A LANDSAT STAND BASAL AREA CLASSIFICATION SUITABLE FOR AUTOMATING STRATIFICATION OF FOREST INTO STATISTICALLY EFFICIENT STRATA

Emily B. Schultz[a], Thomas G. Matney[a], David L. Evans[a], and Ikuko Fujisaki[b]

[a]Forest and Wildlife Research Center, Department of Forestry, Mississippi State University, Box 9861, Mississippi State, MS 39762 USA – (eschultz, tmatney)@cfr.msstate.edu; dle@sitl.cfr.msstate.edu
[b] Fort Lauderdale Research and Education Center, University of Florida, 3205 College Ave, Davie, FL 33314 USA - ikuko@ufl.edu

KEY WORDS: Accuracy, Estimation, Forestry, Inventory, Modeling, Multispectral, Neural, Sampling

ABSTRACT:

Operational large-scale inventories, such as the State of Mississippi's inventory system, must have algorithms to rapidly categorize the forest resource into statistically efficient strata. These procedures must be repeatable and automated for timely and cost effective implementation of short-cycle regional inventory estimation. In addition to longer-term stand structural changes, Landsat Thematic Mapper (TM) data allows the detection of change in forest strata caused by forest management (harvests and thinnings) and natural disasters (2005 hurricane Katrina) on a yearly basis. These classifications are necessary for 1) construction of the inventory sampling frame, 2) display of the current spatial distribution of forest resources and calculation of land areas by strata, and 3) calculation of total area volumes. In order to achieve significant gains through stratification in large-scale inventories, the forest resource must be classified not only by broad timber type but also by a measure of size per unit area. In the past, large-scale inventories have stratified only on the basis of crude GIS timber types such as pine, mixed-pine-hardwood, and hardwood. These strata are not separated on the basis of the variables of interest, volume or size; hence, they are not adequate for achieving precision gains due to stratification which result in decreased required sample size and measurement cost.

Neural networks and logistic regression approaches were used to estimate forest stand basal area per hectare from Landsat Enhanced Thematic Mapper Plus (ETM+) image data of a 227,873 hectare forested area in four counties of Mississippi. Basal area was estimated by forest timber type (pine, mixed-pine-hardwood, and hardwood) into distinct size classes to generate strata yielding a minimum of 25% gain in the precision of stratified random sampling volume estimators resulting in a very significant reduction of sample size and inventory cost. These modeling methods readily lend themselves to the automation necessary for repeated short-cycle inventory assessments, when suitable training data sets are available for the current inventory area.

1. INTRODUCTION

Operational large-scale inventories, such as the State of Mississippi's inventory system, must have algorithms to rapidly categorize the forest resource into statistically efficient strata. These procedures must be repeatable and automated for timely and cost-effective implementation of short-cycle regional inventory estimation.

In order to achieve significant gains through stratification in large-scale inventories, the forest resource must be classified not only by broad timber type for administrative, reporting, and visualization purposes, but also by a measure of timber size per unit area. In the past, large-scale inventories have stratified only on the basis of crude GIS timber types such as pine, mixed-pine-hardwood, and hardwood (Bauer et al., 1994). These strata are not separated on the basis of the variables of interest, volume or size; hence, they are not adequate for achieving precision gains due to stratification which result in decreased required sample size and cost.

Statistically efficient gains from stratified sampling frames are dependent upon separation of the means and variances of the response variable of interest.

Cochran (1977) gives the following formula for estimation of statistical gains from optimal allocation of simple random samples to strata.

$$V_{ran} = V_{opt} + \frac{\sum_{h=1}^{h=L} N_h\left(S_h - \overline{S}\right)^2}{nN} + \frac{\sum_{h=1}^{h=L} N_h\left(\overline{Y}_h - \overline{Y}\right)^2}{nN} \quad (1)$$

where:

V_{ran} is the variance of a simple random sample of size n,
V_{opt} is the variance with the n samples optimally allocated to the strata,
h is the strata index,
L is the number of strata,
$N_h, \overline{Y}_h, and\ S_h$ are the population size, mean, and standard deviation of the h th stratum,
$\overline{S}\ and\ \overline{Y}$ are the strata-weighted mean, and standard deviation, and
n and N are the total sample, and population size.

Similar gain estimation equations can be derived for cluster, multi-stage, and two-phase sampling designs.

The two far right hand side terms of equation 1 are the reduction in the variance of the mean due to the separation of strata means and standard deviations in an optimally allocated sample of total size n. Thus, to maximize gains in a stratified

inventory, the stratification ancillary variable must be chosen to provide maximum separation of strata means and variances of the response (inventory design) variable to be estimated.

Developing statistically efficient stratified sampling frames for large-scale forest inventories from remotely sensed data, thus, requires an equation to predict a measure of size variable strongly related to targeted volume response variables from the remotely sensed data. This equation can than be used to stratify the image into strata that have enough separation in mean and variance in the targeted volume response variables to yield significant efficiency gains due to an optimum sample allocation. Stratification on the basis of a simple stand composition categorization like pine, mixed-pine-hardwood, and hardwood are only useful for administrative, reporting, and visualization purposes because there is little or no separation in means or variance. Stratification always produces gains but minimal gains, generally, will not pay the cost of stratification.

The objectives of this paper are to:

1. Apply linear regression and neural network modeling methods to multi-spectral (Landsat ETM+) data to estimate a measure of size variable (basal area per unit area) highly correlated to the primary response variable (total stem outside bark cubic volume per unit area), and

2. Post stratify the inventory study area into typical size class based strata, and estimate the gains obtained by the stratifications.

2. BACKGROUND

Recent advances in change detection technology and classification methodologies and procedures are making it possible to use multi-spectral data to provide accurately stratified images, not only by crude GIS forest types but also by density classes (Berryman, 2004; McCombs et al., 2003) and growth stages (Fujisaki et al., 2005). Berryman (2004) used multi-spectral Landsat EMT+ (30m) data and light detection and ranging (lidar) data to classify images of natural southeastern US pine stands by number of trees per unit area and height. McCombs et al. (2003) used small footprint lidar and high resolution multispectral (0.61m) data to identify individual trees, mean height, and trees per unit area in pine plantations. Fujisaki et al. (2005) characterized natural forest stands into regeneration-immature, intermediate, and mature growth classes using bands 1 – 4 of Landsat TM data. Large-area inventories (Parker et al., 2005) can use these advanced technologies to develop stratified sampling schemes resulting in increased precision with fewer required plots to meet specified sampling error at stated confidence levels. Fewer field plots translate into reduced costs, which are significant for large inventories. Statistically efficient stratification of the resource will also result in improved estimation of merchantable volumes, biomass, and sequestered carbon at reduced costs.

The forest products industry is a major component of Mississippi's economic base. Timber is one of Mississippi's most valuable crops and accounts for more than $1 billion of harvested forest products annually (Munn and Tilley, 2005). The State has 6.8 million forested hectares and over 200,000 landowners. The amount of pine and hardwood stumpage utilized in 2001 resulted in $801 million in payments to Mississippi landowners.

The Mississippi Institute for Forest Inventory (MIFI) is meeting the need for an accurate and spatially based inventory of the State's forest resources. Mississippi has been divided into five forest regions, and one region is inventoried each year, on a rotating basis, to generate current volume estimates by species type, age class, and land ownership. Allocation of sample plots is designed to achieve a 10 – 15% sampling error for total cubic foot volume outside bark with 95% confidence at the county area level.

The operational objective of this research is to reduce the cost while improving the estimates of current and future inventories of timber volume. Because of cooperatively funded research and development efforts between MIFI and the Mississippi State University Forest and Wildlife Research Center (FWRC), the results of this work are uniquely positioned to have direct impact on a large-area forest inventory and the forest products economic sector of Mississippi.

3. METHODS

The data utilized in this study are associated with the 1999 Mississippi Forest Inventory Pilot Program (Parker et al., 2005) that directly preceded MIFI's initial operational inventory in 2004. Random plots were allocated to a 227,873 hectare forested area in four counties of east central Mississippi using stratified random sampling criteria based on GIS forest cover types (pine, mixed-pine-hardwood, and hardwood) derived from 1999 Landsat ETM+ leaf-on, 30-meter resolution, multispectral satellite imagery. Determination of forest cover types was based on the standard normalized difference vegetation index, NDVI, (Sader, et al., 2003). US Geological Survey (USGS) imagery was obtained from the Global Land Cover Facility Web site, http://glcf.umiacs.umd.edu/index.shtml, supported by the National Aeronautical and Space Administration (NASA) and the University of Maryland. The number of plots allocated to each county in order to achieve a +- 10% sampling error at the 95% confidence level was estimated from the variability for total cubic volume. Plots were allocated optimally based on forest cover type, but because the mean and variance separation of the cover types were minimal, the allocation for all practical purposes was proportional. Global positioning system (GPS) units were used to navigate to plots where field inventory data were collected.

A one-hidden layer, four-element back propagation neural network and multi-linear regression analysis (Table 1) were applied to the Landsat ETM+ multispectral data to develop the best possible system for estimating basal area. Mulitspectral bands 1-5, and 7, and two derived variables, normalized difference vegetation index (NDVI) and normalized difference moisture index (NDMI) (Sader et al., 2003), were utilized. Only procedures that can be readily automated were investigated. Unless a system can be automated, it is not efficient for large-area inventories that require extensive image processing for rotating regional inventories on an annual basis or inventories monitoring frequent change.

Considerable effort was expended developing models that included interactions of raw band multispectral data and the derived variables. Past researchers have often not devoted sufficient time to increasing classification accuracy by including interaction terms in models. Neural networks were

investigated, in addition to regression analyses, because of their native capability of detecting complex interaction relationships.

GIS forest type	Regression coefficients			N	$S_{y.x}$	R^2
	a	b	c			
Pine 1	51.768	.2795	-.50342	128	5.933	44.5
Mix 2	83.995	-.7732	-.38015	251	6.523	40.6
Hard wood 3	90.814	-.5783	-.47936	109	7.584	29.5

Table 1. Regression coefficients and fit statistics (no. of observations, standard error of prediction, and coefficient of determination) for predicting basal area (BA) per hectare by pine, mixed-pine-hardwood, and hardwood forest GIS types using leaf-on Landsat ETM+ multi-spectral bands 2, 3, and 5 for four counties in Mississippi, USA, in 1999, where BA = a + b (band 3) + c ((band 2 * band 5) / (band 3)).

4. RESULTS

Tables 2 and 3 summarize for the neural network and regression basal area per unit area prediction models the

1. Indices of fit, I^2, (one minus the quantity of the error sum of squares divided by the total sum of squares) and root mean squared errors ,RMSE, by GIS forest type and

2. User and producer errors associated with the use of the models to post classify (stratify) the 488 plots into three basal area classes.

Producer error is the number of plots out of the total number of plots for a forest type (pine, mixed, and hardwood) that were incorrectly identified by basal area class. User error is defined as the number of one basal area class identified, as that basal area class, that were identified incorrectly. The overall percent error was calculated as 100 times the number of incorrectly classed plots divided by the total number of plots. The percent accuracy is 100 minus the percent error.

The basal area limits were chosen to represent what the authors considered to be reasonable low (0 – 19.5 m^2/ha], medium (19.5 – 26.4 m^2/ha], and high basal (26.4 – ∞ m^2/ha) area per area for the study area.

The strength of the relationships between basal area and the band data were better than anticipated. As expected, the pine models were the best and the hardwood models were worst models. The enormous number of hardwood forest types in the study area increases the experimental noise tremendously. In terms of the ability of the neural network and regression models to predict basal area and accurately post stratify an image into basal area classes, there was no practical difference. The neural network model produced slightly higher indices of fit and lower root mean squared errors than the regression models. However, the regression model had a lower classification error rate. The neural network was expected to outperform the regression model because of its native ability to model complicated interactions. The regression probably performed as well as the neural network because a great deal of effort was put into developing its functional form including interaction terms.

Assessment of the capability of the basal area prediction system to create strata that would produce large gains in precision are shown in Table 4. These data demonstrate that stratification systems based on basal area prediction can yield a 25% gain in precision and about a 25% reduction in the number of plots necessary to meet the required precision and confidence level of the inventory. This gain is relatively large compared to the meager 1% precision gain obtained when stratification is done on simple GIS forest types.

Table 4 was prepared by determining for the GIS forest type stratification the number of plots (150) when optimally allocated to the strata that would yield an estimated volume within +- 10% of the true means volume at the 95% confidence level. The 150 plots were then optimally allocated to the stratifications of basal area, GIS forest type, and basal area within GIS forest type. Using these allocations the gains in precision due to stratification were calculated according to Equation 1. Potential reduction in sample size due to stratification were determined by calculating the numbers of samples required to estimated the mean volume within +- 10% of the true mean volume at the 95% confidence level.

Table 5 shows the design statistics and optimal allocation for each classification (stratification) shown in Table 4.

5. DISCUSSION AND CONCLUSIONS

The basal area per unit area prediction strength of the regression and neural network models were for all practical purposes identical. In both models for all GIS types, the significant variables were bands 2, 3, and 5. The interaction term (band 2 * band 5) / (band 3) accounted for most of the variation in basal area per unit area. The model term, band 3, while accounting for a small amount of the variation, acts as a correction term to compensate for overshooting by the interaction term. After inclusion of the raw band information in the regression and network models, the derived variables, NDVI and NDMI, did not enter into the models.

The indices of fit for the models and the ability of the models to accurately post stratify the image into basal area classes, were surprisingly good. These equations allowed a stratified efficiency gain of 25% and reduction of required sample size of 20% for the 3-basal area classification example chosen. While the gains in precision are good, there is great opportunity for even larger gains by improving the relationship of remotely sensed data and measures of stand density. We are currently investigating the use of lidar data and change detection variables as possible means of improving the accuracy of stand density estimates.

Results were only reported for a typical basal area classification that would be implemented by the MIFI inventory system. Trials with numerous 3- and 4-class basal area classifications (different class limits) yielded similar gains. It is worth noting that in these trials we observed that classification accuracy was very sensitive to number of classes and class limits. Many of the classifications we examined had better accuracies and larger

Pine GIS forest type							
		Estimated basal area class			Total no.	Producer	% Producer
		1	2	3	plots	error	error
Observed	1	22	16	2	40	18	45.0
basal area class	2	4	32	7	43	11	25.6
	3	2	18	25	45	20	44.4
Total no. plots		28	66	34	128		
User error		6	34	9		49[a]	
% User error		21.4	51.5	26.5			38.3[b]
RMSE		5.76			I^2	0.46	
Mixed GIS forest type							
		Estimated basal area class			Total no.	Producer	% Producer
		1	2	3	plots	error	error
Observed	1	67	44	1	112	45	40.2
basal area class	2	23	40	9	72	32	44.4
	3	6	41	20	67	47	70.1
Total no. plots		96	125	30	251		
User error		29	85	10		124[a]	
% User error		30.2	68.0	33.3			49.4[b]
RMSE		6.36			I^2	0.43	
Hardwood GIS forest type							
		Estimated basal area class			Total no.	Producer	% Producer
		1	2	3	plots	error	error
Observed	1	23	22	1	46	23	50.0
basal area class	2	9	19	5	33	14	42.4
	3	4	20	6	30	24	80.0
Total no. plots		36	61	12	109		
User error		13	42	6		61[a]	
% User error		36.1	68.0	50.0			56.0[b]
RMSE		7.44			I^2	0.30	
All GIS forest types							
		Estimated basal area class			Total no.	Producer	% Producer
		1	2	3	plots	error	error
Observed	1	112	82	4	198	86	43.4
basal area class	2	36	91	21	148	57	38.5
	3	12	79	51	142	91	64.1
Total no. plots		160	252	76	488		
User error		48	161	25		234[a]	
% User error		30.0	63.9	32.9			48.0[b]
RMSE		6.47			I^2	0.41	

Table 2. Error matrices, producer and user errors, overall error ([a]) and overall percent error ([b]) for classifying basal area per hectare by pine, mixed-pine-hardwood, hardwood, and all (pine, mixed, and hardwood) forest GIS types using leaf-on Landsat ETM+ multi-spectral bands 1 - 5 and 7 and neural network modeling root mean squared error (RMSE) and index of fit (I^2) for four counties in Mississippi, USA, in 1999.

Pine GIS forest type							
		Estimated basal area class			Total no. plots	Producer error	% Producer error
		1	2	3			
Observed basal area class	1	20	16	4	40	20	50.0
	2	3	33	7	43	10	23.3
	3	2	18	25	45	20	44.4
Total no. plots		25	67	36	128		
User error		5	34	11		50[a]	
% User error		20.0	50.7	30.6			39.1[b]
RMSE		5.85			I^2	0.45	
Mixed GIS forest type							
		Estimated basal area class			Total no. plots	Producer error	% Producer error
		1	2	3			
Observed basal area class	1	64	47	1	112	48	42.9
	2	20	44	8	72	28	38.9
	3	5	42	20	67	47	70.1
Total no. plots		89	133	29	251		
User error		25	89	9		123[a]	
% User error		28.1	66.9	31.0			49.0[b]
RMSE		6.47			I^2	0.41	
Hardwood GIS forest type							
		Estimated basal area class			Total no. plots	Producer error	% Producer error
		1	2	3			
Observed basal area class	1	23	22	1	46	23	50.0
	2	8	18	7	33	15	45.5
	3	4	16	10	30	20	66.7
Total no. plots		35	56	18	109		
User error		12	38	8		58[a]	
% User error		34.3	67.9	44.4			53.2[b]
RMSE		7.48			I^2	0.30	
All GIS forest types							
		Estimated basal area class			Total no. plots	Producer error	% Producer error
		1	2	3			
Observed basal area class	1	107	85	6	198	91	46.0
	2	31	95	22	148	53	35.8
	3	11	76	55	142	87	61.3
Total no. plots		160	256	83	488		
User error		42	161	28		231[a]	
% User error		28.2	62.9	33.7			47.3[b]
RMSE		6.57			I^2	0.40	

Table 3. Error matrices, producer and user errors, overall error ([a]) and overall percent error ([b]) for classifying basal area per hectare by pine, mixed-pine-hardwood, hardwood, and all (pine, mixed, and hardwood) forest GIS types using leaf-on Landsat ETM+ multi-spectral bands 2, 3, and 5 and regression modeling root mean squared error (RMSE) and index of fit (I^2) for four counties in Mississippi, USA, in 1999.

Classification	Percentage gain: Due to mean difference	Percentage gain: Due to variance difference	Percentage gain: Total	Required sample size (n)*
GIS forest type	0.40	0.41	0.81	150
Basal area (BA)	25.13	0.14	25.27	113
GIS forest type and BA	27.97	0.59	28.56	109

* Required samples (n) optimally allocated to produce an estimated mean total cubic volume outside bark within +-10% of the true mean at the 95% confidence level.

Table 4. Estimated gains over simple random sampling due to optimum allocation of a sample of 150 plots to GIS forest type, basal area, and GIS type and basal area classifications for the study area. The 150 plots are the required number when optimally allocated to the GIS forest type. Classes estimate the total cubic foot volume outside bark within +-10% of the true mean at a 95% confidence percent.

GIS Forest type stratification					
Strata	Area (ha)	Number of plots	Mean volume (m^3/ha)	Std. of volume (m^3/ha)	Required samples(n)*
Pine	43511	128	154.7	89.0	25
Mixed	143855	251	163.1	104.3	97
Hardwood	40506	109	175.9	108.5	28
Total or mean	227873	488	163.8	102.1	150
Basal area (BA) only stratification					
Low BA	72513	149	93.0	83.4	45
Medium (Med) BA	119812	256	181.2	88.3	79
High (Hi) BA	35548	83	237.1	94.9	25
Total or mean	227873	488	163.8	87.7	113
Basal area (BA) within GIS forest type stratification					
Pine-Low BA	8498	25	76.9	91.5	4
Pine-Med BA	22775	67	157.3	73.7	9
Pine-Hi BA	12238	36	203.9	77.0	5
Mixed-Low BA	51008	89	94.4	81.6	23
Mixed-Med BA	76226	133	186.8	90.8	38
Mixed-Hi BA	16620	29	265.4	93.8	9
Hardwood-Low BA	13007	35	101.0	82.7	6
Hardwood-Med BA	20810	56	196.4	93.9	11
Hardwood-Hi BA	6689	18	257.9	112.1	4
Total or mean	227873	488	163.8	87.0	109

*Number of sample plots required for the combined sample mean to be within +-10% of the true mean at the 95% confidence level.

Table 5. Schemata and design statistics for calculating statistical efficiency gains of optimal allocation of samples to strata for the 227,873 ha of forested area in four counties of east central Mississippi USA.

efficiency gains than the typical basal area classification that we reported on here. Considerable effort should, thus, be expended on choosing class numbers and limits to yield maximum statistical efficiency gains due to stratification. The authors are now investigating procedures for calculating optimum class limits for given numbers of classes.

The results of this promising research will be employed in subsequent MIFI inventories to stratify by basal area within GIS forest type. The current inventory reporter software that uses simple GIS forest type stratification is available on the Web at www.mifi.ms.gov/mission.htm. At a minimum, using the models reported here, a reduction in the number of plots required to meet inventory precision specifications should lower the cost of plot measurement by at least 25%. With the expectation of additional model improvement, increased gains and further reduction of costs are anticipated.

These procedures are reproducible and automatable for timely and cost effective implementation of short-cycle regional inventory estimation. Even though these equations were not calibrated against additional images and would produce biased estimates of basal area, the basal area classes derived by using these equations would still exhibit the required mean separation. Gains are achieved by relative separation of means and are not dependent upon an extremely accurate estimated of basal area, thus, the basal area classification is readily automatable.

6. REFERENCES

Bauer, M.E., T.E. Burk, A.R. Ek, P.R. Coppin, S.D. Lime, T.A. Walsh, D.K. Walters, W. Befort, and D.F. Heinzen, 1994. Satellite Inventory of Minnesota Forest Resources. Photogrammetric Engineering & Remote Sensing, 60(3), pp. 287-298.

Berryman, B.N. 2004. Investigation of Relationships Between Landsat ETM+ Data and Ground-Adjusted LiDAR Measurements in Southern Pine Stands. MS Thesis, Mississippi State University, Mississippi State, MS USA, 50 p.

Cochran, W.G. 1977. Sampling Techniques. John Wiley and Sons, Inc, New York, pp. 99–101.

Fujisaki, I, P.D. Gerard, and D.L. Evans. 2005. Classification of forest growth stage using Landsat TM data. In proceedings of Optics & Photonics: Remote Sensing and Modeling of Ecosystems for Sustainability II. San Diego, California, 31 July – 4 August, 2005.

McCombs, J.W., S.D. Roberts, and D.L. Evans, 2003. Influence of fusing lidar and multispectral imagery on remotely sensed estimates of stand density and mean tree height in a managed loblolly pine plantation. Forest Science, 49(3), pp. 457-466.

Munn, I.A. and B.K. Tilley. 2005. Forestry in Mississippi: The impact of the forest products industry on the Mississippi economy, an input output analysis. Forest and Wildlife Research Center, Research Bulletin FO301, Mississippi State University. 26 pp.

Parker, R.C., P.A. Glass, H.A. Londo, D.L. Evans, K.L. Belli, T.G. Matney, and E.B. Schultz. 2005. Mississippi's Forest Inventory Pilot Program: Use of Computer and Spatial Technologies in Large Area Inventories. FWRC Research Bulletin No. 274, Mississippi State University, Mississippi State, MS USA, 22 p.

Sader, S.A., M. Bertrand, and E.H. Wilson. 2003. Satellite change detection of forest harvest patterns on an industrial forest landscape. Forest Science, 49(3), pp. 341-353.