non-existent. Therefore, this feature was not used as auxiliary data for the modeling of forest attributes from NFI data.

The alternating bands of brighter and darker pixels, visible in some regions of the study area, appear to be caused by the presence of SLC-off data gaps. Figure 2.9(a) is a subset of Figure 2.2 depicted at a finer scale that clearly illustrates this alternating pattern. The bands are not visible in overlap zone 4, bounded by red pixels on the western edge and blue pixels on the eastern edge. They are clearly visible in overlap zone 2, with blue pixels on the western edge and red pixels on the eastern edge. These artifacts are periodic in nature in the spatial domain. Therefore, a power spectrum of these spatial characteristics of the image can be estimated via a Fast Fourier Transform (FFT).

Figure 2.9(b) shows the power spectrum of the image in (a), using the ImageJ image processing and analysis software (Abràmoff et al., 2004). Periodic features in the spatial domain appear as bright spots, seen in the bottom half, with lower frequency features being closest to the center. The direction of the vector from the center of the power spectrum to the feature corresponds with the direction of the spatial pattern in the original image. Because the power spectrum is symmetrical about the center, only half of these features need to be masked, as shown in the top half. By using the corresponding inverse FFT on the masked power spectrum, the spatial artifacts can be filtered from the original image.

Figure 2.9(c) shows the difference between the original and filtered images, with larger differences appearing as brighter pixels. The difference image clearly shows that the largest differences, due to the periodic spatial patterns induced by the SLC-off gaps, occur primarily in overlap zone 2 and are mostly absent in overlap zone 4. It is not immediately clear whether similar results would be found in a comparable analysis of the images in Figures 2.3 or 2.5-2.8.

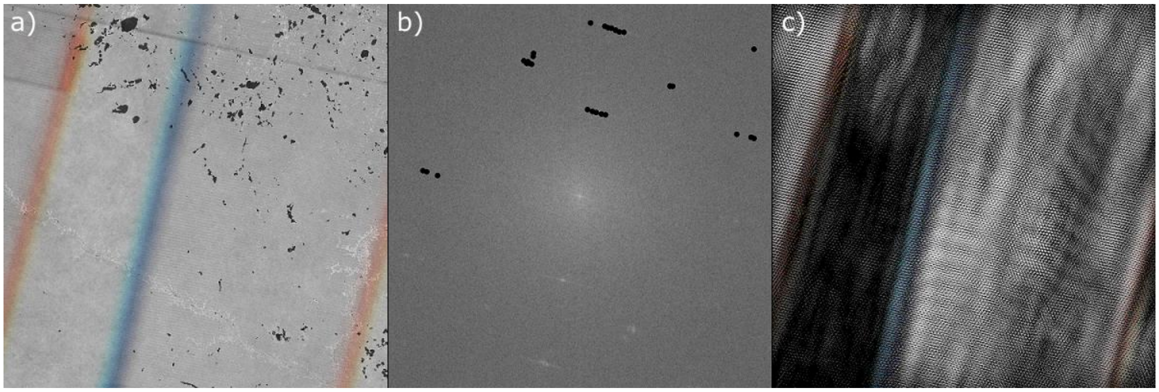

Figure 2.9. (a) Subset of Figure 2.2, (b) its power spectrum showing masked artifacts in black, and (c) the difference between the original and filtered images. Larger differences appearing brighter.

The images in Figure 2.3 show that much of the improvement in fit with increasing series order for greenness and wetness appears to be for cropland pixels (in cyan), which have much larger RMSE values for these two metrics than other pixels in the 1st order image. Improvement in the fit for brightness appears to be uniformly distributed across the study area, with RMSE for brightness remaining much larger than RMSE for greenness and wetness for water pixels (in red). This is to be expected, since overall scene luminance is typically the largest source of variability across an image. Likewise, the aggregate study area RMSE values depicted in the graphs in Figure 2.4 show that greenness and wetness improved only marginally beyond the 2nd order, while brightness exhibited steady improvement with increasing series order. Similar results are seen for maximum values, while the results for minimum values are similar only for brightness and greenness. Minimum wetness exhibited steady improvement with increasing series order.

Figure 2.10(a) shows a subset of the 3rd order image in Figure 2.3, displayed at a finer scale. Figure 2.10(b) and (c) show the results a power spectrum analysis comparable to what was done in Figure 2.9. The difference image suggests that the artifacts caused by the SLC-off gaps in the composite images of TCT metrics are also present in the images generated from the harmonic regression procedure. Furthermore, these artifacts are again most visible in overlap zone 2.

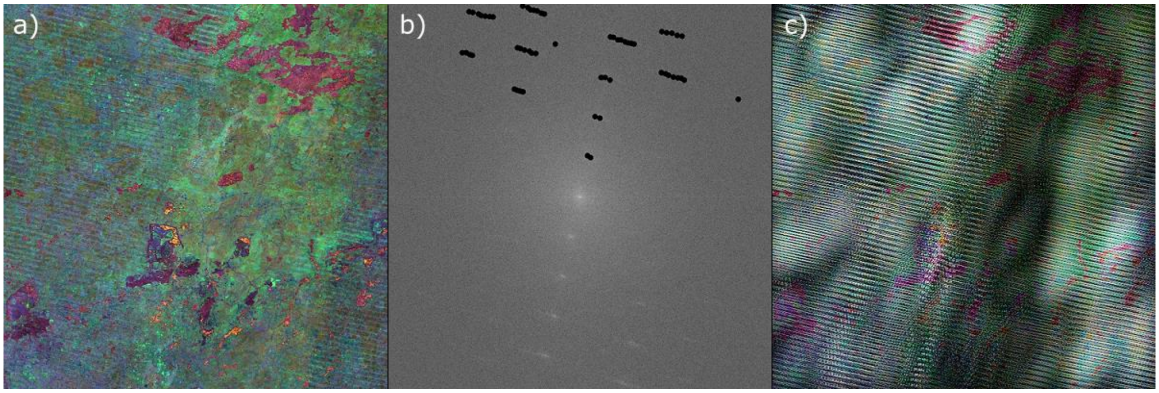

Figure 2.10. (a) Subset of the 3rd order image in Figure 2.3, (b) its power spectrum showing masked artifacts in black, and (c) the difference between the original and filtered images. Larger differences appearing brighter.

Based on the landscape patterns seen in Figures 2.5 through 2.8, harmonic regression of time series of Landsat imagery captures information that is correlated with known historical land cover patterns throughout the state. In the case of Figure 2.5, croplands correspond well with pixels appearing along the color spectrum from yellow to brown, while forest land corresponds with the spectrum from cyan to light purple. Other land cover types, such as developed land (urban areas) and water bodies, also correspond to distinctive spectral patterns in the image. This correspondence holds, though perhaps to a lesser degree, for the higher order images as well.

The cloud and snow filters effectively remove most observations contaminated by clouds and snow. Yet, there remain some visible image artifacts regardless of series order, as noted in the results for Figure 2.3. The image artifacts are less noticeable for forest land pixels with increasing series order. However, the opposite effect is noted for some water and cropland pixels.

Closer examination of these artifacts at finer scales suggests that they are also associated with SLC-off gaps that fall on cropland or water pixels.

Figure 2.11 shows a small subset of Figure 2.5, which includes a boundary between overlap zones 2 and 4 (i.e. zones with only sidelap and no overlap, respectively). The SLC-off artifacts become faintly visible in the 3rd order image, but are clearly visible in the 4th order image. The artifacts are present only in overlap zone 2, on the left-hand side of each image in the figure. Further examination across the study area at finer scales indicates that these artifacts are also present in overlap zone 4, but generally only for water pixels, or for cropland pixels that fall within the areas of differential missing data across bands seen as colored lines in Figures 2.1 and 2.2.

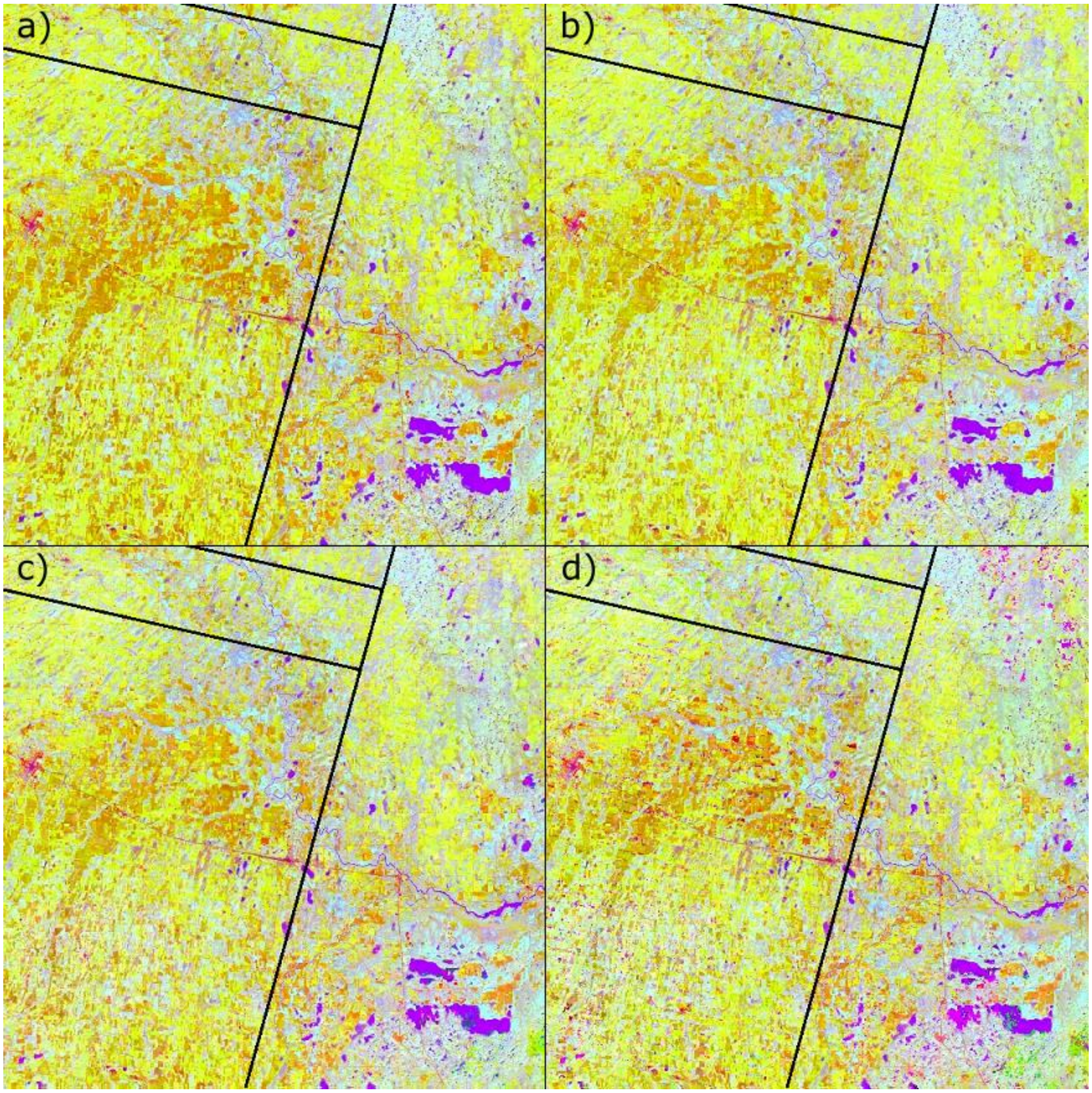

Figure 2.11. Subset of the 1st to 4th order (a-d) images in Figure 2.5. The black lines represent WRS-2 scene boundaries.

It is suggested here that these SLC-off artifacts are due to the composite images being constructed from different pools of observations for neighboring pixels in areas affected by the SLC failure, leading to discontinuities in the resultant harmonic regression coefficient images. In the case of the study area of Minnesota, located at northerly latitudes, the issue is mitigated to some degree by the width of overlap zone 2 (those areas with sidelap). The further north a Landsat scene is located, the more sidelap there is with adjacent scenes, and the greater the likelihood that an SLC-off gap will be filled by a pixel observation from another scene.

While these artifacts are clearly visible in the higher order images of the Fourier series coefficients, they are less conspicuous in the lower order images. This is particularly noticeable for SLC-off gaps falling on cropland and water pixels. This result, along with the results presented in Figures 2.3 and 2.4, suggests that higher order harmonics of the 3rd and 4th order series are fitting extraneous noise rather than meaningful signal, i.e. the spurious oscillations noted in earlier studies. In the case of water pixels, the noise is likely the result of specular reflectance, from the mirror-like surface of calm water bodies, which varies with the view angle of the sensor. In the case of croplands, the noise is likely due to variable cropping patterns from year to year. Since the harmonic regression was conducted on composite images constructed from data collected over the study timeframe (five years), many cropland pixels will not exhibit regular periodic behavior because of crop rotation.

As was done earlier in Figures 2.9 and 2.10, a power spectrum analysis could be used, along with the corresponding inverse FFT, to correct these spatial artifacts. However, that process would involve a considerable amount of image interpretation and manipulation of the power spectrum image by the analyst that would be difficult to automate over large areas. More important, it is not immediately clear whether these artifacts would necessarily cause similar results in maps of forest attributes developed from the *k*NN and RF models. For these reasons, the spatial FFT correction was left for a future study. As a more efficient means of mitigating the potential impacts of these artifacts, the uncorrected 3rd order images were selected for use as auxiliary data for the modeling of forest attributes from NFI data.

Models of forest attributes using k-nearest neighbors and Random Forests

There are a few observations to note in the results of the regression models shown in Tables 2.1 and 2.2. For both *k*NN and RF modeling approaches, the Fourier models demonstrated marked improvement over the composite models, in terms of the ratios of coefficients of determination. Also noted for the *k*NN Fourier models, regardless of the response variable, R^2 increased while the optimal value of k decreased relative to the composite models. This means that as the correlation between the predictor and response variables increased, the predicted values were based on fewer neighbors, suggesting they were more local in terms of the feature space of predictor variables.

Another interesting result for all regression models, regardless of modeling approach or predictor feature space, is that R^2 values for those models using the response variables for only live broadleaf trees (broadleaf tree models) were all smaller than the corresponding models for all live trees (all tree models). While one might expect this result, it is notable that the ratios for the broadleaf tree models are all larger than those of the corresponding all tree models. The larger relative improvement in the broadleaf tree models, as measured by the ratios of the coefficients of determination, suggests that estimated Fourier coefficients enable differentiation between broadleaf and coniferous trees.

The classification models show a similar pattern of results for both the *k*NN and RF modeling approaches. For the results of the two-class models, shown in Tables 2.3-2.6, there is improvement in overall accuracy of approximately 11 percentage points in the Fourier models relative to the composite models. For the eight-class models, the improvement in overall accuracy amounts to almost 20 percentage points. Interestingly for the *k*NN models, while the two-class case saw a substantial decrease in the optimal value of *k* (33 vs. 12) when using the Fourier rather than composite models, the eight-class case showed essentially no difference (eight vs. nine).

Regardless of modeling approach or feature space used, for the two-class models the class accuracies (both producer's and user's accuracy) were greater for the 'non-forest' class than the 'forest' class. It also should be noted that there were 12,013 'non-forest' units in the reference set vs. 5,330 'forest' units, so it would be expected that even by random assignment the 'non-forest' class would have the greater class accuracy. The 'non-forest' class accuracies were approximately 7-11% greater, while the 'forest' class accuracies were approximately 22-32% greater using the Fourier models relative to the composite models.

For the eight-class models, the individual class results shown in Tables 2.7-2.10 are less straightforward to summarize. Using the composite feature space with both the *k*NN and RF approaches, the 'cropland' class had the greatest accuracies by a substantial margin, followed by the 'broadleaf forest', 'coniferous forest', and 'water' classes. Using the Fourier feature space with both the *k*NN and RF approaches, the class results differ depending on whether producer's or user's accuracies are considered. Based on producer's accuracy, the 'cropland' class again had the greatest accuracy, followed closely by the 'water' and 'broadleaf forest' classes, then the 'coniferous forest' class. Based on user's accuracy, the 'water' class had the greatest accuracy, followed closely by the 'cropland' class, then the 'broadleaf forest' and 'coniferous forest' classes.

Regardless of the modeling approach and feature space used, the rarely sampled 'grassland' and 'non-stocked forest' classes had the lowest classification accuracies of all classes, with no reference units being correctly classified. The results for the 'settlement' and 'wetland' classes were mixed, with intermediate accuracies. Using the composite feature space with both the *k*NN and RF approaches, these classes had greater user's accuracies than producer's accuracies, with the 'settlement' class showing slightly greater accuracies than the 'wetland' class. The same pattern holds for the Fourier models.

All classes showed greater individual producer's and user's accuracies using the Fourier models rather than the composite models, some markedly so. The pattern of improvement was consistent, whether using the *k*NN or RF modeling approaches. Relative to the class accuracies achieved using the composite models, the Fourier models showed improvements ranging from approximately 10-20% for the 'cropland' class, 35-70% for the two stocked forest classes, 40-165% for the 'settlement' class, 85-150% for the 'water' class, and 85-370% for 'wetland' class.

Although it was not a formal objective of the study, it should be noted that on balance the RF models produced slightly greater R^2 values and classification accuracies than the comparable *k*NN models, though this was not universally true. For the regression models using the composite feature space, the *k*NN models resulted in R^2 values that were marginally greater than those of the corresponding RF models. However, the converse was true for all regression models using the Fourier feature space. From the perspective of comparing modeling approaches, the classification results were once again less straightforward. For the two-class models using either the composite or Fourier feature space, the overall accuracies were nearly identical and there were no clear patterns in the individual class results. For the 8-class models, the overall accuracies were approximately one to two percentage points better using the RF approach. For the composite models using the RF approach, the user's accuracies of almost all classes were at least marginally greater than the class results using the *k*NN approach. The corresponding producer's accuracies were less conclusive. A similar pattern can be seen in the results for the eight-class Fourier models. It is possible that even these small differences between the two modeling approaches may become negligible with additional optimization of the *k*NN models, particularly through the use of feature selection or feature weighting during development of the feature space of predictor variables.

The objective of the study was to compare alternative feature spaces derived from dense Landsat time series imagery, and not necessarily to produce the best possible models under either alternative. However, even without extensive efforts at optimization of the *k*NN models, the results compare favorably with similar earlier studies. McRoberts (2009b) used the *k*NN algorithm with data collected on FIA sub-plots and 12 spectral features derived from multi-date TM and ETM+ imagery to predict tree volume, density, and basal area for a study area in Northern Minnesota corresponding to WRS-2 Path 27 Row 27. The mean R^2 value for all three models was 0.11 using individual sub-plots and pixels, and 0.24 when aggregating FIA observations to the plot level and using a 3x3 pixel about each plot center. Likewise, the 3-class (non-forest, coniferous forest, and broadleaf forest) and 4-class (non-forest, coniferous forest, broadleaf forest, and mixed forest) classification models had overall accuracies of 72% and 65% respectively, with individual class accuracies ranging from 62-91% and 10-89% respectively. McRoberts (2012), in a study using the same set of predictor and response variables for the same study area as the 2009 study, determined that the optimal value of *k* that minimized RMSE for tree volume models was 25 with a feature space consisting of a subset of 6 of the original 12 spectral features.

The RF regression model results are also comparable to those from an earlier study using FIA data and Landsat time series imagery. Zhu and Liu (2015) conducted a study in southeast Ohio for mapping and estimation of live tree aboveground biomass. The study used measurements from all sub-plots on a subset of 161 homogeneous FIA plots from a predominantly high biomass (over 100 metric tons/ha) 11-county study area, along with terrain-corrected NDVI values derived from six cloud-free Landsat scenes from one growing season. Predictions were made using six modeling methods, with R^2 values ranging from 0.48 for RF to 0.54 for artificial neural networks.

Conclusions

There are a number of conclusions to be drawn from this study. First, harmonic regression using Fourier series fitted to composite images derived from dense time series of Landsat imagery was shown to be an effective means of handling missing data in the imagery due to the presence of clouds and snow. Second, the artifacts in the ETM+ image record due the SLC-off condition were mitigated to a great extent both through supplementation with concurrent TM imagery during compositing and by use of a low-order Fourier series. Third, and most important, the estimated Fourier coefficients developed by harmonic regression of TCT time series were correlated with land cover, including tree cover, based on a qualitative assessment of the imagery and knowledge of land cover patterns in the study area.

The strength of this correlation was quantifiably demonstrated using two non-parametric modeling approaches and a range of response variables derived from NFI data. Regression models using a feature space of estimated Fourier coefficients showed a two- to three-fold increase in explained variance for a small set of continuous response variables, relative to comparable models using monthly image composites. Similarly, the overall accuracies of the two-class and eight-class classification models using the Fourier feature space were approximately 10 to 20 percentage points higher than the models using the composite feature space. The corresponding individual class accuracies likewise were approximately six to 21 percentage points higher in the two-class case and 10 to 45 percentage points higher in the eight-class case when using the Fourier coefficients versus the composite images as predictor variables.

The ultimate utility of this study relates to the role that these models might play in improving the precision of population estimates of land cover and forest attributes. Comprehensive national forest inventories, such as the one conducted by FIA in the US, are expensive by nature. In order to collect the information used in this study, particularly the continuous response variables, field crews had to travel to and make tree measurements on each of the forested plots used in the analysis, amounting to approximately 6,000 plots in this case. This represents a substantial investment by the USFS. Yet, even at the FIA sampling intensity of one plot per 2,400 hectares, the sample size required for reliable estimation would limit the application of most design-based methods to geographic areas, or populations, that are on the order of multiple counties within a state. By incorporating models such as the ones used in this study into model-assisted or model-based approaches, the scale of application of FIA data could approach what is needed for local forest management decisions.

Chapter 3: Boosting and model-based optimization of *k*-nearest neighbors: applications for small area estimation and national forest inventory

Summary

A variant of the *k*-nearest neighbors (*k*NN) algorithm is proposed for modeling continuous response variables from sample survey data with auxiliary population data that provides a unified framework for global optimization of *k*NN while simultaneously selecting feature variables and correcting for prediction bias. An empirical study was conducted to test the small area estimation performance of the proposed estimator using national forest inventory data with dense time series of Landsat imagery. Features were extracted from all Landsat scenes acquired during the study timeframe 2009-2013 for one ecological unit in the state of Minnesota by means of harmonic regression. The locations of 1138 plots collected by the United States Forest Service Forest Inventory Analysis (FIA) program within the study area and timeframe were used to sample from a simulated population of tree canopy cover (TCC). Bamboo *k*NN, which stands for boosting and model-based optimization (MBO) of *k*-nearest neighbors, was used to construct and optimize a nonparametric model for predicting TCC using a feature space of estimated harmonic regression coefficients. The assumed but unknown bias of the *k*NN smoother is estimated by recursively fitting residuals, known as L_2 boosting, with the *k*NN smoother used to make the initial predictions. The proposed MBO algorithm is a Markov chain Monte Carlo method for generating candidate solutions to the boosted *k*NN model using a parameterized probabilistic model for selecting values of *k* and subsets of feature variables from the solution space. This model is adaptively modified, using earlier candidate solutions, to concentrate the search in the most promising regions of the solution space. Guidelines are suggested for

determining the appropriate order of recursion for boosting, as well as length of the chain. At the end of the chain, a small sample of candidate solutions was drawn from the solution space using the updated probability weights. These candidate *k*NN models were used to construct predictive intervals for spatial domains over a range of sizes. Coverage tests were conducted by determining the proportion of spatial domains, for each domain size tested, whose predictive intervals contained the actual TCC value observed in the simulated population. The results showed that the coverage proportion approached the theoretical value when using a 4th order boost, for spatial domains as small as the area represented by an FIA sample unit, with the unboosted model coverage proportion smaller than the theoretical value.

Introduction

A national forest inventory (NFI), such as the one conducted by the United States Forest Service (USFS), is designed to provide information on the status of the forests of a nation as a whole, or perhaps for large domains within it. An NFI is therefore often characterized as being supportive of national or strategic forest management, as opposed to local or tactical decision-making. It is common practice in NFI to use direct estimators, which rely solely upon the sample units drawn from the domain and are design-unbiased or approximately so, to make statistical inferences about the domain population. This reliance upon the domain sample limits the minimum size of a domain for which valid inferences can be made. Indirect estimators effectively increase the size of the domain sample by borrowing strength from all sample units, including those outside of the domain. They typically produce estimates with smaller variance, though potentially larger bias, than direct estimators and can be used for small area estimation (SAE) problems where the domain sample is too small to make valid or useful inferences with a direct estimator (Rao, 2003). SAE techniques, therefore, have the potential to extend the utility of NFI data to a scale of application approaching what is needed to inform local forest management decisions.

Indirect estimators are typically based on the use of models that define the relationship between auxiliary data, available either as unit-level data for all population units or as area-level data for well-defined areas within the population, in conjunction with survey data available for the sample units. Unit-level models have been used extensively in NFI because of the widespread availability of remote sensing imagery that can be gathered inexpensively for the entire population, relative to the cost of collecting sample survey data in the field. One modeling approach that has been used extensively in this context is the *k*-nearest neighbors (*k*NN) algorithm, due to its relative simplicity and multivariate potential to provide predicted values of all attributes observed for sample units. Some examples of its application in NFI include the work of Tomppo (1990), Tokola et al. (1996), and Katila and Tomppo (2001) in Finland, and Moeur and Stage (1995), Franco-Lopez et al. (2001), Ohmann and Gregory (2002), and McRoberts et al. (2002) in the United States. McRoberts et al. (2007) were the first to develop a formal model-based approach to SAE using *k*NN. McRoberts (2012) also provided a thorough review of the use of *k*NN in NFI with auxiliary variables, as well as an overview of modeling choices and diagnostics for assessing model performance.

Overview of the *k*NN algorithm

Fix and Hodges (1951) were the first to propose the *k*NN classification rule. Simply stated, the rule assigns to any unclassified population unit the modal class among its *k*-nearest neighbors, with proximity measured in the space of auxiliary variables used for classification, from the set of classified sample units. Cover (1968) was one of the first to examine the *k*NN rule for estimation, proving that for a range of probability distributions, the large sample risk of the estimator is no worse than twice the minimum expected squared-error loss (i.e. Bayes risk).

In the regression case, the kNN estimator is a type of kernel smoother similar to that proposed by Nadaraya (1964) and Watson (1964). A variety of kernels can be used to determine the weights associated with the sample units that fall within the kernel used to produce a weighted average estimate. The primary difference between the two approaches is that the degree of smoothing in the Nadaraya-Watson estimator is determined by the bandwidth of the kernel, whereas for the kNN estimator it is determined by the value of k. In effect, the former uses a fixed width but variable density kernel, while the latter uses a variable width but fixed density kernel.

Following the notation used by McRoberts et al. (2010), let Y denote a vector of response variables with known values for a sample of size n from a finite population. Let X denote a vector of p auxiliary variables with known values for all population units. The set of population units for which values of both response and auxiliary variables are known is designated the reference set. The set of population units for which predictions of the response variable are desired is designated the target set. The space defined by the auxiliary variables X is designated the feature space. For regression problems, where the values of Y are continuous, the kNN prediction for the ith element of the target set is:

$$\widetilde{y_i} = \frac{1}{W_i}\sum_{j=1}^{k} w_{ij}y_{ij}, \qquad [\ 1\ ]$$

where $\{y_{ij}, j = 1, 2, \ldots, k\}$ are the values of the response variable associated with the k units in the reference set that are nearest in feature space X with respect to distance metric d to target unit i. The weight assigned to each neighbor y_{ij} of target unit i is w_{ij}, with $W_i = \sum_{j=1}^{k} w_{ij}$.

Despite the simplicity of the kNN algorithm, there are still a number of modeling decisions that impact its performance. These include the choice of the number of neighbors k, distance metric d, weighting scheme w, and feature space X. As mentioned earlier, the value of k determines the degree of smoothing of the estimator. This value defines the balance of the trade-off between the variance and bias of the estimator, with smaller values minimizing bias and larger values minimizing variance (Cover, 1968). Loftsgaarden and Quesenberry (1965) provided some early guidance on the choice of k, suggesting $n^{\frac{1}{2}}$ as a value likely to give good results. Based on a simulation study, Enas and Choi (1986) arrived at values ranging from $n^{\frac{1}{4}}$ to $n^{\frac{3}{8}}$ depending upon the sample proportion and underlying covariance structure for small samples. Fukunaga and Hostetler (1973) derived an expression for the optimal choice of k, in terms of minimizing the approximated mean-square error, as a function of n, p, and the underlying distribution, with optimal k increasing as n and p increase.

Prasath et al. (2017) provide an extensive review of 54 distance metrics, which they group into eight major distance families, that have been proposed for use with the kNN algorithm. One widely applied family of distance metrics is Minkowski distance, defined by the equation:

$$d = \sqrt[m]{\sum_{i=1}^{p} |x_{1i} - x_{2i}|^{m}}, \qquad [\,2\,]$$

where x_1 and x_2 are points in feature space with p dimensions and $m \geq 1$ is the order of the distance metric d. The most commonly used cases of $m = 1$ or $m = 2$ correspond to Manhattan (“taxicab”) and Euclidean (“straight-line”) distances respectively, while the limiting case of $m = \infty$ corresponds to Chebyshev (“chessboard”) distance. Several studies of the *k*NN algorithm for both classification and regression have shown that model performance using Manhattan distance is comparable to if not better than that using most alternative metrics, including Euclidean and Chebyshev distances (Ooi et al., 2013; Chomboon et al., 2015; Hu et al., 2016; Prasath et al., 2017). Aggarwal et al. (2001) demonstrated the superiority of smaller m particularly for higher dimensional feature spaces.

Royall (1966) was the first to formalize a weighted version of the *k*NN estimator, with larger weights w assigned to sample units nearer in feature space to the target unit. Dudani (1976) demonstrated the admissibility of the distance-weighted *k*NN rule for classification problems, with Stone (1977) and Devroye (1978) providing generalized proofs in the regression case using a variety of weighting functions. However, Bailey and Jain (1978) proved that, for classification problems where the number of training samples is large, the probability of error of a simple majority rule will be no greater than any distance-weighted rule for *k*NN. Macleod et al. (1987) reexamined the distance-weighted *k*NN rule and demonstrated that a carefully weighted rule can outperform the unweighted rule for some classification problems with a finite training set. Studies have also documented that the optimal value of k increases with the use of non-uniform weights, thereby confounding the relative impact of the choice of k and w (Dudani, 1976; Hechenbichler & Schliep, 2004; Batista & Silva, 2009).

The number of dimensions p of feature space X also affects the performance of the *k*NN algorithm. Assuming a uniformly distributed sample in X, the side of a hypercube needed to contain a fraction f of the sample must have length $f^{\frac{1}{p}}$. Therefore, as $p \to \infty$, then $f^{\frac{1}{p}} \to 1$. From the perspective of the *k*NN algorithm, this means that in order to keep k fixed as p increases, the search neighborhood for finding neighbors becomes larger, approaching a neighborhood as large as the entire feature space. Bellman (1961) dubbed the issue ‘the curse of dimensionality’. In the case of the kNN estimator, this leads to increased bias due to over-smoothing (Friedman, 1997; Bengio at al., 2005).

One of the most intuitive approaches to addressing the curse of dimensionality is to eliminate extraneous features, otherwise known as feature selection. In the context of machine learning, Blum and Langley (1997) identified three approaches to feature selection: filter, wrapper, and embedded methods. In supervised machine learning, an induction algorithm (e.g. the kNN algorithm) must induce a classifier from a set of training instances (i.e. the reference set). Using the filter method, individual features are selected using a preprocessing step, before being passed through to the induction algorithm. Typical examples include using p-values or other criteria, such as the Mallow’s C_p statistic or the Akaike Information Criterion, to evaluate features independently and determine those to be passed to the induction algorithm. Using the wrapper method, feature selection is accomplished using the induction algorithm to perform the evaluation and takes into account interactions among features by selecting subsets of features. Typical examples include hill climbing (e.g. forward-selection and backward-elimination), simulated annealing, and genetic algorithms. Kohavi and John (1997) demonstrated significant improvement in classification accuracy using the wrapper method instead of the filter method with some commonly used classifiers.

Embedded methods are similar to wrapper methods, with the key difference being that feature selection is embedded directly within the induction algorithm and lack the generality of both filter and wrapper methods. Embedded methods are widely considered to be less computationally intensive and less prone to overfitting than wrapper methods. Typical examples include regularization techniques (e.g. ridge regression) and decision tree methods. One promising embedded method for use with the *k*NN algorithm is the Random *k*NN for Feature Selection (RKNN-FS) procedure proposed by Li et al. (2011). Their method was developed to address the common problem in bioinformatics of having high dimensional feature data (i.e. tens of thousands of genes from microarray data) with small reference samples (i.e. limited patient diagnoses or outcomes). It is similar in principle to Random Forests (Breiman, 2001), but uses an ensemble of random *k*NN models instead of forest of random decision trees. Their study of 21 microarray gene expression datasets suggested that RKNN-FS provided better classification results than Random Forests with increasing data complexity.

Another approach to reducing the bias of a smoother is through boosting. Schapire (1990) first proved that a weak learning algorithm could be converted into a stronger one by recursively boosting it by a small but significant amount. Tukey (1977) first proposed such a method, dubbing it ‘twicing’, whereby the residuals of the smoother are used to estimate its bias by feeding them back into the smoother. Di Marzio and Taylor (2004) proposed an iterative re-weighting algorithm for boosting kernel density estimates and provided justification for its use in bias reduction. Cornillon et al. (2008) showed that iterative bias reduction schemes, such as the one proposed by Di Marzio and Taylor, correspond to L_2 boosting (Friedman, 2001), which is simply the repeated least squares fitting of residuals. Under this approach, a reasonably large value of the smoothing parameter is chosen in order to intentionally oversmooth the data, resulting in an estimator with relatively small variance but substantial bias that can be estimated from the residuals to correct the initial smoother. Park et al. (2009) demonstrated that L_2 boosting is superior for reducing bias to the use of higher-order kernels, which corresponds to larger weights given to nearer neighbors.

Global optimization is the process of finding the point in a solution space where an objective function attains an extremum (maximum or minimum). Global optimization algorithms can be grouped into deterministic and stochastic approaches. While deterministic approaches provide a theoretical guarantee of finding the global extremum, they are generally slower than stochastic approaches that converge only in probability to the global extremum (i.e. heuristics). Devroye (1978) showed that, under a variety of noise conditions, the *k*NN estimator converges in probability as the sample size approaches infinity. He further suggested that this property could be exploited to design a random search algorithm for optimization of the regression function.

Zlochin and Dorigo (2002) classified heuristic algorithms, such as selecting the best choice of parameters for the *k*NN algorithm, into instance-based and model-based search. Instance-based search methods generate new candidate solutions solely from the current solution, or the current population of solutions. Typical examples include genetic algorithms and simulated annealing. Model-based search methods generate candidate solutions using a parameterized probabilistic model that is adaptively modified, using earlier candidate solutions, to concentrate the search in the most promising regions of the solution space. Typical examples include stochastic gradient descent, ant colony optimization, and estimation of distribution algorithms. Bartz-Beielstein and Zaefferer (2017) provide an excellent overview of stochastic search algorithms as well as a comprehensive taxonomy of model-based optimization (MBO) approaches.

Proposal of Bamboo *k*NN

In this manuscript a variant of *k*NN regression is proposed and designated Bamboo *k*NN, for boosting and model-based optimization of *k*-nearest neighbors. This methodology leverages the bias reduction properties of recursive L_2 boosting, as well as a model-based search heuristic for global optimization of the regression function. The heuristic includes a method of feature selection embedded within the *k*NN induction algorithm. Both the more mundane and novel characteristics of the algorithm will be discussed in detail here, including the choice of distance metric and weighting scheme, as well as the use L_2 boosting and model-based optimization with embedded feature selection.

Distance metric

Based on the findings of earlier studies into the impacts of the choice of distance metric on the performance of the *k*NN algorithm, the Manhattan distance metric (MD) is used with Bamboo *k*NN. While Prasath et al. (2017) documented better performance when using other distance metrics by some measures of classification accuracy, MD was among the top seven of 54 performers for the 28 noise-free datasets they examined. Furthermore, as the level of noise in each of the datasets increased, the relative ranking of MD improved, with it being ranked the top performer overall at the highest noise level. While the Euclidean distance metric also exhibited a similar trajectory to MD with increasing noise, it trailed behind it by every measure, supporting the findings of Aggarwal et al. (2001) that smaller values of m for the Minkowski distance metric are preferred with high-dimensional or noisy data.

Weighting scheme

Because of the lack of conclusive evidence suggesting the universal superiority of distance-weighted over unweighted (i.e. equal-weighted) schemes, as well as the interrelationships of the choices of w, k, and p with the *k*NN algorithm, an unweighted scheme is used with Bamboo *k*NN. This is further justified by the use of L_2 boosting in Bamboo *k*NN and the aforementioned results of Park et al. (2009) showing that L_2 boosting is superior to the use of higher-order kernels that impart a distance-weighting scheme on the *k*NN algorithm.

L₂ boosting

Once again following the notation introduced earlier, first consider the case where the target set is also the reference set. The unboosted *k*NN prediction for the *i*th element of this target set, using an unweighted scheme ($w_i = \frac{1}{k}$, $W_i = 1$ in Eq. [1]) and $\{i = 1, \dots, n\}$, is:

$$\widetilde{y_i} = \frac{1}{k}\sum_{j=1}^{k} y_{ij},$$

with the additional criterion that unit i is excluded from the reference set when determining the *k*-nearest neighbors of target unit i, i.e. the ‘leave one out’ criterion. If the 1st order residual for the *i*th element is $e_i^1 = y_i - \widetilde{y_i}$, then the 1st order boosted *k*NN prediction is:

$$\tilde{y}_i^1 = \frac{1}{k}\sum_{j=1}^{k}(y_{ij} + e_{ij}^1),$$

which could also be referred to as the ‘twiced’ *k*NN prediction. Clearly, this procedure could be applied recursively and the *b*th order boosted *k*NN prediction would be:

$$\tilde{y}_i^b = \frac{1}{k}\sum_{j=1}^{k}(y_{ij} + e_{ij}^1 + \cdots + e_{ij}^b), \qquad [\,3\,]$$

where the *b*th order residual is $e_i^b = y_i - \tilde{y}_i^{b-1}$ and $\tilde{y}_i^0 = \widetilde{y_i}$, or simply the unboosted *k*NN prediction for unit i. The proposed stopping criteria for determining the order of recursion are minimization of the squared-error loss and estimated bias using the reference set. Once the order of recursion has been determined, as well as the corresponding set of residuals for all reference units, *k*NN predictions can be made for any target units not contained in the reference set using Eq. [3] by removing the ‘leave one out’ criterion.

Model-based optimization with embedded feature selection

Using a model-based mode of inference, the vector of response values Y is assumed to be generated by some stochastic superpopulation model, with the value of the *i*th population unit expressed as:

$$y_i = \mu_i + \varepsilon_i,$$

where ε_i is the random deviation of observation y_i about its mean μ_i. With Bamboo kNN, we further assume that the superpopulation model can be approximated using the boosted kNN estimator of μ_i defined in Eq. [3], given a sample from Y of size n (i.e. the reference set), a vector X of p auxiliary variables with known values for all N population units (i.e. the feature space), and the Minkowski distance metric d (i.e. $m = 1$ in Eq. [2]).

The MBO algorithm proposed here is a Markov chain Monte Carlo (MCMC) method for generating candidate solutions of Eq. [3], using a parameterized probabilistic model for selecting values of k and subsets of X from the solution space, that is adaptively modified using earlier candidate solutions. It is based upon an algorithm similar to the one introduced in RKNN-FS for classification for determining the measure of support provided by a set of feature variables, but is modified for regression and extended to include support provided by values of k.

The algorithm can be visualized using a set of three urns. Urn #1 contains $\{k_{max} = round(bn^{\frac{1}{2}})\}$ balls, labelled ‘1’ to k_{max}, representing possible values of k in Eq. [3], given sample size n and order of recursion b. Urn #2 contains p balls, labelled ‘1’ to p, representing the possible number of dimensions of subsets of X. Urn #3 also contains p balls, labelled ‘x_1 to ‘x_p’, representing individual feature variables in X. Initially, all balls within an urn have an equal probability of being selected. The process of fitting and evaluating a random boosted *k*NN model follows a number of steps:

1) A ball is randomly drawn from urn #1 to determine the value of k.

2) A ball is randomly drawn for urn #2 to determine the value of s, the number of dimensions in the feature space.

3) Then s balls are randomly drawn, without replacement, from urn #3 to determine the feature variables used to construct an s-dimensional subset of X.

4) The *k*-nearest neighbors of each unit in the reference set are determined, using the 'leave one out' criterion and the Manhattan distance.

5) Eq. [3] is used to make a *b*th order boosted *k*NN prediction for each unit in the reference set.

6) The observations and predictions of Y are used to calculate the coefficient of determination (R^2) of the model.

Steps 1-6 are repeated c times, generating a set of c candidate solutions. The value of c should be chosen to provide as large a sample of points in the solution space as possible while remaining practicable within the constraints of the available computing resources.

Next, the R^2 values of the candidate solutions are used to update the probability weights of all of the balls in each of the three urns. This is accomplished by calculating the mean R^2 value of all candidate solutions that were constructed using a given ball. If a ball was not used to construct any of the candidate solutions, it is assigned a probability weight of zero. Balls assigned a larger mean R^2 value will be given larger probability weight for the next step in the Markov chain, where a new set of c candidate solutions will be generated from the urns. Probability weights are normalized within each urn to sum to one.

The probability weights of the balls in each urn are updated, using each ball's assigned mean R^2 values from the current set of c candidate solutions, by a two-fold process. First, the R^2 values are converted to likelihoods by assuming that the squared residuals follow an exponential distribution. Given the exponential probability density function $f(x) = \lambda e^{-\lambda x}$, then setting $x = (y_i - \tilde{y}_i)^2$ and $\lambda = 1$ results in $f(x) = e^{-(y_i - \tilde{y}_i)^2}$, with the natural logarithm of the function being $-(y_i - \tilde{y}_i)^2$. Given $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n} {y_i}^2}$, then the log-likelihood can be expressed as $\ln(L) = -\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = (R^2 - 1)\sum_{i=1}^{n} {y_i}^2$, with $\sum_{i=1}^{n} {y_i}^2$ being a constant. Second, Bayes rule is applied to update the probability weights of the balls used to construct the current set of candidate solutions (i.e. the prior) by multiplication with likelihood e^{R^2-1} to get the probability weights for generating a new set candidate solutions at the next step in the Markov chain (i.e. the posterior). The chain should be run to sufficient length l to suggest minimization of the squared-error loss and estimated bias using the reference set. Separate chains should be run for each order of recursion b included in the solution space.

Empirical study of Bamboo *k*NN

Materials

A study was conducted using national forest inventory data and satellite imagery to test the performance of the proposed algorithm for SAE problems. The study area was the portion of the Western Superior Uplands Section (212K) of the hierarchical framework of ecological units developed by the USFS (Cleland, Avers, et al. 1997; Cleland, Freeouf, et al. 2007; McNab et al. 2007) that falls within the boundary of the state of Minnesota, an area of approximately 1.38 million hectares. This ecological section contains a mixture of land covers/uses that is predominantly forest and agriculture. The landscape consists of relatively level glacial drift plains of poorly drained loam soils. The forest types are primarily aspen-birch, maple-beech-birch, and spruce-fir.

The USFS Forest Inventory and Analysis (FIA) program conducts the national forest inventory of the US. For a comprehensive description of the FIA program, see Bechtold and Patterson (2005). Within the study area and during the five-year period of study timeframe (2009-2013), FIA collected a sample of 1138 cluster plots across all land covers/uses, including water. The sampling intensity is approximately one plot per 1200 hectares, or twice the FIA national baseline level. A core set of tree and forest attributes was measured at sample locations.

However, in order to test the proposed estimator over a range of spatial scales, it is necessary to have a population for which all units have known values for attribute of interest Y, in addition to auxiliary variables X. Because Y values for actual population units are known only at plot locations, the 2011 National Land Cover Database Tree Canopy Cover (TCC) dataset was used as a simulated population for the study. TCC was developed using a Random Forests model with Landsat-5 satellite imagery, topographic data, and manually interpreted sample points using National Agriculture Imagery Program aerial photography (Coulston et al., 2012; Homer at al., 2015). The dataset is provided as a raster image with a pixel resolution of 30 meters, and pixel values represent the predicted percentage of tree canopy cover at the pixel location. The predicted tree canopy cover value of the pixel at the location of the central sub-plot was assigned to each plot in the sample as the observed value. This amounted to a 0.0074% sample of the population.

The auxiliary variables X were derived from Landsat satellite imagery collected during the study timeframe. Wilson et al. (2018) demonstrated the utility of harmonic regression for feature extraction from dense time series of Landsat imagery for modeling a variety of tree and forest attributes from national forest inventory data. The estimated harmonic regression coefficients of a 3rd order Fourier series individually fitted to each of the tasseled cap transformation (TCT) components of brightness, greenness, and wetness for all Landsat-5 and Landsat-7 scenes collected between 2009 and 2013, after masking out pixels contaminated by clouds and snow, were used as the 21-dimensional feature space. For more information on the methods used for feature extraction, see Wilson et al. (2018). The X values of the auxiliary variables were extracted from the image pixels at the location of the central sub-plots in the FIA sample. This dataset, in conjunction with the TCC data described earlier, formed the reference set that was used to optimize the predictive kNN models. Because of the amount of computer processing required to do feature extraction, feature variables were also derived for only a purposive sample of 15 Public Land Survey System (PLSS) townships, each approximately 9300 hectares in size, representing the range of tree canopy cover conditions across the study area, to test the proposed algorithm's SAE performance.

Methods

The Bamboo *k*NN algorithm, described fully in the previous section, was implemented using the R statistical computing language (R Core Team, 2016), along with the packages 'snow' for parallel processing (Tierney et al., 2016), 'rgdal' for processing geospatial raster data (Bivand et al., 2016), and 'RANN.L1' for fast nearest neighbor search using Manhattan distance (Arya et al., 2015). The algorithm was initialized using eight chains, representing the set $\{b = 0, \dots, 7\}$ orders of recursion, with $c = 100$ candidate solutions drawn at each step of the Markov chain and each chain being of length $l = 3500$ steps. For each chain and at each step, the mean R^2 value and estimated bias of all current candidate solutions were computed.

At the end of each chain, a sample of 10 candidate solutions was drawn from the solution space using the updated probability weights for the balls in each of the three urns to produce predictions for each target unit (pixel) in the sample of 15 PLSS townships. Because TCC is expressed as a percentage, and is therefore a bounded variable, predicted values less than zero and greater than 100 were set to zero and 100 respectively. These predictions, along with the tally of reference units identified as nearest neighbors of target units within spatial domains of various sizes (i.e. small areas), were used to construct predictive intervals to test the SAE properties of Bamboo *k*NN over a range of spatial scales. Each township was subdivided into aliquot parts of $\frac{1}{4}$ (~2300 ha), $\frac{1}{9}$ (~1000 ha), $\frac{1}{36}$ (a section or ~260 ha), $\frac{1}{144}$ (a quarter-section or ~65 ha), and $\frac{1}{576}$ (a quarter-quarter-section or ~16 ha). Coverage tests were conducted at each level of subdivision, as well as for the individual reference units. Predictive intervals of 75, 90, and 95% were constructed from the candidate solutions for each of the domain sizes (i.e. aliquot parts) and each reference unit. Using these predictive intervals, the proportion of spatial domains, and reference units, whose predictive interval contained the observed value was calculated at each level of subdivision, along with the mean length of the predictive intervals.

Results and discussion

The mean R^2 values of the current candidate solutions at each step of the Markov chain for each order of recursion are shown in Figure 3.1. The short gap in the record for the unboosted model is due to a loss of data, though the trend remains clear even without it. Once beyond approximately step 600 in the chain, the unboosted model had the smallest mean R^2 values, by a substantial margin, of all eight models tested. Although there is some indication that the value would continue to increase in the short run with additional steps in the chain, the apparent rate of increase is modest. The effect of boosting can be seen clearly in the figure, with higher orders of recursion achieving successively larger mean R^2 values up to the 4th order model. From the 5th order onward, the boosts appear to plateau and fail to reach the level of the 4th order model.

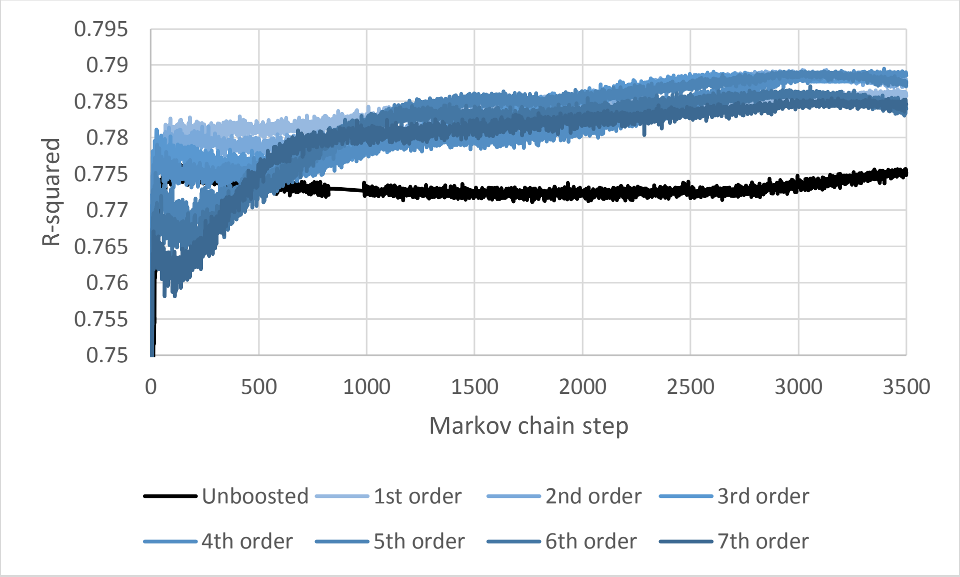

Figure 3.1. Mean R^2 value of the current candidate solutions at each step of the Markov chain for each order of recursion.

The estimated biases based on the current candidate solutions at each step of the Markov chain for each order of recursion are shown in Figure 3.2. There is a similar gap in the data record for the unboosted model, but again the trend is clear. The disparity in performance between the unboosted and boosted models is even more pronounced in this figure, with the unboosted model having the greatest estimated bias. Unlike the previous results, there is no indication that the estimated bias of the unboosted model would improve with additional steps in the chain. As with the mean R^2 values, the estimated bias of the model improves with increasing order of recursion up to the 4th order, with no clear evidence of improvement beyond that level. The results presented in both of these figures suggest that the benefit of estimating the bias of the model with higher orders of recursion is offset by increasing variance in the estimate, which is the inverse of the pattern seen with increasing values of k. In the case of the study population, the optimal value of b appears to be four.

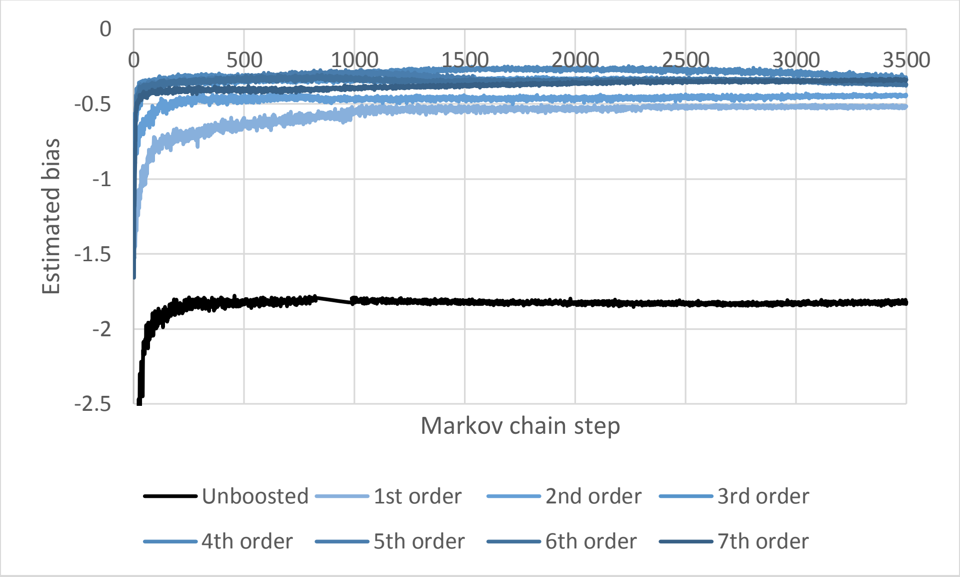

Figure 3.2. Estimated bias of the current candidate solutions at each step of the Markov chain for each order of recursion.

For the eight models tested, the probability weights of the balls in each urn after the final step in the corresponding Markov chain are depicted in Figures 3.3, 3.4, and 3.5. Figure 3.3 shows that as the order of recursion increases, the mean of the distribution of k increases from 14 for the unboosted model to 168.6067 for the 7th order boosted model. Figure 3.4 shows that as the order of recursion increases, the mean of the distribution of p decreases from 20.26709 for the unboosted model to 16.87689 for the 7th order boosted model. Although there is some variability across models in the probability weights of the individual feature variables shown in Figure 3.5, there is consistency in the ranking of those features with the largest and smallest probability weights, particularly among the boosted models. For all models, feature variables #1, 3, and 5 (i.e. mean annual brightness, mean annual wetness, and annual variability in greenness respectively) have the largest and #16 (i.e. 4-month variability in brightness) has the smallest probability weights.

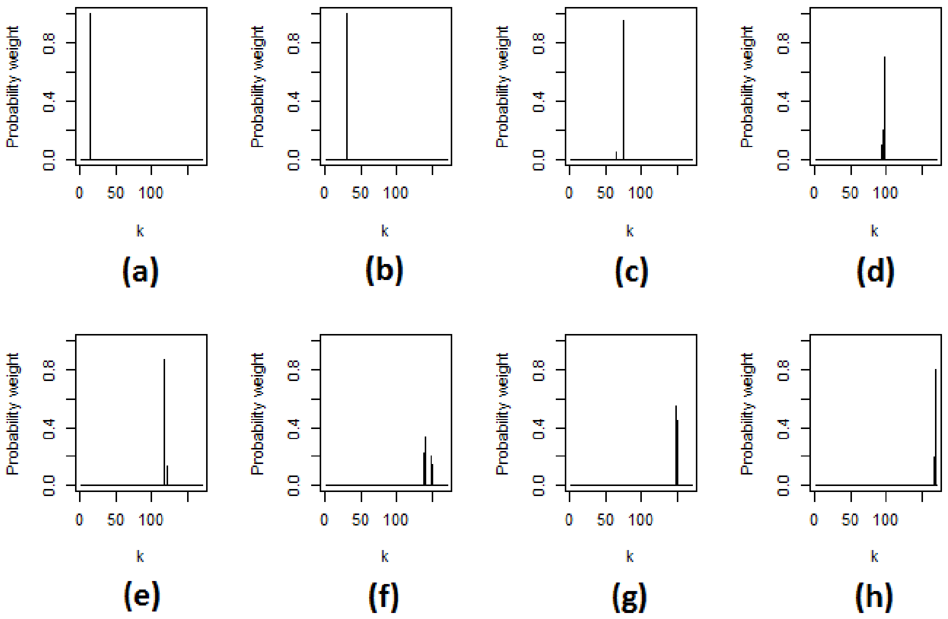

Figure 3.3. Probability weights of values of k (i.e. balls in urn #1) for the (a) unboosted and (b-h) 1st through 7th order boosted models.

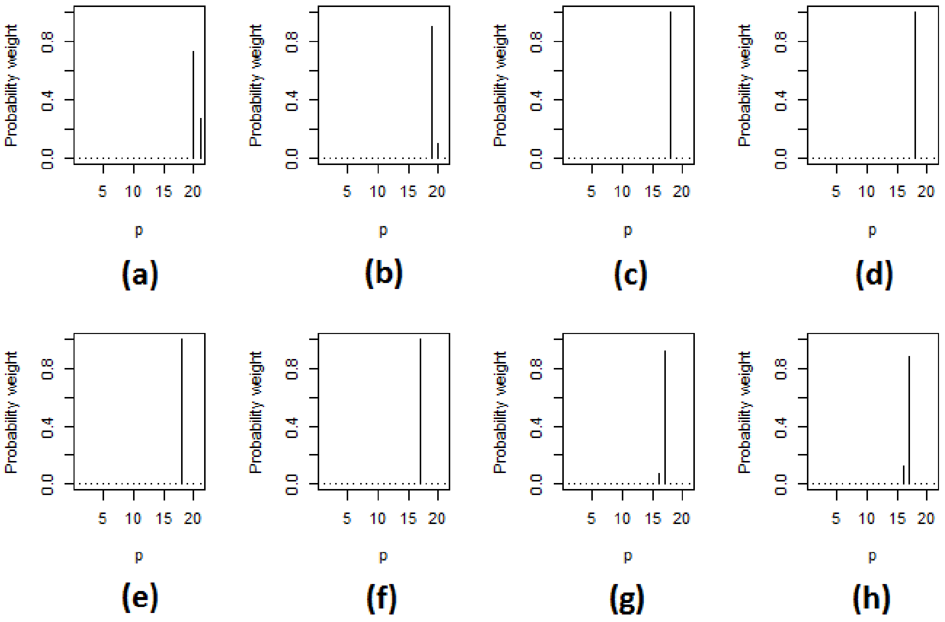

Figure 3.4. Probability weights of values of p (i.e. balls in urn #2) for the (a) unboosted and (b-h) 1st through 7th order boosted models.

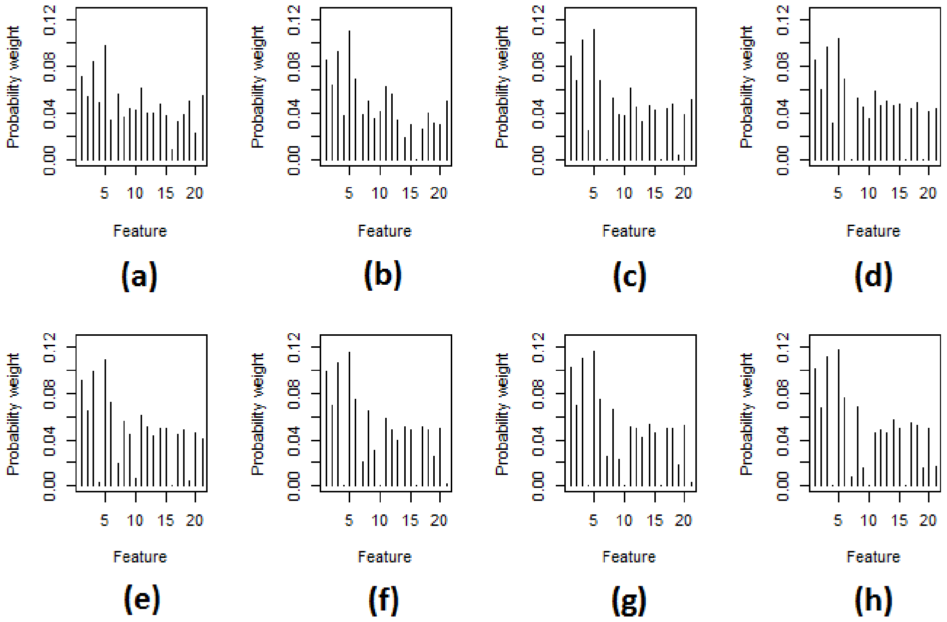

Figure 3.5. Probability weights of individual feature variables (i.e. balls in urn #3) for the (a) unboosted and (b-h) 1st through 7th order boosted models.

The coverage test results, by order of recursion, of the individual reference units are presented in Figure 3.6. The figure shows that the proportion of predictive intervals containing the observed value is smaller than the theoretical value for the unboosted model, but that it rapidly approaches the theoretical value by the 2nd order of recursion and hovers near that level as order increases. While there is some evidence supporting the previous suggestion of the benefit of higher order boosts for bias correction being offset by increasing variance, particularly for the 95% predictive intervals that appear to be slightly too conservative, it is inconclusive for the 75 and 90% predictive intervals. The corresponding lengths of the 75, 90, and 95% predictive intervals, by order of recursion, are presented in Figure 3.7. The pattern for each interval is roughly the same, with increasing interval length up to the 3rd order of recursion, a dip at the 4th order, then a slightly higher plateau at higher orders. Figures 3.6 and 3.7 also support the conclusion that the optimal value of b for the study population appears to be four.

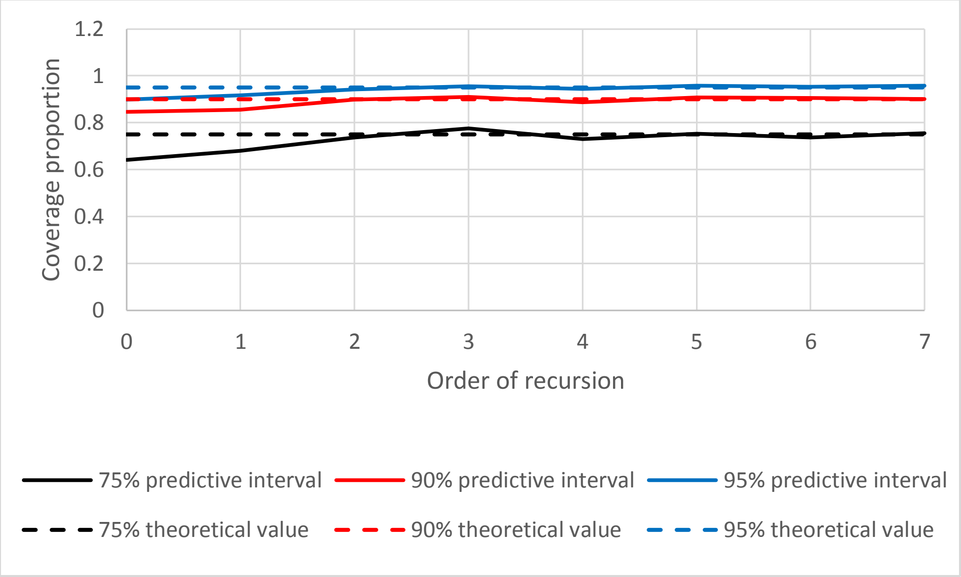

Figure 3.6. Coverage proportion of 75, 90, and 95% predictive intervals by order of recursion of the model. Theoretical values are shown as dashed lines.

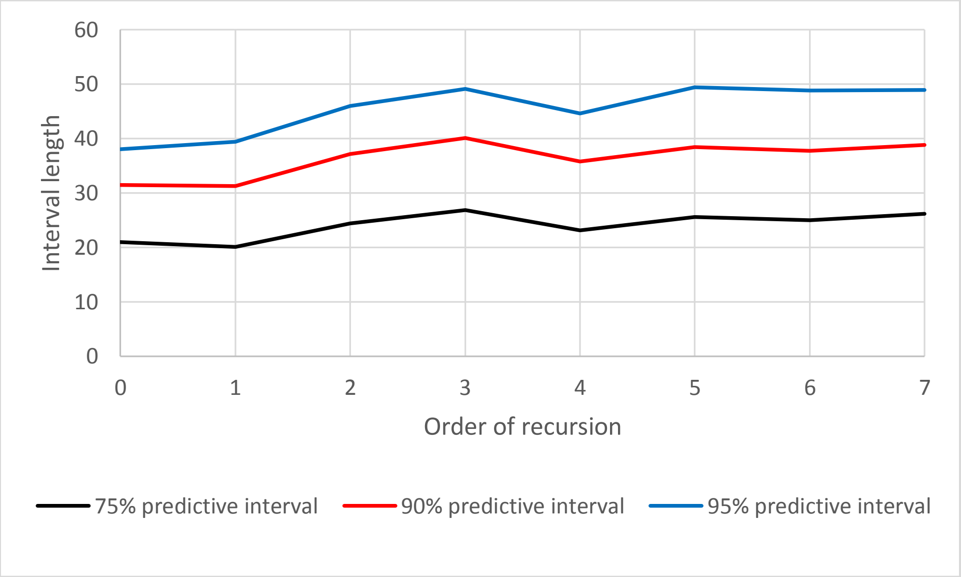

Figure 3.7. Length of 75, 90, and 95% predictive intervals by order of recursion of the model.

The coverage test results of spatial domains for the 75, 90, and 95% predictive intervals are presented in Figures 3.8, 3.9, and 3.10 respectively. In each of these figures, the proportion of predictive intervals containing the observed value is graphed against domain area as a series of connected line segments, with one series for each of the eight models tested. In each figure, the theoretical value of the predictive interval is depicted as a horizontal dashed line. It is noted that in the case of the largest domain sizes, townships, the sample size is only 15 and provides a poor estimate of the true coverage proportion and predictive interval length for the population of all townships in the study area. For all other domain sizes, the sample size is at least 60, providing more reliable estimates of the true coverage proportions and predictive interval lengths. The patterns are similar for each predictive interval, with the unboosted model having coverage proportions far below the theoretical value. With increasing order of recursion, again approximately up to the 4th order, the boosted models generally approach the theoretical value of the predictive interval. Beyond that level, the coverage proportions begin to exceed the theoretical value, particularly for the larger domain sizes.

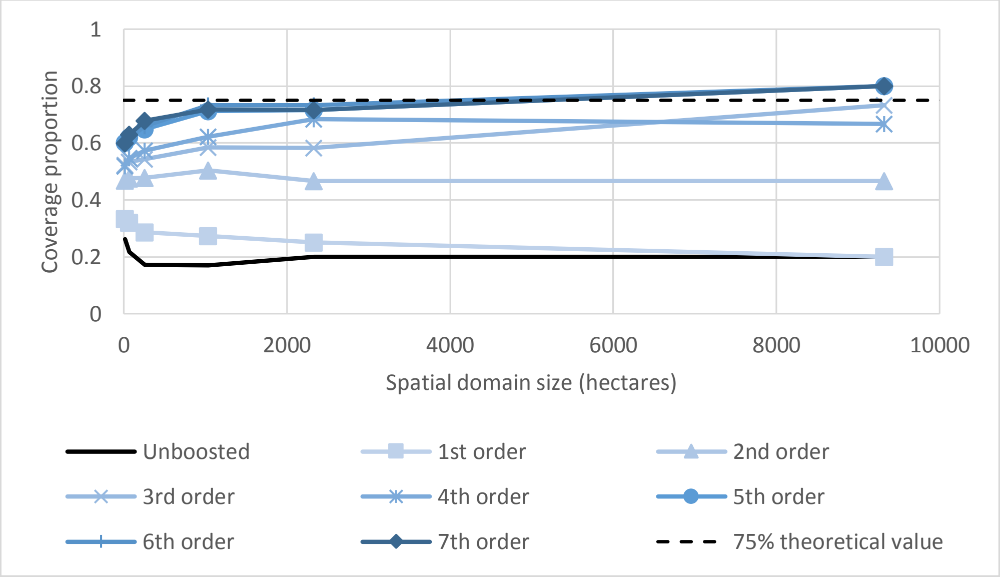

Figure 3.8. Coverage proportion of 75% predictive interval by spatial domain size and order of recursion of the model. Theoretical value is shown as dashed line.

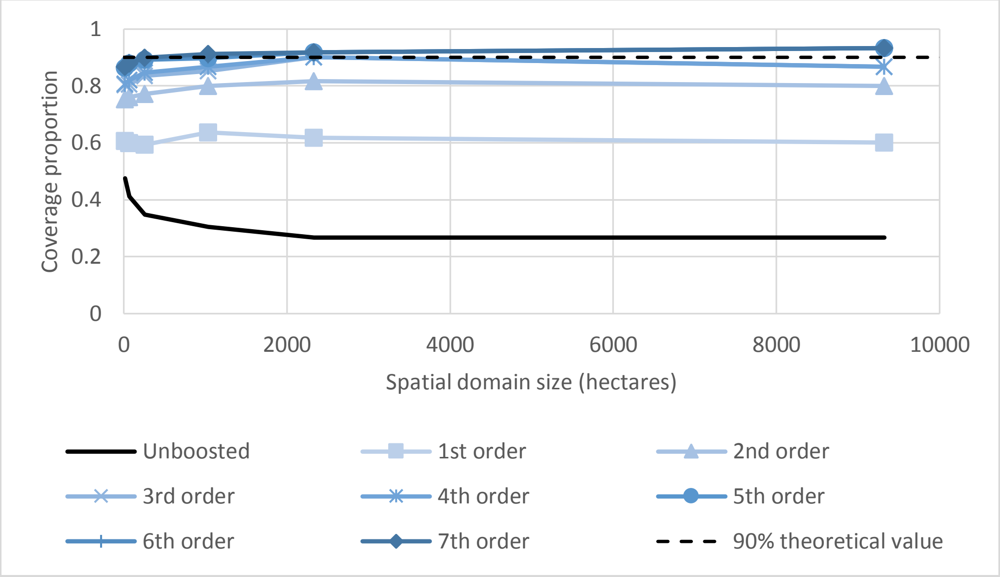

Figure 3.9. Coverage proportion of 90% predictive interval by spatial domain size and order of recursion of the model. Theoretical value is shown as dashed line.

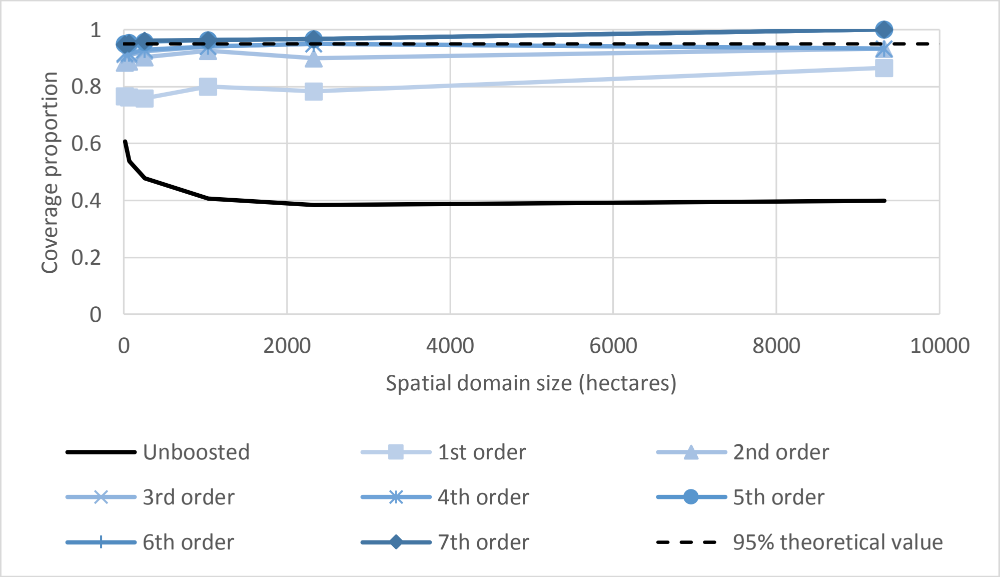

Figure 3.10. Coverage proportion of 95% predictive interval by spatial domain size and order of recursion of the model. Theoretical value is shown as dashed line.

It is noted that, while the pattern of the estimated coverage proportions approaching the theoretical value holds in general, the figures also suggest an effect due to domain size. With decreasing domain area, there is an exponential decrease in estimated coverage proportion. The effect is most pronounced for the 75% predictive intervals. One possible explanation for this is spatial autocorrelation of the residuals for spatial domains less than approximately 1200 ha in size, the area represented by a reference unit given the sampling intensity of the study, and corresponds to lag distances less than approximately 3.5 kilometers. However, without a sample that permits calculation of yet shorter lag distances, it is not possible to use an empirical spatial model to correct for this effect.

The corresponding lengths of the 75, 90, and 95% predictive intervals for spatial domains, by order of recursion, are presented in Figures 3.11, 3.12, and 3.13 respectively. In each of these figures, the length of the predictive interval is graphed against domain area as a series of connected line segments, again with one series for each of the eight models tested. The figures clearly show that estimated predictive interval lengths increase with both increasing theoretical value and order of recursion. For all theoretical values, lengths increase most rapidly between the 1st and 2nd order of recursion, with generally diminishing rates of increase beyond that level. The hypothesized spatial autocorrelation of the residuals, noted previously in the results of coverage proportions, appears to have the inverse effect on predictive interval lengths, i.e. exponentially increasing length with decreasing domain area below approximately 1200 ha.

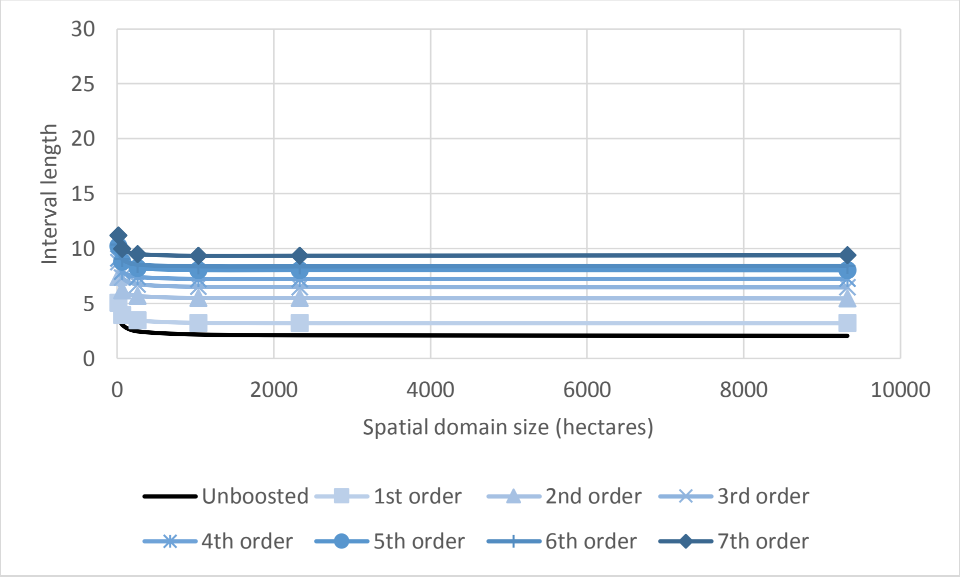

Figure 3.11. Length of 75% predictive interval by spatial domain size and order of recursion of the model.

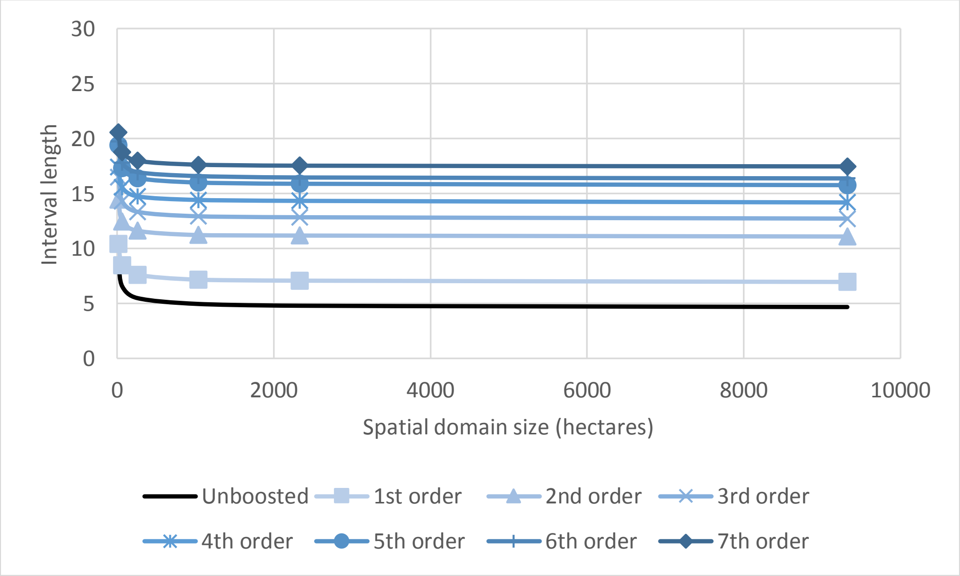

Figure 3.12. Length of 90% predictive interval by spatial domain size and order of recursion of the model.

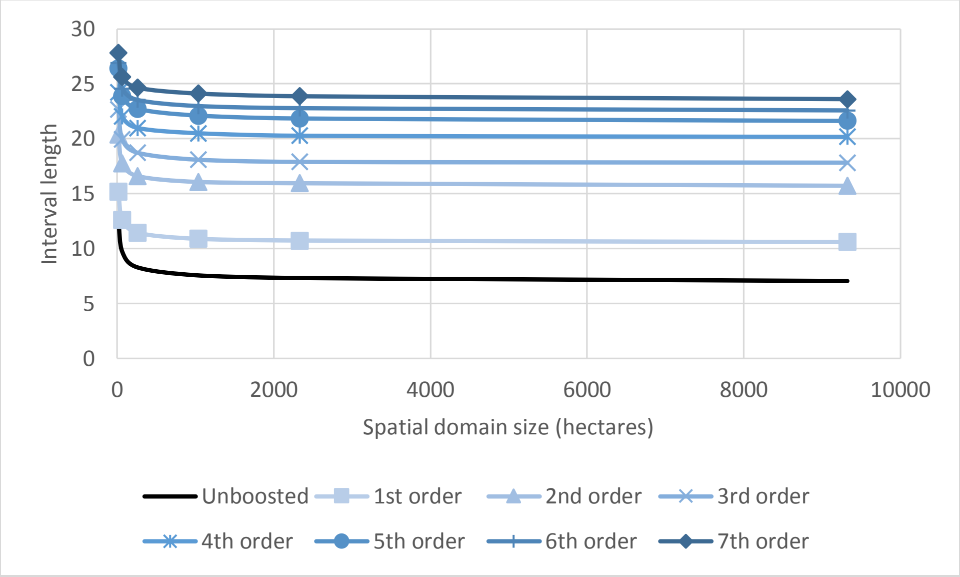

Figure 3.13. Length of 95% predictive interval by spatial domain size and order of recursion of the model.

Taken together the coverage proportion and predictive interval length results suggest that, conditional upon the sample, approximately valid predictive intervals can be produced for spatial domains as small as the area represented by a reference unit, i.e. 1200 ha in the current study. By choosing an appropriate order of recursion for the boosted model, i.e. 4th order in the current study, bias can be estimated adequately without incurring a penalty of extraneous variance in the estimate. With decreasing domain area less than this threshold, it is postulated that spatial autocorrelation of the residuals results in estimated coverage proportions that are smaller than the theoretical values. Because this occurs despite increasing estimated predictive interval length, it is further postulated that these errors are dominated by residual bias that cannot be estimated from the sample alone due to boundary effects (Hastie and Loader, 1993).

Although correcting for boundary bias is beyond the scope of this study, a metric for use with Bamboo *k*NN is proposed for estimating proximity to the boundary in order to identify spatial domains where estimated predictive intervals would lead to invalid inferences. The proposed metric, neighborliness, is based on the fact that the reference sample units define a convex hull in feature space, designated the reference hull, and that the boundary of feature space is assumed to fall on or outside of it. Using the 4th order boosted model and the reference units as the target set, the nearest neighbors were identified for the set of $\{c_i, i = 1, \ldots, 10\}$ candidate solutions, resulting in a set of k_i x n matrices of nearest neighbors. The neighborliness of each reference unit was calculated by tallying the total number of occurrences of the reference unit in all 10 matrices and dividing this total by the mean tally of all reference units. Units with smaller neighborliness values are expected to be located in feature space nearer to the reference hull boundary, while units with larger neighborliness values are expected to be located within its interior. The neighborliness values assigned to the reference units were then used to estimate the neighborliness of all pixels in the target set, with the same 4th order model used to make predictions of TCC percentage.

Figure 3.14 depicts the observed TCC percentage, predicted TCC percentage, and estimated neighborliness, respectively, of all pixels in a representative sample township from the study area, T40N R20W. The color ramp applied to the observed and predicted TCC percentage pixel values has a range from white (0) to black (100). Forested areas in the township are shaded gray or black, while non-forested areas are shaded white. The color ramp applied to the estimated neighborliness pixel values has a range from red (0.5) to white (1) to blue (1.5). Most forested areas in the township are tinted dark blue, suggesting that they are well interior of the reference hull boundary in feature space. Most non-forest areas are tinted light blue, suggesting that they are also interior of the reference hull, though slightly closer to the boundary and the mean neighborliness value of the reference set. Pixels that appear tinted red are closest to, or in some cases likely beyond, the reference hull boundary. Those pixels that have the darkest red tint correspond to water bodies, which were problematic when using harmonic regression for feature extraction due to the numerous Landsat sensor observations removed by the snow filter (Wilson et al., 2018). Pixels that are tinted light red are land covers/uses that occur infrequently in the study area, such as rotating crops, ephemeral wetland vegetation, and harvested forests.

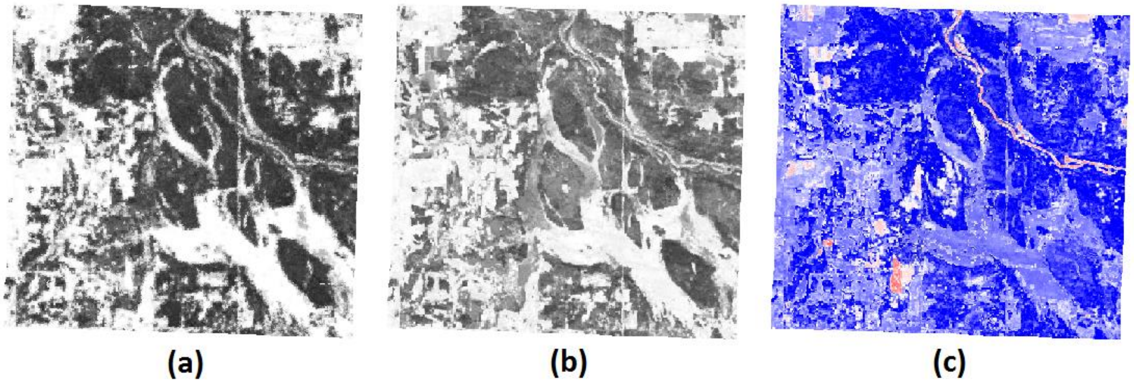

Figure 3.14. Images for township T40N R20W of (a) observed and (b) predicted tree canopy cover percentages, along with (c) estimated neighborliness. The color ramp applied to the pixel values in (a) and (b) has a range from white (0) to black (100), while the one used in (c) has a range from red (0.5) to white (1) to blue (1.5).

Conclusions

There are a few conclusions to be drawn from this study. First, the proposed Bamboo *k*NN algorithm provides a unified framework for global optimization of the *k*NN model while simultaneously selecting feature variables and correcting for smoother bias. This is accomplished through its use of a stochastic, model-based optimization approach with embedded feature selection, as well as the recursive fitting of residuals, or L_2 boosting. Second, the empirical study using features extracted from time series of satellite imagery and a simulated population of tree canopy cover demonstrated that the Bamboo *k*NN estimator produced valid predictive intervals, following the suggested guidelines for determining the length of the Markov chain and the level of recursion, for spatial domains approaching the area represented by a reference unit in the national forest inventory sample. Third, the study showed that the unboosted *k*NN model produced predictive intervals that were far too narrow and had corresponding coverage proportions smaller than the theoretical value. It is noted, however, that alternative weighting schemes to the unweighted one employed by Bamboo *k*NN were not tested with the unboosted model, and that these might mitigate the problem to some degree. Finally, the proposed neighborliness metric suggests a method for identifying spatial domains within the population where valid inferences cannot be made without supplementation of the reference sample.

Bibliography

Abràmoff, Michael D., Paulo J. Magalhães, and Sunanda J. Ram. 2004. “Image Processing with ImageJ.” *Biophotonics International* 11 (7). Laurin Publishing: 36–42.

Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. 2001. “On the Surprising Behavior of Distance Metrics in High Dimensional Space.” In *International Conference on Database Theory*, 420–34. Springer. doi:10.1007/3-540-44503-X_27.

Arya, Sunil, David Mount, Samuel E Kemp, Gregory Jefferis, and Kirill Müller. 2015. “RANN.L1: Fast Nearest Neighbour Search (Wraps ANN Library) Using L1 Metric.” https://cran.r-project.org/package=RANN.L1.

Badhwar, G.D., J.G. Carnes, and W.W. Austin. 1982. “Use of Landsat-Derived Temporal Profiles for Corn-Soybean Feature Extraction and Classification.” *Remote Sensing of Environment* 12 (1): 57–79. doi:10.1016/0034-4257(82)90007-4.

Baffetta, Federica, Lorenzo Fattorini, Sara Franceschi, and Piermaria Corona. 2009. “Design-Based Approach to K-Nearest Neighbours Technique for Coupling Field and Remotely Sensed Data in Forest Surveys.” *Remote Sensing of Environment* 113 (3): 463–75. doi:10.1016/j.rse.2008.06.014.

Baig, Muhammad Hasan Ali, Lifu Zhang, Tong Shuai, and Qingxi Tong. 2014. “Derivation of a Tasselled Cap Transformation Based on Landsat 8 at-Satellite Reflectance.” *Remote Sensing Letters* 5 (5). Taylor & Francis: 423–31. doi:10.1080/2150704X.2014.915434.

Bailey, T., and A. K. Jain. 1978. “A Note on Distance-Weighted K-Nearest Neighbor Rules.” *IEEE Transactions on Systems, Man, and Cybernetics* 8 (4): 311–13. doi:10.1109/TSMC.1978.4309958.

Banerjee, Sudipto, and Andrew Finley. 2007. “Bayesian Multi-Resolution Modeling for Spatially Replicated Data Sets with Application to Forest Biomass Data.” *Journal of Statistical Planning and Inference* 137 (10): 3193–3205. doi:http://dx.doi.org/10.1016/j.jspi.2006.05.024.

Bannari, A., D. Morin, F. Bonn, and A. R. Huete. 1995. “A Review of Vegetation Indices.” *Remote Sensing Reviews* 13 (1–2). Taylor & Francis: 95–120. doi:10.1080/02757259509532298.

Bartz-Beielstein, Thomas, and Martin Zaefferer. 2017. “Model-Based Methods for Continuous and Discrete Global Optimization.” *Applied Soft Computing* 55: 154–67. doi:https://doi.org/10.1016/j.asoc.2017.01.039.

Batista, Gustavo E. A. P. A., and Diego Furtado Silva. 2009. "How K-Nearest Neighbor Parameters Affect Its Performance." In *Argentine Symposium on Artificial Intelligence*, 95–106.

Bayr, Caroline, Heinz Gallaun, Ulrike Kleb, Birgit Kornberger, Martin Steinegger, and Martin Winter. 2016. "Satellite-Based Forest Monitoring: Spatial and Temporal Forecast of Growing Index and Short-Wave Infrared Band." *Geospatial Health; Vol 11, No 1 (2016): Valencia Issue,* April. http://www.geospatialhealth.net/index.php/gh/article/view/310.

Bechtold, William A., and Paul L. Patterson. 2005. "The Enhanced Forest Inventory and Analysis Program-National Sampling Design and Estimation Procedures."

Bellman, Richard. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Bengio, Yoshua, Olivier Delalleau, and Nicolas Le Roux. 2005. "The Curse of Dimensionality for Local Kernel Machines." *Technical Report No. 1258,* Département d'informatique et recherche opérationnelle, Université de Montréal.

Bivand, Roger, Tim Keitt, and Barry Rowlingson. 2016. "Rgdal: Bindings for the Geospatial Data Abstraction Library." https://cran.r-project.org/package=rgdal.

Blum, Avrim L., and Pat Langley. 1997. "Selection of Relevant Features and Examples in Machine Learning." *Artificial Intelligence* 97 (1–2): 245–71. doi:10.1016/S0004-3702(97)00063-5.

Bolstad, William M. 2007. *Introduction to Bayesian Statistics*. 2nd ed. Hoboken, N.J.: Hoboken, N.J. : John Wiley.

Bradley, Bethany A., Robert W. Jacob, John F. Hermance, and John F. Mustard. 2007. "A Curve Fitting Procedure to Derive Inter-Annual Phenologies from Time Series of Noisy Satellite NDVI Data." *Remote Sensing of Environment* 106 (2): 137–45. doi:10.1016/j.rse.2006.08.002.

Breidenbach, Johannes, and Rasmus Astrup. 2012. "Small Area Estimation of Forest Attributes in the Norwegian National Forest Inventory." *European Journal of Forest Research* 131 (4): 1255–67. doi:10.1007/s10342-012-0596-7.

Breidenbach, Johannes, Arne Nothdurft, and Gerald Kändler. 2010. "Comparison of Nearest Neighbour Approaches for Small Area Estimation of Tree Species-Specific Forest Inventory Attributes in Central Europe Using Airborne Laser Scanner Data." *European Journal of Forest Research* 129 (5): 833–46. doi:10.1007/s10342-010-0384-1.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1). Kluwer Academic Publishers: 5–32. doi:10.1023/A:1010933404324.

Brooks, Evan B., Valerie A. Thomas, Randolph H. Wynne, and John W. Coulston. 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). IEEE: 3340–53.

Byrne, G. F., P. F. Crapper, and K. K. Mayo. 1980. "Monitoring Land-Cover Change by Principal Component Analysis of Multitemporal Landsat Data." *Remote Sensing of Environment*. doi:10.1016/0034-4257(80)90021-8.

Chen, Jin, Per Jönsson, Masayuki Tamura, Zhihui Gu, Bunkei Matsushita, and Lars Eklundh. 2004. "A Simple Method for Reconstructing a High-Quality NDVI Time-Series Data Set Based on the Savitzky-Golay Filter." *Remote Sensing of Environment* 91 (3–4): 332–44. doi:10.1016/j.rse.2004.03.014.

Chen, Jin, Xiaolin Zhu, James E. Vogelmann, Feng Gao, and Suming Jin. 2011. "A Simple and Effective Method for Filling Gaps in Landsat ETM+ SLC-off Images." *Remote Sensing of Environment* 115 (4): 1053–64. doi:10.1016/j.rse.2010.12.010.

Chomboon, Kittipong, Pasapitch Chujai, Pongsakorn Teerarassamee, Kittisak Kerdprasop, and Nittaya Kerdprasop. 2015. "An Empirical Study of Distance Metrics for K-Nearest Neighbor Algorithm." In *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*, 280–85. doi:10.12792/iciae2015.051.

Cleland, David T., Jerry A. Freeouf, James E. Keys, Greg J. Nowacki, Constance A. Carpenter, and W. Henry McNab. 2007. "Ecological Subregions: Sections and Subsections for the Conterminous United States."

Cleland, David T., Peter E. Avers, W. Henry McNab, Mark E. Jensen, Robert G. Bailey, Thomas King, and Walter E. Russell. 1997. "National Hierarchical Framework of Ecological Units." *Ecosystem Management Applications for Sustainable Forest and Wildlife Resources*. Yale University Press: New Haven, CT, USA, 181–200.

Cochran, William G. 1977. *Sampling Techniques*. 3rd ed. Hoboken, N.J.: Hoboken, N.J. : John Wiley.

Cohen, Warren B., Maria Fiorella, John Gray, Eileen Helmer, and Karen Anderson. 1998. "An Efficient and Accurate Method for Mapping Forest Clearcuts in the Pacific Northwest Using Landsat Imagery." *Photogrammetric Engineering and Remote Sensing* 64 (4). Citeseer: 293–99.

Collet, Pierre, and Jean-Philippe Rennard. 2008. "Stochastic Optimization Algorithms." *Intelligent Information Technologies*, 1121–37. doi:10.4018/978-1-59140-984-7.

Cornillon, Pierre-André, Nicolas Hengartner, and Eric Matzner-Løber. 2008. "Recursive Bias Estimation and L2 Boosting." *arXiv Preprint arXiv:0801.4629*, 1–33.

Coulston, John W., Gretchen G. Moisen, Barry T. Wilson, Mark V. Finco, Warren B. Cohen, and C. Kenneth Brewer. 2012. "Modeling Percent Tree Canopy Cover: A Pilot Study."

Photogrammetric Engineering & Remote Sensing 78 (7): 715–27. http://www.treesearch.fs.fed.us/pubs/40860.

Cover, Thomas M. 1968. “Estimation by the Nearest Neighbor Rule.” *IEEE Transactions on Information Theory* 14 (1). IEEE: 50–55. doi:10.1109/TIT.1968.1054098.

Crist, Eric P., and Richard C. Cicone. 1984a. “A Physically-Based Transformation of Thematic Mapper Data---The TM Tasseled Cap.” *IEEE Transactions on Geoscience and Remote Sensing*, no. 3. IEEE: 256–63.

Crist, Eric P., and Richard C. Cicone. 1984b. “Comparisons of the Dimensionality and Features of Simulated Landsat-4 MSS and TM Data.” *Remote Sensing of Environment* 14 (1–3): 235–46. doi:10.1016/0034-4257(84)90018-X.

Curran, Paul. 1980. “Multispectral Remote Sensing of Vegetation Amount.” *Progress in Physical Geography* 4 (3). SAGE Publications: 315–41. doi:10.1177/030913338000400301.

Datta, Gauri Sankar, Bannmo Day, and Ishwar Basawa. 1999. “Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation.” *Journal of Statistical Planning and Inference* 75 (2): 269–79. doi:http://dx.doi.org/10.1016/S0378-3758(98)00147-5.

Dawid, A. P. 2006. “Inference, Statistical—I.” In *Encyclopedia of Statistical Sciences*. 2nd ed. (eds. S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic and N. L. Johnson). John Wiley. doi:10.1002/0471667196.ess1236.pub2.

DeFries, Ruth, Matthew Hansen, and John Townshend. 1995. “Global Discrimination of Land Cover Types from Metrics Derived from AVHRR Pathfinder Data.” *Remote Sensing of Environment* 54 (3): 209–22. doi:10.1016/0034-4257(95)00142-5.

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 87 (418). American Statistical Association: 376–82. doi:10.2307/2290268.

Devroye, Luc P. 1978. “The Uniform Convergence of Nearest Neighbor Regression Function Estimators and Their Application in Optimization.” *IEEE Transactions on Information Theory* 24 (2). IEEE: 142–51. doi:10.1109/TIT.1978.1055865.

Di Marzio, M., and C. C. Taylor. 2004. “Boosting Kernel Density Estimates: A Bias Reduction Technique?” *Biometrika* 91 (1). Oxford University Press, Biometrika Trust: 226–33. http://www.jstor.org/stable/20441091.

Dudani, Sahibsingh A. 1976. “The Distance-Weighted K-Nearest-Neighbor Rule.” *IEEE Transactions on Systems, Man and Cybernetics* SMC-6 (4): 325–27. doi:10.1109/TSMC.1976.5408784.

Dymond, Caren C., David J. Mladenoff, and Volker C. Radeloff. 2002. "Phenological Differences in Tasseled Cap Indices Improve Deciduous Forest Classification." *Remote Sensing of Environment* 80 (3): 460–72. doi:10.1016/S0034-4257(01)00324-8.

Enas, Gregory G., and Sung C. Choi. 1986. "Choice of the Smoothing Parameter and Efficiency of K-Nearest Neighbor Classification." *Computers and Mathematics with Applications* 12 (2 PART A): 235–44. doi:10.1016/0898-1221(86)90076-3.

Eskelson, Bianca N. I., Hailemariam Temesgen, Valerie Lemay, Tara M. Barrett, Nicholas L. Crookston, and Andrew T. Hudak. 2009. "The Roles of Nearest Neighbor Methods in Imputing Missing Data in Forest Inventory and Monitoring Databases." *Scandinavian Journal of Forest Research* 24 (3). Taylor & Francis: 235–46. doi:10.1080/02827580902870490.

Finley, Andrew O., Sudipto Banerjee, and David W. MacFarlane. 2011. "A Hierarchical Model for Quantifying Forest Variables over Large Heterogeneous Landscapes with Uncertain Forest Areas." *Journal of the American Statistical Association* 106 (493). Taylor & Francis: 31–48.

Finley, Andrew O., Sudipto Banerjee, and Ronald E. McRoberts. 2009. "Hierarchical Spatial Models for Predicting Tree Species Assemblages across Large Domains." *The Annals of Applied Statistics* 3 (3). NIH Public Access: 1052.

Finley, Andrew O., Sudipto Banerjee, Bruce D. Cook, and John B. Bradford. 2013. "Hierarchical Bayesian Spatial Models for Predicting Multiple Forest Variables Using Waveform LiDAR, Hyperspectral Imagery, and Large Inventory Datasets." *International Journal of Applied Earth Observation and Geoinformation* 22 (0): 147–60. doi:http://dx.doi.org/10.1016/j.jag.2012.04.007.

Fix, Evelyn, and J. L. Hodges. 1951. "Discriminatory Analysis - Nonparametric Discrimination Consistency Properties." *USAF School of Aviation Medicine, Randolph Field, Texas*. DTIC Document. doi:10.2307/1403797.

Franco-Lopez, Hector, Alan R. Ek, and Marvin E. Bauer. 2001. "Estimation and Mapping of Forest Stand Density, Volume, and Cover Type Using the K-Nearest Neighbors Method." *Remote Sensing of Environment* 77 (3). Elsevier: 251–74.

Friedman, Jerome H. 1997. "On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality." *Data Mining and Knowledge Discovery* 1 (1). Springer: 55–77. doi:10.1023/A:1009778005914.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5). The Institute of Mathematical Statistics: 1189–1232. doi:10.1214/aos/1013203451.

Fukunaga, Keinosuke, and Larry D. Hostetler. 1973. "Optimization of K-Nearest-Neighbor Density Estimates." *IEEE Transactions on Information Theory* 19 (3): 320–26. doi:10.1109/TIT.1973.1055003.

Geerken, Roland A. 2009. "An Algorithm to Classify and Monitor Seasonal Variations in Vegetation Phenologies and Their Inter-Annual Change." *ISPRS Journal of Photogrammetry and Remote Sensing* 64 (4): 422–31. doi:https://doi.org/10.1016/j.isprsjprs.2009.03.001.

Gelman, Andrew, John B. Carlin, Hal Steven Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. Texts in Statistical Science. Boca Raton: CRC Press.

Ghosh, M., and J. N. K. Rao. 1994. "Small Area Estimation: An Appraisal." *Statistical Science* 9 (1). Institute of Mathematical Statistics: 55–76. doi:10.2307/2246284.

Goerndt, Michael E., Vicente J. Monleon, and Hailemariam Temesgen. 2011. "A Comparison of Small-Area Estimation Techniques to Estimate Selected Stand Attributes Using LiDAR-Derived Auxiliary Variables." *Canadian Journal of Forest Research* 41 (6). NRC Research Press: 1189–1201. doi:10.1139/x11-033.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2016. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment*, July 6. doi:10.1016/j.rse.2017.06.031.

Goward, Samuel N., Compton J. Tucker, and Dennis G. Dye. 1985. "North American Vegetation Patterns Observed with the NOAA-7 Advanced Very High Resolution Radiometer." *Vegetatio* 64 (1): 3–14. doi:10.1007/BF00033449.

Gregoire, T. G. 1998. "Design-Based and Model-Based Inference in Survey Sampling: Appreciating the Difference." *Canadian Journal of Forest Research* 28 (10). NRC Research Press: 1429–47. doi:10.1139/x98-166.

Hall, Dorothy K., George A. Riggs, and Vincent V. Salomonson. 1995. "Development of Methods for Mapping Global Snow Cover Using Moderate Resolution Imaging Spectroradiometer Data." *Remote Sensing of Environment* 54 (2): 127–40. doi:10.1016/0034-4257(95)00137-P.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, et al. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." *Science* 342 (6160): 850 LP-853. http://science.sciencemag.org/content/342/6160/850.abstract.

Hastie, Trevor, and Clive Loader. 1993. "Local Regression: Automatic Kernel Carpentry." *Statistical Science* 8 (2). Institute of Mathematical Statistics: 120–29. http://www.jstor.org/stable/2246148.

Healey, Sean P., Warren B. Cohen, Yang Zhiqiang, and Olga N. Krankina. 2005. “Comparison of Tasseled Cap-Based Landsat Data Structures for Use in Forest Disturbance Detection.” *Remote Sensing of Environment* 97 (3): 301–10. doi:10.1016/j.rse.2005.05.009.

Hechenbichler, Klaus, and Klaus Schliep. 2004. “Weighted k-nearest-neighbor techniques and ordinal classification.” *Discussion paper // Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München No. 399*. http://nbn-resolving.de/urn:nbn:de:bvb:19-epub-1769-9.

Helmer, E. H., S. Brown, and W. B. Cohen. 2000. “Mapping Montane Tropical Forest Successional Stage and Land Use with Multi-Date Landsat Imagery.” *International Journal of Remote Sensing* 21 (11). Taylor & Francis: 2163–83. doi:10.1080/01431160050029495.

Henderson, C. R. 1975. “Best Linear Unbiased Estimation and Prediction under a Selection Model.” *Biometrics* 31 (2). [Wiley, International Biometric Society]: 423–47. doi:10.2307/2529430.

Hermance, J. F. 2007. “Stabilizing High-Order, Non-Classical Harmonic Analysis of NDVI Data for Average Annual Models by Damping Model Roughness.” *Internatiional Journal of Remote Sensing* 28 (12). Bristol, PA, USA: Taylor & Francis, Inc.: 2801–19. doi:10.1080/01431160600967128.

Hird, Jennifer N., and Gregory J. McDermid. 2009. “Noise Reduction of NDVI Time Series: An Empirical Comparison of Selected Techniques.” *Remote Sensing of Environment* 113 (1): 248–58. doi:10.1016/j.rse.2008.09.003.

Homer, Collin G., Jon Dewitz, Limin Yang, Suming Jin, Patrick Danielson, George Xian, John Coulston, Nathaniel Herold, James Wickham, and Kevin Megown. 2015. “Completion of the 2011 National Land Cover Database for the Conterminous United States-Representing a Decade of Land Cover Change Information.” *Photogrammetric Engineering and Remote Sensing* 81 (5): 345–54.

Horvitz, Daniel G., and Donovan J. Thompson. 1952. “A Generalization of Sampling without Replacement from a Finite Universe.” *Journal of the American Statistical Association* 47 (260). Taylor & Francis Group: 663–85.

Hu, Li-Yu, Min-Wei Huang, Shih-Wen Ke, and Chih-Fong Tsai. 2016. “The Distance Function Effect on K-Nearest Neighbor Classification for Medical Datasets.” *SpringerPlus* 5 (1). Cham: Springer International Publishing: 1304. doi:10.1186/s40064-016-2941-7.

Huang, C., B. Wylie, L. Yang, C. Homer, and G. Zylstra. 2002. “Derivation of a Tasselled Cap Transformation Based on Landsat 7 at-Satellite Reflectance.” *International Journal of Remote Sensing* 23 (8). Taylor & Francis: 1741–48. doi:10.1080/01431160110106113.

James, W., and Charles Stein. 1961. “Estimation with Quadratic Loss.” In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 361–79. Fourth Berkeley Symposium on

Mathematical Statistics and Probability. Berkeley, Calif.: University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512173.

Jin, Suming, and Steven A. Sader. 2005. “Comparison of Time Series Tasseled Cap Wetness and the Normalized Difference Moisture Index in Detecting Forest Disturbances.” *Remote Sensing of Environment* 94 (3): 364–72. doi:10.1016/j.rse.2004.10.012.

Jönsson, Per, and Lars Eklundh. 2002. “Seasonality Extraction by Function Fitting to Time-Series of Satellite Sensor Data.” *IEEE Transactions on Geoscience and Remote Sensing* 40 (8): 1824–32. doi:10.1109/TGRS.2002.802519.

Ju, Junchang, and David P. Roy. 2008. “The Availability of Cloud-Free Landsat ETM+ Data over the Conterminous United States and Globally.” *Remote Sensing of Environment* 112 (3): 1196–1211. doi:10.1016/j.rse.2007.08.011.

Kangas, Annika, and Matti Maltamo. 2006. *Forest Inventory: Methodology and Applications*. Vol. 10. Springer Science & Business Media.

Karlson, Martin, Madelene Ostwald, Heather Reese, Josias Sanou, Boalidioa Tankoano, and Eskil Mattsson. 2015. “Mapping Tree Canopy Cover and Aboveground Biomass in Sudano-Sahelian Woodlands Using Landsat 8 and Random Forest.” *Remote Sensing* 7 (8). Multidisciplinary Digital Publishing Institute: 10017–41.

Katila, Matti. 2006. “Empirical Errors of Small Area Estimates from the Multisource National Forest Inventory in Eastern Finland.” *Silva Fennica* 40 (4). The Finnish Society of Forest Science: 729.

Katila, Matti, and Erkki Tomppo. 2001. “Selecting Estimation Parameters for the Finnish Multisource National Forest Inventory.” *Remote Sensing of Environment* 76 (1): 16–32. doi:10.1016/S0034-4257(00)00188-7.

Kauth, Richard J., and G. S. Thomas. 1976. “The Tasselled Cap--a Graphic Description of the Spectral-Temporal Development of Agricultural Crops as Seen by Landsat.” In *LARS Symposia*, 159.

Kohavi, Ron, and George H. John. 1997. “Wrappers for Feature Subset Selection.” *Artificial Intelligence* 97 (1–2). Elsevier: 273–324. doi:10.1016/S0004-3702(97)00043-X.

Lazar, Radu, Glen Meeden, and David Nelson. 2008. “A Noninformative Bayesian Approach to Finite Population Sampling Using Auxiliary Variables.” *Survey Methodology* 34 (1): 51.

LeMay, Valerie, and Hailemariam Temesgen. 2005. “Comparison of Nearest Neighbor Methods for Estimating Basal Area and Stems per Hectare Using Aerial Auxiliary Variables.” *Forest Science* 51 (2). Society of American Foresters: 11. http://www.ingentaconnect.com/content/saf/fs/2005/00000051/00000002/art00003.

Li, Shengqiao, E. James Harner, and Donald A. Adjeroh. 2011. “Random KNN Feature Selection - a Fast and Stable Alternative to Random Forests.” *BMC Bioinformatics* 12 (1). BioMed Central: 450. doi:10.1186/1471-2105-12-450.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22.

Little, Roderick J. 2004. “To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling.” *Journal of the American Statistical Association* 99 (466). American Statistical Association: 546–56. doi:10.2307/27590409.

Loftsgaarden, D. O., and C. P. Quesenberry. 1965. “A Nonparametric Estimate of a Multivariate Density Function.” *The Annals of Mathematical Statistics* 36 (3). Institute of Mathematical Statistics: 1049–51. doi:10.1214/aoms/1177700079.

Macleod, James E. S., Andrew Luk, and D. Michael Titterington. 1987. “A Re-Examination of the Distance-Weighted K-Nearest Neighbor Classification Rule.” *IEEE Transactions on Systems, Man and Cybernetics* 17 (4): 689–96. doi:10.1109/TSMC.1987.289362.

Magnussen, Steen, Erkki Tomppo, and Ronald E. McRoberts. 2010. “A Model-Assisted K-Nearest Neighbour Approach to Remove Extrapolation Bias.” *Scandinavian Journal of Forest Research* 25 (2). Taylor & Francis: 174–84.

Magnussen, Steen, Ronald E. McRoberts, and Erkki O. Tomppo. 2009. “Model-Based Mean Square Error Estimators for K-Nearest Neighbour Predictions and Applications Using Remotely Sensed Data for Forest Inventories.” *Remote Sensing of Environment* 113 (3): 476–88. doi:10.1016/j.rse.2008.04.018.

Maxwell, S. K., G. L. Schmidt, and J. C. Storey. 2007. “A Multi-scale Segmentation Approach to Filling Gaps in Landsat ETM+ SLC-off Images.” *International Journal of Remote Sensing* 28 (23). Taylor & Francis: 5339–56. doi:10.1080/01431160601034902.

McNab, W. H., D. T. Cleland, J. A. Freeouf, J. E. Keys Jr., G. J. Nowacki, and C. A. Carpenter. 2007. “Description of ‘Ecological Subregions: Sections of the Conterminous United States’ (First Approximation).”

McRoberts, Ronald E. 2009a. “Diagnostic Tools for Nearest Neighbors Techniques When Used with Satellite Imagery.” *Remote Sensing of Environment* 113 (3): 489–99. doi:http://dx.doi.org/10.1016/j.rse.2008.06.015.

McRoberts, Ronald E. 2009b. “A Two-Step Nearest Neighbors Algorithm Using Satellite Imagery for Predicting Forest Structure within Species Composition Classes.” *Remote Sensing of Environment* 113 (3): 532–45. doi:http://dx.doi.org/10.1016/j.rse.2008.10.001.

McRoberts, Ronald E. 2011. “Satellite Image-Based Maps: Scientific Inference or Pretty Pictures?” *Remote Sensing of Environment* 115 (2): 715–24. doi:10.1016/j.rse.2010.10.013.

McRoberts, Ronald E. 2012. “Estimating Forest Attribute Parameters for Small Areas Using Nearest Neighbors Techniques.” *Forest Ecology and Management* 272 (0): 3–12. doi:http://dx.doi.org/10.1016/j.foreco.2011.06.039.

McRoberts, Ronald E, and Erkki O. Tomppo. 2007. “Remote Sensing Support for National Forest Inventories.” *Remote Sensing of Environment* 110 (4). Elsevier: 412–19.

McRoberts, Ronald E., Erkki O. Tomppo, and Erik Næsset. 2010. “Advances and Emerging Issues in National Forest Inventories.” *Scandinavian Journal of Forest Research* 25 (4). Taylor & Francis: 368–81. doi:10.1080/02827581.2010.496739.

McRoberts, Ronald E., Erkki O. Tomppo, Andrew O. Finley, and Juha Heikkinen. 2007. “Estimating Areal Means and Variances of Forest Attributes Using the K-Nearest Neighbors Technique and Satellite Imagery.” *Remote Sensing of Environment* 111 (4): 466–80. doi:10.1016/j.rse.2007.04.002.

McRoberts, Ronald E., Mark D. Nelson, and Daniel G. Wendt. 2002. “Stratified Estimation of Forest Area Using Satellite Imagery, Inventory Data, and the K-Nearest Neighbors Technique.” *Remote Sensing of Environment* 82 (2). Elsevier: 457–68.

McRoberts, Ronald E., Steen Magnussen, Erkki O. Tomppo, and Gherardo Chirici. 2011. “Parametric, Bootstrap, and Jackknife Variance Estimators for the K-Nearest Neighbors Technique with Illustrations Using Forest Inventory and Satellite Image Data.” *Remote Sensing of Environment* 115 (12): 3165–74. doi:10.1016/j.rse.2011.07.002.

McRoberts, Ronald E., Warren B. Cohen, Erik Næsset, Stephen V. Stehman, and Erkki O. Tomppo. 2010. “Using Remotely Sensed Data to Construct and Assess Forest Attribute Maps and Related Spatial Products.” *Scandinavian Journal of Forest Research* 25 (4). Taylor & Francis: 340–67. doi:10.1080/02827581.2010.497496.

Moeur, Melinda, and Albert R. Stage. 1995. “Most Similar Neighbor: An Improved Sampling Inference Procedure for Natural Resource Planning.” *Forest Science* 41 (2): 337–59. http://dx.doi.org/10.1093/forestscience/41.2.337.

Moody, Aaron, and David M. Johnson. 2001. “Land-Surface Phenologies from AVHRR Using the Discrete Fourier Transform.” *Remote Sensing of Environment* 75 (3). Elsevier: 305–23.

Muja, Marius, and David G. Lowe. 2009. “Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration.” *VISAPP (1)* 2 (331–340): 2.

Nadaraya, E. A. 1964. “On Estimating Regression.” *Theory of Probability & Its Applications* 9 (1). SIAM: 141–42. doi:10.1137/1109020.

Næsset, Erik, Terje Gobakken, Svein Solberg, Timothy G. Gregoire, Ross Nelson, Göran Ståhl, and Dan Weydahl. 2011. “Model-Assisted Regional Forest Biomass Estimation Using LiDAR and InSAR as Auxiliary Data: A Case Study from a Boreal Forest Area.” *Remote Sensing of Environment* 115 (12): 3599–3614. doi:10.1016/j.rse.2011.08.021.

Ohmann, Janet L., and Matthew J. Gregory. 2002. “Predictive Mapping of Forest Composition and Structure with Direct Gradient Analysis and Nearest- Neighbor Imputation in Coastal Oregon, U.S.A.” *Canadian Journal of Forest Research* 32 (4). NRC Research Press Ottawa, Canada: 725–41. doi:10.1139/x02-011.

Ooi, Hui-Lee, Siew-Cheok Ng, and Einly Lim. 2013. “ANO Detection with K-Nearest Neighbor Using Minkowski Distance.” *International Journal of Signal Processing Systems,* 208–11. doi:10.12720/ijsps.1.2.208-211.

Opsomer, Jean D., F. Jay Breidt, Gretchen G. Moisen, and Göran Kauermann. 2007. “Model-Assisted Estimation of Forest Resources with Generalized Additive Models.” *Journal of the American Statistical Association* 102 (478). Taylor & Francis: 400–409.

Park, B. U., Y. K. Lee, and S. Ha. 2009. “L2 Boosting in Kernel Regression.” *Bernoulli* 15 (3). International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability: 599–613. doi:10.3150/08-BEJ160.

Pickell, Paul D., Txomin Hermosilla, Ryan J. Frazier, Nicholas C. Coops, and Michael A. Wulder. 2016. “Forest Recovery Trends Derived from Landsat Time Series for North American Boreal Forests.” *International Journal of Remote Sensing* 37 (1). Taylor & Francis: 138–49. doi:10.1080/2150704X.2015.1126375.

Potgieter, A. B., A. Apan, P. Dunn, and G. Hammer. 2007. “Estimating Crop Area Using Seasonal Time Series of Enhanced Vegetation Index from MODIS Satellite Imagery.” *Australian Journal of Agricultural Research* 58 (4): 316–25. http://dx.doi.org/10.1071/AR06279.

Prasath, V. B. Surya, Haneen Arafat Abu Alfeilat, Omar Lasassmeh, and Ahmad B. A. Hassanat. 2017. “Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier - A Review.” *arXiv Preprint arXiv:1708.04321*. http://arxiv.org/abs/1708.04321.

Pringle, M. J., M. Schmidt, and J. S. Muir. 2009. “Geostatistical Interpolation of SLC-off Landsat ETM+ Images.” *ISPRS Journal of Photogrammetry and Remote Sensing* 64 (6): 654–64. doi:10.1016/j.isprsjprs.2009.06.001.

R Core Team. 2016. “R: A Language and Environment for Statistical Computing.” Vienna, Austria. https://www.r-project.org/.

Rao, J. N. K. 2003. *Small Area Estimation*. Hoboken, N.J.: Hoboken, N.J. : John Wiley.

Rao, J. N. K. 2011. “Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal.” *Statistical Science* 26 (2). Institute of Mathematical Statistics: 240–56. http://projecteuclid.org/euclid.ss/1312204015.

Reams, Gregory A., William D. Smith, Mark H. Hansen, William A. Bechtold, Francis A. Roesch, and Gretchen G. Moisen. 2005. “The Forest Inventory and Analysis Sampling Frame.” In *The Enhanced Forest Inventory and Analysis Program—National Sampling*

Design and Estimation Procedures (Gen. Tech. Rep. SRS-80), 11–26. U.S. Department of Agriculture Forest Service, Southern Research Station.

Reed, Bradley C., Jesslyn F. Brown, Darrel VanderZee, Thomas R. Loveland, James W. Merchant, and Donald O. Ohlen. 1994. "Measuring Phenological Variability from Satellite Imagery." *Journal of Vegetation Science* 5 (5). Wiley: 703–14. doi:10.2307/3235884.

Roerink, G. J., M. Menenti, and W. Verhoef. 2000. "Reconstructing Cloudfree NDVI Composites Using Fourier Analysis of Time Series." *International Journal of Remote Sensing* 21 (9). Taylor & Francis Ltd: 1911–17. http://10.0.4.56/014311600209814.

Roy, David P., Junchang Ju, Kristi Kline, Pasquale L. Scaramuzza, Valeriy Kovalskyy, Matthew Hansen, Thomas R. Loveland, Eric Vermote, and Chunsun Zhang. 2010. "Web-Enabled Landsat Data (WELD): Landsat ETM+ Composited Mosaics of the Conterminous United States." *Remote Sensing of Environment* 114 (1): 35–49. doi:10.1016/j.rse.2009.08.011.

Roy, David P., Junchang Ju, Philip Lewis, Crystal Schaaf, Feng Gao, Matt Hansen, and Erik Lindquist. 2008. "Multi-Temporal MODIS-Landsat Data Fusion for Relative Radiometric Normalization, Gap Filling, and Prediction of Landsat Data." *Remote Sensing of Environment* 112 (6): 3112–30. doi:10.1016/j.rse.2008.03.009.

Royall, Richard M. 1966. "A Class of Non-Parametric Estimates of a Smooth Regression Function." Dept. of Statistics, Stanford University.

Sakamoto, Toshihiro, Masayuki Yokozawa, Hitoshi Toritani, Michio Shibayama, Naoki Ishitsuka, and Hiroyuki Ohno. 2005. "A Crop Phenology Detection Method Using Time-Series MODIS Data." *Remote Sensing of Environment* 96 (3–4): 366–74. doi:10.1016/j.rse.2005.03.008.

Schapire, Robert E. 1990. "The Strength of Weak Learnability." *Machine Learning* 5 (2). Springer: 197–227.

Scharlemann, Jörn P. W., David Benz, Simon I. Hay, Bethan V. Purse, Andrew J. Tatem, G. R. William Wint, and David J. Rogers. 2008. "Global Data for Ecology and Epidemiology: A Novel Algorithm for Temporal Fourier Processing MODIS Data." *PLoS ONE* 3 (1). San Francisco, USA: Public Library of Science: e1408. doi:10.1371/journal.pone.0001408.

Sellers, P. J., C. J. Tucker, G. J. Collatz, S. O. Los, C. O. Justice, D. A. Dazlich, and D. A. Randall. 1994. "A Global 1 Deg by 1 Deg NDVI Data Set for Climate Studies. Part 2: The Generation of Global Fields of Terrestrial Biophysical Parameters from the NDVI." *International Journal of Remote Sensing* 15 (17). Taylor & Francis: 3519–45. doi:10.1080/01431169408954343.

Sellers, Piers J., Sietse O. Los, Compton J. Tucker, Christopher O. Justice, Donald A. Dazlich, G. James Collatz, and David A. Randall. 1996. "A Revised Land Surface Parameterization (SiB2) for Atmospheric GCMS. Part II: The Generation of Global Fields of Terrestrial Biophysical Parameters from Satellite Data." *Journal of Climate* 9 (4). American

Meteorological Society: 706–37. doi:10.1175/1520-0442(1996)009<0706:ARLSPF>2.0.CO;2.

Skakun, Robert S., Michael A. Wulder, and Steven E. Franklin. 2003. “Sensitivity of the Thematic Mapper Enhanced Wetness Difference Index to Detect Mountain Pine Beetle Red-Attack Damage.” *Remote Sensing of Environment* 86 (4): 433–43. doi:10.1016/S0034-4257(03)00112-3.

Steinberg, Joseph, National Center for Health Statistics (U.S.), National Institute on Drug Abuse, and Workshop on Synthetic Estimates. 1979. *Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion*. Rockville, Md.: Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute on Drug Abuse, Division of Research ; Washington: Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute on Drug Abuse, Division of Research.

Stone, Charles J. 1977. “Consistent Nonparametric Regression, with Discussion.” *Annals of Statistics* 5. JSTOR: 595–645.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics* 9 (1). BioMed Central Ltd: 307. http://www.biomedcentral.com/1471-2105/9/307.

Stueve, Kirk M., Ian W. Housman, Patrick L. Zimmerman, Mark D. Nelson, Jeremy B. Webb, Charles H. Perry, Robert A. Chastain, Dale D. Gormanson, Chengquan Huang, and Sean P. Healey. 2011. “Snow-Covered Landsat Time Series Stacks Improve Automated Disturbance Mapping Accuracy in Forested Landscapes.” *Remote Sensing of Environment* 115 (12). Elsevier: 3203–19.

Tierney, Luke, A. J. Rossini, Na Li, and H. Sevcikova. 2016. “Snow: Simple Network of Workstations.” https://cran.r-project.org/package=snow.

Tokola, T., J. Pitkänen, S. Partinen, and E. Muinonen. 1996. “Point Accuracy of a Non-Parametric Method in Estimation of Forest Characteristics with Different Satellite Materials.” *International Journal of Remote Sensing* 17 (12). Taylor & Francis: 2333–51.

Tomppo, Erkki. 1990. “Satellite Image-Based National Forest Inventory of Finland.” *Photogrammetric Journal of Finland* 12 (1): 115–20.

Tukey, John W. 1977. *Exploratory Data Analysis*. Vol. 2. Reading, Mass.

Vogt, J. T. and W. Brad Smith. 2017. “Forest Inventory and Analysis Fiscal Year 2016 Business Report.” *FS-1075*. Washington, DC: U.S. Department of Agriculture, Forest Service, Washington Office. 74 p.

Wand, M. P., and M. C. Jones. 1995. *Kernel Smoothing*. London ; New York: London ; New York : Chapman and Hall.

Watson, Geoffrey S. 1964. “Smooth Regression Analysis.” *The Indian Journal of Statistics* 26 (4). Springer: 359–72. doi:10.2307/25049340.

Wilson, B. Tyler, Andrew J. Lister, and Rachel I. Riemann. 2012. “A Nearest-Neighbor Imputation Approach to Mapping Tree Species over Large Areas Using Forest Inventory Plots and Moderate Resolution Raster Data.” *Forest Ecology and Management* 271 (0): 182–98. doi:10.1016/j.foreco.2012.02.002.

Wilson, Barry T., Joseph F. Knight, and Ronald E. McRoberts. 2018. “Harmonic Regression of Landsat Time Series for Modeling Attributes from National Forest Inventory Data.” *ISPRS Journal of Photogrammetry and Remote Sensing* 137 (2018): 29-46. doi:10.1016/j.isprsjprs.2018.01.006.

Woodcock, Curtis E., Richard Allen, Martha Anderson, Alan Belward, Robert Bindschadler, Warren Cohen, Feng Gao, et al. 2008. “Free Access to Landsat Imagery.” *Science* 320 (5879): 1011 LP-1011. http://science.sciencemag.org/content/320/5879/1011.1.abstract.

Wulder, Michael A., Jeffrey G. Masek, Warren B. Cohen, Thomas R. Loveland, and Curtis E. Woodcock. 2012. “Opening the Archive: How Free Data Has Enabled the Science and Monitoring Promise of Landsat.” *Remote Sensing of Environment* 122: 2–10. doi:10.1016/j.rse.2012.01.010.

Yates, F., and P. M. Grundy. 1953. “Selection without Replacement from within Strata with Probability Proportional to Size.” *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 253–61.

Yuan, Fei, Kali E. Sawaya, Brian C. Loeffelholz, and Marvin E. Bauer. 2005. “Land Cover Classification and Change Analysis of the Twin Cities (Minnesota) Metropolitan Area by Multitemporal Landsat Remote Sensing.” *Remote Sensing of Environment* 98 (2–3): 317–28. doi:10.1016/j.rse.2005.08.006.

Zhang, C., W. Li, and D. Travis. 2007. “Gaps-fill of SLC-off Landsat ETM+ Satellite Image Using a Geostatistical Approach.” *International Journal of Remote Sensing* 28 (22). Taylor & Francis: 5103–22. doi:10.1080/01431160701250416.

Zhang, X., M.A. Friedl, C.B. Schaaf, A.H. Strahler, J.C.F. Hodges, F. Gao, B.C. Reed, and A. Huete. 2003. “Monitoring Vegetation Phenology Using MODIS.” *Remote Sensing of Environment* 84 (3). doi:10.1016/S0034-4257(02)00135-9.

Zheng, Sheng, Chunxiang Cao, Yongfeng Dang, Haibing Xiang, Jian Zhao, Yuxing Zhang, Xuejun Wang, and Hongwen Guo. 2014. “Retrieval of Forest Growing Stock Volume by Two Different Methods Using Landsat TM Images.” *International Journal of Remote Sensing* 35 (1). Taylor & Francis: 29–43. doi:10.1080/01431161.2013.860567.

Zhu, Xiaolin, and Desheng Liu. 2015. “Improving Forest Aboveground Biomass Estimation Using Seasonal Landsat NDVI Time-Series.” *ISPRS Journal of Photogrammetry and*

Remote Sensing 102 (Supplement C): 222–31. doi:https://doi.org/10.1016/j.isprsjprs.2014.08.014.

Zhu, Zhe, Curtis E. Woodcock, and Pontus Olofsson. 2012. “Continuous Monitoring of Forest Disturbance Using All Available Landsat Imagery.” *Remote Sensing of Environment* 122: 75–91. doi:10.1016/j.rse.2011.10.030.

Zlochin, Mark, and Marco Dorigo. 2002. “Model-Based Search for Combinatorial Optimization: A Comparative Study.” In *Parallel Problem Solving from Nature*, 651–61. Springer. doi:10.1007/3-540-45712-7_63.