

Mapping Ecological Systems with a Random Forest Model: Tradeoffs between Errors and Bias

Emilie Grossmann¹, Janet Ohmann², James Kagan³, Heather May¹ and Matthew Gregory¹

¹ Forest Ecosystems and Society, Oregon State University

² Pacific Northwest Research Station, USDA Forest Service

³ Institute for Natural Resources, Oregon State University

Methods for generating vegetation maps from remotely sensed data have advanced greatly within the last three decades since the LANDSAT program originated. They range from supervised and unsupervised classifications of single images, to classifications based on multitemporal imagery, to the integration of ancillary information with remotely sensed imagery (Holmgren and Thureson 1998). The latter techniques allow more detailed and accurate estimations of plant community composition and they were essential for creating the 2000 update for the USGS GAP vegetation layer. The level of specificity of Nature Serve's Ecological Systems (Systems) with respect to species composition makes many of them impossible to differentiate based on imagery alone. This is a common problem with remote sensing of vegetation (Kalliola and Syrjanen 1991). However, the combination of imagery and ancillary information on climate, landform and soil often provides enough information to map the Systems across the landscape at 30m resolution.

Classification trees (CART) and their extensions are a family of modeling techniques that are often used in ecological analysis (De'ath and Fabricius 2000, Cutler et al. 2007). CART models are also used to build predictive vegetation maps, based on relationships between vegetation, imagery and ancillary environmental data (e.g., Franklin 2002). Single CART models are built through recursive partitioning, wherein the response variable is iteratively divided into groups sequentially with group 'purity' increasing with each division (Breiman et al. 1984). Divisions are based on thresholds within explanatory variables. CART models have been popularized for mapping through the See5/C5.0 module for ERDAS Imagine software, and have been used to build the GAP vegetation layer in other regions (Lowry 2005). CART models, however, are prone to over-fitting data, which can lead to predictive errors.

Random forest (RF) models are an extension of CART that limits the over-fitting problem. Rather than building a single predictive tree model from all available data, RF builds hundreds of tree models, using randomized subsets of plot data and explanatory variables to build each tree. This process of internal cross-validation prevents the over-fitting problem inherent to a single CART model (Breiman 2001), hence they are becoming more popular for vegetation mapping (Prasad et al. 2006, Iverson et al. 2008, Evans and Cushman 2009). However, RF models can exhibit bias problems especially when plot-samples are unbal-

anced among the classes (Chen et al. 2004). Because Ecological Systems seldom occupy equal areas across any given region, their representation within systematic plot samples is normally unbalanced.

In our work mapping Multi-Resolution Land Characteristics Consortium (MRLC) mapzones 2 and 7, we explored the implications of mapping methods in the GAP mapping process, focusing on RF as a promising technique because it is known for making accurate classification predictions from noisy, non-normal data (Breiman 2001). Here, we present two contrasting maps of forested Ecological Systems across the West Cascades ecoregion (Figure 1) in Western Oregon based on: a) RF and b) RF with an associated bias adjustment procedure (RF_Adj). We contrast their differences, strengths and weaknesses, and make some recommendations for future GAP vegetation mapping efforts. Note that the maps presented here are not final GAP

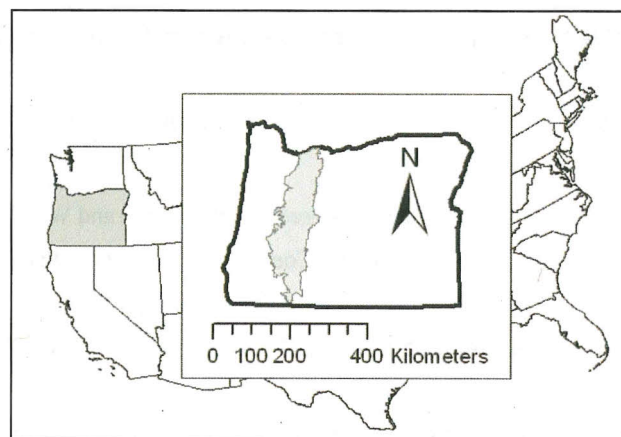


Figure 1. The West Cascades ecoregion within Oregon (Omernik 1987).

vegetation maps. Additional data and steps were used to map the less common forested Ecological Systems, and non-forest Systems for the actual GAP vegetation layer.

Methods

Study Area

We modeled the portion of the West Cascades ecoregion that falls within Oregon (Figure 1). The resultant model region stretches nearly across the state from North to South,

and is bounded by the Willamette Valley and Klamath Mountains on the West and the East Cascades ecoregion in the East. It encompasses most of the Cascade mountain range in the state, which ranges in elevation from nearly sea level to 11,249 feet (3,428 m). The climate ranges from moist and moderate in the foothills, to cold and snowy near the crest, becoming drier and colder near the eastern border of the region. A wide array of conifer forests and a few oak woodlands make up the majority of its forested Systems (Table 1).

Data

Our models were built from vegetation plot data (species and cover) obtained from a variety of sources, including the Forest Inventory and Analysis (FIA) Annual and Periodic surveys, Bureau of Land Management's Current Vegetation Survey and the US Forest Service Region 6's Current Vegetation Survey. The plots were classified to Ecological Systems (see Grossmann et al. 2008 for details). For the comparison presented here, Ecological Systems represented by fewer than 50 plots were dropped from the models. Each plot was intersected with spatial grids for 93 environmental variables that fall into four categories: 1) LANDSAT imagery and derivatives, 2) DAYMET modeled climate variables, 3) elevation and landform and 4) soil parent material (see Grossmann et al. 2008 for details).

Random Forest Modeling

We built RF models using the randomForest package (Liaw and Wiener 2002) for R statistical software (R Development Core Team 2006). Random forest models are built through a multi-step process (for details, see Breiman 2001). First, a bootstrap sample is selected from the plot data and a classification tree is built from the sample. Each node within the tree is constructed by selecting a random subset of the environmental variables and determining which variable yields the most effective split for maximizing purity in the two resultant groups. Nodes are continuously added to the tree until there is one plot per leaf. This process is repeated until the desired number of trees has been built (1000 in this analysis). To obtain a prediction from the forest of classification trees, each tree is allowed one 'vote' for the model prediction. Whichever Ecological System receives the most 'votes' from all of the trees in the random forest becomes the model prediction.

Variable selection

We selected environmental variables from the original pool of 93 using a randomization procedure, followed by sequential reverse and forward variable selection.

Table 1: Ecological Systems mapped within the West Cascades, and plot counts for modeling and accuracy assessment.

Code	Ecological System	Number of Plots	
		Modeling	Accuracy
4205*	EC Mesic Montane Mixed-Conifer Forest and Woodland	95	10
4214*	MC Dry-Mesic Mixed Conifer Forest and Woodland	238	26
4215	MC Mesic Mixed Conifer Forest and Woodland	446	50
4217*	MC Lower Montane Black Oak-Conifer Forest and Woodland	61	7
4219*	MC Red Fir Forest	58	6
4222*	NP Dry Douglas-fir Forest and Woodland	92	10
4224	NP Maritime Dry-Mesic Douglas-fir-Western Hemlock Forest	624	69
4226*	NP Maritime Mesic-Wet Douglas-fir-Western Hemlock Forest	320	36
4228	NP Mountain Hemlock Forest	156	17
4240*	NRM Ponderosa Pine Woodland and Savanna	105	12
4267*	RM Poor-Site Lodgepole Pine Forest	71	8
4272*	NP Dry-Mesic Silver Fir-Western Hemlock - Douglas-fir Forest	178	20
4333*	NP Lowland Mixed Hardwood Conifer Forest and Woodland	177	20
Total Plots:		2621	291

* Systems adjusted by bias-correction procedure.

Geographic abbreviations: EC = Eastern Cascades, MC = Mediterranean California, NRM = Northern Rocky Mountain, NP = North Pacific, RM = Rocky Mountain.

The randomization step consisted of building 1000 RF models, each containing 10 environmental variables selected randomly from the original 93. Reverse variable selection was applied to the best of the random models, eliminating variables that did not improve model accuracy. We then applied forward selection to this model, choosing helpful variables from the original 93 that were not already included. This multi-step process was used instead of simple forward or reverse selection in order to allow the inclusion of potentially interactive variables in groups. Simple forward variable selection yielded smaller models with lower accuracy than the best of the randomized models, and simple reverse selection yielded very large models with accuracy no better than the best of the randomized models.

Bias Adjustment

Our plot sample is heavily biased in favor of some Ecological Systems (especially 4224 and 4215, Table 1). In cases like this, RF often exhibits a predictive bias in favor of the over-sampled classes (Chen et al. 2004). Other researchers have used the approach of down-sampling over-sampled classes to limit the bias problem (e.g., Evans and Cushman 2009). However, because many of our Systems were represented by only a few plots (Table 1), we chose an alternate approach, correcting for model bias after the mapping process. Ecological Systems needing bias-corrections (Table 1) were identified from the RF's out-of-box error-matrix by determining whether the number of plots pre-

dicted in a specific class (column-totals in Table 3) were significantly less than the number of observations in that class (row-totals in Table 3). For these Systems, we generated maps analogous to probability surfaces, by mapping the number of votes for each class from the RF model, at each pixel. We converted these maps to presence-absence maps by selecting pixels that were above a threshold, which was set to yield the correct total area for that class. We estimated the correct total area for each System from the FIA Annual plots, which are a systematic sample of the region.

The Ecological System presence-absence maps were layered within ArcInfo GIS, with the rarest Systems on top, and the unadjusted RF prediction as the base layer. Our method is analogous to using the 'cutoff' feature within the R implementation of RF (Liaw and Wiener 2002). The primary differences between our method and R's 'cutoff' feature are: 1) our vote cutoff values are set by target map areas for the Systems; and 2) our mapped predictions are prioritized according to the target areas as well.

Accuracy Assessment

We assessed the accuracy of the models for local-scale predictions using an independent set of plot data, created by withholding 10% of each class from the model-building process (Table 1). From the withheld data, we generated an error-matrix relating model predictions to the correct values, and calculated percent accuracy overall and

Table 2: Variables used in the random forest model. Note: the bias-adjusted map was constructed from the original RF model, and so the variables are the same for each.

Variables
<u>Climate (Daymet)</u>
DIFTMP Difference between maximum August temperature and December minimum temperature
SMRTP Growing season moisture stress (ratio of temperature to precipitation from May-September)
<u>Elevation/Topography</u>
PRR Potential relative radiation
SLPPCT Slope (percent)
MLI McComb's Landform Index
<u>Location</u>
Y location (latitude)
<u>Soil Parent Material*</u>
SILICIC Contains rocks with minerals high in silica.
<u>Imagery</u>
MR5700 Median-filtered, ratio of LANDSAT TM(TM) band 5 to TM band 7
MTM500 Median-filtered, TM band 5
STDTM500 Standard deviation texture measure of TM band 5
ADR5700 Absolute difference texture measure of the ratio of TM bands 5 and 7
ADTC100 Absolute difference texture measure of Tasseled Cap transformation, band 1
ADTC200 Absolute difference texture measure of Tasseled Cap transformation, band 2

*Information based on composite map of SSURGO, SRI, and Oregon Geology

for each class, and Kappa statistics (Cohen 1960) from that error-matrix. We also assessed accuracy with respect to fuzzy class designations (shown in Table 3) to ascertain whether our errors were major or minor misclassifications (Gopal and Woodcock 1994, see Grossmann et al. 2008 for logic behind fuzzy class designations).

At the model-region scale, we assessed our maps for areal bias, comparing the area represented by each Ecological System in the maps with a sample-based estimate for the area (from FIA Annual plots) of each System in the landscape.

Results and Discussion

Models

Our final RF model contained 13 explanatory variables (Table 2). For the adjusted RF map, 10 of the 13 classes were corrected for bias during the mapping process (see Table 1). Despite the different methods, the ecoregion-wide spatial distributions of the Ecological Systems within the two maps were strikingly similar (Figure 2).

Plot-level Accuracy Assessment

Both methods had similar levels of overall accuracy (Kappa = .498 and .469 for RF and RF_adj, Table 3),

but had different strengths and weaknesses. In both models, the dominant System (4224, North Pacific Dry-Mesic Western Hemlock - Douglas-fir Forest) was quite well-predicted (Table 3). Those Systems whose plot-samples were relatively small (< 100, see Table 1) were less reliably predicted, and their accuracy statistics varied more widely between the two methods (Table 3). The adjustment process often improved accuracy for the relatively minor Systems at the expense of the dominant Systems, but had only a minor impact on overall accuracy. Fuzzy accuracy was very similar for both methods because most of the bias-corrections shifted predictions among fuzzy-similar Systems. However, in the case of System 4267 (RM Poor-site Lodgepole Pine Forest), the bias adjustment introduced significant errors, converting nearly all of 4267 to 4240 (NRM Ponderosa Pine Forest and Woodland). In the course of building the final GAP layers, these types of problems from the bias-adjustment were screened for and corrected (see Grossmann et al. 2008 for details).

Regional-level Accuracy of Areal Representation

The unadjusted RF model severely overestimated the area of Systems 4224 and 4215, while it underestimated the areas of most of the minor classes (Figure 3). The adjustment process improved the areal estimates, but did not completely eliminate the bias problem. However,

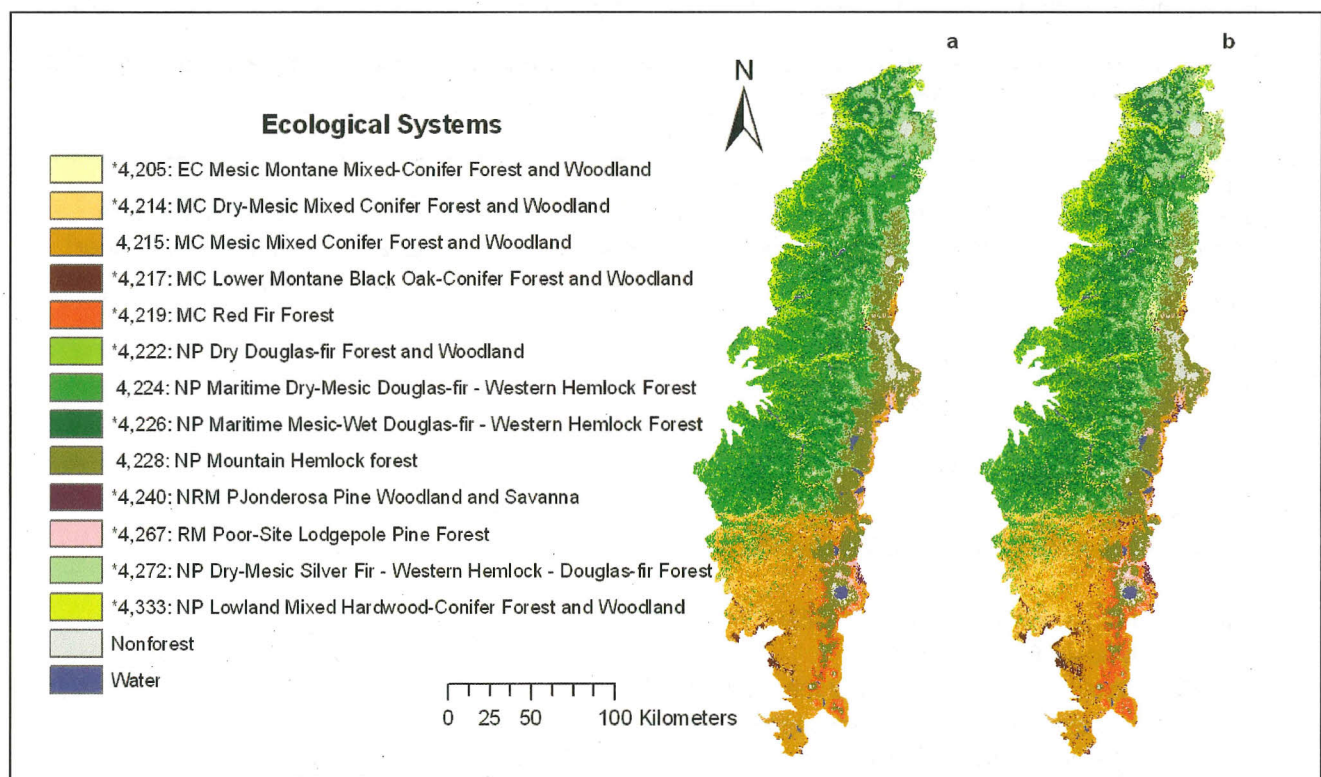


Figure 2. Regional-scale maps of the West Cascades ecoregion, based on a random forest model (a) and a bias-adjusted mapping of the same random forest model (b). Systems marked with a * were enhanced within the second

Table 3: Point-level accuracy statistics for independent data. Diagonal (dark grey) shading indicates correct classification. Light grey, off-diagonal shading indicates similar classes, or 'fuzzy-correct' classifications. Ecological System names for observed and predicted codes are listed in Table 1.

Random Forest														Row	%	% Fuzzy
observed	predicted													Totals	Correct	Correct
	4205	4214	4215	4217	4219	4222	4224	4226	4228	4240	4267	4272	4333			
4205	5	0	0	0	0	0	2	0	0	0	0	2	1	10	50.0%	70.0%
4214	0	7	6	1	0	0	7	1	0	1	1	1	1	26	26.9%	88.5%
4215	0	2	45	0	0	0	2	0	0	0	1	0	0	50	90.0%	98.0%
4217	0	1	3	3	0	0	0	0	0	0	0	0	0	7	42.9%	57.1%
4219	0	0	2	0	2	0	1	0	0	0	0	1	0	6	33.3%	66.7%
4222	0	1	1	0	0	1	6	0	0	0	0	0	1	10	10.0%	80.0%
4224	0	2	4	0	0	1	52	4	0	0	0	1	5	69	75.4%	98.6%
4226	0	1	1	1	0	0	18	14	0	0	0	0	1	36	38.9%	91.7%
4228	0	0	1	0	1	0	0	0	10	0	0	5	0	17	58.8%	94.1%
4240	0	0	2	0	0	1	0	0	0	8	1	0	0	12	66.7%	83.3%
4267	0	0	1	0	0	0	0	0	0	0	7	0	0	8	87.5%	87.5%
4272	0	0	0	0	0	0	7	1	4	0	0	8	0	20	40.0%	100%
4333	0	2	0	0	0	1	6	6	0	0	0	0	5	20	25.0%	85.0%
Column Totals	5	16	66	5	3	4	101	26	14	9	10	18	14			
														Total %		
														Correct:	57.4%	91.4%
														Kappa:	0.498	0.774

Random Forest, Adjusted														Row	%	% Fuzzy
observed	predicted													Totals	Correct	Correct
	4205	4214	4215	4217	4219	4222	4224	4226	4228	4240	4267	4272	4333			
4205	6	0	0	0	0	0	1	0	0	0	0	2	1	10	60.0%	80.0%
4214	0	7	3	2	0	1	7	1	0	3	0	1	1	26	26.9%	88.5%
4215	0	4	40	0	0	0	2	0	0	3	1	0	0	50	80.0%	98.0%
4217	0	1	3	3	0	0	0	0	0	0	0	0	0	7	42.9%	57.1%
4219	0	0	2	0	2	0	1	0	0	0	0	1	0	6	33.3%	66.7%
4222	0	0	1	0	0	2	6	0	0	0	0	0	1	10	20.0%	80.0%
4224	1	4	3	0	0	1	48	5	0	0	0	2	5	69	69.6%	98.6%
4226	2	1	1	1	0	0	12	17	0	0	0	0	2	36	47.2%	91.7%
4228	0	0	1	0	1	0	0	0	8	0	0	7	0	17	47.1%	94.1%
4240	0	0	1	0	0	1	0	0	0	10	0	0	0	12	83.3%	91.7%
4267	0	0	1	0	0	0	0	0	0	7	0	0	0	8	0.0%	0.0%
4272	1	0	0	0	0	0	3	2	4	0	0	10	0	20	50.0%	100%
4333	0	2	0	0	0	1	4	8	0	0	0	0	5	20	25.0%	85.0%
Column Totals	10	19	56	6	3	6	84	33	12	23	1	23	15			
														Total %		
														Correct:	54.3%	89.7%
														Kappa:	0.469	0.738

in this case it is likely that the remaining bias in favor of 4224 for the adjusted map can be partially accounted for by our exclusion of several minor Systems in the models. The area estimated for 'Other' Systems by the inventory plots likely accounts for a portion of the remaining apparent positive biases within Systems 4224, 4215 and 4228.

Conclusions

The unadjusted RF method was better with respect to point-level accuracy, with the highest values for percent correct, and for Kappa (absolute and fuzzy), but the high point-level accuracy was partially achieved by over-predicting the dominant class (Table 3). This tendency is highlighted most strongly in the areal statistics summary (Figure 3). The post-modeling adjustment procedure cor-

rected for a significant portion of this bias, at the cost of losing some plot-level accuracy. In general, users who need to work locally may need to aggregate fuzzy Systems to boost plot-level accuracy to acceptable levels.

As is often the case, there is no unequivocal answer to the question of which method is the best. Rather, each method comes with its own strengths and weaknesses. Modelers should carefully consider these trade-offs in conjunction with their goals and objectives for a given project. Map users need to understand the methods used to create a map in order to make informed assessments of appropriate and inappropriate uses (Fassnacht et al. 2006).

In the process of developing models for all of the ecoregions within MRLC mapZones 2 and 7 for the USGS GAP vegetation layer, we concluded that the RF model with the bias adjustment of the under-represented classes

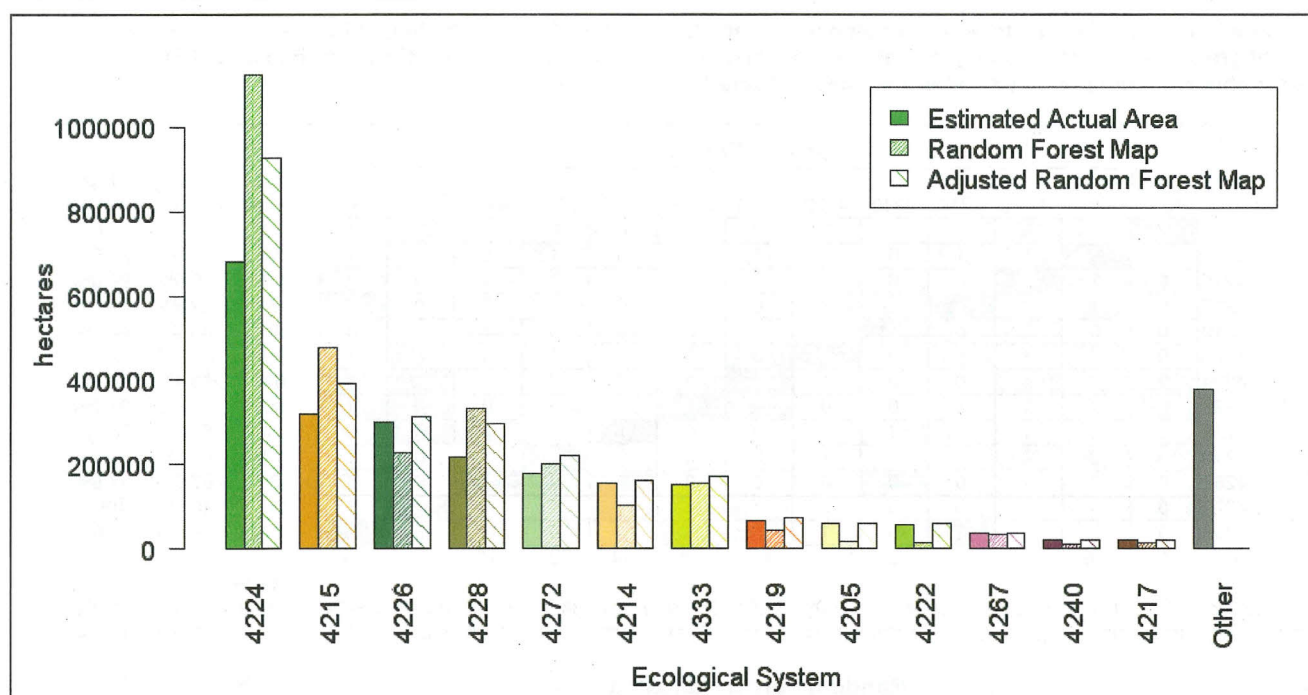


Figure 3. Map areal bias at the ecoregion scale. Actual area estimates are derived from the FIA annual plots, a systematic sample of the region. The 'Other' category is shown here because our models excluded several Ecological Systems that are present in the region, but were poorly represented in this plot data set. Color-coding corresponds to the legend in Figure 2. See Table 1 for Ecological System names.

struck the best balance between plot-level accuracy and regional-scale class-representation. Looking towards future updates to the USGS GAP vegetation layer, additional methods-comparison research will undoubtedly be needed as modeling methods will continue to evolve and improve (Ahn et al. 2007).

Literature Cited

- Ahn, H., H. Moon, M. J. Fazzari, N. Lim, J. J. Chen and R. L. Kodell. 2007. Classification by ensembles from random partitions of high-dimensional data. *Computational Statistics & Data Analysis* 51:6166-6179.
- Breiman, L. 2001. Random forests. *Machine Learning* 45:5-32.
- Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone. 1984. *Classification and Regression Trees*. Chapman & Hall/CRC, New York.
- Chen, C., A. Liaw and L. Breiman. 2004. *Using random forest to learn imbalanced data*.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37-46.
- Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson and J. J. Lawler. 2007. Random forests for classification in ecology. *Ecology* 88:2783-2792.
- De'ath, G., and K. E. Fabricius. 2000. Classification and Regression Trees: A powerful, yet simple technique for ecological data analysis. *Ecology* 81:3178-3192.
- Evans, J., and S. Cushman. 2009. Gradient modeling of conifer species using random forests. *Landscape Ecology* 24:673-683.
- Fassnacht, K. S., W. B. Cohen and T. A. Spies. 2006. Key issues in making and using satellite-based maps in ecology: A primer. *Forest Ecology and Management* 222: 167-181.
- Franklin, J. 2002. Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral. *Applied Vegetation Science* 5:135-146.
- Gopal, S., and C. Woodcock. 1994. Theory and Methods for Accuracy Assessment of Thematic Maps Using Fuzzy-Sets. *Photogrammetric Engineering and Remote Sensing* 60:181-188.

-
- Grossmann, E. B., J. S. Kagan, J. L. Ohmann, H. K. May, M. Gregory and C. Tobalske. 2008. *Final Report on Land Cover Mapping Methods, Map Zones 2 and 7, PNW ReGAP*. Institute for Natural Resources, Oregon State University, Corvallis, OR.
- Holmgren, P., and T. Thuresson. 1998. Satellite remote sensing for forestry planning - A review. *Scandinavian Journal of Forest Research* 13:90-110.
- Iverson, L. R., A. M. Prasad, S. N. Matthews and M. Peters. 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management* 254:390-406.
- Kalliola, R., and K. Syrjanen. 1991. To What Extent Are Vegetation Types Visible in Satellite Imagery. *Annales Botanici Fennici* 28:45-57.
- Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. *R News* 2: 18-22.
- Lowry, J. 2005. A brief overview of the Southwest Regional GAP land cover mapping effort. *GAP Analysis Bulletin* 13:3-5.
- Omernik, L. M. 1987. Ecoregions of the Conterminous United States. *Annals of the Association of American Geographers* 77: 118-125.
- Prasad, A., L. Iverson and A. Liaw. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* 9: 181-199.
- R Development Core Team. 2006. R: A language and environment for statistical computing. in. *R Foundation/or Statistical Computing*, Vienna, Austria.