

# LAND COVER

## Identifying Longleaf Ecosystems Using Polytomous Logistic Regression

John S. Hogland and Mark D. Mackenzie  
School of Forestry and Wildlife Sciences, Auburn University, Alabama

Southeast, PLR would be well suited to differentiate forested ecosystem types.

### Introduction

Longleaf ecosystems, one of the most species-rich plant ecosystems outside of the tropics, are estimated to occupy  $1.2 \times 10^6$  ha across the Southeastern United States, a mere 5 percent of the  $24.3 \times 10^6$  ha pre-European settlement estimate (Outcalt and Sheffield 1996). This dramatic loss of habitat has had a substantial impact on numerous plants and animals, and is the primary reason that many Southeastern species have been listed as threatened or endangered (Tuldge 1999). These findings indicate a strong need for the conservation and restoration of these critically endangered ecosystems (Noss et al. 1995).

While conservation and restoration efforts have begun, they have been limited, in part, by the lack of information depicting the current location of these ecosystems. Long-term studies, such as the Forest Inventory Analysis, have been useful in identifying trends in longleaf ecosystem decline (Kelly and Bechtold 1990; Outcalt and Sheffield 1996), but are ill-suited to provide meaningful information at fine spatial scales. Due to the coarse nature of these data sets (e.g., 20 km grain), organizations have had to take a broad-based approach toward longleaf ecosystem management, monitoring, and restoration, often limiting the efficacy of their efforts. To become more effective, these organizations need accurate, fine-scale data sets that identify forested ecosystem types and depict the current location and distribution of longleaf ecosystems.

Remotely sensed data provide a unique opportunity to generate such a data set by linking fine-grain (30 m) spectral information with spatially explicit examples of forested ecosystem types. Few analysts, however, have successfully differentiated longleaf ecosystems from other coniferous ecosystems using common classification techniques (e.g., maximum likelihood classifiers, clustering, classification trees, and artificial neural networks) due to the amount of spectral overlap among coniferous ecosystems in the Southeast. Using polytomous logistic regression (PLR), which allows for  $> 2$  response variables, we demonstrated the flexibility and utility of probabilistic classifiers when substantial spectral overlap among land cover types exists (Hogland et al. *in progress*). Given the similarities between longleaf and other coniferous ecosystems in the

### Methodology

To identify the current distribution of longleaf ecosystems, we employed an iterative hierarchical classification scheme (IHCS) that utilized PLR (Agresti 2002), digital elevation models (DEMs), and multitemporal Landsat enhanced thematic mapper plus (ETM+) imagery. Each Landsat ETM+ scene was preprocessed by the Multi-Resolution Land Cover Consortium to Level 1T standards (NASA 2005) and grouped into one of three seasons based on the acquisition date: leaf on spring, leaf on fall, and leaf off winter. Due to the inherent variability among multitemporal Landsat ETM+ scenes, all scenes were normalized and merged, by season, to a common radiometric scale using a newly developed normalization procedure (Hogland and MacKenzie *in progress*). PLR, statistical analyses, and accuracy assessments were performed using SAS version 8.2 (SAS® 2005). Model implementation was performed using ARCGIS version 8.3 and ESRI's Spatial Analyst extension (ESRI® 2005).

Our IHCS is a multistage classification that constrains the conditional probabilities of one PLR classification by the specific classes of a more general PLR classification. The benefits of IHCS include fewer field samples, the preservation of modeling and classification errors, a hierarchically organized classification, and the ability to account for confounding temporal features (TF).

Stage 1, iteration 1 of our IHCS identified generalized land cover types (after Homer et al. 2004), and seasonal TF (Table 1) using training data, normalized multitemporal ETM+ imagery, and a maximum likelihood allocation rule (MLAR). To account for the effects of TF, pixels categorized as clouds, smoke, or burn areas in the first iteration of stage 1 were allocated to land cover types by restricting the explanatory variables of the second iteration PLR models to seasonal ETM+ imagery that did not have a given season's TF (Table 1). Land cover types identified in each iteration of stage 1 were then merged to produce our final land cover map. Land cover training data, used to develop our stage 1 classification model, were collected through image and photo interpretation. To assess the accuracy of our stage 1 land cover map, we used a cross-validation technique that estimated the level of agreement (kappa), on a scale of -1 to 1, between observed and predicted land cover types (SAS® 2005).

Table 1. Land cover types, cross-validated MLAR accuracies, and the number of training points for stage 1 land cover classification.

Land Cover Types	Training Points	Cross-Validated User Accuracy (%)	Cross-Validated Producer Accuracy (%)
Winter Burn *	355	95	95
Winter Smoke *	63	98	100
Fall Burn *	244	85	82
Fall Clouds *	98	92	98
Fall Smoke *	82	89	89
Spring Burn *	637	92	90
Spring Clouds *	176	88	88
Spring Smoke *	132	80	83
Fields	547	99	98
Bare Ground / Urban	144	99	99
Deciduous	341	91	92
Evergreen	341	82	85
Water	438	99	99
Wet Vegetated Areas	265	86	87

\* Confounding seasonal TF. Once identified, these TF were reanalyzed using a subset of ETM+ bands representing seasons not affected by the season of the TF and classified as one of the six remaining land cover types.

Stage 2 of our IHCS generated a series of forested ecosystems probability distributions using field-interpreted samples, ETM+ spectral values, and DEMs. Similar to stage 1, TF had confounding effects on forested ecosystem probabilities. To account for these effects, an iterative scheme, as described in stage 1, was used to generate a series of ecosystem PLR models. The probability distribution of each ecosystem iteration was constrained to Deciduous and Evergreen land cover types using corresponding stage 1 iterations. Forested ecosystem probabilities were then merged, using each iteration of stage 2, to produce a final probability distribution for each forested ecosystem.

Forested ecosystem types were defined for our project as systems composed of primarily one overstory species (i.e., one species makes up at least 75 percent of the overstory, Table 2). Longleaf ecosystem types were split into two basic subgroups, Coastal Plain Longleaf ecosystems and Mountain Longleaf ecosystems, based on density, topography, species composition, and moisture availability (after Peet and Allard 1993).

Field data were collected for each of the ecosystem types and related to ETM+ imagery and DEMs using ground coordinates collected with a global positioning system (GPS). Due to access availability and the presence of large, contiguous, coniferous, and deciduous stands on public lands (Outcalt and Sheffield 1996), field data were primarily collected in national forests,

national wildlife refuge areas, and military installations. Map accuracy, kappa estimates, and model validation were assessed using independent field samples and an MLAR (Hogland et al. *in progress*).

To simplify our PLR models, redundant and insignificant explanatory variables were removed from each stage/iteration of our IHCS using a stepwise procedure (SAS 2005). Thresholds for variables entering and staying in each PLR model were set at a significance level of 0.15 and  $\leq 0.05$ , respectively.

## Results

We developed a statistically significant PLR model for the first iteration of stage 1 in our IHCS ( $X^2_{df=234} = 17,011.4$ , p-value  $< 0.0001$ ;  $\tilde{R}^2 = 0.9953$ ). Overall accuracy for this model, using an MLAR, was 92 percent with a mean kappa score of 0.91 (95% CI; 0.90, 0.92). In this model, all Landsat ETM+ bands contributed significantly to our ability to distinguish land cover and TF types ( $\alpha \leq 0.05$ ). For pixels categorized as one of the TF types, we developed statistically significant PLR models with high overall accuracies and kappa scores (Table 3). In these models, some ETM+ bands did not significantly contribute to our ability to distinguish land cover types (at  $\alpha \leq 0.05$ ) and subsequently were removed (Table 4). Using an MLAR, land cover types were assigned to each pixel across our study area (Figure 1).

Table 2. Ecosystem types, validated MLAR accuracies, predicted accuracies, and the number of samples for the first iteration ecosystem classification.

Ecosystem Types		User* Validated Accuracy (%)	Lower (95% CI) Predicted User Accuracy (%)	Upper (95% CI) Predicted User Accuracy (%)	Producer* Validated Accuracy (%)	Training Sample Size	Validation Sample Size
Slash		58	39	78	60	127	36
Hardwoods		75	72	98	83	94	186
Mixed		35	32	84	41	89	147
Longleaf	Mountain Longleaf	100	56	98 <sup>+</sup>	79	86	15
	Coastal Plain Longleaf	66	41	78	62	130	187
Loblolly		64	31	82	69	96	25

\* User and producer accuracies were adjusted for unequal sample size and refer to the probability of accurately classifying an observed ecosystem type versus the probability of accurately classifying a predicted ecosystem type, respectively.

<sup>+</sup>Observed value not within 95 percent confidence interval.

Table 3. Stage 1 land cover model statistics. Model naming convention identifies ETM+ imagery used in each PLR model (i.e., minus winter indicates that all the winter imagery was removed from that PLR model, thereby removing the confounding winter TF).

Model Iteration	Model Name	Chi-Square	Degrees of Freedom	p-value	$\tilde{R}^2$	Overall Accuracy (%)	Mean Kappa
1	All ETM+ Imagery	17011.4	234	< 0.0001	0.9953	92	0.91
2	Minus Winter	6487.23	55	< 0.0001	0.9879	95	0.94
	Minus Fall	6559.06	45	< 0.0001	0.9894	95	0.94
	Minus Spring	6606.11	55	< 0.0001	0.9903	95	0.95

For stage 2, iteration 1 of our IHCS, we generated a statistically significant PLR model that accurately predicted forested ecosystem probability distributions ( $X^2_{df=60} = 1241.57$ , p-value < 0.0001;  $\tilde{R}^2 = 0.89$ ). In this model, not all Landsat ETM+ bands significantly contributed to our ability to distinguish among ecosystem types (Table 4). Overall independent classification accuracy for this model, using an MLAR, was 66 percent, with a mean kappa score of 0.60 (95% CI, 0.56, 0.63). Based on an independent measure of model fit, this model accurately predicted pixel probabilities for all but the Mountain Longleaf ecosystem type (Table 2), indicating a good model fit. Subsequent ecosystem PLR models, for field samples occurring in TF types, indicated similar trends. Applying these ecosystem models back to the imagery produced an accurate depiction (i.e.,

good model fit) of the probability distribution of each ecosystem type across our study area (Figure 2). Using an MLAR, the most probable ecosystem type was assigned to each pixel across our study area (Figure 3).

### Discussion

We accurately mapped longleaf and other coniferous and deciduous ecosystem probability distributions across portions of the Southeast using PLR and an IHCS. These data sets can be used to identify the most probable locations of longleaf ecosystems, to identify potential longleaf ecosystem restoration sites, and to incorporate ancillary data sets to prioritized restoration locations. By weighting the area of each pixel

Table 4. Landsat ETM+ bands that were removed from each PLR model in our IHCS. ETM+ bands (rows) that have an “x” in one of the stage 1 or 2 PLR models (columns) were removed from that analysis.

Season / Band	PLR Land Cover Models				PLR Ecosystem Models			
	Initial	Minus Winter	Minus Fall	Minus Spring	Initial	Minus Winter	Minus Fall	Minus Spring
Winter Band 1		x				x		
Winter Band 2		x			x	x	x	x
Winter Band 3		x	x			x		
Winter Band 4		x				x		
Winter Band 5		x				x		x
Winter Band 7		x		x	x	x		
Fall Band 1			x				x	
Fall Band 2			x				x	
Fall Band 3			x				x	
Fall Band 4			x				x	
Fall Band 5			x		x		x	
Fall Band 7			x		x	x	x	x
Spring Band 1			x	x				x
Spring Band 2				x				x
Spring Band 3		x		x				x
Spring Band 4				x				x
Spring Band 5			x	x	x		x	x
Spring Band 7				x	x		x	x

by the probability of each ecosystem, managers can obtain a spatially explicit estimate of the amount of ecosystem area for a predefined location. If managers want a certain level of assurance of area estimates, they can incorporate model error into their area calculation, producing area confidence intervals.

PLR was chosen as our classification technique based on its flexibility and modeling assumptions (Hosmer and Lemeshow 1989; Agresti 2002; Johnson and Wichern 2002; Hogland et al. *in progress*). The flexibility in the PLR methodology comes from its focus on directly modeling class probabilities, the way in which it estimates means and variances (i.e., multinomial distribution), and the way it estimates slope parameters (i.e., maximum likelihood estimation). This allows for categorical and continuous explanatory variables, and maintains useful model building tools that help assess issues of parsimony, overparameterization, and model fit. In addition, PLR estimates model error, thus providing a way to determine a level of confidence in modeled probabilities.

While PLR is a very flexible and robust classification technique, there are potentially a few drawbacks. The first deals with

the efficiency of PLR (Efron 1975). With today’s computers, though, this is no longer an issue. Second, PLR cannot solve a maximum likelihood estimate (MLE) for beta values when there is no overlap in class explanatory values. While an unsolvable MLE may be troubling in terms of mathematic complexity and model fit estimates (Agresti 2002), viewed from a classification perspective this situation means that some of the class types can, with 100 percent accuracy, be separated from the rest of the class types given a set of rules. In this situation, a probabilistic classification is not required. Instead, class types can be assigned using Boolean operators.

### Summary

Using multitemporal Landsat ETM+ imagery, DEMs, PLR, and an IHCS, we accurately depicted the current distribution of longleaf ecosystems. By presenting these data sets in terms of probabilities, we provide users with a more accurate representation of our classification and the flexibility needed to answer fine-scale longleaf ecosystem questions. Finally, in light of our success with PLR and our IHCS, we are incorporating these data sets and methods into the Alabama Gap Analysis Project.

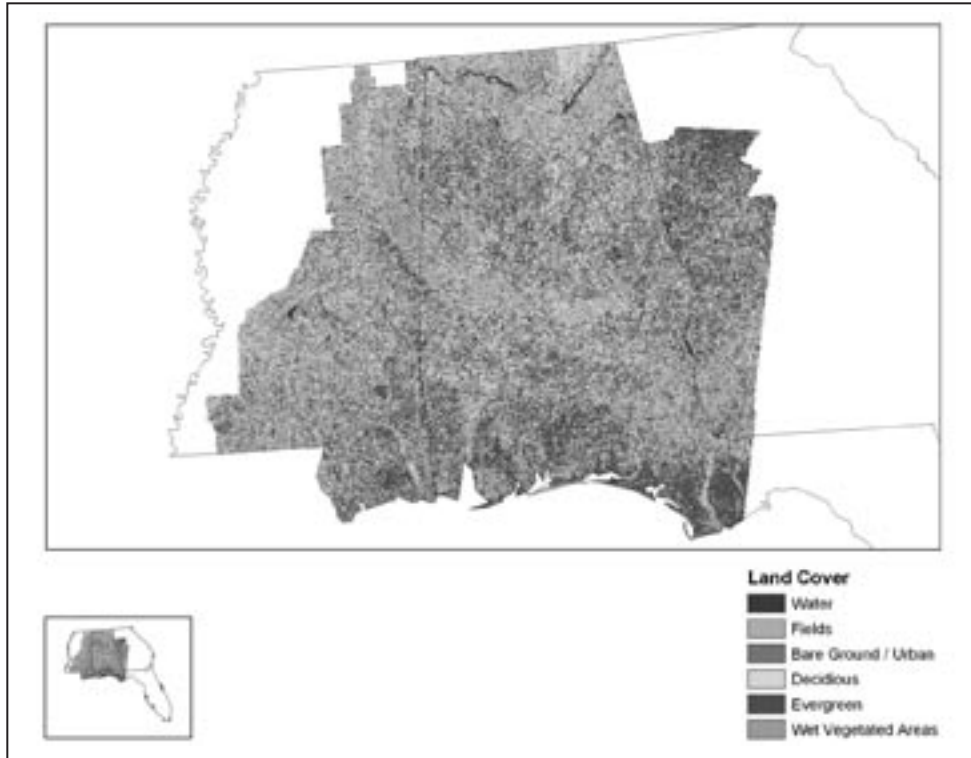


Figure 1. Final land cover map after adjusting for confounding seasonal variables.

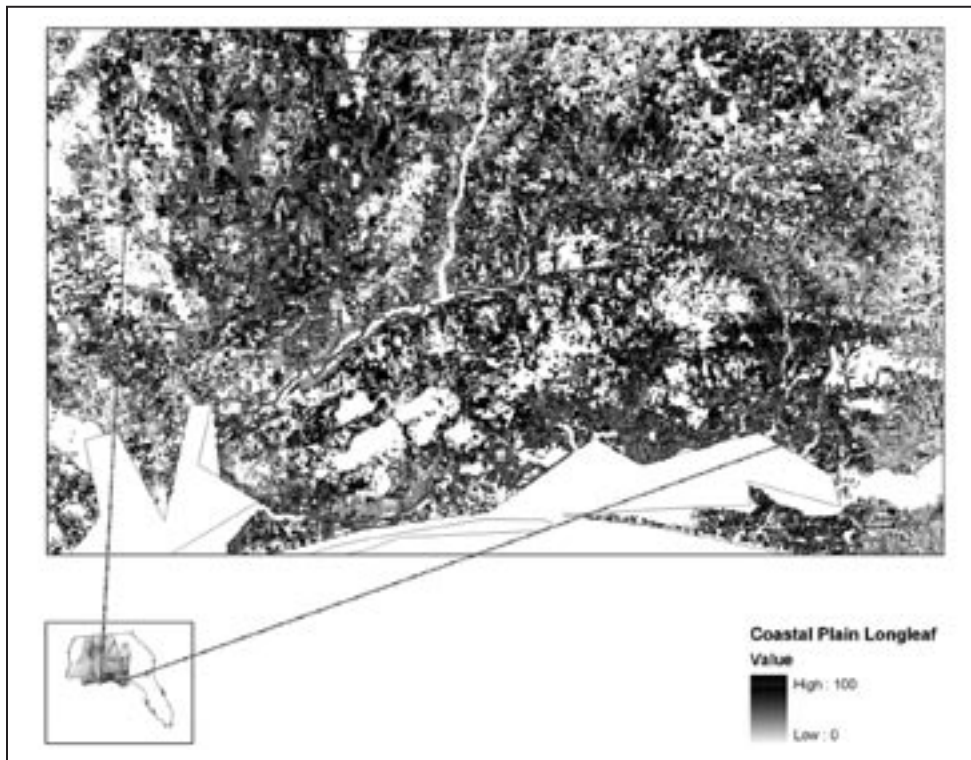


Figure 2. Example of one forested ecosystem probability distribution for Blackwater State Forest and Eglin Air Force Base, located in the panhandle of Florida. As color transitions from white to black, the probability of finding the Coastal Plain Longleaf ecosystem increases from 0 percent to 100 percent.

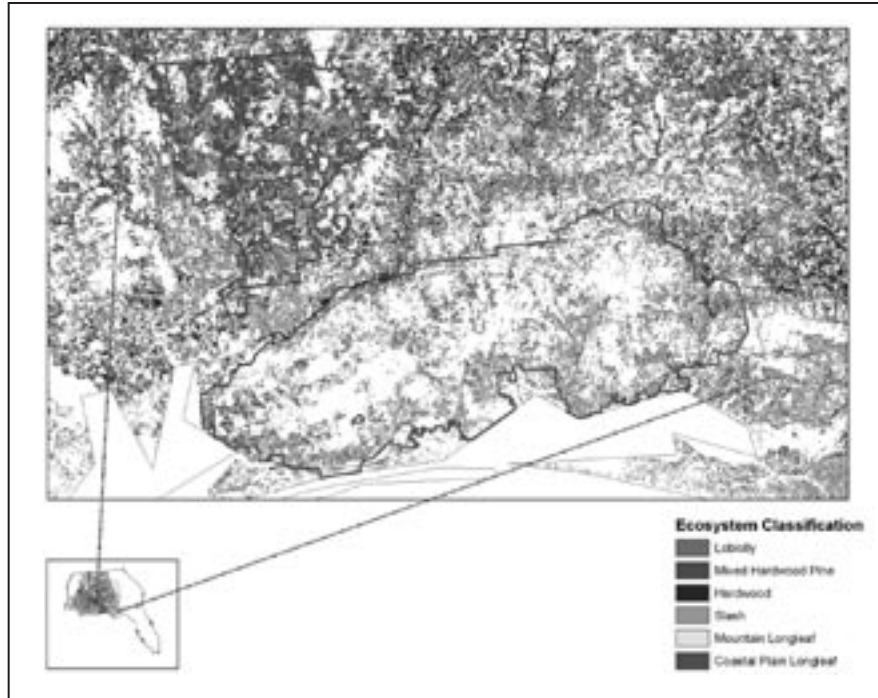


Figure 3. The most probable forested ecosystem type for Blackwater State Forest and Eglin Air Force Base, located in the panhandle of Florida, based on an MLAR. Unclassified areas (areas in white) represent a land cover type other than Evergreen or Deciduous.

### Literature Cited

- Agresti, A. 2002. *Categorical data analysis*. Hoboken, N.J.: Wiley-Interscience. 710 pp.
- Efron, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* 70: 892–98.
- ESRI®. 2005. ArcGIS Desktop Help. Available from <<http://webhelp.esri.com/arcgisdesktop/9.1/index.cfm?TopicName=welcome>>. Accessed June 7, 2005.
- Hogland, J. S., N. Billor, and M. D. MacKenzie. In progress. Comparing polytomous logistic regression and discriminant analysis: A remote sensing perspective.
- Hogland, J. S., and M. D. MacKenzie. In progress. Bringing images to a common radiometric scale using Aggregate No Change Regression: A comparison between radiometric normalization techniques.
- Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan. 2004. Development of a 2001 National Landcover Database for the United States. *Photogrammetric Engineering and Remote Sensing* 70 (7): 829–40.
- Hosmer, D. W., and S. Lemeshow. 1989. *Applied logistic regression*. New York: Wiley-Interscience. 307 pp.
- Johnson, R. A., and D. W. Wichern. 2002. *Applied multivariate statistical analysis*. Upper Saddle River, N.J.: Prentice Hall. 767 pp.
- Kelly, J. F., and W. A. Bechtold. 1990. The longleaf pine resource. In proceedings of symposium on the management of longleaf pine, April 4–6, 1989, Long Beach, Miss., 11–22. Department of Agriculture, Forest Service, Southern Forest Experiment Station.
- NASA. 2005. Landsat 7 science data users handbook. Available at <[http://ftpwww.gsfc.nasa.gov/IAS/handbook/handbook\\_toc.html](http://ftpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html)>. Accessed January 5, 2005.
- Noss, R. F., E. T. LaRoe, and J. M. Scott. 1995. Endangered ecosystems of the United States: A preliminary assessment of loss and degradation. *Biological Report* 28. National Biological Service, Washington, D.C.
- Outcalt, K. W., and R. M. Sheffield. 1996. The longleaf pine forest: Trends and current conditions. Resource Bulletin SRS-9. Asheville, N.C.: U.S. Department of Agriculture, Forest Service, Southern Research Station. 23 pp.
- Peet, R. K., and D. J. Allard. 1993. Longleaf pine vegetation of the southern Atlantic and eastern gulf coast regions: A preliminary classification. In *The longleaf pine ecosystem: Ecology, restoration and management*, proceedings of the Tall Timbers Fire Ecology Conference no. 18., Tallahassee, Fla.
- SAS®. 2005. “SAS OnlineDoc Version Eight,” available from <<http://v8doc.sas.com/sashtml/>>. Accessed June 20, 2005.
- Tuldge, C. 1999. Plant (protecting endangered species). *Index on Censorship* 28: 164–66.