Modeling Percent Tree Canopy Cover: A Pilot Study

John W. Coulston, Gretchen G. Moisen, Barry T. Wilson, Mark V. Finco, Warren B. Cohen, and C. Kenneth Brewer

Abstract

Tree canopy cover is a fundamental component of the landscape, and the amount of cover influences fire behavior, air pollution mitigation, and carbon storage. As such, efforts to empirically model percent tree canopy cover across the United States are a critical area of research. The 2001 national-scale canopy cover modeling and mapping effort was completed in 2006, and here we present results from a pilot study for a 2011 product. We examined the influence of two different modeling techniques (random forests and beta regression), two different Landsat imagery normalization processes, and eight different sampling intensities across five different pilot areas. We found that random forest out-performed beta regression techniques and that there was little difference between models developed based on the two different normalization techniques. Based on these results we present a prototype study design which will test canopy cover modeling approaches across a broader spatial scale.

Introduction

Tree canopy cover is a primary structural characteristic that is important both within a forest setting and in non-forest land covers such as urban lands. Tree canopy cover is the proportion of an area covered by the vertical projection of tree crowns (Jennings *et al.* 1999). The amount of tree canopy cover is directly related to biomass and carbon (Suganuma *et al.* 2006, Kelndorfer *et al.* 2006) as well as air pollution mitigation (Nowak *et al.* 2006) and stream water temperatures (Webb and Crisp, 2006). Additionally, tree canopy cover is a major component of forest fire behavior and fuel models (Rollins and Frame, 2006) and a critical aspect of forest management activities (Jennings *et al.* 1999). Because of the importance of tree canopy cover, the Multi-Resolution Land Characteristic consortium (MRLC) developed and distributed a map-based product of percent tree canopy cover as part of the 2001 National Land Cover Database (NLCD) (Homer *et al.* 2007). The scope of this paper is to lay the foundation for development of the 2011 version of the NLCD percent tree canopy map.

The 2001 NLCD percent tree canopy map is a freely available, 30 m raster geospatial dataset covering the coterminous United States, coastal Alaska, Hawaii, and Puerto Rico. These data contain percent tree canopy estimates, as a continuous variable, for each pixel across all land covers and types. Because percent tree canopy cover is not directly calculable from the spectral information found in Landsat imagery, fine-scale data were used to develop the response variable and coarser (30 m) data derived from the 1992 NLCD, Landsat-5 and -7, and digital elevation models were used as the explanatory variables. The response data were generally derived from 1 m, panchromatic, digital orthophoto quarter-quadrangles (DOQQ's). Initially one to two DOQQ's (approximately 6 to 8 km^2) were selected for each Landsat scene. As the database development progressed, three to four DOQQ's between 1 and 4 km^2 were used (Homer *et al.* 2007). The DOQQ's were typically selected in a purposive fashion to have coverage across the Landsat scene including scene edges. To develop the response data, each DOQQ was classified as "tree cover" or "no tree cover" at the 1 m pixel level using a classification tree algorithm and then resampled to 30 m to calculate percent tree canopy cover. Multi-season top-of-atmosphere Landsat-5 and -7 data and indices (e.g., tasseled cap), along with digital elevation models and derivates (e.g., slope), and other ancillary data (e.g., 1992 NLCD land-cover) were used as the explanatory variables. Regression trees, implemented through Cubist ©, were then developed to model the empirical relationship between the response and explanatory variables and create the map products. The final step was to develop and apply a "liberal" forest mask to avoid commission errors in, for example, agricultural areas (Huang *et al.* 2001).

Accuracy assessments of the 2001 NLCD percent tree canopy cover dataset were provided by Homer *et al.* (2004), Greenfield *et al.* (2009), and Nowak and Greenfield (2010). Homer *et al.* (2004) provide model fit statistics for three mapping zones (Homer and Gallant, 2001) in the eastern (Zone 60), north central (Zone 41), and interior west (Zone 16) United States. Based on cross-validation, the correlation coefficient between observed and predicted percent canopy cover was 0.88, 0.78, and 0.93 in zones 16, 41, and 60, respectively. The mean absolute error was 9.9 percent, 14.1 percent, and 8.4 percent in zones 16, 41, and 60, respectively. Greenfield *et al.* (2009) used imagery available

John W. Coulston is with the US Forest Service, 4700 Old Kingston Pike, Knoxville, TN 37919, (jcoulston@fs.fed.us).

Gretchen G. Moisen is with the US Forest Service, 507 25th Street, Ogden UT 84401.

Barry T. Wilson is with the US Forest Service, 1992 Folwell Avenue, St. Paul, MN 55108.

Mark V. Finco is with Red Castle Resources, 2222 West 2300 South, Salt Lake City, UT 84119.

Warren B. Cohen is with the US Forest Service, 3200 SW Jefferson Way, Corvallis, OR 97331.

C. Kenneth Brewer is with the US Forest Service, 1601 N. Kent Street, 4th floor, Arlington, VA, 22209.

Photogrammetric Engineering & Remote Sensing
Vol. 78, No. 7, July 2012, pp. 715–727.

0099-1112/12/7807–715/$3.00/0

through Google Earth™ to estimate percent tree canopy cover for randomly selected counties and designated places across the United States. These estimates were made by manually interpreting 200 photo-points as tree cover or non-tree cover within each county or place. The percent tree cover from the 200 photo points was then compared to the average percent tree canopy cover from NLCD 2001. Greenfield *et al.* (2009) found that the average percent tree canopy cover estimate for the selected areas was consistently higher based on photo interpretation as compared to the NLCD canopy cover estimates. They reported a range of differences from less than 1 percent to 25 percent. Nowak and Greenfield (2010) employed a similar photo interpretation-based approach but sampled all mapping zones. They found that canopy cover was underestimated in 64 of 65 mapping zones and that the mean underestimation was 9.7 percent nationally. Despite the challenges associated with creating a national percent tree canopy cover map, there is substantial interest on the part of the MRLC and the US Forest Service to update these data for the 2011 NLCD.

Because of the high costs of broad-scale inventory, monitoring, and mapping programs, it makes sense to leverage existing programs and infrastructures across agencies, to develop products such as the NLCD percent tree canopy cover map. The USDA Forest Service Forest Inventory and Analysis (FIA) program provides both *in situ* and remotely sensed data on forest and land conditions across the United States using a probability-based sample that covers all lands. As Homer *et al.* (2004) described, the FIA data can be used as training data for broad-scale mapping efforts. Examples of such efforts are described by Blackard *et al.* (2008) and Ruefenacht *et al.* (2008) for creating biomass maps and forest-type maps respectively. The FIA program is currently implementing measurement protocols to quantify percent tree canopy cover on all lands using the same equal probability sampling design. Clearly this provides a synergistic opportunity for the MRLC and the US Forest Service to develop the 2011 NLCD percent tree canopy cover map.

The objectives of this study were to (a) investigate the use of FIA percent tree canopy cover data for constructing empirical models to estimate percent tree canopy cover at unmeasured locations, (b) examine the impacts of sampling intensity and Landsat scene normalization on the development of percent tree canopy cover models, (c) compare model performance between random forest models and beta regression models, and (d) identify design specifications for a prototype study where percent tree canopy cover will be modeled over much broader geographic regions.

Methods

Study Areas

Five pilot areas across the United States were selected for this study (Plate 1). They were located in Georgia (GA), Kansas (KS), Michigan (MI), Oregon (OR), and Utah (UT). Each of the study areas were the size of approximately one Landsat scene, but were positioned to span multiple path rows in order to examine the impact of seam lines across multiple scenes. Additionally, each pilot area was selected to cross local ecological gradients. For example, the GA study area ranged from the piedmont region in the south, through the Atlanta metropolitan area, to the heavily forested Appalachian Mountains in the north. The OR study area was slightly larger than the other study areas so that it would intersect with some existing lidar datasets for future research.

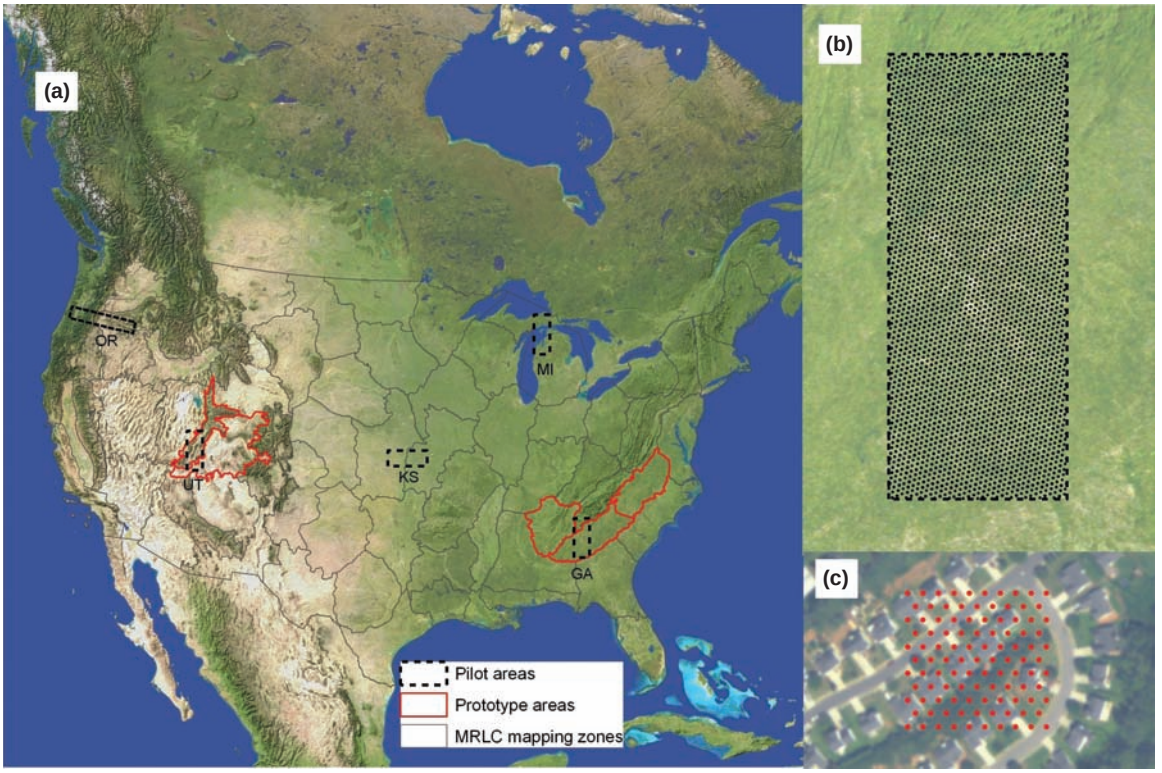

Plate 1. (a) Location of pilot study areas and proposed prototype study areas, (b) Equal probability grid identifying the approximate locations of the primary sampling units, and (c) Example of a secondary sampling unit consisting of 105 points.

Sample Design and Photo Interpretation

The FIA sample design was adopted for this study (Bechtold and Patterson, 2005). FIA uses a quasi-systematic sample based on White *et al.* (1992) where the nominal sampling intensity was approximately one sample location per 2,400 ha across all land covers and types. This design is assumed to produce a random equal probability sample (McRoberts *et al.*, 2006) and is implemented in a panel design where each of the five panels in GA, MI, and KS contain 20 percent of the sample locations, and each of the ten panels in UT and OR contain 10 percent of the sample locations. For the purposes of this study, the base FIA sample was intensified to 4X intensity (one plot per 600 ha) using the procedures described by White *et al.* (1992) (Plate 1). The sample was intensified so that we could examine the impact of sampling intensity (from 20 percent to 400 percent of the base sampling intensity) on model development. We then used a simple dot grid approach to estimate canopy cover for each sample location. Manual photo-interpretation of dot grids are a standard technique for determining area (Slama *et al.* 1980) and are particularly well suited for estimating canopy area of a given location of known size (Nowak *et al.*, 1996). For our purposes, we used a 105 point triangular-grid that filled a 90 m by 90 m (0.81 ha) area surrounding the sample location and then registered the 0.81 ha square to the NLCD 2001 dataset (Plate 1). Manual photo-interpretation was used to classify each of the 105 points, per sample location, as either "tree canopy" or "no tree canopy." In most cases the photo interpreters had detailed ground-based knowledge of the area they were photo interpreting. The overall design was considered a two-stage sampling design where the 0.81 ha area was the primary sampling unit (PSU), and each of the 105 points within PSUs were the secondary sampling units. The design based estimators of proportion canopy cover in each PSU, mean proportion canopy cover in each study area, and the standard error of the estimate were obtained following Cochran (1977).

Imagery, Vegetation Indices, and Ancillary Data

We used imagery provided by the National Agriculture Imagery Program (NAIP) (USDA 2009) to develop the response data (percent tree canopy cover). Landsat-5 data, digital elevation data, and derivatives were used to develop the explanatory data. The NAIP data were 1 m resolution and available as either natural color (KS, MI, UT study areas), or natural color and color-infrared (GA, OR study areas). All NAIP imagery was collected during the growing season in 2009, and therefore was leaf-on imagery. The six reflective bands from Landsat-5 were used. The Landsat data were also leaf-on and from either 2009 or 2008, depending on cloud cover. Initially, the Landsat data were converted to top-of-atmosphere reflectance and then to surface reflectance using a simple dark object subtraction. These data were considered calibrated, but "non-normalized" because no additional adjustments were made to compensate for bidirectional reflectance distribution function (BRDF) effects in the imagery. The Landsat data were also "normalized" using Model II regression techniques described in Beaty *et al.* (2010). This normalization technique used the overlap areas among Landsat scenes to radiometrically match each scene in a mosaic to each other. This calibration was performed on each of the reflectance bands to minimize differences in reflectance values due to BRDF effects. From both the normalized and non-normalized Landsat data, the normalized difference vegetation index (NDVI) and tasseled cap values were calculated. The Elevation Derivatives for National Applications data were used for elevation, slope, cosine of aspect, and compound topographic index and were also 30 m resolution. Other ancillary data included the 2001 NLCD land cover and percent tree canopy cover.

Percent Tree Canopy Cover Models and Comparisons

For modeling purposes, recall that the percent tree canopy cover was estimated for each 0.81 ha PSU, and note that the explanatory variables (Landsat bands, vegetation indices, elevation and derivatives, and 2001 NLCD products) were 30 m (0.09 ha) resolution. The response variable (percent tree canopy cover) was taken directly from the estimate for each PSU. The explanatory variables for modeling were developed by calculating the mean and standard deviation of each variable for each PSU. Because the PSU was registered to the NLCD base, the means and standard deviations for each variable were simply calculated using 3 $\times$ 3 pixel window focal statistics.

Numerous modeling approaches are available to develop empirical models of percent tree canopy cover. Here we used both the random forests algorithm developed by Breiman (2001) and beta regression (Ferrari and Cribari-Neto, 2004). Generally speaking, the random forest modeling approach is a non-parametric technique in that there are no distributional assumptions. The term random forest may be confusing particularly given the context of this research. "Random" in this case refers to random bootstrap resampling of the data, and the term "forest" refers an ensemble of regression trees (i.e., forest). The beta-regression approach is a parametric technique that is analogous to multiple regression except that rather than assuming the residuals are normally distributed the implementation of this techniques assumes that the residuals are beta-distributed which is a more tenable assumption for modeling proportions. Random forests is relatively resistant to multicollinearity whereas beta-regression is not. However, prediction error is more directly estimated using beta-regression because it is a parametric approach whereas additional bootstrap techniques must be used to estimate prediction error from random forest models. More information about each technique follows.

Random forests is an ensemble method that uses bootstrap sampling to develop multiple models to improve prediction. In the following example adopted from Liaw and Wiener (2002), we assume that we have a dataset with 100 observations, and we set the forest size *a priori* to 500 regression trees. To construct the ensemble we draw 500 bootstrap samples. The bootstrap samples are selected with replacement and each bootstrap sample has on average 63 observations (63 percent of observations). For each bootstrap sample, a regression tree is developed, but instead of determining the best split across all explanatory variables, a predetermined number of explanatory variables (say five) are randomly selected and the best split among those variables is selected. Predicted values are then obtained by averaging the predictions from each of the 500 individual trees. Typically, model fit statistics are then developed from the out-of-bag (OOB) data. Recall that each bootstrap sample contains about 63 percent of the observations. The remaining 37 percent of the observations are the OOB data for a given tree. At each bootstrap iteration, estimates for the OOB data are made based on models developed from the bootstrap sample. The OOB predictions are then aggregated across all samples to estimate mean square error (MSE) and pseudo R^2. For modeling we used the R version 2.12 (R Development Core Team, 2010) random forest library (Liaw and Wiener, 2002) to construct empirical models of percent tree canopy cover.

Beta regression was also used to develop empirical models of percent tree canopy cover. This approach was used by Korhonen *et al.* (2007) to develop models of canopy cover in Finland. Beta regression is an extension of generalized linear models and is specifically well suited for estimating parameters for empirical models of rates and

proportions. This technique assumes the errors are beta distributed (rather than normally distributed, as assumed under ordinary least squares solutions) and a link function is used to transform the estimated values into the appropriate range. Beta regression is valid for continuous response variables in the standard unit interval [0,1]. Because the response typically also included numerous observations of 0 and 1, we transformed percent tree canopy cover for each observation as recommended by Smithson and Verkuilen (2006): $((cc_i(n-1) + 0.5)/n)$ where n was the number of observations. To construct beta regression models, the full set of potential explanatory variables was reduced by removing highly correlated variables ($r > 0.7$) by study area. This was done to minimize multicollinearity. Models were constructed manually based on the 4X sample for each study area, and the parameters were estimated individually for each sampling intensity (0.2X – 4X). The final models only included explanatory variables that were significant at $\alpha = 0.10$. For modeling we used the R version 2.12 (R Development Core Team, 2010) betareg library (Cribari-Neto and Zeileis, 2010) to construct empirical models of percent tree canopy cover.

For each study area 32 empirical models were developed. A separate model was developed for each modeling approach (beta regression, random forest), for each sampling intensity (0.2X, 0.4X, 0.6X, 0.8X, 1X, 2X, 3X, and 4X) using the normalized and non-normalized Landsat data as part of the explanatory data. These sampling intensities were selected because, from an FIA programmatic viewpoint, they are easily implemented as part of FIA standard data collection protocols. We examined and compared model performance based on recommendations provided by Duane *et al.* (2010). Model accuracy was examined based on pseudo R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(cc_i - \widehat{cc_i})^2}{\sum_{i=1}^{n}(cc_i - \overline{cc})^2} \quad (1)$$

where R^2 = pseudo R^2 (percent variance explained), cc_i = observed percent canopy cover of the i^{th} observation, $\widehat{cc}$ = predicted percent canopy cover of the i^{th} observation, and $\overline{cc}$ = mean percent canopy cover across n observations.

Equation 1 is essentially one minus the MSE divided by the variance of cc where the variance is calculated with n in the denominator rather than $n - 1$. In addition to pseudo R^2, root mean square error (RMSE) was also calculated, but it should be noted that in this case, RMSE was inversely proportional to pseudo R^2 because the numerator in equation 1 was the sum squared prediction error and the denominator was fixed for each study area. To make comparisons across the different sampling intensities and different modeling approaches, the pseudo R^2 and RMSE were calculated based on all observations (i.e., the 4X sample). For example, suppose a random forest model was constructed using the 1X sampling intensity (approx 1,000 observations). In this case, about 3,000 observations were not used to develop the model. Out-of-bag predictions were retained for the 1,000 observations from the 1X sample, and model predictions were also made for the remaining 3,000 observations. Equation 1 was then used to calculate the pseudo R^2 based on n = 4,000 observations. The ability of each model to replicate the sample distribution was examined using cumulative distribution functions (CDF) constructed using 1 percent increments from 0 percent to 100 percent. Model bias was assessed by examining the slope and intercept of the observed versus predicted linear regression line for each model by study area and sampling intensity (0.2X to 3X) for the normalized and non-normalized datasets (Pineiro *et al.* 2008). In this case, we used the remaining sample (approximately 1,000 observations) as a hold-out data set for fitting the linear regression model and estimating the parameters. To examine map bias, percent tree canopy cover maps were produced from each model for each study area and sampling intensity for the normalized and non-normalized datasets. Recall that the models were developed based on focal means (and standard deviations) in the 3 × 3 window around each sample location. For creating the maps, the original 30 m pixel values were used rather than the focal mean of each explanatory variable. However, the focal standard deviation explanatory variables were still based on a 3 × 3 window. The mean map-based percent tree canopy cover estimate was then compared to the design-based estimate of the mean from the 4X sample, which has the smallest sampling error, for the following classes: all land covers, urban land cover, and forest land cover. Additionally, CDFs for each map were constructed and compared to the observed CDF.

Having determined the appropriate sample size for each study area (0.2X to 4X), the best performing modeling approach (beta regression or random forests), and whether normalized Landsat data were required, we then examined how robust the solution was across broader geographic areas. To accomplish this we combined the modeling data for the eastern study areas (GA, KS, and MI) and developed one single empirical model; likewise with the western study areas (OR and UT). We also constructed one single US model across all study areas. For example, suppose that the beta regression model outperformed the random forest model for each of the two western study areas and that the pseudo R^2 etc. stabilized at the 1X sampling intensity (approx 1,000 observations). Also assume that the models required the normalized Landsat data. Based on this example we would fit a beta regression model for the eastern study areas where one-half of the observations (500) were selected from each study area. The "western model" would then be assessed based on pseudo R^2, CDF, and slope and intercept of the observed verses predicted regression line. However, while the model would be fit across several studies areas, the model performance would be assessed for each study area based on the entire 4X sample.

Results

Based on the photo interpretation of the 4X sample, the average percent canopy cover (across all 2001 NLCD land cover classes) ranged from 66 percent in the GA study area to 12.8 percent in the KS study area (Table 1). The standard error of these estimates ranged from 0.69 percent in the MI study area to 0.40 percent in the KS study area. When considering the forest land cover class (as defined by NLCD 2001), the MI study area had the highest average percent canopy in the forest class (90.5). Conversely, the UT study area had the lowest average percent canopy cover in the forest class (52 percent). The same pattern was observed when considering average percent canopy cover in urban land covers where again MI had the highest (53.5 percent), and UT had the lowest (9.1 percent). Typically, the standard errors of the estimates were highest when considering urban land cover classes (Table 1). For example, the standard error for average canopy cover in urban classes in MI was 2.69 percent which means that the 95 percent confidence interval was 48.1 percent to 58.8 percent.

Overall, the empirical percent tree canopy cover models had pseudo R^2 ranging from 0.53 to 0.90 and the best model fits were observed in the MI study area (Figure 1) based on the 4X sample. The random forest models consistently outperformed the beta regression models. The degree to which the random forest model outperformed the beta regression model was related to how good the model was.

TABLE 1. MEAN PERCENT TREE CANOPY COVER ($\bar{cc}$) AND THE STANDARD ERROR OF THE MEAN (S.E.($\bar{cc}$)) FOR EACH STUDY AREA FOR NLCD 2001 FOREST LAND COVER, URBAN LAND COVER, AND ACROSS ALL LAND COVER CLASSES BASED ON THE 4X SAMPLE

	Sample-based	
Study Area	$\bar{cc}$	s.e ($\bar{cc}$)
GA		
Forest	84.1	0.45
Urban	41.1	0.94
All	66.0	0.53
KS		
Forest	71.0	1.57
Urban	14.5	1.28
All	12.8	0.40
Ml		
Forest	90.5	0.59
Urban	53.5	2.62
All	40.5	0.69
OR		
Forest	66.5	0.60
Urban	26.4	2.33
All	41.6	0.51
UT		
Forest	52.0	0.68
Urban	9.1	1.68
All	27.4	0.47

For example, the smallest difference between the two modeling approaches was observed in the MI study area where the pseudo R^2 were on average 0.89 and 0.87 for the random forest model and beta regression model respectively. Conversely, the UT study had the largest difference where, on average, the random forest and beta regression models had pseudo R^2 of 0.68 and 0.59, respectively. Regardless of modeling technique, the use of normalized Landsat imagery did not significantly improve model fit based on the pseudo R^2 (Figure 1). The largest gain was observed in the OR study area where the pseudo R^2 increased from about 0.80 to 0.81 when the normalized data were used. Pseudo R^2 was relatively stable across sampling intensities, however the stability was more pronounced for the beta regression modeling approach as compared to the random forest models. The KS study areas was the exception where there was a relatively large increase (0.06) in pseudo R^2 between the 0.2X and 0.4X sampling intensities (Figure 1).

As expected, RMSE behaved inversely proportional to pseudo-R^2, and for this reason RMSE results are presented in brevity. Using the normalized data did not significantly decrease RMSE for either modeling approach across study areas therefore only the results from the non-normalized data are presented in Figure 2. Across all study areas, as sampling intensity increased RMSE decreased. However, these decreases were marginal, generally 1 percent to 3 percent (Figure 2). Also, the random forest models had lower RMSE than the beta regression models (generally 1

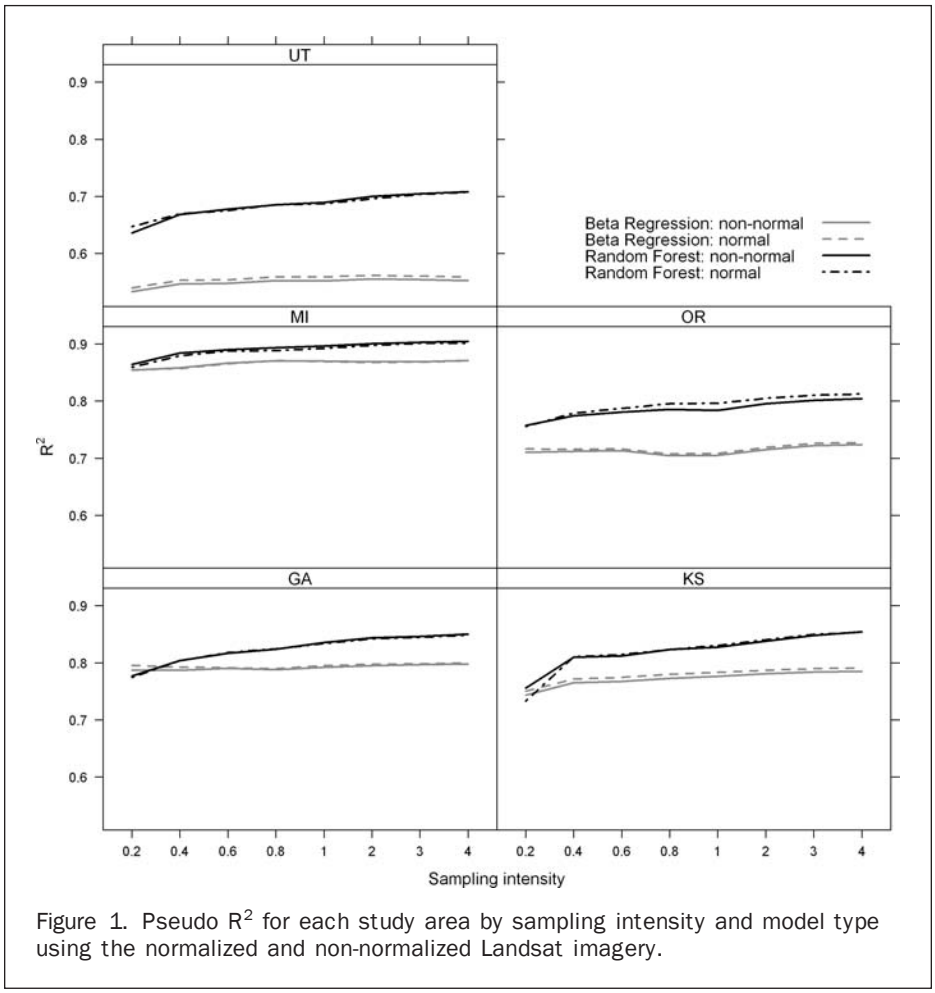

Figure 1. Pseudo R^2 for each study area by sampling intensity and model type using the normalized and non-normalized Landsat imagery.

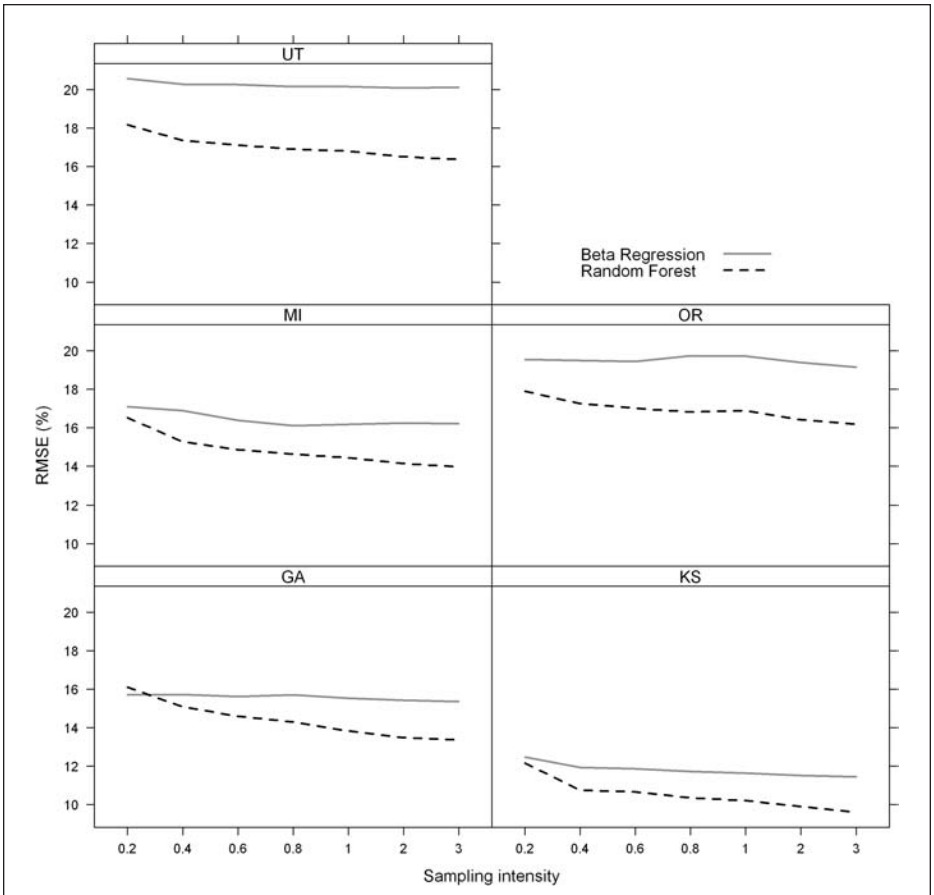

Figure 2. Root Mean Square Error (RMSE) for each study area by sampling intensity and model type using the non-normalized Landsat imagery.

percent to 3 percent). Models with the highest RMSE were observed in the UT study area, and models with the lowest RMSE were observed in the KS study areas (Figure 2).

We examined the slope and intercept of the observed versus predicted regression line for each model (0.2X to 3X); the desired values of the slope and intercept were one and zero, respectively. There was little difference between models developed using the normalized data as compared models developed using the non-normalized data, so Figures 4 and 5 display only results based on the non-normalized data. When considering the beta regression model, slopes were typically greater than one and intercepts were less than zero (Figures 4 and 5). This pattern was consistent across sampling intensities. Although in the KS study area, slopes and intercepts of the beta regression models did move towards one and zero, respectively, as sample intensity increased. Intercept values less than zero indicated a tendency to over-predict canopy cover at the low end of the distribution. Slope values greater than one can indicate under-prediction at the upper end of the distribution but this was dependent upon the intercept. When considering the random forest models, slopes were generally closer to one when compared to the beta regression models. In most of the study areas the slope approached one at sampling intensities greater than 0.6X (Figure 3). The KS study area was the exception where the slope stabilized around a 2X sampling intensity. With the exception of the GA study area, the intercepts of observed versus predicted regression lines for random forest models were approximately zero at sampling intensities at least 0.8X (Figure 4). In the GA study area, the intercept of the observed versus predicted regression line stabilized at approximately -2 at the 0.6X sampling intensity.

CDFs were used to examine whether each model was able to reproduce the distribution of percent tree canopy cover observed in the sample data (Figure 5). There was no appreciable difference between results from the same modeling approach when comparing normalized and non-normalized data therefore only results based on the non-normalized data are presented in Figure 5. In general, both beta regression models and random forest models under-estimated the proportion of observations in the tails of the distribution but the magnitude of the underestimation varied by study area and modeling approach. The beta regression model using the non-normalized data for the MI study area provides an example where approximately 40 percent of the observations had 0 percent canopy, but based on the beta regression model about 1 percent of the predictions had 0 percent canopy cover. In the upper tail of the distribution, the observed data had about 75 percent of the observations with less than 100 percent canopy cover (i.e., 25 percent with 100 percent canopy cover). However, the beta regression model using the non-normalized data had 95 percent of the observations with less than 100 percent canopy cover (i.e., 5 percent with 100 percent canopy cover). For the MI study area based on the random forest model, the results

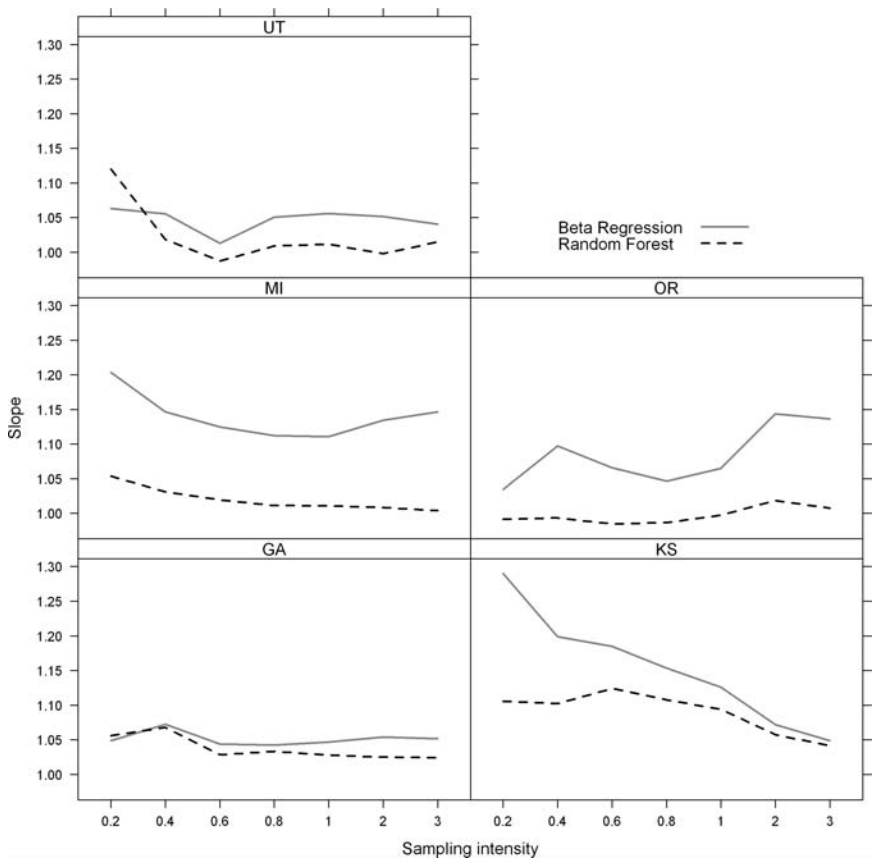

Figure 3. Slope of the observed versus predicted regression line for each study area by sampling intensity and model type using the non-normalized Landsat imagery.

improved. Approximately 35 percent of the predicted values were 0 percent canopy cover (as compared to 40 percent of the observed values) and about 13 percent of the predicted values were 100 percent canopy (as compared to 25 percent of the observed values). In general, the random forest models performed better when comparing observed and predicted CDFs. There was also no appreciable difference in CDFs between models developed with different sampling intensities. When examining CDFs, the clear pattern was that the random forest model performed better than the beta regression models and that sampling intensity and data normalization did not influence the model behavior with respect to CDF.

We examined the map-based CDF for each modeling approach, study area, sampling intensity, and data normalization option. These results were the same as the model-based results (i.e., Figure 5 was the same regardless of using the model output for each observation or the map output). This indicated that the sample did capture the variability in the explanatory maps. While there was evidence of overestimation in the lower tail of the distribution and underestimation at the upper tail of the distribution, the map-based means (across all 2001 NLCD land cover classes) were within two standard errors of the sample-based means (Figure 6) for both modeling approaches and using either the normalized or non-normalized data. Results are only shown for the 0.6X sampling intensity and non-normalized data because sampling intensity and data normalization displayed little influence. The results were more variable when considering the forest land cover class based on the beta regression models, where map-based estimates were more than two standard errors away from the sample-based estimate for the KS and MI study areas. The random forest models performed slightly better for the forest land cover class, where only the map-based estimate for the MI study area was more than two standard errors away from the sample mean (Figure 6). With respect to the urban land cover class, the beta regression model performed slightly better than the random forest model because three of the five study areas had map-based means that fell within two standard errors of the sample mean whereas the random forest model only had two (Figure 6). The largest difference was observed in the MI study area where the sample based estimate was 53.5 percent (standard error 2.62 percent) and the map-based estimates were 43.4 percent and 39.2 percent based on the beta regression model and random forest model, respectively, using the non-normalized data.

In summary, using the normalized data did not significantly improve models and overall the random forest models tended to outperform the beta regression models. The sampling intensity influenced the slope and intercept of the observed versus predicted regression lines and to a lesser extent pseudo-R^2 and RMSE. For the random forest models, these impacts tended to stabilize at sampling intensities of 0.6X for all study areas except KS where the

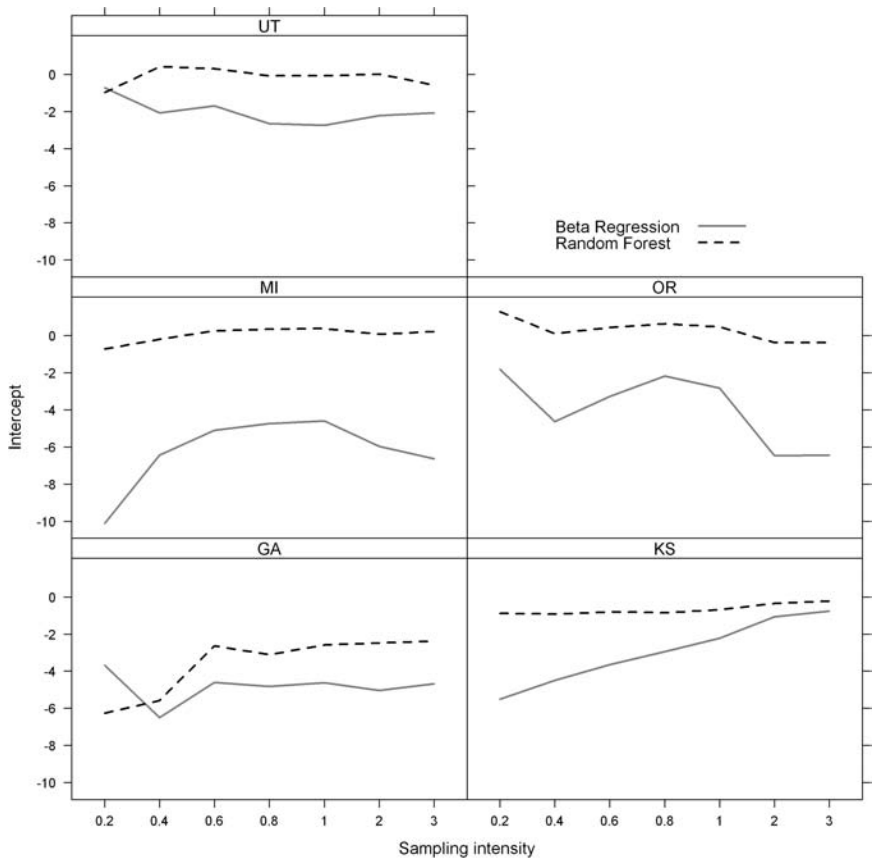

Figure 4. Intercept of the observed versus predicted regression line for each study area by sampling intensity and model type using the non-normalized Landsat imagery.

slope stabilized at 2X. The 0.6X and 2X sampling intensities equated to approximately 600 and 2,000 observations, respectively. The sampling intensity had little influence on CDFs or on the differences between map-based and sample-based estimates of mean percent tree canopy cover. These summary findings narrowed the options for the final stage of the analysis which was to combine data across study areas to construct models. To accomplish this we randomly drew 1,000 samples for the eastern US study areas (333 samples from each GA, KS, and MI) and 1,000 samples from the western study areas (500 samples from each OR and UT). We then constructed two regional random forest models using the non-normalized data. Likewise, we drew 1,000 samples across all study areas (200 sample from each) to construct a single US random forest model using the non-normalized data. Overall, the results from the regional models were comparable to those results observed based on individual models constructed for each pilot area for the 1X study (Table 2). For example, the pseudo R^2 values were typically the same regardless of whether the 1X pilot models or the regional models were used. The largest difference was observed in the GA study area where the pseudo R^2 based on the regional and individual pilot area (1X sample) were 0.80 and 0.83, respectively. When considering the US model, pseudo R^2 values were slightly less than either the regional models or the pilot area models (Table2). The magnitudes of these differences in pseudo R^2 were typically -0.02. The CDFs based on the regional and US models were comparable to those observed based on individual pilot area models (Figure 7). These results suggest that suitable models can be constructed over relatively broad and diverse geographic areas based on relatively few samples.

Discussion

Our approach to modeling percent tree canopy cover had some similarities and some differences to that described by Homer *et al.* (2004) for the NLCD 2001 product. We developed our response data using a probabilistic design, where the approach used for the 2001 NLCD percent tree canopy cover models can best be described as purposive. Also, our response data were developed from manual photo-interpretation of natural and false color images, whereas an automated approach using panchromatic images was used for the 2001 effort. The response data for the 2001 NLCD production and this pilot study were both developed from 1m resolution imagery. Even though the sampling approaches differed, the overall sampling intensity (on an areal basis) was similar between the 2001 effort and the 1X sample used in our research. We used similar predictor variables as were used in the past; however we only used leaf-on Landsat imagery rather than multi-season imagery used for the 2001 effort. We also used the 2001 percent tree canopy cover estimates as an explanatory variable. The use of 2001 dataset did not improve overall model fits but did help with areas of no canopy cover. The empirical models we developed had

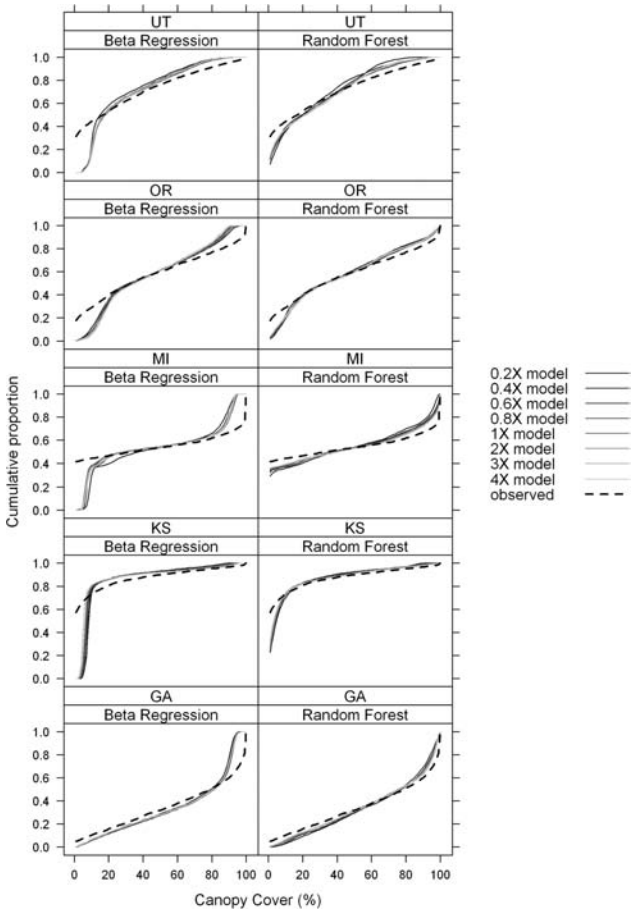

Figure 5. Cumulative distribution function of the observed canopy cover and predicted canopy cover for each sampling intensity by study area and model type using the non-normalized Landsat data.

similar or in some cases improved fit statistics to those reported by Homer *et al.* (2004). However, a true comparison of model fit between this pilot effort and the 2001 effort could not be made because of the purposive sampling approach used in the 2001 effort (see Duane *et al.*, 2010 for example).

Greenfield *et al.* (2009), and Nowak and Greenfield (2010) provided accuracy assessments of the 2001 NLCD percent tree canopy cover product and, as described earlier, they found that map-based estimates were consistently lower than sample-based estimates from photo interpretation. Our results indicated that sample-based and map-based estimates were within sampling error when all land cover classes were considered. For all study area estimates except MI, map and sample-based estimates were comparable for the forest land cover class. The MI study area also exhibited differences between map-based and sample-based estimates in the urban land cover class. Part of the issue with the MI study areas was the distribution of observed canopy cover. Over 30 percent of the MI study area was water and the northern portion of the study area was heavily forested (typically approaching 100 percent canopy cover). These "heavy-tailed" distributions of the response data were difficult to model as indicated particularly in the MI and KS study areas (Figure 5). Because the CDFs were virtually the same when the empirical models were applied to the sample data and the map data, this indicated that the sample covered the variability in the maps of explanatory variables. However, the choice of explanatory variables did not explain all the variation in the response. This concept was key, because the overestimation of canopy cover in the low-end of the distribution and the underestimation of canopy cover in the high-end of the distribution is characteristic of most parametric modeling approaches because they will tend to estimate toward the central tendency (i.e., the mean) as the percent variance explained by the model decreases. In the case of the random forest models, recall that the final estimates were the average estimate over all trees in the forest. This type of bootstrap estimate was in fact a smoothing algorithm which dampened the thickness of the tails. With future canopy-cover modeling projects, we recommend that canopy cover samples only be drawn for land, and that water, particularly large water bodies, be excluded. This would partially alleviate some of the issues observed in the MI study area.

Cost is a primary concern when developing national-scale geospatial products. With respect to percent tree canopy cover, the major costs are associated with the sampling intensity of the response data and the normalization of the Landsat imagery. Our results suggest that about 1,000 samples were sufficient to develop empirical models of percent tree canopy cover across broad and diverse geographic areas. In fact, little gain in model performance was observed by intensifying the base FIA sample. In a separate simulation analysis presented by Tipton *et al.* (in press), similar results were observed regarding sampling intensity and Moisen *et al.* (in press) observed that study areas could be combined with minimal influence on model fit statistics. The reason for observing only modest gains in model performance at increased sample size is because model performance is only partially related to the sample size. The important question, when considering model development, is whether the variability in the explanatory variables accounts for the variability in the response variable. When there are strong relationships between the response and explanatory variables, fewer samples are required for model development. Alternatively, when the variability in the explanatory variables does not explain all the variability in the response, increasing the sample size again may have little influence on the model fit. Regardless of which scenario is the case for this study, the small gains in model performance were not worth the additional cost of collecting those data.

Normalizing the Landsat scenes within each study area was also a high cost component of this research. The initial hypothesis was that different atmospheric conditions between and among path / rows in each study area would create explanatory variables that differed in scale in different parts of the study areas and therefore model performance would decline. The normalization process actually had little influence on model performance. The likely reason for this behavior was that the minor changes in scale in the reflectance values were small in relation to the overall variability of the reflectance values across the study area. On the surface, these minor gains in model performance do not warrant the additional cost of the normalization process. However, we must also consider the visual appearance of the final geospatial product. Visually perceptible differences between reflectance values of adjacent scenes (i.e., seam lines), while not statistically significant, resulted in noticeable artifacts in the final tree canopy cover product. The Landsat data for the UT study area had the most pronounced boundary. When a modeled map of percent tree canopy cover was created using the non-normalized data, the boundary was clearly evident (Plate 2). However, the same boundary was not observed when using the normalized data

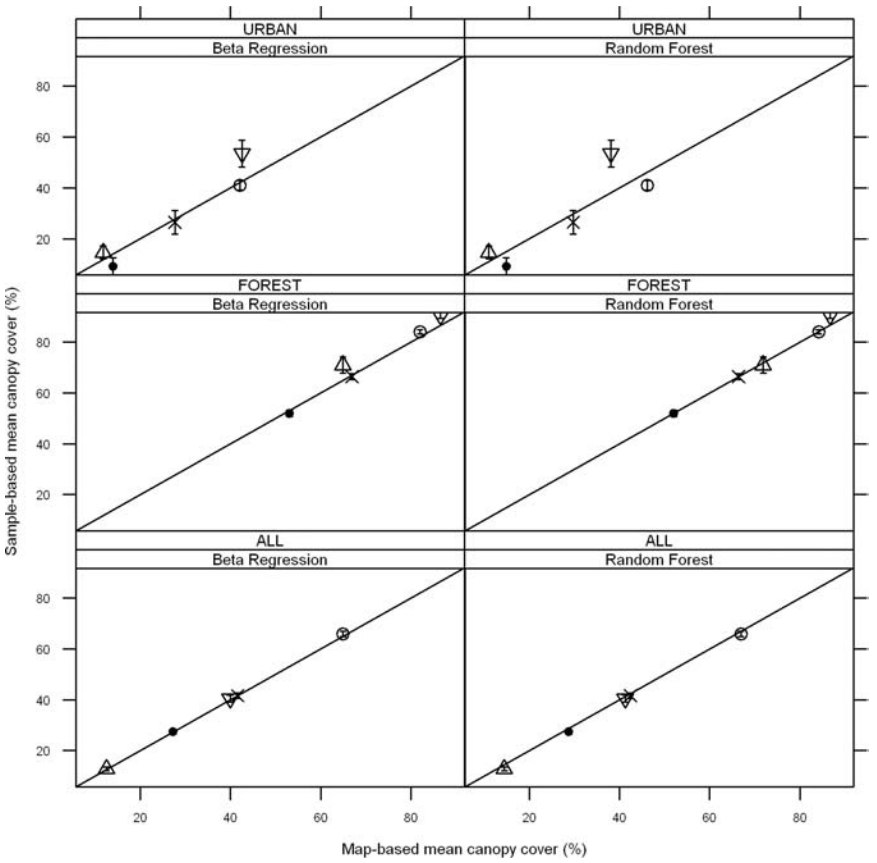

Figure 6. Sample-based mean percent canopy cover versus map-based mean percent canopy cover for the 0.6X sampling intensity by study area, model type, and 2001 land cover class using the non-normalized Landsat imagery. ○ = GA, Δ = KS, ∇ = MI, X = OR, • = UT. The vertical line denotes two standard errors and the diagonal line is the 1:1 line.

(Plate 2). In reality, each map has about the same accuracy but as a practical matter geospatial data are also judged on appearance and overlay with higher resolution imagery and the emergence of arbitrary seam lines causes general concern among potential users.

As with all modeling applications, estimates of percent tree canopy cover have error. However, unlike a regression scenario outside the spatial domain, modeled maps produced at 30 m can be visually inspected by overlaying finer resolution imagery. When this is done, one inevitably finds locations where low values of modeled percent canopy cover where the value should actually be zero (e.g., agricultural fields, shore lines). In the previous national canopy cover mapping effort Homer *et al.* (2004) created a liberal forest mask to force estimates of canopy cover to zero in areas that clearly had no trees. This approach, while straight forward, introduces additional error into the final map because both the mask and the original percent canopy cover map have error, and it is difficult to know how these errors propagate when the two datasets are combined. This masking procedure may also have contributed to the underestimation of tree canopy cover reported by Nowak and Greenfield (2010). A seemingly more appropriate approach is to provide both a map of the percent canopy cover estimates and the standard errors for each estimate as a separate dataset or map layer. This poses some technical challenges because the development of the standard error estimator for random forests is just beginning to be examined (Sexton and Laake, 2009). Freeman *et al.* (2010) provide an alternative that may be applicable for producing estimates of uncertainty around percent canopy cover estimates from random forests models. Other ensemble modeling techniques, such as stochastic gradient boosting, may also allow greater flexibility that could reduce the need for masking while maintaining adequate model fits. One area of further research is to develop alternatives to masking. This may be accomplished by producing uncertainty estimates or using alternative modeling techniques.

In this pilot study we used 1 m resolution NAIP imagery to develop our response variable. This has several implications. Canopy cover estimates are scale dependent. In other words an estimate of cover depends on the ability to resolve a pattern that changes as the resolution changes (Coulston *et al.* 2010, Jennings *et al.* 1999). That means that our estimates of cover will likely differ from estimates made *in situ* or from 0.5 m or 5 m imagery, for example (Frescino and Moisen, In press). These differences can arise for several reasons. One reason is that at finer scales the ability to identify gaps in the canopy (and distinguish gaps from shadows) will generally improve. Another reason is that with finer scale imagery the photo-interpreters' ability to distinguish woody plants that can achieve tree-form and

Table 2. Pseudo R^2 and the Slope and Intercept of the Observed versus Predicted Regression Line for Each Study Area based on the Regional and US Model

	Model extent	
Study Area	Regional	US
GA		
R^2	0.80	0.78
intercept	−4.73	2.68
slope	1.06	0.97
KS		
R^2	0.83	0.81
intercept	−0.99	0.26
slope	0.98	0.96
Ml		
R^2	0.89	0.87
intercept	0.43	2.54
slope	1.04	1.01
OR		
R^2	0.78	0.75
intercept	−0.68	−0.52
slope	1.01	0.99
UT		
R^2	0.69	0.65
intercept	0.50	2.44
slope	1.02	1.05

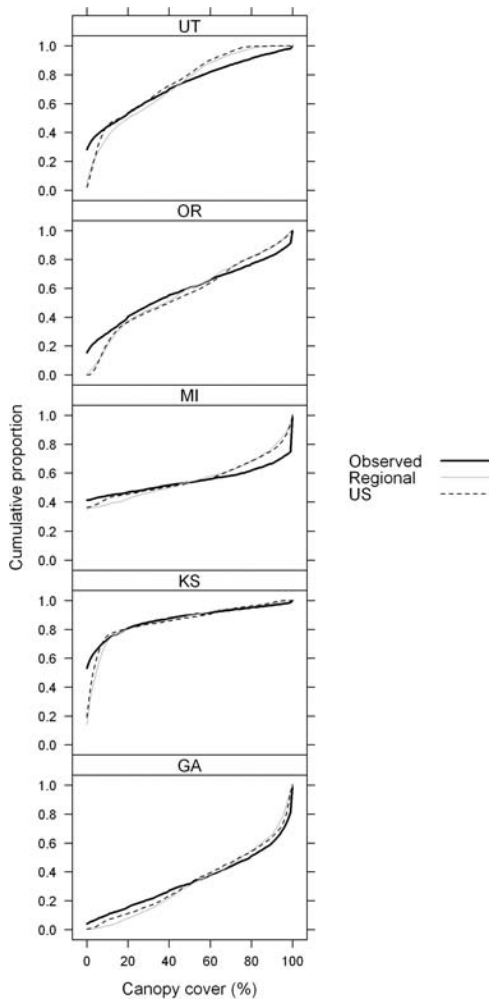

Figure 7. Observed and predicted cumulative distribution functions for each study are based on the regional and US random forest model.

those that cannot achieve tree-form will also increase. Another potential issue arose because the NAIP imagery was not available in stereo. Canopy cover is defined as the vertical projection of the crown, and this can only be assessed at nadir, although Avery and Burkhart (1994) note that angles less than 3 degrees may be considered vertical for all practical purposes. Further research is needed to examine the influence of off-nadir photo interpretation on estimates of percent canopy cover, particularly in steep terrain. Regardless of the potential challenges of using NAIP imagery for developing canopy cover model, NAIP imagery provides a consistent set of information across the coterminous United States. While other data may provide better estimates of tree canopy cover (e.g., lidar), it is currently cost-prohibitive to acquire these data for national-scale products.

Conclusions and Prototype Design

The approach for modeling percent tree canopy cover presented here yielded encouraging results. First, the sampling approach for developing the response data was appropriate for model development at relatively low sampling intensities. Second, the random forest modeling technique outperformed the beta regression approach, which will be beneficial in a production environment because the random forest modeling approach does not require the practitioner to test individual explanatory variables for significance as is common with traditional regression approaches. From a statistical perspective, model fits using the normalized Landsat imagery were equivalent to model fits using the non-normalized imagery. However, from an end-user perspective, the normalization process minimized the emergence of seam lines. Based on the findings presented here, the US Forest Service plans to move forward with a broader-scale prototype study. The prototype study covers five MRLC mapping zones (Homer and Gallant, 2001): three in the southern US and two in the western US (Plate1). The southern prototype area generally falls in the piedmont physiographic zone and the western prototype area generally falls in the Colorado plateaus physiographic zone. The prototype areas have substantial overlap with the GA and UT pilot study areas presented here. This will allow for a more direct comparison between pilot effort results and prototype results. Each prototype area will be sampled at the 0.2X sampling intensity described in this paper. This sampling intensity yields approximately 1,550 and 1,800 observations in the western and southern prototype areas, respectively. The results from the prototype study will set the specifications for the final production environment and the timely development of the 2011 NLCD percent tree canopy cover layer.

Acknowledgments

Support for this project was provided by the USDA Forest Service Forest Inventory and Analysis national techniques research group and the USDA Forest Service Remote Sensing Applications Center. The author's also thank Chris King (US Forest Service) for a technical edit of this manuscript.

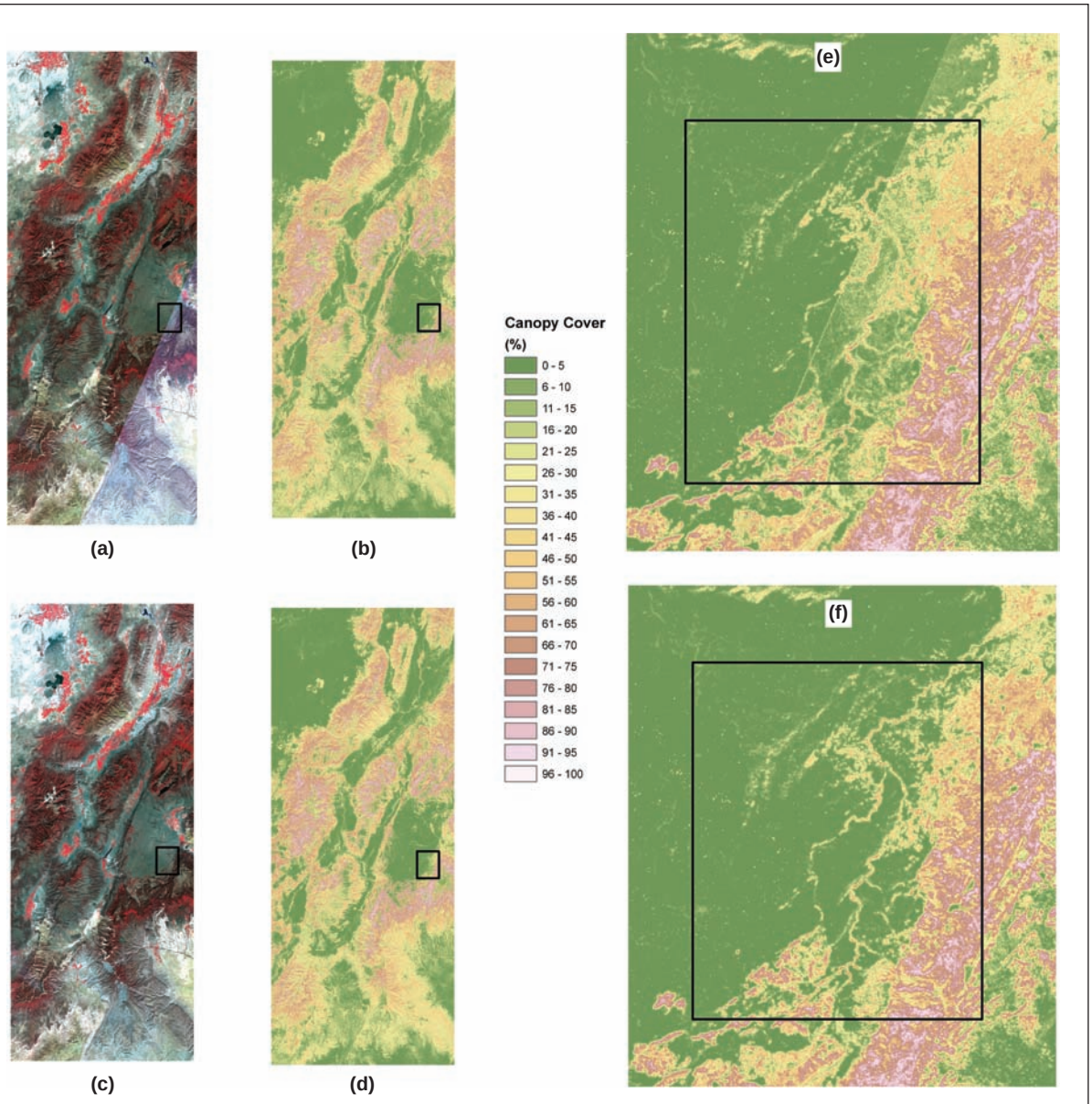

Plate 2. Example of map seam lines in UT study area: (a) is the non-normalized Landsat data, (b) is the normalized Landsat data, (c) and (d) are the modeled percent tree canopy cover maps using (a) and (b), respectively, (e) displays the influence of the using non-normalized Landsat data for the area of interest denoted by the black box, and (f) displays how the normalization process improves the visual appearance.

References

Avery, T.E., and H.E., Burkhart, 1994. *Forest Measurements*, Fourth edition, McGraw-Hill, New York, 408 p.

Beaty, M., M. Finco, and K. Brewer, 2010. *Using Model II Regression to Radiometrically Normalize Landsat Scenes for the Purpose of Mosaicing*, USDA Forest Service, RSAC-10012-RPT1, Remote Sensing Applications Center, Salt Lake City, UTAH, 6 p.

Bechtold, W.A., and P.L. Patterson (editors), 2005. *The Enhanced Forest Inventory and Analysis Program - National Sampling Design and Estimation Procedures*, USDA Forest Service, General Technical Report SRS-80, Southern Research Station, Asheville, North Carolina, 85 p.

Blackard, J.A., M.V. Finco, E.H. Helmer, G.R. Holden, M.L. Hoppus, D.M. Jacobs, A.J. Lister, G.G. Moisen, M.D. Nelson, R. Riemann, B. Ruefenacht, D. Salajanu, D.L. Weyermann, K.C. Winterberger, T.J. Brandeis, R.L. Czaplewski, R.E. McRoberts, P.L. Patterson, and R.P. Tymcio, 2008. Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information, *Remote Sensing of Environment*, 112(4): 1658–1677.

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

Cribari-Neto, F., and A. Zeileis, 2010. Beta regression in R, *Journal of Statistical Software*, 34(2):1–24.

Cochran, W.G., 1977. *Sampling Techniques*, Third edition, John Wiley & Sons, New York, 428 p.

Coulston, J.W., S.N. Oswalt, A.B. Carraway, and W.B. Smith, 2010. Assessing forestland area based on canopy cover in a semi-arid region: A case study, *Forestry,* 83(2):143–151.

Duane, M.V., W.B. Cohen, J.L. Campbell, T. Hudiburb, D.P Turner, and D.L. Weyermann, 2010. Implications of alternative field sampling designs for Landsat-based mapping of stand age and carbon stocks in Oregon forests, *Forest Science*, 56(4):405–416.

Freeman, E.A., T.S. Frescino, and G.G. Moisen, 2010. ModelMap: An R package for model creation and map production, URL: *http://cran.r-project.org/web/packages/ModelMap/vignettes/VModelMap.pdf* (last date accessed: 29 March 2012).

Ferrari, S., and F. Cribari-Neto, 2004. Beta regression for modeling rates and proportions, *Journal of Applied Statistics*, 31(7): 799–815.

Frescino, T.S., and G.G. Moisen, in press. Comparing alternative tree canopy estimates derived from digital aerial photography and field-based assessments, *Proceedings of the 2010 Forest Inventory and Analysis Symposium*, 12-14 October 2008, Knoxville, Tennessee, USDA Forest Service, General Technical Report SRS-xxx, Southern Research Station, Asheville, North Carolina.

Greenfield, E.J., D.J. Nowak, and J.T. Walton, 2009. Assessment of 2001 NLCD percent tree and impervious cover estimates, *Photogrammetric Engineering & Remote Sensing*, 75(11): 1279–1286.

Homer, C., J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J.N. VanDriel, and J. Wickham, 2007. Completion of the 2001 National Land Cover Database for the conterminous United States, *Photogrammetric Engineering & Remote Sensing*, 73(4):337–341.

Homer, C.G., and A. Gallant, 2001. Partitioning the conterminous United States into mapping zones for Landsat TM land cover mapping, US Geological Survey, URL: *http://landcover.usgs.gov/pdf/homer.pdf* (last date accessed: 29 March 2012).

Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan, 2004. Development of a 2001 National Land Cover Database for the United States, *Photogrammetric Engineering & Remote Sensing*, 70(7):829–840.

Huang, C., L. Yang, B. Wylie, and C. Homer, 2001. A strategy for estimating tree canopy density using Landsat 7 ETM and high resolution images over large areas, *Proceedings of the Third International Conference on Geospatial Information in Agriculture and Forestry*, 05-07 November, Denver, Colorado, unpaginated CD-ROM.

Jennings, S.B., N.D. Brown, and D. Sheil, 1999. Assessing forest canopies and understory illumination: Canopy closure, canopy cover and other measures, *Forestry,* 72(1):59–73.

Kellndorfer, J.M., W. Walker, E. LaPoint, M. Hoppus, and J. Westfall, 2006. Modeling height, biomass, and carbon in US forests from FIA, SRTM, and ancillary national scale data sets, *Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 31 July - 04 August 2006, Denver, Colorado, pp. 3591–3594.

Korhonen, L., K.T. Korhonen, P. Stenberg, M. Maltamo, and M. Rautiainen, 2007. Local models for forest canopy cover with beta regression, *Silva Fennica,* 41(4):671–685.

Liaw, A., and M. Wiener, 2002. Classification and Regression by random Forest, *R News*, 2(3):18–22.

McRoberts, R.E., G.R. Holden, M.D. Nelson, G.C. Liknes, and D.D. Gormanson, 2006. Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service, *Canadian Journal of Forest Research,* 36:2968–2980.

Moisen, G.G., J.W. Coulston, B.T. Wilson, W.B. Cohen, and M.V. Finco, In press. Choosing appropriate subpopulations for modeling tree canopy cover, *Proceedings of the 2010 Forest Inventory and Analysis Symposium*, 12-14 October 2008, Knoxville, Tennessee, (USDA Forest Service, General Technical Report SRS-xxx, Southern Research Station, Asheville, North Carolina).

Nowak, D.J., D.E. Crane, and J.C. Stevens, 2006. Air pollution removal by urban trees and shrubs in the United States, *Urban Forestry and Urban Greening*, 4(3-4):115–123.

Nowak, D.J., and E.J. Greenfield, 2010. Evaluating the National Land Cover Database tree canopy and impervious cover estimates across the coterminous United States: A comparison with photo-interpreted estimates, *Environmental Management*, 46(3):378–390.

Nowak, D.J., R.A. Rowntree, E.G. McPherson, S.M. Sisinni, E.R. Kerkmann, and J.C. Stevens, 1996. Measuring and analyzing urban tree cover, *Landscape and Urban Planning*, 36(1):49–57.

Pineiro, G., S. Perelman, J.P. Guerschman, and J.M. Paruelo, 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed, *Ecological Modeling*, 216(3-4):316–322.

R Development Core Team, 2010. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL *http://www.R-project.org* (last date accessed: 29 March 2012).

Rollins, M.G., and C.K Frame (editors), 2006. *The LANDFIRE Prototype Project: Nationally Consistent and Locally Relevant Geospatial Data for Wildland Fire Management*, USDA Forest Service, General Technical Report RMRS-GTR-175, Rocky Mountain Research Station, Fort Collins, Colorado, 416 p.

Ruefenacht, B., M.V. Finco, M.D. Nelson, R. Czaplewski, E.H. Helmer, J.A. Blackard, G.R. Holden, A.J. Lister, D. Salajanu, D. Weyermann, and K. Winterberger, 2008. Conterminous US and Alaska forest type mapping using forest inventory and analysis data, *Photogrammetric Engineering & Remote Sensing*, 74(11):1379–1388.

Sexton, J., and P. Laake, 2009. Standard errors for bagged and random forest estimators, *Computational Statistics and Data Analysis*, 53(3):801–811.

Slama, C.C, 1980. *Manual of Photogrammetry*, Fourth edition, American Society of Photogrammetry, Falls Church, Virginia, 1,056 p.

Smithson , M., and J. Verkuilen, 2006. A better lemon squeezer?, Maximum-likelihood regression with beta-distributed dependent variables, *Psychological Methods*, 11(1):54–71.

Suganuma, H., Y. Abe, M., Taniguchi, H. Tanouchi, H. Utsugi, T. Kojima, and K. Yamada, 2006. Stand biomass estimation method by canopy cover for application to remote sensing in an arid area of Western Australia, *Forest Ecology and Management*, 222(1-3):75–87.

Tipton, J., G. Moisen, P. Patterson, T.A. Jackson, and J. Coulston, in press. Sampling intensity and normalizations: Exploring cost-driving factors in nationwide mapping of tree canopy cover, *Proceedings of the 2010 Forest Inventory and Analysis Symposium*, 12-14 October 2008, Knoxville, Tennessee, (USDA Forest Service, General Technical Report SRS-xxx, Southern Research Station, Asheville, North Carolina).

U.S. Department of Agriculture, 2009. National Agriculture Imagery Program, URL: *http://www.apfo.usda.gov/FSA/apfoapp?area=home&subject=prog&topic=nai*, USDA Farm Service Agency, Aerial Photography Field Office, Salt Lake City, Utah (last date Accessed: 29 March 2012).

Webb, B.W., and D.T. Crisp, 2006. Afforestation and stream temperature in a temperate maritime environment, *Hydrological Processes*, 20(1):51–66.

White, D., J. Kimerling, and S.W. Overton, 1992. Cartographic and geometric components of a global sampling design for environmental monitoring, *Cartographic and Geographic Information Systems*, 19(1):5–22.

(Received 06 August 2011; accepted 28 October 2011; final version 06 December 2011)