CHOOSING APPROPRIATE SUBPOPULATIONS FOR MODELING TREE CANOPY COVER NATIONWIDE

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

ABSTRACT

In prior national mapping efforts, the country has been divided into numerous ecologically similar mapping zones, and individual models have been constructed for each zone. Additionally, a hierarchical approach has been taken within zones to first mask out areas of nonforest, then target models of tree attributes within forested areas only. This results in many models nationwide, which reduces the number of training points per model, increases the cost of the process, results in numerous seam lines, and complicates validation efforts. Consequently, we use response data based on photo-interpreted aerial photography and spatially continuous predictor data (Landsat imagery, topographic and other ancillary data) in five pilot areas across the country to explore the effect of the choice of modeling subpopulation on models of tree canopy cover. Using Random Forests as our predictive tool, we explore the consequences of modeling pilot areas alone, modeling groups of pilot areas, and modeling hierarchically within each pilot area. Recommendations are made for appropriate modeling subpopulations to be used in a nationwide tree canopy cover map.

INTRODUCTION

The Multi-Resolution Land Characteristics (MRLC, http://www.mrlc.gov/) consortium has developed plans for the 2011 National Land Cover Dataset (NLCD) which will include an approximate Anderson Level II classification, percent impervious surface, and percent tree canopy cover. Because it is central to its business needs, the US Forest Service, Forest Inventory and Analysis (FIA) program has assumed responsibility for the latter, and will be developing this Tree Canopy Cover (TCC) layer. Recently a national pilot project was launched to test the use of high resolution photography acquired though the National Agriculture Imagery Program (NAIP) coupled with extensive ancillary data layers through alternative sampling and modeling methodologies in support of this commitment. A number of studies have resulted from initial pilot analyses answering questions about alternative means to observe tree canopy cover (Frescino and others 2011), relationship between photo-based tree canopy cover and canopy modeled from FIA plots (Toney and others 2011), repeatability in photo-interpretation (Jackson and others 2011), efficient sampling strategies (Jackson and others 2011), and, in this paper, choice of appropriate subpopulations over which to construct predictive models.

Tree canopy cover in the conterminous U.S. is remarkably diverse. Previous nationwide mapping efforts, like that of the US Forest Biomass map (Blackard and others 2008) , nationwide forest type and forest type group maps (Ruefenacht and others 2008), as well as Landfire (Rollins and Frame 2006) have tried to accommodate this diversity by using 66 different mapping zones (Homer and Gallant 2001, Figure 1). In these efforts, mapping zones were modeled independently and in some cases forest masks were first developed, then models developed solely for the areas predicted to be forest. With most FIA mapping efforts, the sampling intensity of the training data is fixed at the nominal sampling intensity of the base FIA program (approximately 1 plot per 6000 ac). Therefore developing models for relatively small mapping zones decreases the number of training points available. Additionally, when small mapping zones are used the number of models increases which results in increased cost, seamline issues, and complicated validation approaches. Consequently, we used photo-interpreted data collected in five pilot areas in the conterminous United States to explore the effect of modeling over larger, more geographically diverse areas, as well as the value of empirically masking non-tree areas prior to modeling tree canopy cover.

METHODS

DATA

Photo-interpreted percent tree canopy cover data from NAIP imagery was collected in 5 diverse pilot areas in the United States, including areas in Oregon, Utah, Kansas, Michigan and Georgia. Photo plots were collected on the

Gretchen G. Moisen, Research Forester, US Forest Service Rocky Mountain Research Station, Ogden, UT, 84401
John W. Coulston, Research Forester, US Forest Service Southern Research Station, Knoxville, TN, 37919
Barry T. Wilson, Research Forester, US Forest Service Northern Research Station, St. Paul, MN, 55108
Warren B. Cohen, Research Forester, US Forest Service Pacific Northwest Research Station, Corvallis, OR, 97331
Mark V. Finco, Contract Leader, Red Castle Resources, US Forest Service Remote Sensing Applications Center, Salt Lake City, UT, 84119

FIA grid, intensified 4-fold, and each photo plot consisted of 105 dots distributed in a 90 m square area (Figure 2). Each dot was characterized as being a tree or not-a-tree, then the proportion of tree dots were summarized for each plot. This percent tree cover was used as the response variable in models described below. Predictor variables included Landsat-5 reflectance bands, 30 m elevation, transformed aspect, slope, topographic positional index, land cover from the 2001 NLCD, and Bailey's ecoregions. Because many of the predictor variables originated from 30 m products, assignment to each 90 m plot was accomplished by taking a focal mean over a 3x3 window for continuous variables, and focal majority for the categorical variables. In addition, the standard deviations for all continuous predictor variables within the 3x3 window were included as predictor variables. Following findings presented in Tipton and others (2011), a subset of the total data available equivalent to the intensity of the FIA grid was used for modeling, and an equal size independent test set used for testing in these analyses.

MODEL

Classification and regression trees (Breiman and others 1984) are flexible and robust tools that are well suited to the task of modeling the relationship between a response and a set of explanatory variables for the purposes of making spatial predictions in the form of a map. These are intuitive methods, often described in graphical or biological terms. Typically shown growing upside down, a classification or regression tree begins at its root. An observation passes down the tree through a series of splits, or nodes, at which a decision is made as to which direction to proceed based on values of the explanatory variables. Ultimately, a terminal node or leaf is reached and predicted response is given, the mean of observations in the node for a continuous response, or a vote for a categorical response. (See De'ath and Fabricius 2000 for a thorough explanation, and Moisen 2008 for a simple overview.)

Although classification and regression trees are powerful tools by themselves, much work has been done in the data mining and machine learning fields to improve the predictive ability of these models by combining separate tree models into what is often called a committee of experts, or ensemble. One such tool, Random Forests (Breiman 2001) is receiving increasing attention in the ecological and remote sensing literature. In this technique, a bootstrap sample of the training data is chosen. At the root node, a small random sample of explanatory variables is selected and the best split made using that limited set of variables. At each subsequent node, another small random sample of the explanatory variables is chosen, and the best split made. The tree continues to be grown in this fashion until it reaches the largest possible size, and is left un-pruned. The whole process, starting with a new bootstrap sample, is repeated 500 or more times. The final prediction is a vote (for categorical responses) or average (for continuous variables) from prediction of all the trees in the collection. All of the following analyses were fit using the "randomForest" library in R (Liaw and Wiener 2002).

SMALL VS. LARGE MAPPING ZONES

Using the training data sets described above, eight models of tree canopy cover were constructed. The first five were individual "pilot area" models for each of Georgia, Michigan, Kansas, Utah, and Oregon which only contained training data from each of their respective areas. The sixth model, called the "East" model used training data from GA, MI and KS, while the seventh "West" model used all the training data from OR, UT, and KS. The eighth model was called a "USA" model and used all the training data in all five pilot areas. These eight models were applied to the test data sets within each of the pilot area and the resulting metrics of the relationship between observed and predicted values in these test sets were compared. The metrics included: correlation, root mean squared error (RMSE), and slope of a regression line. Density plots of observed and predicted, which are like a continuous version of histograms reflecting the relative number of plots by tree canopy cover class, were also compared.

NO-TREE MASK

This analysis involves building two models for each pilot area. First, a binary response of "trees present" versus "no trees present" was modeled as a function of all the predictor variables, again using Random Forests. The probability of tree presence was predicted over the test data and these probabilities were then converted to binary "trees present" or "no trees present" using the prevalence of treed land in each area as the threshold. (See Freeman and Moisen 2008a for a discussion of thresholding options). Using these predictions over the test data, assessments were made of the tree mask using the PresenceAbsence library (Freeman and Moisen 2008b) in R. This first tree presence model was then applied to the training data so that only those training plots predicted to have trees present were included in a continuous model of tree canopy cover, the second model. To validate the effectiveness of combining the first and second models, test data plots predicted to have "no trees present" from the first model were simply given a predicted tree canopy cover of zero, while test data plots predicted to have "trees present" were then assigned tree canopy cover predictions from the second model. This, in effect, empirically masks out areas thought to have no trees present at all. Comparisons of the final predicted versus observed tree canopy values in each pilot area (including treed and non-treed land) were done using metrics as above.

RESULTS AND DISCUSSION

SMALL VS. LARGE MAPPING ZONES

Figure 3 illustrates the effect of increasing mapping unit size on map accuracy metrics, including correlation, root RMSE, and slope of a line fitted between predicted vs. observed values in the independent test set (with intercept term.) Interestingly, little difference between accuracy metrics is noted between the individual pilot area models, and models built for larger areas (East, West and USA models). Note that models built for large areas naturally included many more training plots. The only exception is in cases where model predictions were made over areas whose data were not included in that particular model. For example, the West model predicted over Michigan, or the East model predicted over Utah. In addition, density plots of the true tree canopy cover values in each pilot area were plotted along with densities obtained by applying the four classes of models (pilot, East, West and USA) to that same training data, as illustrated in Figure 4. As with the accuracy metrics, there was little difference in the densities obtained under the four modeling scenarios except in cases where no data from that particular pilot area was used in the model.

NO-TREE MASK

Figure 5 illustrates the results from the tree presence model in UT which were typical of the other pilot areas. The first graph (Figure 5a) is a Receiver Operator Curve (ROC Plot) indicating a strong model fit and high Area Under the Curve (AUC) value of 0.94. Here, sensitivity, or proportion of correctly predicted positive observations, reflects a model's ability to detect a presence, given at least one tree actually occurs at a location. Specificity, or proportion of correctly predicted negative observations, reflects a model's ability to predict an absence where trees do not exist. The second graph in this figure (Figure 5b) illustrates how measures of map accuracy change with different threshold values. In UT, approximately 70 percent of the land area had trees present. This graph illustrates how using prevalence as a threshold to convert probability predictions in to a presence-absence map resulted in maximizing map accuracy.

Plots exploring the effect of first creating a tree presence model prior to modeling tree canopy cover are illustrated in Figures 6 and 7. Figure 6 illustrates the effect of using a no-tree mask on map accuracy metrics, including correlation, root RMSE, and slope of a line fitted between predicted vs. observed values in the independent test set (with intercept term.) Little difference between accuracy metrics is noted between the unmasked, and masked approaches. In Figure 7a, predicted tree canopy cover from a single unmasked model is plotted against the tree canopy cover response from the photo interpretation illustrating the tendency to predict canopy where no trees exist at all (the zero line on the x-axis). Next in Figure 7b, predicted tree canopy cover from the tree presence model followed by the masked tree canopy model is plotted against the tree canopy cover response from the photo interpretation illustrating a slight reduction in the number observations where canopy was predicted over no-tree areas, but also an increase in errors of false negative (the zero line on the y-axis). Finally in Figure 7c, the predicted probability of having trees present from the tree presence model is plotted against predicted tree canopy cover with no masking, illustrating the strong relationship between masked and unmasked scenarios suggesting most of the necessary information may be contained in a single model. That is, an empirical mask constructed prior to modeling tree canopy cover may not be that effective in improving the final tree canopy cover map. Also shown in 7c is the prevalence-based threshold in blue (~70 percent of the Utah pilot area is treed) above which plots are predicted to have trees. In addition, the pink vertical line illustrates a threshold a user might impose by applying a 10 percent cover threshold to the predicted tree canopy cover. Interestingly, these two thresholding criteria applied to two different models identify very closely to the same sets of plots, again indicating not a lot of additional information is gained by hierarchically modeling tree/no-tree followed by tree canopy cover in an empirical fashion.

CONCLUSION

Random Forests is a flexible and robust tool for mapping tree canopy cover over large geographic areas. Although past nationwide mapping efforts have delineated many small mapping zones across the country, the analyses conducted here suggest that modeling over much larger zones does not compromise model fit. This provides an opportunity to decrease the cost of the mapping process, reduce the numerous seam lines, and simplify validation efforts. Still to be investigated, however, is the effect of modeling over larger units with decreased sampling intensity. This could further reduce sampling costs. In addition, modeling hierarchically by creating an empirical tree presence model prior to modeling tree canopy cover does not completely alleviate the problem of predicting tree canopy cover where no trees exist, and does tend to mask treed areas as no-tree erroneously. However, this does not diminish the importance of applying a variety of regionally-specific masks, such as water and impervious surface masks, to the final product.

Naturally, results from these pilot tests as well as those described in Tipton and others 2011, and Jackson and others 2011 need to be confirmed over larger geographic areas. The NLCD Tree Canopy Cover project is entering the prototype phase. In this prototype, photo interpreted as well as ancillary data are being collected in two diverse areas, one approximately 49 million acres in size in the Interior Western U.S., the other approximately 59 million acres in

size in the Southeastern U.S. Prototype tests will be run to provide yet stronger basis for production mapping which is scheduled to begin in the fall of 2011.

ACKNOWLEDGEMENTS

We are grateful for the tremendous effort given by numerous photo-interpreters in the FIA units. Without them, the NLCD 2011 Tree Canopy Cover pilot project would not have been possible. We also appreciate the staff at the Remote Sensing Application Center for helping us meet very tight timelines. Finally, we extend our thanks the entire pilot team for all the good energy and banter.

LITERATURE CITED

Blackard, J.; Finco, M; Helmer , E. [and others]. 2008. Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. Remote Sensing of Environment. 112: 1658-1677.

Breiman, L., Friedman, R. A., Olshen, R. A. and Stone, C. G. 1984. Classification and regression trees. Pacific Grove, CA, USA: Wadsworth.

Breiman, L. 2001. Random Forests. Machine learning. 45:5-32.

De'ath, G.; Fabricius, K. E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 81: 3178-3192.

Freeman, E.A.; Moisen, G.G. 2008a. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and Kappa. Ecological Modelling. 217:48-58.

Freeman, E. A.; Moisen, G.G. 2008b. PresenceAbsence R Library. Journal of Statistical Software. 23(11).

Homer, C. G.; Gallant, A. 2001. Partitioning the conterminous United States into mapping zones for Landsat TM land cover mapping. USGS White Paper available at http://landcover.usgs.gov.

Jackson, T.; Moisen, G.G.; Patterson, P.L.; Tipton, J. 2011 Repeatability in photo-interpretation of tree canopy cover and its effect on predictive mapping. In McWilliams, W., Moisen, G.G.; Roesch, F. (eds.). Forest Inventory and Analysis (FIA) Symposium 2010; October 5-7, 2010; Knoxville, TN. Proc. SRS-CD. Knoxville, TN: U.S. Department of Agriculture Forest Service, Southern Research Station. 1 CD.

Liaw, A.; Wiener, M. 2002. Classification and Regression by randomForest. R News. 2(3): 18-22.

Moisen, G. G. Classification and regression trees. 2008. In: Sven Erik Jørgensen and Brian D. Fath (Eds.), Ecological Informatics, Encyclopedia of Ecology. Volume 1: 582-588.

Rollins, M.G.; Frame, C.K., tech. eds. 2006. The LANDFIRE Prototype Project: nationally consistent and locally relevant geospatial data for wildland fire management. Gen. Tech. Rep. RMRS-GTR-175. Fort Collins: U.S. Department of Agriculture Forest Service, Rocky Mountain Research Station. 416 p.

Ruefenacht, B.; Finco, M.V.; Nelson, M.D. [and others]. 2008. Conterminous U.S. and Alaska forest type mapping using forest inventory and analysis data. Photogrammetric Engineering and Remote Sensing. 74(11):1379-1388.

Tipton, J.; Moisen, G.G.; Patterson, P.L. [and others]. 2011. Sampling intensity and normalizations: exploring cost-driving factors in nationwide mapping of tree canopy cover. In McWilliams, W., Moisen, G.G.; Roesch, F. (eds.). Forest Inventory and Analysis (FIA) Symposium 2010; October 5-7, 2010; Knoxville, TN. Proc. SRS-CD. Knoxville, TN: U.S. Department of Agriculture Forest Service, Southern Research Station. 1 CD.

Figure 1—Mapping/modeling zones (Homer & Gallant, 2001) used in previous NLCD mapping efforts.

Figure 2—Five pilot areas including one each in Georgia, Michigan, Kansas, Utah, and Oregon. Photo-based sample plots were interpreted at 4 times the FIA grid intensity within each plot area. Each photo plot consisted of 105 photo points used to estimate percent tree canopy cover on the plots.

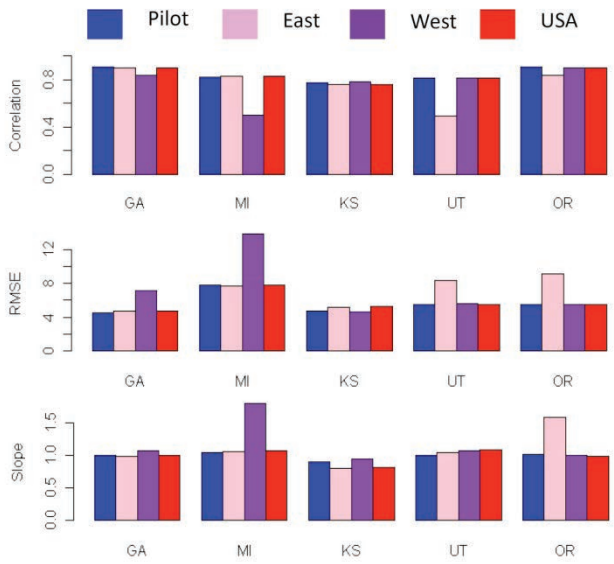

Figure 3—Correlation, RMSE, and slopes obtained in each of the five pilot areas when applying five different tree canopy models to independent test data. "Pilot" models (blue) included only training data from each individual pilot area. The "East" model (pink) included data from Georgia, Michigan, and Kansas. The "West" model (purple) included data from Oregon, Utah, and Kansas. And the "USA" model (red) included data from all the pilot areas.

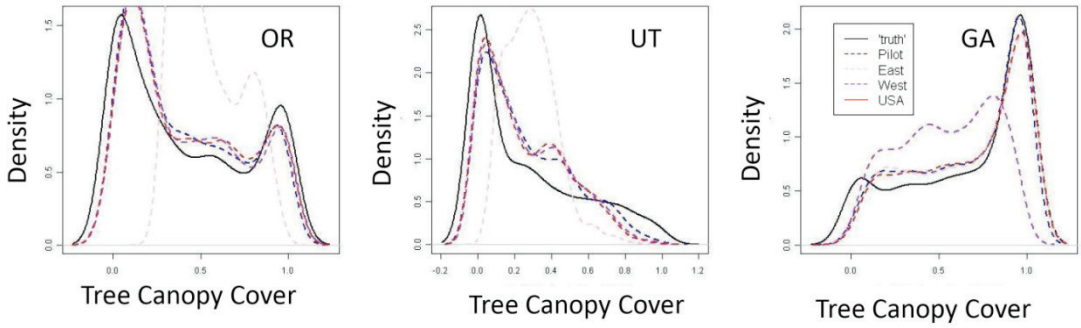

Figure 4—Density plots of tree canopy cover in independent test sets in three pilot areas, a. Oregon, b. Utah, and c. Georgia. Solid black lines reflect the "truth" from photo-interpreted data. Dotted blue lines reflect prediction from the individual pilot area models, then dotted pink, purple, and red from East, West, and USA models respectively.

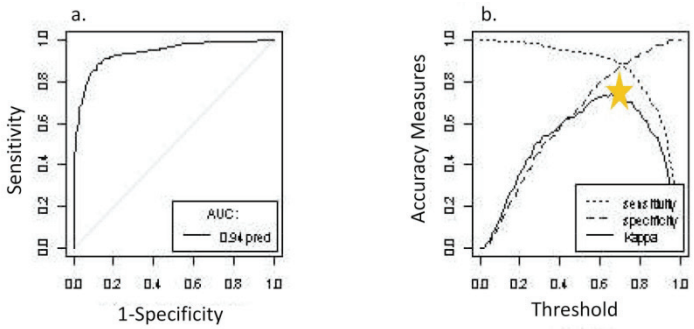

Figure 5—Results from the tree/not-tree model in UT. Plot a) is a Receiver Operator Curve (ROC Plot). Plot b) illustrates how using prevalence (indicated by the yellow star) as a threshold to convert probability predictions into a presence-absence map resulted in maximizing map accuracy.

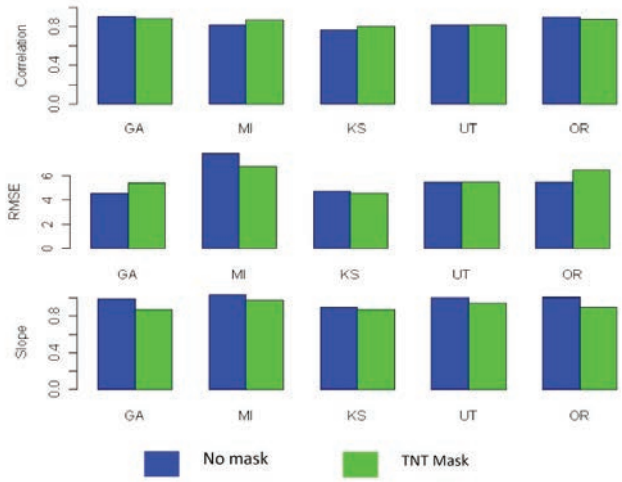

Figure 6—Correlation, RMSE, and slopes obtained in each of the five pilot areas when applying a tree canopy model without a mask (blue) versus a tree canopy model with an empirical mask (green) to an independent test set.

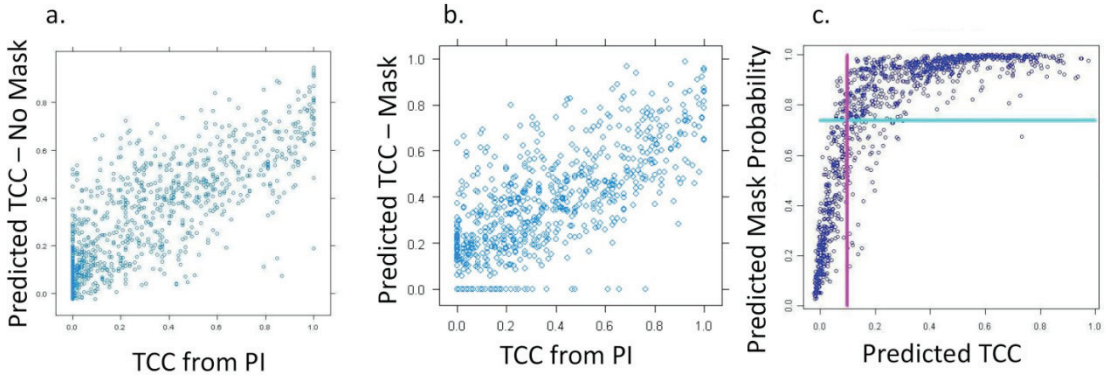

Figure 7—Scatter plots exploring effect of first creating a tree/no-tree mask prior to modeling tree canopy cover. In 7a, predicted tree canopy cover from a single unmasked model is plotted against the tree canopy cover response from the photo interpretation. In 7b, predicted tree canopy cover from the tree/no-tree model followed by the masked tree canopy model is plotted against the tree canopy cover response from the photo interpretation. In 7c, the predicted probability of having trees present from the tree/no-tree model is plotted against predicted tree canopy cover with no masking, with the prevalence-based threshold shown horizontally in blue and threshold a user might impose by applying a 10 percent cover threshold to the predicted tree canopy cover shown vertically in purple.