Ecological Modelling 233 (2012) 1–10

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

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

# Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in Random Forest models of tree species distributions in Nevada

Elizabeth A. Freeman*, Gretchen G. Moisen, Tracey S. Frescino

*USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA*

ARTICLE INFO

*Article history:*
Received 30 August 2011
Received in revised form 2 March 2012
Accepted 3 March 2012
Available online 5 April 2012

*Keywords:*
Random forests
Species distributions
Down sampling
Species prevalence

ABSTRACT

Random Forests is frequently used to model species distributions over large geographic areas. Complications arise when data used to train the models have been collected in stratified designs that involve different sampling intensity per stratum. The modeling process is further complicated if some of the target species are relatively rare on the landscape leading to an unbalanced number of presences and absences in the training data. We explored means to accommodate unequal sampling intensity across strata as well as the unbalanced species prevalence in Random Forest models for tree and shrub species distributions in the state of Nevada. For the unequal sampling intensity issue, we tested three modeling strategies: fitting models using all the data, down-sampling the intensified stratum; and building separate models for each stratum. We explored unbalanced species prevalence by investigating the effects of down-sampling the more prevalent response (presence or absence), and by optimizing the cutoff thresholds for declaring a species present. When modeling species presence with stratified data that was collected with different sampling intensities per stratum, we found that neither down-sampling the intensified stratum, nor fitting individual strata models, improved model performance. We also found that balancing the number of presences and absences in a training data set by down-sampling did not improve predictive models of species distributions, and did not eliminate the need to optimize thresholds. We then apply our final choice of model to the full raster layers for Nevada to produce statewide species distribution maps.

Published by Elsevier B.V.

## 1. Introduction

Maps of tree species presence and silvicultural metrics like basal area are needed throughout the world for a wide variety of forest land management applications. Knowledge of the probable location of certain key species of interest as well as their spatial patterns and associations to other species are vital components to any realistic land management activity. Mapping vegetation characteristics over broad geographic areas has received considerable attention in the US. Here, extant ground-based measurements collected by the US Forest Service Forest Inventory and Analysis program (Bechtold and Patterson, 2005; Gillespie, 1999) are often used as the response in predictive models of tree species distributions and other forest attributes using a variety of modeling techniques (Ohmann and Gregory, 2002; Moisen and Frescino, 2002; Blackard et al., 2008a,b). One such technique, Random Forests (Breiman, 2001) has proved very effective for predictive mapping of ecological attributes from climactic and topographic data (Attorre et al., 2011; Cutler et al., 2007; Garzón et al., 2006; Iverson et al., 2004, 2008; Prasad et al., 2006; Rehfeldt et al., 2006; Scarnati et al., 2009) as well as remotely sensed data (Baccini et al., 2008; Chan and Paelinckx, 2008; Evans and Cushman, 2009; Gislason et al., 2006; Ham et al., 2005; Lawrence et al., 2006; Powell et al., 2010).

While the use of Random Forests for mapping species distributions can be relatively straightforward for many applications, challenges arise when data are collected in stratified designs that involve different sampling intensity per stratum. The modeling process is further complicated if some of the target species are relatively rare on the landscape leading to an unbalanced number of presences and absences in the training data. Both of these problems presented themselves in the state of Nevada where a recent photo-based inventory pilot was conducted statewide to fill a gap in the nationwide forest inventory information conducted by FIA (Frescino et al., 2009). This Nevada Photo-based Inventory Pilot (NPIP) involved acquisition, processing, and photo-interpretation of large scale aerial photographs for numerous purposes. One purpose was the production of forest and nonforest species distribution maps for the state, constructed by modeling species presence as a function of remotely sensed and bioclimatic predictor layers. The challenges posed by this data set were, first, data collection was stratified, with unequal sampling intensity across strata. Areas of Nevada that had been predetermined as forested were sampled with 10 times the intensity as the non-forest areas because of heightened interest in tree species. Second, the species have a wide

* Corresponding author. Tel.: +1 801 625 5377; fax: +1 801 625 5723.
*E-mail address:* eafreeman@fs.fed.us (E.A. Freeman).

0304-3800/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.ecolmodel.2012.03.007

range of prevalence, ranging from 1% to 87%. Classification algorithms that minimize overall error rates can cause problems for species with unbalanced prevalence. By seeking to minimize overall error, such algorithms can increase the error rate for the rare class.

Chen et al. (2004) provide an algorithm for down-sampling within Random Forest that could potentially address both of these issues. In a down-sampled Random Forest, instead of a bootstrap sample from the entire dataset, each tree is built from a bootstrap sample from the rare class or less intensive stratum, along with a sub-sample of the same size from the more common class or more intensive stratum. They found that this approach improved the prediction accuracy of the rare category, with the added benefit of improved computation times for very large datasets. Using this algorithm, Chen et al. (2004) found down-sampling comparable to weighted Random Forest, in which observations in the rare category are given proportionally more weight than those in the more common category. Other existing techniques for dealing with rare categories include up-sampling (Japkowicz and Stephen, 2002) and cost-sensitive learning (Elkan, 2001). Up-sampling involves duplicating the rare class of less intensive stratum in the training data to ensure equal proportion of presence and absences are used for each tree, while cost-sensitive learning assigns different costs to errors of omission and errors of commission. McCarthy et al. (2005) compared down-sampling, up-sampling and cost-sensitive learning in Random Forests, and found comparable performance between the three techniques, though they found cost sensitive learning to have a slightly advantage in very large datasets (greater than 10,000 examples). Drummond and Holte (2003) found that down-sampling outperformed up-sampling in a decision tree learner. Evans and Cushman (2009) found down-sampling to perform well when mapping the presence of four conifer species in Northern Idaho, USA.

In this study, we explore means to accommodate the unequal sampling intensity across strata as well as the unbalanced species prevalence in Random Forest models for trees and shrubs in the state of Nevada. For the unequal sampling intensity issue, we test three modeling strategies. First, we fit models using all the data, ignoring the different probability of selection in each strata; second, we model with down-sampling of the intensified stratum; and third, we build separate models for each stratum. We then explore unbalanced species prevalence by investigating the effects of down-sampling the more prevalent response (presence or absence), and by optimizing the cutoff thresholds for declaring a species present. We then apply our final choice of model to the full raster layers for Nevada to produce statewide species distribution maps.

## 2. Methods

### 2.1. Response data

The presence or absence of the tree and shrub species was derived from a Photo-Based Inventory Pilot that was conducted by FIA in the years 2004–2005 within the state of Nevada (Frescino et al., 2009).

The distribution of photo plots for NPIP relied on the sampling design of the national FIA program (Reams et al., 2005). FIA conducts a comprehensive inventory of forest lands across all ownerships in the United States. Permanently established ground plots are measured annually based on a systematic sample of regularly spaced hexagons, each representing approximately 6000 acres. The plots are delineated into 5 panels; each panel is 20% of the data, measured on an annual cycle. In the West, each panel is divided again into subpanels: one subpanel measured every year over 10 years.

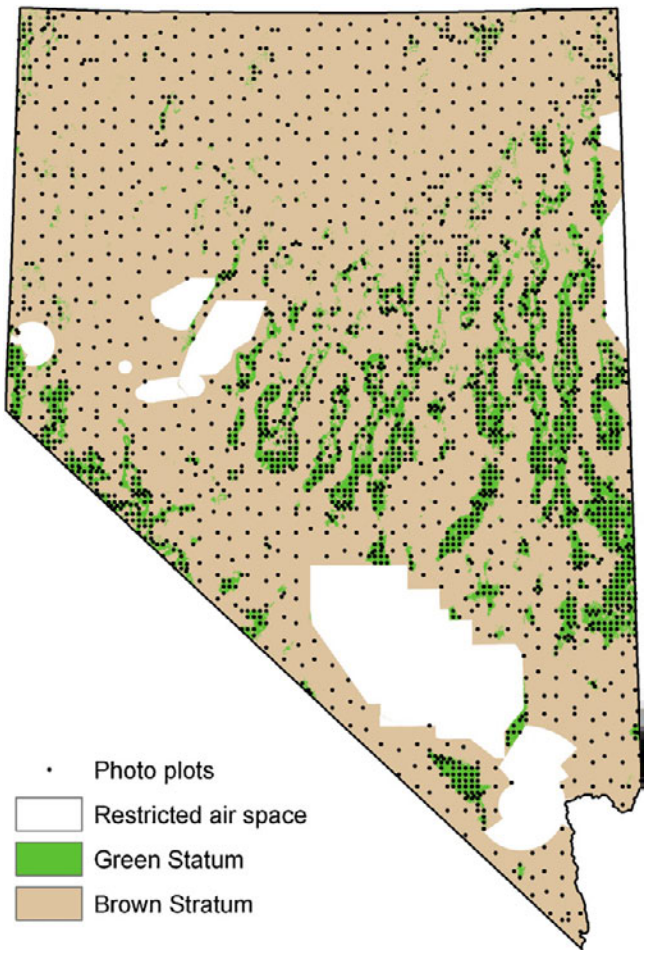


**Fig. 1.** Map of Nevada Photo-Based Inventory Pilot study showing approximate photo plot locations with sampling intensification in Green Stratum.

Data for the NPIP study was collected on FIA plot locations using a stratified sample, with unequal sampling intensity across strata (Fig. 1). The state was pre-stratified into 3 initial strata using a pixel-based, 250-m resolution map of predicted timberland forest, woodland forest, and non-forest areas (following the methods described in Blackard et al., 2008a,b). The combined timberland and woodland stratum is hereafter referred to as the Green Stratum, and the non-forest stratum is referred to as the Brown Stratum. Photo data were collected on all FIA locations (i.e., all 10 subpanels) within the Green Stratum, with exception of areas with restricted air space, totaling 1455 plots. However, only one-tenth of the FIA locations (i.e., one subpanel) within the Brown Stratum were photo-sampled, totaling 877 plots, resulting in an overall total of 2332 plots.

Each photo plot consisted of a grid of points distributed within a 250 m radius circle covering approximately 20 ha (48 acres) of land. There were a total of 49 points per plot representing about an acre each with the center point straddling the FIA field plot center. Using 6-in resolution color photography, each point was assigned a value identifying the object in the photograph the point fell on. If the object was a live tree, a species or species group attribute was assigned as well. If the object was a shrub, the shrub was identified as either sagebrush or not sagebrush. A species was classified as "present" on a photo plot if it was identified at any of the 49 points.

In addition to the challenge of having different sampling intensities within each stratum, many of the species in this data set also exhibited unbalanced prevalence. That is, the ratio of presences

**Table 1**
Species groups prevalence on NPIP plots. Stratum prevalence of 1% or less indicated in bold.

| | Number of plots | | | Percent of plots | | Estimated observed prevalence (%) |
|---|---|---|---|---|---|---|
| | Total | Green | Brown | Green | Brown | |
| Other shrub | 1909 | 1130 | 779 | 78 | 89 | 87 |
| Sage | 1446 | 931 | 515 | 64 | 59 | 60 |
| Juniper | 1238 | 1155 | 83 | 79 | 9 | 19 |
| Pinyon pine | 1184 | 1133 | 51 | 78 | 6 | 16 |
| Mountain mahogany | 249 | 240 | **9** | 17 | **1** | 3 |
| White fir | 127 | 124 | **3** | 9 | **0** | 2 |
| Aspen | 78 | 68 | **10** | 5 | **1** | 2 |
| Limber pine | 49 | 47 | **2** | 3 | **0** | 1 |
| Total plots | 2330 | 1454 | 876 | | | |

to absences of a particular species in the sample data was highly skewed. With the extremely rare species (prevalence of 1% or less) there were fewer than 20 presences in the entire dataset, leading to erratic models and error estimates. Therefore for this paper we concentrate on 8 species and species groups with 3% or greater prevalence in at least one of the strata. These 8 species have estimated prevalence across the landscape (weighted combination of Green and Brown Strata) of 1% to 87% (Tables 1 and 3).

The study includes four species with highly unbalanced prevalence (1–3% landscape prevalence). While these four species have 3% or greater prevalence in the Green Stratum, they have 1% or less in the Brown Stratum. This means that while they have 47 or more observed presences in the Green Stratum, they have only 2–10 presences in the Brown (Table 1). With so few observed presences, even prevalence independent error statistics such as AUC become unreliable (DeLong et al., 1988). Therefore, while all 8 species are included in the figures and tables, decisions about how best to handle unequal sampling intensity and unbalanced prevalence were based on all 8 species in the Green Stratum, but only the 4 more prevalent species in the Brown Stratum.

### 2.2. Predictor data

The predictor data set included 16 raster layers of multi-temporal, remotely sensed imagery and digital topographic data.

We used 250-m resolution, 16-day, cloud-free, composites of Moderate Resolution Imaging Spectroradiometer (MODIS) imagery for spring, summer, and fall of 2005. These included visible-red (RED) and near-infrared (NIR) bands and 2 vegetation indices: normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI). The RED and NIR bands are commonly used for discriminating vegetation by the sensitivity in reflectance values, or spectral signatures. In general, healthy, green vegetation absorbs visible-red light and reflects near-infrared light. NDVI is a ratio of RED and NIR bands and acts to accentuate live vegetation cover by reducing the multiplicative noise in the bands. EVI further reduces noise by incorporating the blue visible spectral band, thereby enhancing sensitivity in higher biomass areas (Huete et al., 2002). Three different dates of each variable were used as a multi-temporal approach to capture phenological differences that may occur among seasons.

The topographic variables originated from the 90-m resolution, National Elevation Dataset (NED) generated by the United Stated Geological Survey (USGS) (Gesch et al., 2009). Elevation, in meters, was re-sampled to a 250-m pixel size using the nearest neighbor algorithm in ArcMap (ESRI, 2009). Slope, in percent, and aspect, in degrees, were derived from this 250-m elevation product using ArcGIS and aspect northing and easting variables were calculated using sine and cosine functions, respectively, to convert aspect from a circular variable to a linear variable.

**Table 2**
Optimized thresholds from test set used for map production. Test set was selected so that the proportion of plots in the Green and Brown Strata reflects the proportion of land area in Nevada. Thresholds optimized to maximize kappa.

| Species | Threshold | | |
|---|---|---|---|
| | Baseline | Stratified | Balanced |
| Other shrub | 0.68 | 0.70 | 0.42 |
| Sage | 0.54 | 0.45 | 0.39 |
| Juniper | 0.54 | 0.50 | 0.49 |
| Pinyon pine | 0.51 | 0.37 | 0.56 |
| Mountain mahogany | 0.42 | 0.40 | 0.74 |
| White fir | 0.38 | 0.25 | 0.67 |
| Aspen | 0.36 | 0.36 | 0.83 |
| Limber pine | 0.03 | 0.04 | 0.81 |

### 2.3. Random forest

Random Forest models are built as an ensemble of classification or regression trees (Breiman et al., 1984). In a Random Forest model, 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 is 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 a large number of times. As the final prediction is a vote or average from prediction of all the trees in the collection.

In addition to the traditional validation technique of predicting over an independent test set, Random Forest also offers an Out-Of-Bag (OOB) validation technique. Each tree is built from a bootstrap sample of the training data, containing approximately 64% of the data points. The tree can then be used to make predictions on the remaining 36% of the data points. This is repeated for all the trees, and then aggregated across the forest.

Our analysis is carried out with the randomForest (Liaw and Wiener, 2002) package in R (R Development Core Team, 2008). When building models with the R randomForest package there are two main user controlled parameters: the number of variables to try at each node, the 'mtry' argument, and the number of trees in the forest, the 'ntree' argument. We used the 'tuneRF()' function from the randomForest package to optimize the number of variables to try at each split. To determine the optimal number of trees for our data, we tested models with varying numbers of trees and plotted model error (in term of AUC) as a function of tree number. For most species and models, the plot of AUC as a function of number of trees approaches a flat line between 500 and 1000 trees; therefore all subsequent models are built with 1000 trees.

To minimize the stochasticity inherent in random forest models, we averaged the errors over 20 examples of each type of model.

### 2.4. Validation

When the training data has varying intensity across strata, both test set and OOB validation techniques must be used with caution for all models. If the independent test set is drawn using simple random selection, it will be biased towards the intensified stratum, and will not reflect the population as a whole. OOB validation conducted on the training data is similarly biased. OOB Error measures can be calculated within each stratum, but not for the entire population. Consequently, we constructed an independent test set representative of the entire state by randomly withholding 5% of the sample plots in the Green Stratum, and 50% of the sample plots in the Brown Stratum.

Using this independent test set, models were assessed in term of Area Under the Curve (AUC), kappa, prevalence, and the ratio of sensitivity to specificity. Good models will have high AUC and kappa, a proportion of predicted presences very close to the proportion of observed presences, and sensitivity will be similar to specificity. Poor models will have AUC near 0.5, kappa near zero, over or under predict the observed prevalence, and have very different rates sensitivity and specificity.

While threshold dependant accuracy measures such as percentage correctly classified (PCC), sensitivity, and specificity have a long history of use in ecology, not only do they depend on the researcher's choice of threshold, but they also can be highly dependent on species prevalence (Manel et al., 2001). The AUC, on the other hand, provides a threshold independent method of evaluating the performance of presence/absence models. To calculate the AUC the true positive rate (sensitivity) is plotted against the false positive rate (1.0-specificity) as the threshold varies from 0 to 1. A good model will achieve a high true positive rate while the false positive rate is still relatively small; thus the plot will rise steeply at the origin, and then level off at a value near the maximum of 1 resulting in an area under the curve near 1. The plot for a poor model (whose predictive ability is the equivalent of random assignment) will lie near the diagonal, where the true positive rate equals the false positive rate for all thresholds resulting in an area under the curve near 0.5.

The AUC is a valuable threshold independent measurement of model accuracy; however, map making still requires a choice of threshold. We optimized our thresholds to maximize kappa as described in Freeman and Moisen (2008b). The kappa statistic summarizes all the available information in the confusion matrix. Kappa measures the proportion of correctly classified units after accounting for the probability of chance agreement. Kappa has been used extensively in map accuracy work (Congalton, 1991). While still requiring a choice of threshold, kappa is more resistant to prevalence than PCC, sensitivity and specificity, and was found by Manel et al. (2001) to be well correlated with the area under the curve (AUC) of ROC plots.

### 2.5. Down sampling for stratification or balance

In Random Forest, the default approach when building each tree is to take a bootstrap sample from the training data as a whole. If the training data is complicated by unequal sampling intensity between strata or unbalanced response category, Random Forest has the option of down sampling the more intensively sampled stratum or more prevalent category by specifying the number of samples to be taken (with replacement) from each strata or response category (Chen et al., 2004). The entire dataset will still be utilized for the forest as a whole, but each individual tree will be built from a sub-sample of the data. The current R implementation of the randomForest package does have some limitations on the options available for down-sampling. The randomForest package offers the down-sampling option for categorical response models but not for continuous response models. Also, it is possible to down-sample by strata or by response category, but not both in the same model.

To investigate the effect of the unequal sampling intensity across strata, we compare 3 modeling strategies: a "Baseline" model constructed on the full dataset, a "Stratified" model constructed on the full dataset with down sampling of the intensified stratum, and, finally, "Separate" models built within each stratum. When building a stratified model to address unequal data collection intensity, the randomForest argument 'strata' specifies the variable used for stratification, and the 'sampsize' argument specifies the number of data points to be randomly selected with replacement (unless specified otherwise) from each strata. A new selection is made for each tree in the forest. For the NPIP Stratified models, we used 'sampsize' to randomly sample (with replacement) all of the points in the Brown Stratum and a fraction of the points in the Green Stratum proportional to the Brown and Green land area in Nevada.

To investigate the effect of the unbalanced species prevalence, we compare 2 modeling strategies: a "Baseline" model on the full dataset as above, and a "Balanced" model with down-sampling of the more prevalent response category so that each tree in the forest is constructed from a balanced sample of presences and absences. When building a balanced model to address unequal prevalence, the 'strata' argument is not given, and the 'sampsize' argument specifies the number of data points to be randomly selected with replacement from each value of the categorical response variable. A new selection is made for each tree in the forest. For the NPIP balanced models, we used 'sampsize' to randomly sample (with replacement) from all of the rare category, and an equal number of points from the common category. We also compare two threshold selection strategies: using the default threshold of 0.5 versus optimizing the threshold to maximize kappa. It has been proposed that building a model with balanced data samples can be used as an alternative to threshold optimization (Evans and Cushman, 2009).

When comparing Baseline, Stratified and Balanced models, we used OOB validation on the full dataset, with validation measures calculated separately for each stratum. To minimize noise from Random Forest's stochasticity, we built 20 models for each strategy and species, and averaged the error. We compare 8 species in the Green Stratum, but only the 4 most prevalent species in the Brown Stratum.

### 2.6. Map creation

The ModelMap (Freeman, 2009) package was used to create the maps. The R software environment offers sophisticated new modeling techniques, but requires advanced programming skills to take full advantage of these capabilities. In addition, spatial data files can be too memory intensive to analyze easily with standard R code. The ModelMap package provides an interface between several existing R packages to automate and simplify the process of model building and map construction. ModelMap uses the randomForest package to construct models and the PresenceAbsence (Freeman and Moisen, 2008a) package to validate binary response models. Finally, ModelMap uses the rgdal (Kiett et al., 2010) package to read and predict over GIS raster data. Large maps are read in sections, to keep memory usage reasonable.

The final model built from the full dataset was used to predict over all 250 m pixels in Nevada. This resulted in a probability surface showing the proportion of trees voting for species presence. Thresholds optimized by maximizing kappa over the representative test data were applied to this surface to create the presence/absence maps. To obtain population error measures, we deliberately selected an independent test set to reflect the ratio of land area in Nevada in the Green and Brown strata. However, due to the high intensification of the Green stratum, an adequately

sized test set (approximately 22% of total data) required setting aside 50% of the data points from the Brown Stratum (while only 5% of the Green Stratum), leaving few Brown data points for model training. This could adversely affect prediction quality in the Brown Stratum.

Thus, for map production we follow the strategy outlined in Fielding and Bell (1997) of using a test set for model validation, but using all available data to create the final models for map construction. The test set model was used to calculate error measures for the population as a whole. As the final model for map production was created from more data than the test model, then it is reasonable to expect it has comparable or better accuracy. In addition, within strata OOB error measures were calculated from the final model itself.

## 3. Results

Results are summarized in Figs. 2 and 3. These figures report the average error for 20 repeats of each model, in terms of 4 error statistics: AUC, kappa, the difference between predicted and observed prevalence and the ratio of sensitivity to specificity. Good models have high AUC and kappa, a predicted prevalence equal to the observed prevalence(a prevalence difference near zero), and similar error rates in both the observed presences and the observed absences (a sensitivity to specificity ratio near 1).

The species are arranged in the figures in order of decreasing prevalence. This may give the impression that lower prevalence is associated with higher model quality. However, we believe that this is a case of correlation rather than causation. The highest prevalence species included in the study are shrubs, followed by woodland species. These species may simply be more difficult to correctly identify on aerial photos, and more difficult to predict from satellite imagery. The lowest prevalence species, on the other hand, are timber species, which may stand out more on photos, and which are strongly associated in Nevada with particular predictor layers, such as elevation. Also, the timber species are very rare in the Brown Stratum, and thus the high Brown Stratum AUC is unreliable.

### 3.1. Unequal sampling intensity across strata

#### 3.1.1. Baseline versus separate models

The Baseline models built from both strata and the Separate models built for individual strata had similar performance, though in the Brown Stratum the Baseline models tended to slightly outperform models built from a single stratum. When predicting within the Green Stratum, the Baseline model had higher AUC than the Separate model for 3 species, tied for 2 species, and had lower AUC for 2 species (Fig. 2a). When looking at kappa (Fig. 2c), the Baseline model performed better for 5 species, tied for 1 species and performed worse for 2 species. The differences were very slight, with the largest difference only 0.01 in both AUC and kappa. In terms of prevalence (Fig. 2e) the Baseline model was more accurate than the Separate model for 2 species, tied for 3 species and less accurate for 3 species. In terms of the ratio of sensitivity to specificity (Fig. 2g) the Baseline model performed better than the Separate Model for 4 species and performed worse for 4 species. Similar to AUC and kappa, the differences in prevalence and the ratio of sensitivity to specificity were slight in the low prevalence species, but did become more apparent in the higher prevalence species (Other Shrub, Sage, and Juniper). However, there was not an obvious pattern in these differences. For example, looking at the two largest differences in predicted prevalence, in Other Shrub, the Baseline model had a more accurate prevalence (under predicting by 3%) than the Separate model (under predicting by 6%), while in Juniper, the Baseline model had a less accurate prevalence (over predicting by 5%) than the Separate model (over predicting by 2%).

When predicting within the Brown Stratum, the Baseline model had a higher AUC than the Separate model for 2 species and tied for the other 2 species (Fig. 2b) and the Baseline model also had a higher kappa for 3 species and tied for 1 species (Fig. 2d), had a more accurate prevalence for 3 species and less accurate for 1 species (Fig. 2f) and performed better for all 4 species in terms of the ratio of sensitivity to specificity (Fig. 2h). The differences in AUC and kappa were also slightly higher in the Brown Stratum, with differences of 0.02 in AUC, 0.08 in kappa, both in Pinyon.

#### 3.1.2. Baseline versus stratified models

Again, the Baseline and the Stratified models had very similar performance. When predicting within the Green Stratum, the Baseline model had a higher AUC than the Stratified model for 2 species, tied for 4 species and had a lower AUC for 2 species (Fig. 2a) and had a higher kappa for 6 species and tied for 2 species (Fig. 2c). The differences were larger than in the Separate model comparison, but still small, with the largest difference 0.02 in AUC and 0.06 in kappa. However, the Baseline model had a more accurate prevalence than the Separate model in only 3 species, tied in 1 species and was less accurate in 4 species (Fig. 2e). When looking at the ratio of sensitivity to specificity (Fig. 2g), the Baseline model performed better than the stratified model in only 2 species, tied in 1 species and performed worse in 5 species.

When predicting within the Brown Stratum, the Baseline model had a higher AUC than the Separate model for 1 species, tied for 2 species and had a lower AUC for 1 species (Fig. 2b). With the Baseline model, all 4 species had a higher kappa (Fig. 2d) and a better ratio of sensitivity to specificity (Fig. 2h) than the Stratified model, and the Baseline model had a more accurate prevalence than the Stratified model for 2 species, tied for 1 species and was less accurate for 1 species (Fig. 2f).

### 3.2. Unbalanced species prevalence

#### 3.2.1. Baseline versus balanced models

First, we take advantage of the AUC to compare the Baseline and Balanced models independent of the choice of threshold. When predicting locations within the Green Stratum, the Baseline model had as a higher AUC than the Balanced model for 2 species, tied for 2 species and had lower AUC for 4 species (Fig. 3a). Within the Brown Stratum, the Baseline model tied the Balanced model for 3 species and had a lower AUC for 1 species (Fig. 3b). In both strata, these differences were very slight. The largest difference was 0.03, and most were less than 0.01.

Next we looked at threshold dependant error measures: kappa, prevalence, and the ratio of sensitivity to specificity.

With thresholds of both models optimized to maximize kappa, the Baseline model had higher kappa values for 3 species, tied for 4 species and had lower kappa for 1 species within the Green Stratum (Fig. 3c), and had higher kappa for 1 species, tied for 2 species and had lower kappa for 1 species in the Brown Stratum (Fig. 3d). The largest difference was for Limber pine in the Green Stratum, where the optimized Baseline kappa was 0.03 higher than the optimized Balanced kappa. In the other species, differences were 0.01 or less. The optimized Baseline model had a more accurate prevalence than the optimized Balanced model for 1 species, tied for 4 species, and had less accurate prevalence for 3 species within the Green Stratum (Fig. 3e), and had more accurate prevalence for 1 species, tied for 2 species and had less accurate prevalence for 1 species within the Brown Stratum (Fig. 3f). The average difference in predicted prevalence between the two models ranged from 0% to 5%, with most species having differences of 1% or less. The ratio of sensitivity and specificity for the optimized Baseline model was better than that of

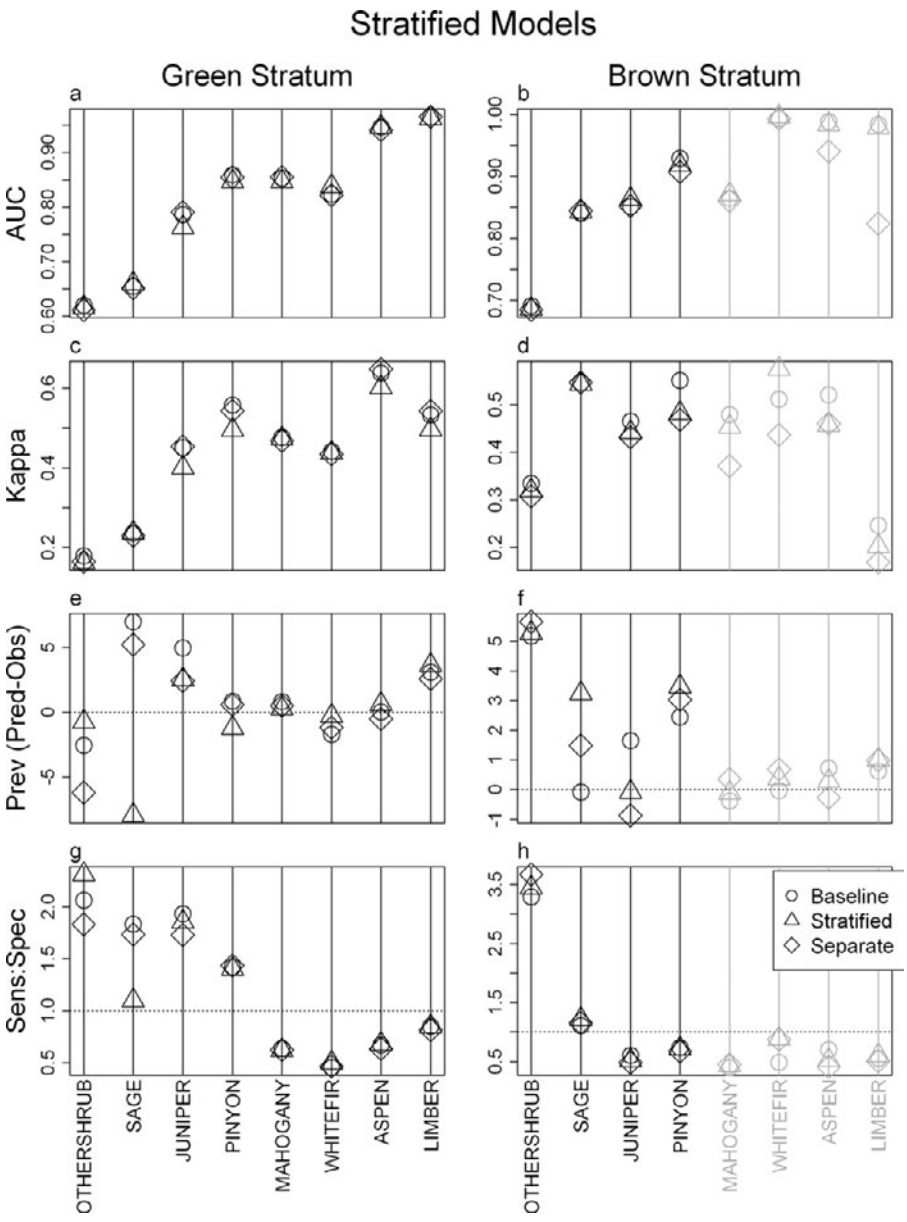


**Fig. 2.** Effect of stratification, averaged over 20 repeats. Baseline Models built from both strata of the training data set, Stratified Model with down sampling of the intensified stratum, and Separate Models from the individual strata. Species arranged in order of decreasing prevalence. Good models have an AUC near 1, high values of kappa, predicted minus observed prevalence near zero, and a ratio of sensitivity to specificity near 1. Prevalence of 1% or less shaded in gray.

the optimized Balanced model for 3 species and worse for 5 species within the Green Stratum (Fig. 3g), and was better for 3 species and tied for 1 species within the Brown Stratum (Fig. 3h).

With the default threshold of 0.5, the differences between the Baseline and the Balanced model were more pronounced than when the thresholds were optimized. The default Baseline model had higher kappa values than the default Balanced model for only 2 species, tied for 2 species and had lower kappa for 4 species within the Green Stratum (Fig. 3c), and in the Brown Stratum kappa was tied for 3 species and lower for 1 species (Fig. 3d). The differences between the default models were also larger than the differences between the optimized models, with differences in kappa of up to 0.14. The default Baseline model had a more accurate prevalence than the default Balanced model for 4 species and a less accurate prevalence for 4 species within the Green Stratum (Fig. 3e), and in the Brown Stratum, the default Baseline model had a more accurate prevalence for 1 species, tied for 2 species and had a less accurate prevalence for 2 species (Fig. 3f). The ratio of sensitivity and specificity for the default Baseline model was worse than that of the default Balanced model for all 8 species within the Green Stratum (Fig. 3g), and was better for 2 species and worse for 2 species within the Brown Stratum (Fig. 3h).

### 3.2.2. *Optimized threshold versus default 0.5 threshold*

As would be expected, optimizing the threshold to maximize kappa produced a higher kappa value than the default threshold of 0.5 for all species for both models and both strata.

With the Baseline model, optimizing the threshold to maximize kappa produced more accurate prevalence than the default threshold of 0.5 in 6 species and less accurate prevalence for 2 species within the Green Stratum (Fig. 3e) and in the Brown Stratum, the optimized baseline model had a more accurate prevalence than the default Baseline model for 2 species and less accurate prevalence for 2 species (Fig. 3f). The difference is prevalence between the optimized and the default Baseline models ranged from 3% to 23%. Optimizing the threshold did as good or better job at

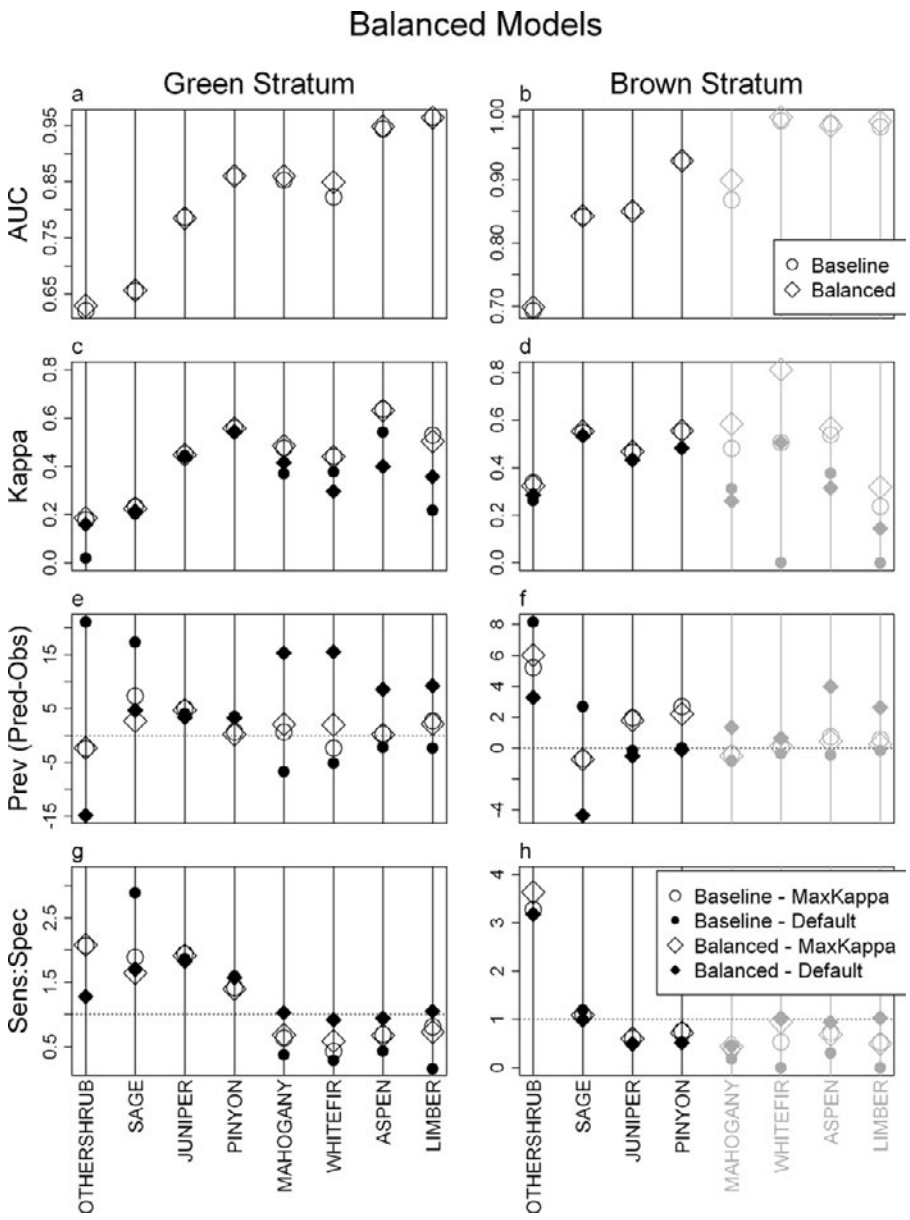


**Fig. 3.** Effect of Balance, averaged over 20 repeats. Baseline Models built from the training data set, Stratified Model with down sampling of the intensified stratum. Baseline and Balanced models shown with default threshold of 0.5 and with threshold optimized to maximize kappa. Species arranged in order of decreasing prevalence. Good models have an AUC near 1, high values of kappa, predicted minus observed prevalence near zero, and a ratio of sensitivity to specificity near 1. Prevalence of 1% or less shaded in gray.

balancing the error rates between the observed presences and observed absences (the ratio of sensitivity to specificity) than the default threshold of 0.5 in 7 species and a worse job in 1 species within the Green Stratum (Fig. 3g). In the Brown Stratum, the optimized Baseline model had a better ratio of sensitivity to specificity than the default Baseline model in all 4 species (Fig. 3h).

With the Balanced model, optimizing the threshold to maximize kappa produced a more accurate prevalence than the default threshold in 7 species and a less accurate prevalence in 1 species within the Green Stratum (Fig. 3e) In the Brown Stratum the optimized Balanced model had a more accurate prevalence in only 1 species, and a less accurate prevalence in 3 species (Fig. 3f). The difference in prevalence between the optimized and default Balanced models ranged from 1% to 14%. When evaluating the Balanced model in terms of the ratio of sensitivity to specificity, in contrast to kappa and prevalence, optimizing the threshold to maximize kappa was less effective than the default threshold of 0.5 and, in fact, the optimized Balanced model had a better ratio of sensitivity to specificity for 2 species and a worse ratio for 6 species within the Green Stratum (Fig. 3g), and in the Brown Stratum, the optimized Balanced model had a better ratio for 2 species and a worse ratio for 2 species (Fig. 3h).

### 3.3. Map creation

The final models for map creation are built from the full dataset, utilizing all available data (both training and test data) as recommended in Fielding and Bell (1997). Maps were constructed with thresholds optimized on the test set to maximize kappa. These optimized thresholds ranged from 0.03 to 0.83 (Table 2). Interestingly, the high and low extremes were both from the lowest prevalence species investigated, Limber pine and Aspen. The optimized thresholds for Limber pine were very low for the baseline model (0.03) and for the stratified model (0.04), but very high for the balanced

**Table 3**
Prevalence estimated from the observed data (strata weighted to account for unequal sampling intensity) and from the final maps. Thresholds for map predictions optimized by maximizing kappa for the test data.

| Strata | Estimated observed prevalence (%) | Map prevalence | | |
|---|---|---|---|---|
| | | Baseline (%) | Stratified (%) | Balanced (%) |
| Other shrub | 87 | 88 | 88 | 92 |
| Sage | 60 | 55 | 66 | 65 |
| Juniper | 19 | 16 | 15 | 17 |
| Pinyon pine | 16 | 14 | 16 | 13 |
| Mountain mahogany | 3 | 2 | 2 | 2 |
| White fir | 2 | 1 | 2 | 2 |
| Aspen | 2 | 2 | 1 | 2 |
| Limber pine | 1 | 7 | 5 | 1 |

model (0.81). Aspen did not have quite as low optimized thresholds for the baseline model (0.36) or the stratified model (0.36), but was very high for the balanced model (0.83).

The population prevalence estimated from the dataset (accounting for unequal probability of selection within strata) was compared to the prevalence in the maps (Table 3). The map prevalence for the Baseline model ranged from under-predicting by 5% to over-predicting by 6%. The map prevalence for the Stratified model ranged from under-predicting by 6% to over-predicting by 7%. The map prevalence for the Baseline model ranged from under-predicting by 3% to over-predicting by 5%.

Overall map accuracy is estimated from the training data model applied to the independent test set. Accuracy for the individual strata is from the OOB predictions from the full dataset models used to construct the map (Table 4). Overall AUC from the test set ranges from 0.70 to 0.98. OOB AUC from the full dataset models used in map production ranges from 0.61 to 0.97 for the Green Stratum and 0.70 to 0.99 for the Brown Stratum. Overall kappa from the test set ranges from 0.05 to 0.67. OOB kappa from the full dataset ranges from 0.14 to 0.60 for the Green Stratum and 0.08 to 0.55 for the Brown Stratum. Overall prevalence error from the test set ranges from 3% under-prediction to 7% over-prediction. OOB prevalence error from the full dataset ranges from 3% under-prediction to 13% over-prediction for the Green Stratum and 1% under-prediction to 5% over-prediction for the Brown Stratum.

## 4. Discussion

When investigating the unequal sampling intensity across strata, first, we found that the Baseline models built from both strata had a slight advantage to the Separate models built from individual strata. In strata where the species were common, the full dataset models had performance comparable to a model built from the stratum alone. And in strata where the species were rare adding the additional presence data from the high prevalence stratum improved the model's ability to predict the presences even in the low prevalence stratum.

When comparing the Baseline and the Stratified models, we found that the Baseline model had slightly better prediction accuracy, as measured by AUC and kappa, than the Stratified model in the majority of species in both the Green and the Brown Strata. However, in the Green Stratum, the Stratified model did have a better ratio of sensitivity to specificity in 5 of the 8 species. This suggests that if your primary concern is overall prediction accuracy, the Baseline model offers a slight advantage, but if your primary concern is keeping the error rate the same between observed presences and observed absences, stratification may better meet this goal. On the other hand, if a uniform error rate is your primary goal, there are other options besides stratification. In this study, thresholds were optimized to maximize kappa. It is also possible to optimize the threshold to equalize sensitivity and specificity.

In the NPIP study the intensification was tied to the species prevalence. The stratification and intensification in the NPIP study was deliberately chosen to increase the numbers of presences in the collected data. Therefore, the Baseline model had more presences in the bootstrap sample used to construct each Random Forest tree than did the stratified model. If the intensification was on strata that were uncorrelated with species presence, for example ownership boundaries, the results may have been very different.

**Table 4**
Error rates for Baseline model. Test set errors from model built from training data with test set selected so that the proportion of plots in the Green and Brown Strata reflects the proportion of land area in Nevada. OOB error rates for each stratum from the full dataset models used in map production. Thresholds in all models optimized to maximize kappa in the test set. Prevalence error is predicted prevalence minus observed prevalence. Prevalence of 1% or less indicated in bold.

| Species | Threshold | AUC | | |
|---|---|---|---|---|
| | | Test set | Full data OOB | |
| | | Both | Green | Brown |
| Other shrub | 0.68 | 0.70 | 0.61 | 0.70 |
| Sage | 0.54 | 0.84 | 0.66 | 0.84 |
| Juniper | 0.54 | 0.92 | 0.78 | 0.85 |
| Pinyon pine | 0.51 | 0.97 | 0.86 | 0.93 |
| Mountain mahogany | 0.42 | 0.94 | 0.86 | **0.85** |
| White fir | 0.38 | 0.95 | 0.82 | **0.99** |
| Aspen | 0.36 | 0.98 | 0.94 | **0.99** |
| Limber pine | 0.03 | 0.94 | 0.97 | **0.99** |

| Species | Threshold | Kappa | | |
|---|---|---|---|---|
| | | Test set | Full data OOB | |
| | | Both | Green | Brown |
| Other shrub | 0.68 | 0.29 | 0.14 | 0.31 |
| Sage | 0.54 | 0.52 | 0.21 | 0.55 |
| Juniper | 0.54 | 0.67 | 0.44 | 0.45 |
| Pinyon pine | 0.51 | 0.81 | 0.55 | 0.49 |
| Mountain mahogany | 0.42 | 0.62 | 0.45 | **0.30** |
| White fir | 0.38 | 0.39 | 0.43 | **0.50** |
| Aspen | 0.36 | 0.62 | 0.60 | **0.41** |
| Limber pine | 0.03 | 0.05 | 0.33 | **0.08** |

| Species | Threshold | Prevalence error | | |
|---|---|---|---|---|
| | | Test set | Full data OOB | |
| | | Both (%) | Green (%) | Brown (%) |
| Other shrub | 0.68 | 7 | 1 | 3 |
| Sage | 0.54 | −3 | 13 | −1 |
| Juniper | 0.54 | 3 | 2 | −1 |
| Pinyon pine | 0.51 | 0 | 3 | 0 |
| Mountain mahogany | 0.42 | 0 | −3 | **−1** |
| White fir | 0.38 | 0 | −2 | **0** |
| Aspen | 0.36 | 0 | −1 | **0** |
| Limber pine | 0.03 | 6 | 11 | **5** |

When comparing the Baseline and Balanced models with optimized thresholds (Fig. 3, hollow points), we found differences between the approaches were negligible. In a larger dataset, the Balanced model might improve computation speed, but in our dataset this improvement was not notable. The largest differences were in prevalence (Fig. 3e, hollow points), but even here the differences between the two optimized models were not large, and the over and under prediction is not clearly related to prevalence for either optimized model.

If for some reason it was desirable to avoid threshold optimization and insist upon the default threshold of 0.5 (Fig. 3, solid points), then the differences between the Baseline and the Balanced models were more dramatic. These differences were also strongly linked to prevalence, particularly the differences in prevalence accuracy (Fig. 3e, solid points), with the default Baseline model more accurate for our 4 lowest prevalence species (estimated observed prevalence 1% to 3%), and the default Balanced model more accurate for our 4 moderate to high prevalence species (estimated observed prevalence of 16–87%). This is interesting because, in theory, Balanced models are proposed as a way of improving predictions in species with very low or very high prevalence. Instead we found that when the threshold is kept at the default of 0.5, Balanced models actually did worse than the Baseline models for our very low prevalence species.

Also, the default Baseline model tended to over predict the high prevalence species and under predict the 4 low prevalence species, while the default Balanced model over predicted all but one of our species and highly over predicted the 4 rare species. The one species group where the default Balanced model under predicted the prevalence was Other Shrub (observed 87% prevalence) where the default Balanced model under predicted the prevalence by 15%.

It has been proposed than an additional benefit of balanced models is that balancing would allow the use of the default threshold criteria of 0.5 and make optimizing thresholds unnecessary. Our data does not support this. We found, contrary to the approach of Evans and Cushman (2009), that the predictive performance in terms of AUC, kappa and prevalence of the Balanced model was considerably improved by threshold optimization. The one exception we found was that if the primary concern was balancing the ratio of sensitivity and specificity, then the default threshold for the Balanced model performed well.

## 5. Conclusion

Final maps were constructed from Random Forest model constructed on the entire dataset, without down sampling for either stratification or balance.

In the case of the Nevada data, both down sampling for stratification, and constructing individual models for each stratum did not substantially improve and, in some cases reduced, performance of species distribution models. The intensification by strata in the NPIP data collection was deliberately chosen to include more presences in training data. Negating this by using a Stratified model was counterproductive.

When thresholds were optimized, down sampling to balance species prevalence did not substantially improve predictive performance of the models. Also, down sampling for balance did not eliminate the need to optimize thresholds. In fact, balancing without threshold optimization actually worsened the predicted prevalence of our rarest species, the very species that balancing is supposed to be helping. While down sampling can increase processing speed, this was not appreciable for this dataset. Therefore we concluded that the slight gains did not justify the more complicated modeling structure.

## References


Attorre, F., Alfò, M., De Sanctis, M., Francesconi, F., Valenti, R., Vitale, M., Bruno, F., 2011. Evaluating the effects of climate change on tree species abundance and distribution in the Italian peninsula. Applied Vegetation Science 14, 242–255.

Baccini, A., Laporte, N., Goetz, S.J., Sun, M., Dong, H., 2008. A first map of tropical Africa's above-ground biomass derived from satellite imagery. Environmental Research Letters 3, 9.

Bechtold, W.A., Patterson, P.L. (Eds.), 2005. The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, 85 p.

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

Blackard, J., Finco, M., Helmer, E., Holden, G., Hoppus, M., Jacobs, D., Lister, A., Moisen, G.G., Nelson, M., Riemann, R., Ruefenacht, B., Salajanu, D., Weyermann, D., Winterberger, K., Brandeis, T., Czaplewski, R., McRoberts, R., Patterson, P., Tymcio, R., 2008b. 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., Stone, C.G., 1984. Classification and Regression Trees. Wadsworth.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Chan, J.C.W., Paelinckx, D., 2008. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotype mapping using airborne hyperspectral imagery. Remote Sensing of Environment 112 (6), 2999–3011.

Chen C., Liaw, A., Breiman, L., 2004. Using random forest to learn unbalanced data. Technical Report 666, Statistics Department, University of California at Berkeley.

Congalton, R.G., 1991. A review of assessing the accuracy of 586 classifications of remotely sensed data. Remote Sensing of Environment 37 (1), 35–46.

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

DeLong, E.R., Delong, D.M., Clarke-Pearson, D.L., 1988. Comparing areas under two or more correlated Receiver Operating Characteristic curves: a nonparametric approach. Biometrics 44 (3), 837–845.

Drummond, C., Holte, R.C., 2003. 4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Workshop on Learning from Imbalanced Data sets II, ICML, Washington DC, 2003.

Elkan, C., 2001. The foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence.

ESRI (Environmental Systems Resource Institute), 2009. ArcMap 9.3. ESRI, Redlands, California.

Evans, J., Cushman, S., 2009. Gradient modeling of conifer species using random forests. Landscape Ecology 24, 673–683.

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

Freeman, E.A., Moisen, G., 2008a. PresenceAbsence: an R package for presence absence analysis. Journal of Statistical Software 23 (11), 1–31, Available from: http://www.jstatsoft.org/v23/i11.

Freeman, E.A., Moisen, G.G., 2008b. 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., 2009. ModelMap: An R Package for Modeling and Map production using Random Forest and Stochastic Gradient Boosting. USDA Forest Service, Rocky Mountain Research Station, 507, 25th street, Ogden, UT, USA.

Frescino, T.S., Moisen, G.G., Megown, K.A., Nelson, V.J., Freeman, E.A., Patterson, P.L., Finco, M., Brewer, K., Menlove, J., 2009. Nevada photo-based inventory pilot (NPIP) photo sampling procedures. Gen. Tech. Rep. RMRSGTR-222, 30 p.

Garzón, M.B., Blazek, R., Neteler, M., Sïanchez de Dios, R., Sainz Ollero, H., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. Ecological Modelling 197, 383–393.

Gesch, D., Evans, G., Mauck, J., Hutchinson, J., Carswell Jr., W.J., 2009. The National Map—Elevation: U.S. Geological Survey Fact Sheet 2009–3053, 4 p.

Gillespie, A.J.R., 1999. Rationale for a national annual forest inventory program. Journal of Forestry 97 (12), 16–20.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognition Letters 27 (4), 294–300.

Ham, J., Chen, Y., Crawford, M.M., Gosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 43, 492–501.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment 83, 195–213.

Iverson, L.R., Prasad, A.M., Liaw, A., 2004. New machine learning tools for predictive vegetation mapping after climate change: bagging and random forest perform better than regression tree analysis. In: Smithers, R. (Ed.), Proceedings, UK-International Association for Landscape Ecology. Cirencester, UK, pp. 317–320.

Iverson, L.R., Prasad, A.M., Matthews, S.N., Peters, M., 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecology and Management 254 (3), 390–406.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intelligent Data Analysis Journal 6 (5), 18–36.

Kiett, T.H., Bivand, R., Pebesma, E., Rowlingson B., 2010. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.6–31 http://CRAN.R-project.org/package=rgdal.

Lawrence, R.L., Wood, S.D., Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman and Cutler classifications (RandomForest). Remote Sensing of Environment 100, 356–362.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News2, 18–22. Available from: http://CRAN.R-project.org/doc/Rnews/.

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

McCarthy, K., Zaber, B., Weiss, G., 2005. Does cost-sensitive learning beat sampling for classifying rare classes? UBDM '05. In: Proceedings of the 1st International Workshop on Utility-Based Data Mining, Chicago, Illinois, pp. 69–77.

Moisen, G.G., Frescino, T.S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling 157, 209–225.

Ohmann, J.L., Gregory, M.J., 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, 725–741.

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

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

R Development Core Team. R. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

Reams, G.A., Smith, W.D., Hansen, M.H., Bechtold, W.A., Roesch, F.A., Moisen, G.G., 2005. The forest inventory and analysis sampling frame. In: Bechtold, W.A., Patterson, P.L. (Eds.), The Enhanced Forest Inventory and Analysis Program-National Sampling Design and Estimation Procedures. Gen. Tech. Rep. SRS-80. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, 15 p.

Rehfeldt, G.E., Crookston, N.L., Warwell, M.V., Evans, J.S., 2006. Empirical analysis of plant–climate relationships for the western United States. International Journal of Plant Sciences 167, 1123–1150.

Scarnati, L., Attorre, F., Farcomeni, A., Francesconi, F., De Sanctis, M., 2009. Modelling the spatial distribution of tree species with fragmented populations from abundance data. Community Ecology 10, 215–224.