Mapping Hemlocks via Tree-Based Classification of Satellite Imagery and Environmental Data

Frank H. Koch[1], Heather M. Cheshire[2], and Hugh A. Devine[3]

[1]Department of Forestry and Environmental Resources, North Carolina State University, USDA FS Forest Health Monitoring Program, Forestry Sciences Laboratory, 3041 Cornwallis Road, Research Triangle Park, NC 27709

(919) 549-4010; fkoch@fs.fed.us

[2]Department of Forestry and Environmental Resources, North Carolina State University

[3]Department of Parks, Recreation, and Tourism Management, North Carolina State University

ABSTRACT

Within the last few years, the hemlock woolly adelgid (HWA) has made significant inroads into the southern Appalachians. Since the region's native hemlock species are not resistant, timely application of control measures is critical to minimizing hemlock mortality. Unfortunately, hemlock stands in the region are incompletely mapped, and general characteristics of their distribution present serious mapping challenges. One approach for improving classification is to integrate medium-resolution satellite imagery (Landsat, ASTER) and ancillary environmental data. We tested such an approach using images from eastern and western study areas in Great Smoky Mountains National Park. First, we created maps for masking out non-evergreen pixels via unsupervised classification (i.e., cluster busting) of winter images. We then applied the masks to corresponding summer images so we could separate hemlock and non-hemlock evergreens under optimal image conditions. We extracted a large (>14,000) random sample of points from the masked images, stratifying the sample according to an aerial photoraphy-derived vegetation map of the park. At each sample point, we recorded the vegetation label as well as image data and values for a suite of topographic, environmental, and proximity variables recorded in a geographic information system (GIS). We applied a series of tree-based classifications to this training data set to create a set of decision rules that most accurately retains the input class of sample points. Our most successful tree had 79 total "leaves" (i.e., distinct decision-rule pathways). We applied these decision rules to the images to develop hemlock maps of the study area. Thematic accuracy assessment of these maps, based on field survey and photo-derived points, indicated 85% overall accuracy in the eastern study area and 69% success at capturing hemlocks in a partial assessment of the western study area. Additional accuracy assessment may offer an opportunity to refine the rules. However, our decision rules can currently be applied elsewhere in the southern Appalachian region for management planning purposes.

KEYWORDS

HWA, Landsat, ASTER, tree-based classification, southern Appalachians.

INTRODUCTION

Hemlock woolly adelgid (*HWA), Adelges tsugae* Annand, is a non-native pest threatening the southern Appalachians. In the past few years, it has made significant inroads into the region, particularly in the Great Smoky Mountains and along the Blue Ridge Parkway. HWA affects both hemlock species native to the region, eastern hemlock (*Tsuga canadensis* [L.] Carr.) and Carolina hemlock (*T. caroliniana* Engelm.). Neither is resistant, though biological control via introduced predator insects holds promise as a method for combating HWA in natural stands. Unfortunately, any counter-measure faces a couple of significant challenges. First, to understand HWA's likely distribution as it spreads throughout the southern Appalachians, the distribution of its host species must be characterized. Second, there must be some way to predict where HWA is most likely to appear next, allowing managers to prioritize their HWA management strategies. Here, we present an approach to the first challenge of mapping hemlocks in the region.

In the northeastern U.S., satellite images have been used successfully for several HWA-related mapping efforts (e.g., Bonneau et al. 1999a and 1999b, Royle and Lathrop 1997 and 2002). These studies have employed time series of Landsat TM data to characterize change in hemlock health through time with considerable accuracy. However, HWA is a relatively recent arrival in the southern Appalachians, so adelgid-induced damage is likely indistinguishable in multispectral satellite imagery. More basically, simply mapping hemlocks from satellite imagery is difficult. Hemlocks are generally found in isolated stands, often in moist valleys, coves, steep ravines, or on north-facing bluffs (Delcourt and Delcourt 2000) and are distributed throughout heterogeneous forests that include other evergreen species from which hemlocks are difficult to separate (Royle and Lathrop 1997).

Nevertheless, given a large region to survey, satellite imagery is still the most feasible tool for hemlock mapping. Incorporation of ancillary data in tree-based classification offers an alternative to the limited distinguishing power of traditional, spectral-based classification. A growing body of literature has highlighted several different approaches for tree-based methods (Murthy 1998). All start with a sample of data representing each level in the class of interest (e.g., the vegetation classes in a vegetation map) and containing values for a number of continuous or categorical variables. This input training sample is subjected to statistical partitioning techniques that identify key variables and split values to most accurately capture the classes of the input training sample points. The resulting product is a tree of decision rules that can then be implemented in GIS or image processing software as an "expert classifier". Tree-based classification approaches have a number of advantageous features. They are typically non-parametric, so conditions of normality are not critical. Furthermore, they algorithmically select variables to be included in the final tree and ignore any extraneous variables. While the commonly used classification and regression tree (CART) method is limited to binary (i.e., two-way) splits, other methods, such as chi-squared automated interaction detection (CHAID), allow for a greater number of splits (Murthy 1998). To develop

our expert classifier, we used SAS Enterprise Miner software, which allows implementation of several tree-based classification techniques. We used ERDAS Imagine and ESRI ArcGIS for image processing and GIS analysis, respectively.

STUDY AREA

We selected images covering portions of Great Smoky Mountains National Park (GRSM) in North Carolina and Tennessee. The park has a significant eastern hemlock presence in many areas (Taylor 2002). Originally, we developed our classifier using an ~482 km2 image subset from the eastern side of the park (Figure 1). To strengthen the classifier—via a process that will be explained below—we added a second, ~108 km2 study area on the western side of the park (Figure 1). This area is characterized by gentler topography than the other study area.

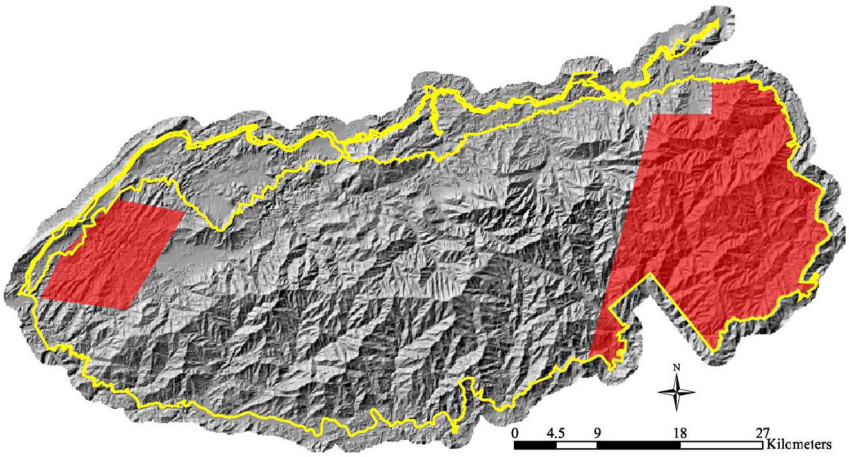

Figure 1. Eastern and western study areas in Great Smoky Mountains National Park.

METHODS

INITIAL ANALYSIS

We acquired a leaf-off October 2001 Landsat ETM+ image from the Global Land Cover Facility (*http://glcf.umiacs.umd.edu/index.shtml*) that covered our eastern study area. We also acquired a leaf-on September 2000 Terra ASTER radiance scene through the NASA Earth Observing System Data Gateway (*http://edcimswww.cr.usgs.gov/pub/imswelcome/*). We chose to use ASTER imagery because of the sensor's good spatial (15-m in the VNIR versus 28.5-m or 30-m for Landsat) and radiometric (12-bit versus 8-bit for Landsat) resolution. After converting the Landsat image to radiance, we fused the 28.5-m multispectral image with its corresponding 14.25-m panchromatic image using an algorithm developed by Halil Cakir (North Carolina State University). This brought the two images closer in spatial resolution. We geometrically corrected the fused Landsat image using a third-order polynomial equation and 92 ground control points (GCPs) collected from color-infrared digital orthophoto quarter quads (DOQQs) of the area (RMSE = 4.1420 m). The ASTER image is

actually a combination of two separate image files, a 15-m resolution, three-band image in the visible and near infrared (VNIR), and a 30-m resolution, six-band image in the short-wave infrared (SWIR), merged into a single 15-m image. Though automatically subjected to preliminary geometric correction upon import to ERDAS Imagine, we further corrected the merged ASTER image using a fourth-order polynomial and 80 GCPs (RMSE = 6.4694 m). This better aligned the ASTER image with the Landsat image. We clipped each image to fit our eastern study area.

A major obstacle when using remotely sensed data of mountainous regions is the highly variable level of ground illumination and radiance due to topographic relief, which can result in loss and alteration of image information (Jensen 1996). A number of topographic normalization equations have been proposed that correct pixel values using models of the Earth's surface to account for terrain-induced variation. Based on the recommendations in several studies, we topographically normalized both images using the C-correction method (Teillet et al. 1982). We calculated coefficients for the C-correction equation by regressing image radiance values on illumination values, which were based on a digital elevation model (DEM) as well as the solar azimuth and elevation at the time of image capture.

The leaf-off Landsat image served as a means to mask out all non-evergreen pixels in the study area. We separated the image into evergreen and non-evergreen vegetation classes via cluster busting (Jensen 1996). The cluster busting process required several iterations of unsupervised classification, where each new iteration focused only on pixels that could not be clearly distinguished at the previous iteration. The results of all iterations were merged into a single evergreen/non-evergreen map. When assessed based on DOQQs, this map had an overall accuracy greater than 85%. After resampling this map from 14.25-m to 15-m resolution, we used it to remove all non-evergreen pixels from the ASTER image.

Summer (or leaf-on) images offer good spectral conditions for species separation, and have less topographic shadowing than winter images (Jensen 1996). We used the masked ASTER image to separate hemlock and non-hemlock evergreen classes. To do this, we created a training data set suitable for application in tree-based classification. Our guiding source for the set was a GIS-based vegetation map for GRSM developed primarily from large-scale (1:12,000) aerial photographs (Welch et al. 2002). This map represented the best available source of information on hemlock distribution. It provided four classes of hemlock presence, recorded as unique polygons: dominant, co-dominant, secondary component, and inclusion. In a GIS, we generated a large random sample of points for each of these classes, as well as a random sample of points from areas outside the hemlock polygons but still in the masked ASTER image (i.e., a non-hemlock evergreen class). Notably, large samples are required for tree-based classifiers to perform well (Murthy 1998). Making sure no image pixel was sampled by more than one point, we scaled the sample sizes to match the proportion of the image each class occupied: ~1,000 points for the co-dominant class, ~1,500 points for the dominant, secondary component, and inclusion classes, and ~3,000 points for the non-hemlock evergreen class. For each sample point, we extracted a number of variables derived from individual raster data layers in the GRSM GIS database (Table 1). We also recorded pixel data from the ASTER image. To minimize potential image-specific bias in the pixel values, we used normalized band ratios rather than per-band pixel values.

We used SAS Enterprise Miner to analyze the input training sample. Accepting that a tree fit to five different classes would be unwieldy, we simplified the training sample by combining the dominant with the co-dominant and the secondary component with the inclusion class. (We used additional trees to subdivide these classes in later iterations, but have not reported on them here.) Exploiting the flexibility of the software, we tested trees with 2-, 3-, 4-, and 5-way splits, then chose the one with the lowest misclassification rate. Tree-based classifiers are susceptible to over-fitting of the training data and thus can be less successful at classifying subsequent data sets (Murthy 1998). To avoid over-fitting, we employed automatic pruning of the trees based on Chi-squared testing at a significance level of $\alpha = 0.20$. We also set a minimum of 10 data points for any output node. Although all four trees tested had very similar misclassification rates, the 3-split tree performed the best, with an initial sample accuracy of 61% and 75 "leaves" (i.e., distinct decision rule pathways). This initial sample accuracy is only a partial reflection of the tree accuracy once applied in the expert classifier, but may be seen as a minimum accuracy threshold under the strictest assessment conditions (i.e., not accounting for mitigating factors such as image resolution and the positional accuracy of field data). Table 1 shows which of the eligible input variables were included in the tree.

The "leaves" from the 3-split tree were imported into the Expert Classifier module of ERDAS Imagine. This module allowed us to construct a set of rules for assigning every evergreen pixel in the study area to an output class based on raster layers for each input variable. We created a final output map by merging the expert classifier result with the non-evergreen pixels identified in the cluster-busted Landsat image.

ENHANCED ANALYSIS

As previously noted, tree-based classifications are susceptible to over-fitting of the training data. While we employed automatic pruning, a single training data set may not represent the full variation of conditions for a phenomenon of interest. To enhance the success of our classifier, we developed supplementary training data points using a study area in the western portion of GRSM. This area is quite different from the eastern study area topographically, but is typical of certain portions of the southern Appalachians.

Our processing methods for the western study area were similar to those for the east. We procured two ASTER radiance images, captured in June 2000 (leaf-on) and November 2003 (leaf-off). The images were geometrically corrected with polynomial equations (39 GCPs and RMSE = 6.1546 for June 2000; 24 GCPs and RMSE = 4.6436 for November 2003), clipped to the study area extent, and topographically normalized via C-correction. We generated an evergreen/non-evergreen mask from the November 2003 image via several iterations of cluster busting, and then applied this mask to the June 2000 image. From the remaining evergreen portion of the image, we generated random sample points in proportion to the area each hemlock class occupied in the western study area: ~800 points from dominant, ~1,500 points from co-dominant, ~50 points from secondary component, and ~500 from inclusion, as well as 2,800 points from non-hemlock evergreen areas. For each sample point, we extracted values for the input variables from raster layers in the GRSM GIS database. We then combined these sample points with our eastern area training data set, yielding a substantially larger set of more than 14,000 sample points across five classes.

Table 1. Variables tested in tree-based classification. Plus signs (+) denote variables actually used in initial classification tree; asterisks (*) denote variables used in enhanced classification tree.

Variable	Resolution (m)	Description
Aspect+*	10	Slope direction based on DEM
Curvature	10	Convexity/concavity based on DEM
Elevation+*	10	Elevation from DEM
Landform Index*	10	Index based on DEM (McNab 1993)
Slope+*	10	Degree slope based on DEM
Topographic Relative Moisture Index+*	10	Dryness-wetness index based on DEM (Parker 1982)
Proximity to Streams+*	10	Grid based on GRSM stream layer
Disturbance History	90	Harvested or cleared land (GRSM data)
Fire Frequency	90	Reoccurring burns, 1920s-80s (GRSM)
Fire History	90	Decades of fires, 1920s-80s (GRSM)
Geology	90	General bedrock formations (GRSM)
ASTER Ratios:		Normalized difference indices to allow generalization of any image-based rules in the output trees. Particular ratios were chosen based on band-to-band correlations. Indices calculated as: (band a - band b)/(band a + band b).
Band 3/Band 1+*	15	
Band 1/Band 2+*	15	
Band 4/Band 5+*	15	
Band 4/Band 6+*	15	
Band 4/Band 7*	15	
Band 4/Band 8*	15	
Band 4/Band 9+*	15	

We repeated testing of 2-, 3-, 4-, and 5-split trees in Enterprise Miner in this enhanced data set, again combining the dominant with co-dominant and secondary component with inclusion hemlock classes to simplify the output tree. We used the same settings for tree-based classification. As in the earlier analysis, the 3-split tree performed best, with an initial accuracy of 62% and 79 “leaves.” In fact, all test runs in this enhanced analysis actually performed better (1-3% higher in initial accuracy) than their counterparts in the first analysis. We imported the “leaves” from the 3-split tree into the Expert Classifier module of ERDAS Imagine and used the resulting decision to classify all evergreen pixels in both study areas. We created final output maps by merging these results with the non-evergreen pixels identified by the cluster-busted October 2001 Landsat (for the eastern study area) and November 2003 ASTER (for the western study area) images.

ACCURACY ASSESSMENT

We completed an accuracy assessment of the eastern study area maps—for the original and enhanced classifiers—using 170 reference points gathered from field surveys or by viewing the CIR DOQQs where appropriate. Based on these data, we were only able to judge hemlock presence/absence, so we simplified our assessment to three classes (hemlock, non-hemlock evergreen, and non-evergreen). We examined map pixel values within a 22.5-meter radius of each reference point (approximately equivalent to a 3 x 3 pixel window). If the map class of the reference point corresponded to any pixel falling within the window, then the classification was judged to be correct. We chose this approach to accommodate positional accuracy limitations of the image geometric correction process (± 7.5 m RMSE) and the reference data points, which were largely recorded with recreational grade GPS units (± 15 m). For each map, we created error matrices and calculated an overall value for Cohen's kappa statistic. Cohen's Kappa statistic indicates how much of an improvement a classification effort is over a completely random classification of the same area (Jensen 1996). It can range from 0 to 1, with 0 being the least possible improvement and 1 being the most. We did not have enough field data to perform a full accuracy assessment of the western study area. However, we did have a small set (n = 32) of hemlock survey points that provided some indication of how the enhanced classifier might perform in this sort of region. We used the same 22.5-m radius window for judging accuracy.

RESULTS

The original and enhanced error matrices (Tables 2 and 3) for the eastern study area are similar, with identical overall accuracies (85.3%; 90% confidence interval of 80.5 to 90.1%). Kappa values for the two matrices were nearly identical: 0.765 for the original and 0.767 for the enhanced classifier. More specifically, the two classifiers performed similarly in mapping hemlocks: the enhanced classifier yielded a higher producer's accuracy but a lower user's accuracy for the hemlock class. The partial accuracy assessment for the western study area indicated that 22 out of 32 hemlock survey points were correctly identified, for an accuracy of 69%. Of the ten misclassified points in the western area, one was mistakenly labeled as non-evergreen; the remaining nine were classified as non-hemlock evergreen.

The hemlock presence maps (Figures 2 and 3) reflected very different levels of hemlock presence. In the western study area, hemlocks appear to be limited to narrow riparian corridors, while they are more broadly distributed in the eastern study area.

Tables 2 and 3. Error matrices based on accuracy assessment of the eastern study area – original classifier (top) and enhanced classifier (bottom).

Classification Totals	Reference Totals: Hemlock	Non-Evergreen	Non-Hemlock Evergreen	Total	User's Accuracy
Hemlock	62	4		66	93.9
Non-Evergreen	10	49		59	83.1
Non-hemlock Evergreen	9	2	34	45	75.6
Total	81	55	34	170	
Producer's Accuracy	76.5	89.1	100.0		**85.3**
					Overall Accuracy

Classification Totals	Reference Totals: Hemlock	Non-Evergreen	Non-Hemlock Evergreen	Total	User's Accuracy
Hemlock	64	6	2	72	88.9
Non-Evergreen	6	49		55	89.1
Non-hemlock Evergreen	11		32	43	74.4
Total	81	55	34	170	
Producer's Accuracy	79.0	89.1	94.1		**85.3**
					Overall Accuracy

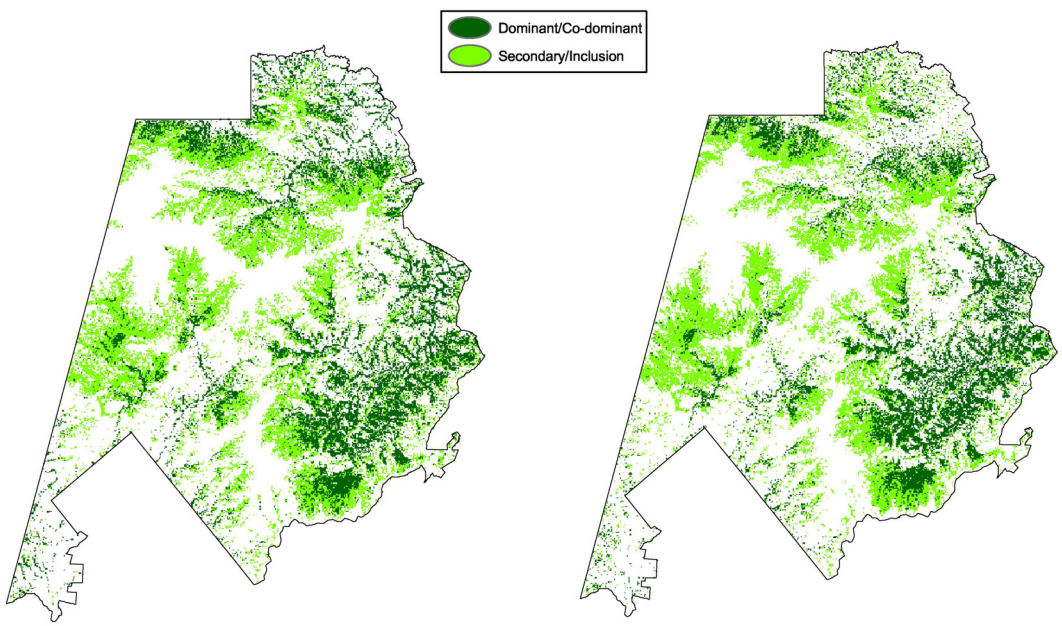

Figure 2. Hemlock distribution maps for the eastern study area – original classifier (right) and enhanced classifier (left).

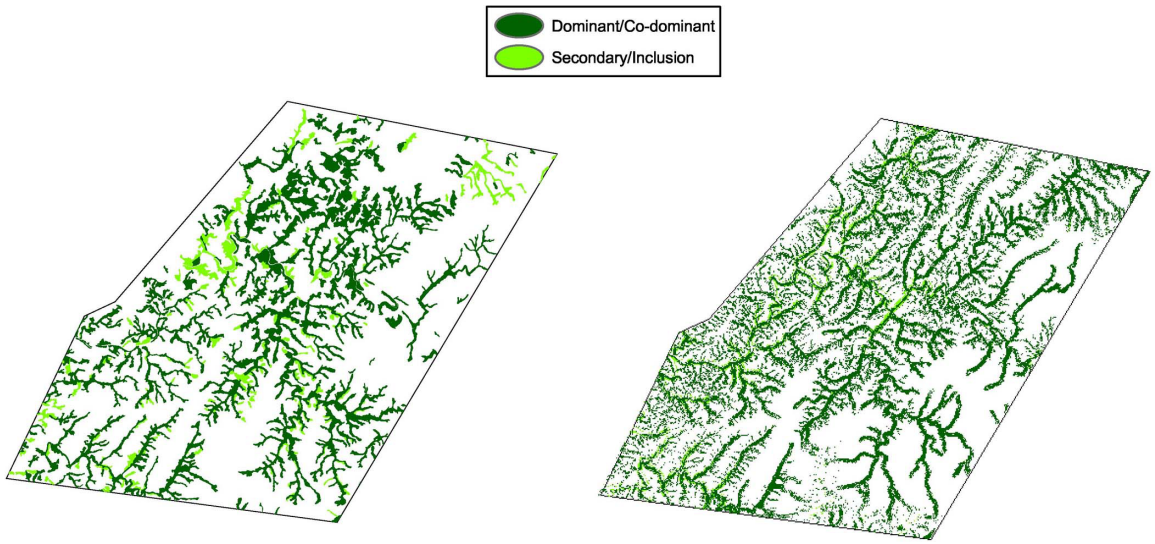

Figure 3. Hemlock distribution map for the western study area (right); photo-derived hemlock polygons used to guide sampling in the western study area (left).

DISCUSSION

The enhanced classifier captured hemlocks fairly well, particularly in the eastern study area. The addition of training data points from the western study area did not substantially alter the enhanced classifier's success in the eastern area, based on the similar accuracies of the original and enhanced classifiers. It is also worth noting that no per-class accuracy value for the enhanced classifier is less than 74.4%, although this assessment only looked at three-class maps. Based on this conservative assessment, the accuracies achieved for the eastern study area meet typical accuracy standards for remote-sensing-derived map products (Congalton and Green 1999). Unfortunately, we do not have enough assessment data to fully judge the enhanced classifier's success in the western study area. However, the hemlock classification accuracy for the west can be reasonably compared to the producer's accuracy for hemlock from the eastern error matrix, and is obviously lower (69% vs. 79%). This may be because so few sample points were available. It may also reflect characteristics of the training data: the photo-derived GRSM vegetation map restricts hemlock to riparian corridors in the western study area, and our enhanced classifier appears to mirror this (Figure 3). This may explain why several of the hemlock survey points in the western area were mistakenly classified as non-hemlock evergreen. Elevation (and more broadly, topography) is less of a factor in the western than in the eastern area, where it more strongly demarcates certain evergreen vegetation types. In locations such as the western study area, hemlocks may be found in small inclusions throughout a forest with numerous other evergreen species. Even a photo-derived vegetation map is likely to miss some of these sparsely distributed inclusions; in fact, they may be smaller than the minimum mapping unit. Such inclusions could be located on the ground and added to the training data set, but this is a potentially expensive proposition that may not result in a significant improvement in accuracy.

We intend to collect additional points in both the eastern and western study areas to expand our accuracy assessment. Ultimately, though, the enhanced classifier will be best served by testing it with data from other parts of the southern Appalachian region. Despite our best efforts, the enhanced classifier may still be over-fitted to the GRSM training data. This can be easily remedied. Though our classifier has too many rules to depict here, its first split logically stratifies the input data into three classes based on elevation. Subsequent splits divide these three elevation-based classes into ever-smaller groups. Notably, any of these finer splits can be manually pruned from the classifier if they are found to be problematic through testing in other parts of the southern Appalachians.

During the testing process and afterwards, the classifier we have described can be used to map hemlock distribution throughout the region. The classifier's rules use topographic and proximity variables that can be calculated from readily available DEMs and stream data. ASTER imagery can be procured for free or at a nominal cost through the Earth Observing System Data Gateway. While the ASTER sensor has only been in service since 1999, much of the southern Appalachian region has already been covered, and it is possible to request satellite tasking that would capture any areas that have not been covered.

The approach outlined here is only one component of an HWA "early warning system" that we have been developing for the southern Appalachians. We have also been working on GIS-based models to predict what areas in the region are at most risk of early HWA infestation. These models yield probability maps that can be used to rank areas at the highest risk of HWA infestation. By overlaying these risk probability maps with classifier-derived maps of hemlock distribution, forest managers can target specific areas for their HWA control efforts, substantially reducing the territory they must cover.

ACKNOWLEDGEMENTS

We would like to thank Kris Johnson, Scott Kichman, Keith Langdon, Mike Jenkins, Tom Remaley, and the many others at GRSM who provided data and general information used in this project. We would also like to thank Chris Ulrey at the Blue Ridge Parkway, as well as numerous people at NCSU and the USDA-FS. This paper describes one component of a project funded by an Evaluation Monitoring grant from the USDA-FS Forest Health Monitoring Program.

REFERENCES

Bonneau, L.R., K.S. Shields, and D.L. Civco. 1999a. Using satellite images to classify and analyze the health of hemlock forests infested by the hemlock woolly adelgid. *Biological Invasions* 1:255-267.

Bonneau, L.R., K.S. Shields, and D.L. Civco. 1999b. A technique to identify changes in hemlock forest health over space and time using satellite image data. Biological Invasions 1: 269-279.

Congalton, R.G. and K. Green. 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Boca Raton, Forida: Lewis Publishers. 137 p.

Delcourt, H.R. and P.A. Delcourt. 2000. Eastern deciduous forests, pp. 357-396. *In* North American Terrestrial Vegetation, ed. M.G. Barbour and W.D. Billings. Cambridge University Press.

Jensen, J.R. 1996. Introductory Digital Image Processing: A Remote Sensing Perspective. Upper Saddle River, New Jersey: Prentice Hall. 318 p.

McNab, W.H. 1993. A topographic index to quantify the effect of mesoscale landform on site productivity. *Canadian Journal of Forest Research* 23:1100-1107.

Murthy, S.K. 1998. Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Mining and Knowledge Discovery* 2:345-389.

Parker, A.J. 1982. The topographic relative moisture index: An approach to soil-moisture assessment in mountain terrain. *Physical Geography* 3:160-168.

Royle, D.D. and R.G. Lathrop. 1997. Monitoring hemlock forest health in New Jersey using Landsat TM data and change detection techniques. *Forest Science* 43(3):327-335.

Royle, D.D. and R.G. Lathrop. 2002. Discriminating *Tsuga canadensis* hemlock forest defoliation using remotely sensed change detection. *Journal of Nematology* 34(3):213-221.

Taylor, G. 2002. Hemlock resources in Great Smoky Mountains National Park, pp. 34-35. *In* Symposium on the Hemlock Woolly Adelgid in Eastern North America, February 5-7, 2002, ed. B. Onken, R. Reardon, and J. Lashomb. East Brunswick, New Jersey.

Teillet, P.M., B. Guindon, and D.G. Goodenough. 1982. On the slope-aspect correction of multispectral scanner data. *Canadian Journal of Remote Sensing* 8:84-106.

Welch, R., M. Madden, and T. Jordan. 2002. Photogrammetric and GIS techniques for the development of vegetation databases of mountainous areas: Great Smoky Mountains National Park. *ISPRS Journal of Photogrammetry and Remote Sensing* 57(1-2):53-68.