ABSTRACT

KOCH, FRANK HENRY, JR. Spatial tools for managing hemlock woolly adelgid in the southern Appalachians (Under the direction of Dr. Heather M. Cheshire)

Native to Asia, the hemlock woolly adelgid (*Adelges tsugae*) has recently spread into parts of the southern Appalachian region. This insect pest attacks both native hemlock species (*Tsuga canadensis* and *T. caroliniana*), has no natural enemies, and can kill hemlock trees within just a few years. While biological control displays promise for combating the pest, such counter-measures are significantly hampered because neither adelgid nor hemlock distribution patterns have been detailed explicitly.

We developed a spatial management system to better target control efforts. The system has two components: (1) a protocol for mapping hemlock stands, and (2) a technique to map areas at risk of imminent hemlock woolly adelgid infestation. To map hemlock stands, we utilized topographically normalized satellite imagery from Great Smoky Mountains National Park. Because hemlocks are difficult to distinguish using just satellite data, we constructed a decision tree classifier that supplemented the imagery with a suite of topographic, environmental, and proximity variables. We then implemented the classifier in a geographic information system and generated hemlock distribution maps. Our final decision tree had 27 terminal nodes and nine variables, with elevation, image band ratios, topographic relative moisture index, and distance to the closest stream among the most important variables. Accuracy assessment—based on field data and aerial photos—of the maps resulting from this tree yielded an overall thematic accuracy of 90% for one study area and 75% accuracy in capturing hemlocks in a second study area.

To map areas at risk, we combined known first-year infestation locations from Great Smoky Mountains National Park and the Blue Ridge Parkway with points from uninfested
hemlock stands, recording a suite of environmental variables for each point. We applied four different techniques (discriminant analysis, \(k\)-nearest neighbor, logistic regression, and decision tree) to generate models from these data in order to predict locations at high risk of imminent hemlock woolly adelgid infestation. We then used the resulting models to generate risk maps of the study region. All techniques performed well, accurately capturing 70-90% of training and validation samples. Discriminant analysis was the most accurate technique, but logistic regression yielded a more practical map from a management standpoint, with large, discrete risk zones. In any case, our results suggest that roads, major trails, and riparian corridors provide an important degree of connectivity enabling long-distance dispersal of the hemlock woolly adelgid, probably by humans or birds.

Both components of our hemlock woolly adelgid management system are built on readily available or easily calculable spatial data. Furthermore, they are constructed generally enough that they should be applicable throughout the southern Appalachians. Overlay of derived maps will allow forest managers to prioritize hemlock stands and allocate resources more efficiently.
SPATIAL TOOLS FOR MANAGING HEMLOCK WOOLLY ADELGID IN THE SOUTHERN APPALACHIANS

by

FRANK HENRY KOCH, JR.

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

FORESTRY

Raleigh

2005

APPROVED BY:

Dr. Heather M. Cheshire
Chair of Advisory Committee

Dr. Hugh A. Devine

Dr. Fred P. Hain

Dr. George R. Hess
DEDICATION

To my family
BIOGRAPHY

I was born in Miami, Florida, but lived there just a few years of my life before my family moved to the Atlanta area. We lived in the northeastern suburbs, in Gwinnett County, which throughout most of my childhood was the fastest growing county in the nation. Obviously, I grew up in a rapidly changing landscape. With the possible exception of that, I can’t think of any particular events or circumstances from adolescence that suggested my eventual interest in natural resources and forest health issues. While I was a good Cub Scout, I was a terrible Boy Scout. I wasn’t particularly into outdoor sports. My family did drive out to the Grand Canyon, and to Yellowstone the year after the big fires, so maybe the seed was planted there.

In high school, I was mostly interested in the humanities. This trend continued until I attended Duke University, where I majored in visual art and philosophy. With these eminently practical majors, gainful employment was surprisingly difficult to find after I graduated in 1993. (There was something of a nationwide recession at the time, so I believe that was partially to blame.) In any case, I had a number of different jobs during the years after graduation, but none of them were the right fit career-wise. Fortunately for me, I also met my wife Amy while at Duke who was supportive of me beginning a new career path—once I found it.

Defining that path took a few years. First, I started to spend a lot of time outside again, hiking, camping, and backpacking. I made trips to the Rockies, to Olympic National Park, and throughout the southern Appalachians. Somewhere along the way, I realized that I needed to get out of my current track and “do something” related to the environment.
(Admittedly, this was only a vague notion at first). I did some reading and research on my own, then enrolled in some natural-resource-related classes at NCSU in 1998. I received some sage advice (i.e., get into the graduate program!) and started working on a Master of Science in natural resources the next year.

My M.S. course of study concentrated on the use of spatial information technologies to characterize and manage natural resources, and my dissertation started as an extension of this work. Initially, I was attracted to the hemlock woolly adelgid issue primarily because I thought it would give me a chance to explore some current remote sensing technologies. However, I have become increasingly interested in the general spatial dynamics of forest pests: their vectors and patterns of spread, their habitats, and the suitable conditions for their survival. I also like the research challenge presented by the broad scales—regional or national—at which forest pests often operate. I am currently working as a research associate in a cooperative work unit of the Department of Forestry and Environmental Resources and the USDA Forest Service Forest Health Monitoring Program. While this has slowed the dissertation writing process somewhat, it has allowed me to apply my graduate research to a number of other critical issues.

At home, I am currently enjoying the opportunity to see my son, Noah, learn to walk and talk, pull the ears of our (remarkably accommodating) dog, and generally create a pleasant state of noisiness and disarray.
ACKNOWLEDGEMENTS

First, I would like to thank the members of my committee—Heather Cheshire, Hugh Devine, Fred Hain, and George Hess—for providing useful guidance on this research. They have given me a lot of latitude on this work. Furthermore, each of them has offered me opportunities that I think most other graduate students don’t get, and have opened a lot of doors for me in terms of making contacts and developing a career.

I enjoyed my time in the NCSU Center for Earth Observation, especially because of the people. Almost everyone who has been involved with the CEO contributed in some way to my experience, but some of them deserve particular mention. Linda Babcock, who runs the CEO ship, provided me with key resources and made sure that both the research and I stayed on track. Many CEO students and staff—some still around, some now elsewhere—offered advice, feedback, and general camaraderie. In particular, I have to mention Dan VanBrunt, Melani Harrell, Mark Smith, Justin Shedd, Sam Taylor, Curtis Belyea, Bill Millinor, Bill Slocumb, Beth Eastman, Joe Knight, and Stacy Nelson. Halil Cakir’s data fusion algorithm had a major impact on how the remote sensing portion of my research proceeded.

Outside the CEO “family”, I have to acknowledge Ashton Drew, Alexa McKerrow (and other folks from the North Carolina Gap Analysis Program), and Robert Jetton for offering resources or insights that I applied in this work. Denise Royle at Virginia Tech has been an incredible sounding board on hemlock woolly adelgid, remote sensing, and many other topics. The folks in the Forest Health Monitoring work unit in RTP have also have been helpful and encouraging during my final push to finish this research.
Conversations with Rusty Rhea and Brad Onken from the USDA Forest Service helped shape this work by identifying areas of focus and tools that would be particularly useful for forest managers confronting the hemlock issue. Kris Johnson, Scott Kichman, and Tom Remaley at Great Smoky Mountains National Park, as well as Chris Ulrey at the Blue Ridge Parkway, provided key data sets that greatly simplified my analyses.

Finally, I have to thank my family—particularly my wife and son, but also my family and my in-laws—for offering encouragement and putting up with my many and odd hours at the lab, on the road, or in front of a computer at home.
TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................ x

LIST OF FIGURES ......................................................................................................... xi

1. INTRODUCTION ..................................................................................................... 1

  1.1 Impacts of Invasive Forest Pests ........................................................................... 1
  1.2 The Hemlock Woolly Adelgid ............................................................................. 3
      1.2.1 Biology ....................................................................................................... 3
      1.2.2 Spread of Hemlock Woolly Adelgid in the U.S. ....................................... 7
  1.3 The Host Species: Eastern and Carolina Hemlock .......................................... 10
  1.4 Challenges for Management ............................................................................... 14
  1.5 Project Objectives ............................................................................................. 17
  1.6 References ......................................................................................................... 19

2. LITERATURE REVIEW ...................................................................................... 25

  2.1 Introduction ......................................................................................................... 25
  2.2 Topographic Normalization of Satellite Imagery from Mountainous Regions ... 26
      2.2.1 Basic Concepts .......................................................................................... 26
      2.2.2 Normalization Methods Based on Digital Elevation Data ......................... 27
      2.2.3 The Minnaert Constant Method ................................................................. 29
      2.2.4 The C-Correction Method ......................................................................... 30
      2.2.5 Advantages and Drawbacks of the C-Correction ................................. 31
      2.2.6 Atmospheric Correction and Topographic Normalization ..................... 33
      2.2.7 Summary ................................................................................................... 34
  2.3 Approaches for Mapping Hemlocks in the Southern Appalachians ................. 35
      2.3.1 Remote Sensing and Hemlock ................................................................. 36
      2.3.2 Decision Trees ......................................................................................... 40
      2.3.3 Examples ................................................................................................... 44
      2.3.4 Summary ................................................................................................... 49
  2.4 Predicting Hemlock Woolly Adelgid Invasion at the Landscape Level ........... 50
      2.4.1 Discriminant Analysis ............................................................................... 52
      2.4.2 Logistic Regression .................................................................................... 57
      2.4.3 Variable Selection in Discriminant Analysis and Logistic Regression ...... 61
      2.4.4 Decision Trees ......................................................................................... 62
      2.4.5 Choice of Method ..................................................................................... 63
      2.4.6 Examples ................................................................................................... 65
      2.4.7 Summary ................................................................................................... 69
  2.5 References ......................................................................................................... 70
3. TOPOGRAPHIC NORMALIZATION OF SATELLITE IMAGERY OF THE SOUTHERN APPALACHIANS VIA THE C-CORRECTION

3.1 Introduction

3.2 Background

3.3 Study Area

3.4 Methods
   3.4.1 Image Acquisition
   3.4.2 Image Pre-processing
   3.4.3 Topographic Normalization
   3.4.4 Assessment of the C-correction

3.5 Results

3.6 Discussion

3.7 Conclusions

3.8 References

4. MAPPING HEMLOCKS VIA TREE-BASED CLASSIFICATION OF SATELLITE IMAGERY AND ENVIRONMENTAL DATA

4.1 Introduction

4.2 Study Area

4.3 Methods
   4.3.1 Image Acquisition and Pre-Processing
   4.3.2 Masking of Non-Evergreen Areas
   4.3.3 Initial Decision Tree Generation
   4.3.4 Enhanced Decision Tree Generation
   4.3.5 Accuracy Assessment

4.4 Results
   4.4.1 Decision Trees: General Characteristics
   4.4.2 Comparison of Output Maps
   4.4.3 Accuracy Assessment Results

4.5 Discussion
   4.5.1 Accuracy
   4.5.2 Generalization

4.6 Conclusions

4.7 References
LIST OF TABLES

Table 3-1. Parameter estimates and C-correction coefficients for the Landsat image subsets. ......................................................................................................................................... 91
Table 3-2. Parameter estimates and C-correction coefficients for the ASTER image subsets. ......................................................................................................................................... 92
Tables 3-3a and 3-3b. Per-band $r^2$ values for the Landsat and ASTER image subsets. ........ 99
Tables 3-4a and 3-4b. Maximum and mean percent per-pixel change (absolute value) between the uncorrected and normalized Landsat and ASTER image subsets. ................. 99
Tables 3-5a and 3-5b. Change in standard deviation (SD) due to normalization for the Landsat and ASTER images. ................................................................................................................. 100
Table 3-6. Change in Landsat image SD within the classes of the binary vegetation map. 100
Table 3-7. Change in ASTER image SD within the classes of the three-class vegetation map. ..................................................................................................................................... 100
Table 3-8. Change in SD due to normalization for the Landsat image subsets. ................... 101
Table 3-9. Change in SD due to normalization for the ASTER image subsets .................... 101
Table 3-10. Error matrix for Landsat sunlit subset. .............................................................. 102
Table 3-11. Error matrix for Landsat sunshade subset. ...................................................... 102
Table 3-12. Error matrix for ASTER sunlit subset ............................................................... 103
Table 3-13. Error matrix for ASTER sunshade subset. ....................................................... 103
Table 4-1. Variables tested for incorporation in the initial and enhanced decision trees. .... 125
Table 4-2. Ranking of variables in the initial decision tree. ................................................ 130
Table 4-3. Ranking of variables in the enhanced decision tree. ........................................... 130
Table 4-4. Error matrix for initial decision tree, accuracy based on immediate pixel ......... 137
Table 4-5. Error matrix for enhanced decision tree, accuracy based on immediate pixel .... 137
Table 4-6. Error matrix for initial decision tree, accuracy based on 22.5-m window. ........... 137
Table 4-7. Error matrix for enhanced decision tree, accuracy based on 22.5-m window. ... 138
Table 5-1. Variables tested for inclusion in classification efforts. ....................................... 165
Table 5-2. Multivariate normality test results for the five training data sets. ..................... 175
Table 5-3. Error rate summaries across the five partitioned data sets. .............................. 175
Table 5-4. Error matrices for the training subset of the median data set: a) discriminant analysis, b) $k$-nearest neighbor, c) logistic regression, d) decision tree. ..................... 178
Table 5-5. Error matrices for the validation subset of the median subset: a) discriminant analysis, b) $k$-nearest neighbor, c) logistic regression, d) decision tree. ..................... 178
Table 5-6. GSMNP total area and hemlock area designated as high risk by each classification technique based on the median data set. .............................................................. 180
Table 5-7. Percentage of the area mapped as high infestation risk by each technique that was also designated high risk by the other three techniques. ................................. 181
Table 5-8. Percentage of surveyed and treated infestation points captured by each classification technique. .............................................................................................................. 183
Table 6-1. Rules list for initial decision tree ...................................................................... 206
Table 6-2. Decision rules for the enhanced tree. .............................................................. 210
LIST OF FIGURES

Figure 1-1. Hemlock woolly adelgid (*Adelges tsugae*), with arrows indicating the adelgid’s stylet bundle.......................................................... 3
Figure 1-2. General life cycle of hemlock woolly adelgid................................................. 5
Figure 1-3. Adelgid ovisacs on eastern hemlock (*Tsuga canadensis*). .................. 5
Figure 1-4. Pattern of hemlock woolly adelgid spread in the eastern U.S. since 1971. ...... 8
Figure 1-5. Geographic range of *T. canadensis*.......................................................... 11
Figure 1-6. Geographic range of *T. caroliniana*.................................................... 12
Figure 2-1. Solar incidence (*i*) and exitance (*e*) angles for inclined terrain. ............... 26
Figure 3-1. Solar incidence angle (*i*) and surface normal for a location on inclined terrain. . 83
Figure 3-2. Great Smoky Mountains National Park, with a close-up of the study area. ...... 87
Figure 3-3. Detail of the Landsat panchromatic image prior to and after low pass filtering.. 89
Figure 3-4. Aspect partitioned subsets for the ASTER image......................................... 91
Figure 3-5. The C-correction topographic normalization model.................................. 92
Figure 3-6. Uncorrected and topographically normalized Landsat image.................... 98
Figure 3-7. Uncorrected and topographically normalized ASTER image..................... 98
Figure 4-1. Eastern and western study areas in Great Smoky Mountains National Park. .... 120
Figure 4-2. Initial decision tree for hemlock classification.......................................... 131
Figure 4-3. Enhanced decision tree for hemlock classification.................................... 132
Figure 4-4. Hemlock distribution maps for the eastern study area: initial and enhanced decision trees................................................................................................. 134
Figure 4-5. Hemlock distribution maps for the western study area: a) initial decision tree; b) enhanced decision tree; c) GSMNP map based on photo interpretation ....................... 135
Figure 5-1. Hemlock woolly adelgid range expansion in the southern Appalachian region.155
Figure 5-2. Great Smoky Mountains National Park (GSMNP) and southern Blue Ridge Parkway (BRP). ................................................................. 162
Figure 5-3. First-year hemlock woolly adelgid infestation sites in GSMNP..................... 163
Figure 5-4. Hemlock woolly adelgid survey and treatment sites recorded after the first year of infestation ........................................................................................................... 173
Figure 5-5. Chi-square Q-Q plots for multivariate normality: infested and uninfested of the full data set................................................................. 174
Figure 5-6. Decision tree for the median data set......................................................... 176
Figure 5-7. Areas at high risk of imminent hemlock woolly adelgid infestation, as defined by classification functions: a) discriminant analysis, b) *k*-nearest neighbor, c) logistic regression; d) decision tree. ..................................................... 179
Figure 5-8. Risk probability surfaces developed with discriminant analysis and logistic regression ........................................................................ 182
Figure 5-9. Overlay of logistic regression risk map and hemlock distribution in GSMNP.. 189
Figure 6-1. SAS Enterprise Miner process diagram for tree construction........................ 204
Figure 6-2. Node index for initial decision tree............................................................... 206
Figure 6-3. Node index for the enhanced decision tree.................................................. 210
1. INTRODUCTION

1.1 Impacts of Invasive Forest Pests

Approximately 50,000 non-indigenous species have been introduced during the course of U.S. history (Pimentel et al. 2000). Not all are harmful; indeed, only a relatively small fraction of these species has resulted in substantial economic and environmental impacts. Still, this “small fraction” costs the U.S. economy as much as $137 billion per year in losses, damages, and control costs (Levine & D’Antonio 2003; Pimentel et al. 2000). More than $4 billion of that total can be attributed to forest pests, including both insects and pathogens (Campbell 2002; Pimentel et al. 2000). In volumetric terms, the annual loss of forest timber in the U.S. due to insects and disease is estimated to be 2.4 billion ft$^3$ (Hof et al. 1997).

Looking specifically at insects, close to 400 non-indigenous species have invaded American forests; roughly 30% are now significant pests (Liebhold et al. 1995; Pimentel et al. 2000). These insects change forest composition, structure, and microenvironments, alter critical ecosystem processes, and increase susceptibility of areas to further invasion and disturbance (Orwig 2002). With the rise in global trade, the number of insects invading U.S. locales is likely to grow substantially. One fairly conservative model projected that 115 new insect species will invade the U.S. by 2020 (Levine & D’Antonio 2003). Not surprisingly, federal land management agencies have devoted a great deal of attention to the invasive forest pest problem. The National Park Service has adopted an agency-wide integrated pest management (IPM) strategy to limit spread of invasives both into and out of national park units (USDI-NPS 2000). In 2002, Dale Bosworth, Chief of the U.S. Forest Service
FS), listed invasive species as one of the four most significant threats to U.S. forest ecosystems (USDA-FS 2004). In response, a team of USDA-FS managers and researchers developed an agency-wide strategic plan with four elements: 1) prevention; 2) early detection and rapid response; 3) control and management; and 4) rehabilitation and restoration (USDA-FS 2004). Moreover, the USDA-FS Forest Health Monitoring Program has focused for some time on the creation of early warning tools for detection and monitoring of particular forest pests (USDA-FS 2003a; 2003b). Unfortunately, such efforts are impeded by a lack of understanding about many invasive species: their patterns of spread, long-term ecological impacts, and effective methods of control (Byers et al. 2002). Often, there is also insufficient focus on how to transfer that knowledge to land managers. For example, while several studies on the ecology of the zebra mussel (Dreissena polymorpha) existed before its proliferation in the lakes and waterways of North America, conservation managers had no real tools to apply these data for control or even to predict the zebra mussel’s spread (Byers et al. 2002).

The hemlock woolly adelgid (Adelges tsugae Annand) is just one of many insect pests that have invaded the forests of eastern North America during the last several decades, but is considered by some to be among the worst, with a potential impact on natural forests equal to Dutch elm disease or chestnut blight (USDA-FS 2002). Many aspects of the hemlock woolly adelgid are still poorly understood, although the pest has already caused significant tree mortality in parts of the eastern U.S. and continues to spread throughout its host species’ range. The USDA-FS and other agencies are currently in the midst of an initiative to learn more about the hemlock woolly adelgid and develop tools to effectively limit the pest’s impact (USDA-FS 2002). This is a significant undertaking for any forest
pest, but a cursory look at the background information on the adelgid and its host species highlights some specific challenges for its monitoring and management.

1.2 The Hemlock Woolly Adelgid

1.2.1 Biology

The hemlock woolly adelgid is a piercing/sucking insect native to parts of China and Japan. In its native range, it is considered an innocuous inhabitant of hemlocks and spruce (McClure et al. 2001). While it occasionally achieves high densities on hemlocks in Japan, it does not appear to do any significant damage to the trees, due either to host tree resistance or possibly predators that keep adelgid populations in check (McClure et al. 2001).

![Figure 1-1. Hemlock woolly adelgid (Adelges tsugae), with arrows indicating the adelgid’s stylet bundle (from McClure et al. 2001).](image)
A small insect (1-2 mm long), the hemlock woolly adelgid (Figure 1-1) normally feeds by inserting a stylet bundle at the base of hemlock needles to extract nutrients stored in the parenchyma cells of the xylem rays (Cheah et al. 2004; McClure et al. 2001). At high densities it can be found on twigs and sometimes even on the trunk (F. Hain, personal communication). This feeding results in desiccation of the hemlock’s needles and, if large numbers of adelgids infest a tree, a cessation of new growth (Cheah et al. 2004). Continued feeding results in visible foliage loss and branch dieback within two to four years. Adelgid infestation can kill a tree within four years, although an infested hemlock may survive for 12 years or longer with sparse foliage at the branch tips and top of the crown (Cheah et al. 2004; McClure et al. 2001; Ward et al. 2004).

The life cycle of the hemlock woolly adelgid (Figure 1-2) is fairly complex, with three parthenogenetic generations that mature on hemlock (McClure & Cheah 2002). The sistens are wingless individuals that begin development in late spring, aestivate during summer, resume development during autumn, and mature by late winter. The progrediens are wingless progeny of the sistens that develop during spring and produce the next sistens generation by late spring. The sexuparae are other progeny of the sistens that develop on hemlock in spring. Adult sexuparae are winged and must leave hemlock to find a spruce (Picea spp.) host on which to lay their eggs to begin the sexuales generation (McClure & Cheah 2002). The three generations each have six development stages: the egg, four nymphal instars, and the adult. Eggs and first instar nymphs, also called crawlers, are quite similar in appearance, but distinguishable by the eyes and legs of the latter (McClure & Cheah 2002). Eggs and crawlers are readily dispersed by wind, deer, birds, and humans (McClure 1991). Eggs and crawlers are most abundant in early spring, though crawlers may
be most active later in spring (McClure 1990). Both sistens and progrediens generations produce the cottony ovisacs—or “woolly masses”—that give the insect its name (Figure 1-3), though the ovisacs of the progrediens generation are somewhat smaller (McClure 1996).

Figure 1-2. General life cycle of hemlock woolly adelgid (from McClure et al. 2001).

Figure 1-3. Adelgid ovisacs on eastern hemlock (Tsuga canadensis). [Credit: M. McClure]
McClure (1991) examined the population dynamics of the hemlock woolly adelgid during the four-year period from initial colonization for three different hemlock stands in New England. Populations exhibited bimodal peaks of abundance, shaped by the availability of new hemlock growth, which the pest strongly prefers to colonize (McClure 1991). Adelgid densities were at their highest during the first year when new hemlock growth was abundant, but declined sharply the second year, primarily because new growth was completely inhibited. Adelgid densities reached a second, subtler peak during the third year, when the hemlocks were able to produce limited new growth because of the previous year’s population decline. Ultimately, the adelgid populations perished after about four years when all trees in the stand had been killed (McClure 1991). As previously noted, other studies indicate that some hemlock stands persist longer than four years, presumably together with some level of hemlock woolly adelgid presence (Colbert et al. 2002).

Hemlock woolly adelgid populations are strongly affected by density-dependent feedback. When infestations are heavy, a large proportion of the sistens progeny become winged sexuparae, which leave hemlock in search of a spruce host. This proportion can be ~90% from heavily infested sites (McClure & Cheah 2002). Presumably these insects die, because there is apparently no suitable spruce host in eastern North America (McClure 1991; McClure & Cheah 2002). Fecundity rates are also shaped by density-dependence; egg production typically ranges from 50 to 175 eggs for a sistens female and from 25 to 125 for a progrediens female under favorable conditions (Cheah et al. 2004; Colbert et al. 2002; McClure 1991). This level of egg production results in explosive population growth if unchecked by predators or other factors (Cheah et al. 2004).
1.2.2 Spread of Hemlock Woolly Adelgid in the U.S.

The hemlock woolly adelgid was introduced to western North America in the 1920s, probably on infested nursery stock (Cheah et al. 2004). It has been known to kill or weaken western hemlock (*Tsuga heterophylla*) trees, but with no noticeable impact on natural forest stands in the region (McClure 1987). With respect to the eastern U.S., it was first observed in a municipal park in Richmond, VA around 1951, with introduction again attributed to nursery stock (Cheah et al. 2004; McClure et al. 2001). For the next few decades, the hemlock woolly adelgid was treated as little more than a nuisance pest of ornamental hemlocks. Unfortunately, the two hemlock species native to the forests of eastern North America—eastern hemlock (*T. Canadensis* (L.) Carr.) and Carolina hemlock (*T. caroliniana* Engelm.)—have since proven to be highly susceptible hosts, exhibiting no measurable resistance to the pest (McClure et al. 2001). By the 1980s, the adelgid had invaded natural stands of eastern hemlock, and was first recorded in New England (Connecticut) in 1985. Regional spread may have been aided by Hurricane Gloria in 1985 or other coastal storms that dispersed adelgids from a relatively small number of existing infestations (McClure 1987). In any case, the pest began to spread rapidly through the Mid-Atlantic U.S., and now ranges from southeastern Maine to Georgia (Ward et al. 2004). Overall, the hemlock woolly adelgid is estimated to currently occupy 26% of the total range of hemlock habitat and 25% of the total host basal area in the eastern U.S. (Morin et al. 2005).

Looking specifically at the southern Appalachians, the hemlock woolly adelgid was detected in Shenandoah National Park, Virginia—the northern end of the region—in May 1988, and has since caused hemlock mortality in excess of 80% within the park (Bair 2002). Isolated infestations were first detected in the Pisgah and Nantahalah National Forests, North
Carolina, during 2001, in Great Smoky Mountains National Park during 2002, and at sites along southern portions of the Blue Ridge Parkway in 2003 (USDI-NPS 2004; SAMAB 2004; C. Ulrey, personal communication). This is consistent with the general rate and pattern of spread for the pest. Estimates of the rate of spread range from 8 to 30 km per year (Evans 2004; McClure 1996; Souto et al. 1996; Ward et al. 2004). Since 1993, the main front of infestation has advanced approximately 17 km per year, with isolated infestations detected far ahead of this front (Cheah et al. 2004) (Figure 1-4). For the last decade, the hemlock woolly adelgid’s range has been expanding most rapidly at the front’s western, southern, and northern fringes (Cheah et al. 2004; Ward et al. 2004).

Figure 1-4. Pattern of hemlock woolly adelgid spread in the eastern U.S. since 1971. Colors indicate year when the pest was first detected in a county.
The hemlock woolly adelgid’s general pattern of spread is shaped by its passive dispersal via a number of different vectors. Although only a single field study (McClure 1990) has attempted to quantify the pest’s dispersal behavior from an infestation point, it was fairly comprehensive regarding local and long-distance dispersal factors. First, monitoring of sticky traps during a 3-month period (May-July) indicated that wind was able to disperse some hemlock woolly adelgid eggs and crawlers to at least the edge of the study zone (1350 m) during a single season, although most eggs and crawlers were found within 600 m of the point of origin. Second, adelgid presence on hemlock seedlings planted in the study zone exhibited a spatial pattern similar to that found on the sticky traps. However, adelgid densities on seedlings browsed by deer were significantly higher than on unbrowsed seedlings, including a majority of the seedlings planted greater than 750 m from the origin. Finally, 3 of 14 birds (21%) captured in a field 2 km away from the nearest hemlock carried 1-15 eggs or crawlers. McClure (1990) further suggested that the importance of birds in the long-range dispersal of the hemlock woolly adelgid could be even greater during bird migration in early spring, when eggs and crawlers are most abundant. Indeed, birds are a major force in its spread, and a likely cause of new, isolated infestations beyond the main front (McClure 1987). In the northeastern U.S., the pest seems to have spread more rapidly along watercourses, and this may be attributable to birds migrating along these riparian corridors (Ward et al. 2004). Humans may be a similarly important dispersal factor, either by introducing infested nursery stock to new areas or by accidentally transporting eggs or crawlers (SAMAB 2004).

Northern expansion of the hemlock woolly adelgid may be limited by climate. Populations have been dramatically reduced by low winter temperatures or sudden
temperature changes during mild winters; for example, adelgid populations declined more than 90% when temperatures fell below –5º F in Connecticut (Ward et al. 2004). Nonetheless, the hemlock woolly adelgid appears sufficiently cold-hardy that winter temperatures are unlikely to impede its spread in southern regions of the U.S. (McClure & Cheah 1999). If forest managers take no action to slow the pest’s spread or reduce its populations significantly, it is quite likely that it will invade all hemlock habitats in the eastern U.S. within the next several decades, except for the very coldest regions (Ward et al. 2004).

1.3 The Host Species: Eastern and Carolina Hemlock

As with most genera in the family Pinaceae (with the exception of Larix), Tsuga species are evergreen. In general, Tsuga species are found in areas with moist climates and low water stresses (Taylor 1993). Of all Pinaceae genera, Tsuga is the most shade-tolerant and least drought-resistant (Farjon 1990). As a result, the species are generally limited to coastal regions or areas where precipitation or moisture is available with regularity, such as the fog belts of high mountains. Tsuga species tend to be in dense forests where air humidity and soil moisture is high.

T. canadensis (Figure 1-5) is an extremely widespread species found throughout the temperate eastern deciduous forest zone (Farjon 1990). Its range extends from northern Alabama and Georgia in the southeastern U.S., through the Mid-Atlantic and New England states, and into Canada as far north as Nova Scotia. It can also be found in Ohio, Indiana, Michigan, and Minnesota (Quimby 1996). In the southeastern U.S., it is primarily limited to mountainous areas (i.e., the Appalachians), although isolated pockets are found outside this
montane zone. Moving north across the continent, the species becomes more generally widespread. *T. canadensis* is typically found at elevations 600-1500 m in the southern Appalachians but as low as sea level in the northern part of its range (Godman & Lancaster 1990). At northern latitudes, *T. canadensis* is a major component of hemlock-white pine-northern hardwood forests, grows locally in pure stands, and can be found at almost any topographic position (Delcourt & Delcourt 2000; Farjon 1990). Distribution of *T. canadensis* is much patchier in the south, though it is still associated with a variety of different forest types (Godman & Lancaster 1990; Quimby 1996). In the southern Appalachians, *T. canadensis* thrives on moist rocky ridges, ravines, and hillsides, and the best stands are found on north- and east-facing slopes where humidity is high and temperatures relatively cool (Delcourt & Delcourt 2000; Godman & Lancaster 1990; Quimby 1996). *T. canadensis* is also commonly associated with riparian zones in the southern Appalachians (Godman & Lancaster 1990).

![Figure 1-5. Geographic range of *T. canadensis* (from Little 1971).](image-url)
*T. caroliniana* (Figure 1-6) has a similar elevation range (700-1200 m) and habitat requirement as *T. canadensis* (usually rocky moist slopes or ridges, occasionally gentle slopes or flat valleys) (Godman & Lancaster 1990; NatureServe 2005). Its geographic range is very narrowly restricted to the southern Appalachian region—from Virginia (and possibly West Virginia) down to Georgia, but primarily western North Carolina (Farjon 1990; Godman & Lancaster 1990). *T. caroliniana* can often be found growing alongside *T. canadensis*, but it tends to grow singly or in very small groups rather than in sizeable stands (Farjon 1990).

Figure 1-6. Geographic range of *T. caroliniana* (from Little 1971).
Hemlocks are extremely long-lived: trees can take 250-300 years to mature and live 800 years (Godman & Lancaster 1990). The high shade tolerance of both *T. canadensis* and *T. caroliniana*—they can survive with as little as 5% of full sunlight—allows their seedlings to develop and establish under dense overstories (Godman & Lancaster 1990). Indeed, hemlocks tend to generate an environment ideal for self-perpetuation, creating a heavily shaded understory that precludes the establishment of seedlings of other species (Godman & Lancaster 1990; Quimby 1996). However, because hemlocks grow so slowly, it generally takes centuries for these late successional/climax species to dominate a stand (Quimby 1996). Moreover, their shallow root systems make them susceptible to wind throw as well as extreme drought and flooding, which can result in tree mortality (Quimby 1996). Hemlocks are also susceptible to several injurious diseases and about two dozen insect pests (Godman & Lancaster 1990). One insect in particular, the elongated hemlock scale (*Fiorinia externa* Ferris), appears to accelerate the decline of some stands also infested by the hemlock woolly adelgid (Danoff-Burg & Bird 2002; McClure 2001).

Hemlocks’ slow growth suggests it will be difficult to regenerate stands that have been devastated by the adelgid. This is exacerbated by the fact that white-tailed deer (*Odocoileus virginianus*) populations have exploded in many parts of eastern North America, and browse preferentially on hemlock seedlings (Ward et al. 2004). At least in the short term, hemlocks are likely to be replaced by deciduous species such as black birch (*Betula lenta*) and red maple (*Acer rubrum*) in the many parts of eastern North America (Orwig 2002; Ward et al. 2004). In the southern Appalachians, there is no obvious tree species replacement for hemlock, but areas it previously occupied may instead be dominated by heavy growth of *Rhododendron* species in the understory (J. Vose, personal communication).
The potential impacts of hemlock replacement are numerous. There will be an inevitable decline in terrestrial habitat diversity with subsequent effects on wildlife; for example, ~120 vertebrate species are strongly associated with *T. canadensis* stands (Ward et al. 2004). Though some bird species may benefit from the potential replacement of hemlocks by hardwoods, a number of other bird species associated with hemlock forests will be negatively affected by their removal (Tingley et al. 2002). The quality of aquatic habitats adjacent to riparian hemlocks will also suffer: increased nitrate leaching is possible due to greater nitrification and inorganic nitrogen availability, and stream temperatures may rise to a level detrimental for many cold-loving fish species (Evans et al. 1996; Jenkins et al. 1999; Ward et al. 2004). Sites in the northeastern U.S. that have already experienced high hemlock mortality have seen an increase in the presence of woody and herbaceous invasives (Orwig & Foster 1998).

Admittedly, given that *T. canadensis* and *T. caroliniana* are late-successional/climax species, it may be more appropriate to examine their outlook over a much longer timeframe. The rehabilitation of hemlock stands is an increasingly important topic in the hemlock woolly adelgid literature (e.g., Ward et al. 2004). Unfortunately, if the hemlock woolly adelgid persists for years after invading all hemlock areas in the eastern U.S., the long-term prospects are unclear, and may remain so unless forest managers develop a practical plan for broad-scale reduction or elimination of the pest.

**1.4 Challenges for Management**

In line with this idea of reducing or eliminating the pest, a major theme of the Third Symposium on Hemlock Woolly Adelgid in the Eastern United States in February 2005 was
the use of biological control. Chemical control via insecticides such as imidacloprid—while highly effective against the hemlock woolly adelgid in small-area or ornamental settings—is infeasible for natural forest stands due to both cost and access (McClure 1992; McClure et al. 2001). While surveys of native or already established North American insect species revealed none that would be effective as predators, a handful of insects from Asia hold some promise for controlling the pest in natural stands (Montgomery & Lyon 1996; Wallace & Hain 2000). Several years ago, a Japanese lady beetle species (Sasajiscymnus tsugae Sasaji and McClure) emerged as the primary candidate, endorsed because it feeds on all adelgid life stages and its life cycle is fairly well synchronized with the pest’s cycle (McClure & Cheah 1999). Since 1995, more than one million individuals of S. tsugae have been released in sites throughout the eastern U.S. (Cheah et al. 2004). In addition, initial trials with several Chinese lady beetles (Scymnus sp.) and a derodontid beetle (Laricobius nigrinus Fender) from British Columbia are currently underway (Cheah et al. 2004; Ward et al. 2004). There has recently been increased research regarding insect-killing fungi as an alternative to predator beetles (Cheah et al. 2004). Little is known about the ultimate degree of effectiveness for any of these agents—initial field releases suggest S. tsugae has some potential for hemlock woolly adelgid reduction (Cheah et al. 2004)—but it has become generally accepted that a suite of predators is more likely to succeed than a single species (Ward et al. 2004).

Unfortunately, application of biological control or any other counter-measure faces a couple of significant difficulties. First, because the hemlock woolly adelgid is dispersed by a wide variety of vectors, it is difficult to predict where the pest will appear next in the nearly 20 million acres of hemlock habitat that can be found in the eastern U.S. (Ward et al. 2004).
There have been some efforts to identify landscape conditions that make certain hemlock stands more susceptible once infested by the pest (e.g., Orwig et al. 2002; Royle & Lathrop 2002b; Young & Morton 2002). Yet, there have been essentially no analyses of landscape-level factors that may predict sites most likely to be invaded in the near future—beyond the aforementioned theory that birds may carry the pest along riparian corridors (Ward et al. 2004). The few prior attempts to quantify hemlock woolly adelgid infestation risk (e.g., Evans 2004; Morin 2005) are not spatially explicit enough for land managers, who need some way to prioritize hemlock stands for adelgid surveying and application of control measures. In particular, predator introductions must be timed as closely as possible with the arrival of the pest, primarily because predators take several generations to establish, resulting in a time lag before they can be effective at controlling adelgid populations (Cheah et al. 2005).

For the purpose of prioritizing stands, it is also critical to characterize the distribution of the host species in a spatially explicit manner—i.e., where the hemlock stands can be found on the ground. While some states in the eastern U.S. have relatively large-scale hemlock distribution maps, others do not, including many of the southern Appalachian states (Ward et al. 2004). State Gap Analysis Programs in the southern Appalachian region and elsewhere are developing Landsat-image-based land cover maps to the alliance level of The Nature Conservancy / United Nations Educational, Scientific, and Cultural Organization (TNC/UNESCO) classification system (Crist & Jennings 2000), but these maps do not focus specifically on hemlock distribution and so have limited thematic accuracy in that regard (A. McKerrow, personal communication). A few researchers in the northeastern U.S. have employed Landsat imagery to map hemlock health decline due to the hemlock woolly
adelgid (e.g., Bonneau et al. 1999a, 1999b; Royle & Lathrop 1997, 2002a), and have done so with reasonably high accuracy, but noted the challenge of mapping hemlocks over a large region. Because hemlocks are often restricted to isolated stands in coves, steep ravines, or on north-facing bluffs, a satellite-based approach for targeting hemlock spatial distribution in the southern Appalachians is especially difficult. These small stands can be obscured by the mountainous terrain or are simply poorly visible at Landsat’s 30-m resolution. Furthermore, the stands are also distributed throughout a heterogeneous forest background that includes a variety of other evergreen-dominated vegetation types. While interpretation of high-resolution aerial photos may be an accurate alternative for delineating hemlock stands locally (Ward et al. 2004; Welch et al. 2002), the cost of capturing and interpreting aerial imagery is substantial. Given a large region that must be surveyed as quickly as possible for the adelgid, and a limited budget, moderate-scale satellite imagery appears to be a more feasible alternative. The incorporation of additional spatial data is the most logical way to improve hemlock mapping success using satellite imagery, but must be done so in a manner that the resulting protocol is easy to use and adapt as necessary.

1.5 Project Objectives

The list of known hemlock woolly adelgid infestations in the southern Appalachians is increasing, but certainly there are active infestation locations that have not yet been detected. In addition, numerous hemlock stands in the region that are likely candidates for imminent infestation (i.e., within the next 1-2 years) have not yet been mapped—and ultimately, all hemlock stands are candidates for invasion. Forest managers are not well equipped to handle these data acquisition challenges. Furthermore, there is no regional
framework for addressing spatial issues, which should be a major concern given the speed
with which the adelgid is spreading while, of course, ignoring jurisdictional boundaries. This
indicates the need for a region-wide “early warning system” to address these challenges, and
is the primary focus of the project described in this document.

We undertook this project with two significant objectives. First, we intended to
develop a protocol for mapping hemlock stands in the southern Appalachian region. The
approach had to be (1) easy to implement region-wide and fairly inexpensive; (2) based on
readily available or calculable data sets; (3) apply mapping techniques easily adaptable to
different geographic areas; and (4) meet reasonable standards for accuracy in locating
hemlock stands. We chose to use satellite data for this objective, due to its low cost and
practicality for regional mapping. This resulted in a noteworthy “sub-objective”: identifying
and implementing an approach for handling the effects of topography on satellite image data.
Remotely sensed images of mountainous areas typically contain deep shadows due to relief,
and many of the shadowed areas are precisely those that contain hemlock stands, so this was
a significant concern for this project.

Our second objective was to use existing data to create models to predict areas at the
highest risk of imminent hemlock woolly adelgid infestation. These models had to be
applicable for generating risk maps that could, in turn, be used by managers to prioritize sites
for biological control and other management efforts. The southern Appalachian region
offered us a good opportunity to perform this type of analysis, because it is the first time
where data has been collected regarding when the adelgid was first detected in a particular
geographic area.
We adopted this two-pronged approach with the idea that the results of these separate objectives could be linked. In other words, by overlaying the hemlock distribution and infestation risk maps that result from the outlined methods, forest managers should be able to prioritize stands and reduce the total area that must surveyed or treated for the hemlock woolly adelgid. A significant advantage of this reduction process is that forest managers should have more time, money, and manpower to devote to control measures, rather than devoting these resources to quantifying the extent of their problem. Notably, the two components of our system are also flexible, so the tools and techniques described here should be applicable—if the hemlock woolly adelgid threat persists—to almost any part of the hemlock range in the eastern U.S., with little modification.

1.6 References


McClure, M.S. 1987. Biology and control of the hemlock woolly adelgid. Connecticut Agricultural Experiment Station, New Haven, CT (USA), Bulletin Number 851.


2. LITERATURE REVIEW

2.1 Introduction

There is an extensive body of literature on the application of remote sensing, geographic information systems (GIS), and spatial analysis for phenomena related to forests, including forest pests. A number of researchers have confronted challenges similar to those facing our hemlock woolly adelgid project and have developed practical solutions on which we can draw. Some research concepts that apply to each of our objectives are discussed in the subsequent sections of this chapter. Section 2.2 provides background on methods to remove topographic effects from satellite imagery—a common problem in mountainous environments—and lays out a case supporting one topographic normalization method, the C-correction, as part of our hemlock mapping protocol. Section 3.3 discusses the challenges of mapping hemlock using satellite data, but also examines how a decision tree classification approach may be a feasible way of doing so for the southern Appalachian region. Finally, Section 3.4 compares a number of statistical classification techniques and suggests how they may be used to map areas at risk of hemlock woolly adelgid infestation. All three sections present benefits and limitations of the highlighted techniques, illustrated by pertinent examples from literature. We believe the examples cited here offer justification for the methods we ultimately adopted for our project.
2.2 Topographic Normalization of Satellite Imagery from Mountainous Regions

2.2.1 Basic Concepts

Topographic slope and aspect can introduce radiometric distortion into the recorded signal of satellite data (Jensen 1996). In satellite images captured over regions of high relief, some areas can be completely in shadow, substantially altering the brightness values of the pixels involved—a phenomenon sometimes labeled *anisotropic reflectance* (Colby & Keating 1998; Jensen 1996). A number of different techniques have been proposed for reducing these effects in multispectral data. In particular, topographic normalization attempts to remove variation in illumination due to terrain, such that two objects with identical reflectance properties will have the same brightness values in the subject image, regardless of their orientation relative to the sun (Jensen 1996).

![Diagram of solar incidence (i) and exitance (e) angles for inclined terrain](image)

Riano et al. (2003) divided topographic normalization techniques into two broad categories: those based on band ratios and those that employ a digital elevation model.
Band ratio corrections are simple techniques, with their chief advantage being that they require no additional input data. A basic assumption of band ratio approaches is that reflectance will increase or decrease proportionally between two ratio bands, thus compensating for topographic effects. Like other correction techniques, this does not account for diffuse irradiance, but the contribution of diffuse irradiance to pixel values varies between spectral bands, contradicting the notion of proportionality. In turn, the use of ratios can result in a loss of information or spectral resolution that might be important in classification (Colby 1991; Riano et al. 2003). For instance, differences in the albedo of ground objects can be hidden by ratio techniques if the objects have similar spectral response curves (Colby 1991). Ekstrand (1996) employed band ratio techniques to normalize satellite imagery used for detecting forest damage, with extremely poor results. Justice et al. (1981) illustrated a different, but equally important, problem with band ratios. In their study, a set of ratios resulted in under-correction of topographic effects primarily because there was not enough variation in spectral values for the band used as the ratios’ denominator.

2.2.2 Normalization Methods Based on Digital Elevation Data

The other family of topographic corrections uses a DEM to calculate the cosine of the incidence angle $i$ (i.e., direct irradiance or illumination) for image pixels according to the formula

$$\cos i = \cos \theta_z \cos \theta_n + \sin \theta_z \sin \theta_n \cos (\phi_s - \phi_n)$$

where $\theta_z$ is the solar zenith angle, $\phi_s$ is the solar azimuth, $\theta_n$ is the surface slope, and $\phi_n$ is the surface aspect, with the latter two values supplied by a DEM of the image area (Hale & Rock 2003). An alternative form of the equation is
\[
\cos i = \cos (90 - \theta_s) \cos \theta_n + \sin (90 - \theta_s) \sin \theta_n \cos (\phi_s - \phi_n).
\]

In this case, \((90 - \theta_s)\), where \(\theta_s\) is the solar elevation value, is substituted for the solar zenith angle \(\theta_z\) in the original equation (Hale & Rock 2003). The latter version of the equation is used because the solar elevation, rather than the solar zenith, is commonly reported in satellite image ephemeris data, along with the solar azimuth value. Values for \(\cos i\) range from -1 (minimum illumination) to 1 (maximum) (Riano et al. 2003). Generally, \(\cos i\) decreases with increasing solar zenith angle (i.e., decreasing solar elevation) or increasing surface slope. This is logical, since illumination tends to be low and shadowing significant in extremely steep terrain or when the sun sits at a low angle relative to a position on the Earth.

DEM-based correction methods can be further divided into two categories: Lambertian and non-Lambertian. Lambertian methods assume that reflectance is independent of observation and reflectance angles, and thus equally bright from all directions. In other words, such methods neglect the typically variable illumination geometry of rugged, real-world surfaces in favor of an isotropic reflectance law (Riano et al. 2003; Richter 1997). Smith et al. (1980) maintained that the Lambertian assumption is only valid for a small range of incidence angles. However, the assumption greatly simplifies calculations and requires fewer inputs than non-Lambertian methods (Jensen 1996; Riano et al. 2003).

Teillet et al. (1982) outlined the Lambertian cosine method for topographic normalization, computed as

\[
\rho_H = \rho_T \left( \frac{\cos \theta}{\cos i} \right)
\]

where \(\rho_H\) is the horizontal surface (i.e., the terrain-corrected image data) radiance and \(\rho_T\) is the inclined surface (i.e., the uncorrected data) radiance. Though a straightforward method,
the model generally results in over-correction, especially in areas of low illumination—commonly northern slopes (Ekstrand 1996; Hale & Rock 2003; Smith et al. 1980). This is particularly problematic with respect to the remote sensing of vegetation types commonly associated with northern aspects—such as hemlocks (Godman & Lancaster 1990). If correction coefficients are empirically derived and applied for each band, results may be somewhat improved (Civco 1989).

2.2.3 The Minnaert Constant Method

Nevertheless, most currently used correction methods accept that surfaces exhibit non-Lambertian behavior. The main non-Lambertian approach is based on the Minnaert constant, a function proposed by Minnaert (1941) to model roughness of the moon’s surface. Teillet et al. (1982) introduced the Minnaert constant to the basic cosine function:

$$\rho_H = \rho_r \left( \frac{\cos \theta_z}{\cos i} \right)^k$$

where k is the Minnaert constant. Calling it a constant is somewhat misleading, since the value must be calculated for each band separately. This can be done after linearization of the previous equation:

$$\ln(\rho_r) = \ln(\rho_H) + k \ln \left( \frac{\cos i}{\cos \theta_z} \right)$$

and using an ordinary linear regression where k and ln(\(\rho_H\)) are the regression coefficients.

Colby and Keating (1998) compared the accuracy of land cover classifications generated from a Landsat Thematic Mapper (TM) image that was corrected with (1) a non-Lambertian Minnaert constant approach, (2) a Lambertian approach, and (3) left uncorrected. They found that the Minnaert constant approach yielded significantly better results than the
others. However, there is a significant possible restriction for its use: Minnaert constant values are cover-type dependent (Teillet et al. 1982). For example, Ekstrand (1996) discovered major inadequacies when one fixed Minnaert constant value was applied for each spectral band in a Landsat TM image of a forested region. Arguably, estimating cover-type specific Minnaert constants requires *a priori* knowledge of the vegetation (Hale & Rock 2003). Unfortunately, the discovery of such information is often the reason the classification effort was initiated in the first place. In any case, such information is generally not known without some form of field sampling or other prior source of vegetation data (Hale & Rock 2003).

### 2.2.4 The C-Correction Method

An alternative to the Minnaert constant method is the statistical-empirical correction method. For each pixel in a satellite image, it is possible to correlate the predicted illumination from the DEM with the image data (Jensen 1996). This correlation is assumed to be linear:

\[
\rho_T = \rho_H + m_k \cos i
\]

where \( m_k \) is the slope of the regression line for band \( k \). Teillet et al. (1982) introduced a variation on the statistical-empirical approach named the C-correction, computed as

\[
\rho_H = \rho_T \left( \frac{\cos \theta_z + c_k}{\cos i + c_{-k}} \right)
\]

where \( c_k = b_k / m_k \), for \( \rho_T = b_k + m_k \cos i \). This model introduces a parameter, \( c_k \), that is the quotient between the slope \( (b_k) \) and intercept \( (m_k) \) of the regression equation \( \rho_T \) versus \( \cos i \) (Riano et al. 2003).
2.2.5 Advantages and Drawbacks of the C-Correction

The C-correction has typically performed well when compared to other normalization approaches. For a study region in southwestern Spain, Riano et al. (2003) compared the C-correction of a Landsat TM image with Minnaert and cosine correction normalization methods. They recorded the success of each method in decreasing heterogeneity of pixel values within ten vegetation classes. Prior to processing, all vegetation classes exhibited significant topographic distortion across the areas in which they were distributed. The C-correction method performed substantially better than the others at increasing intraclass homogeneity, with the Minnaert correction performing worst. Low illumination areas were problematic for all, leading the authors to suggest smoothing of the original slope values prior to implementation. McDonald et al. (2000) compared several different correction methods using Landsat TM imagery of New South Wales, Australia. Tested approaches included the cosine correction, the Minnaert correction, the statistical-empirical method, and the C-correction. The Minnaert and C-correction approaches both performed well statistically and visually, while the others resulted in substantial under- or over-correction. For example, the authors performed a canonical variate analysis (CVA) using the C-corrected image, and determined that pixels with the same land cover class, but occurring at different incidence angles, exhibited quite similar spectral responses after the correction. The authors recommended the C-correction over the Minnaert approach primarily because the parameters are more easily obtained from the data.

This is not to say that the C-correction is without criticism. The approach, like the Minnaert-based correction requires that new correction coefficients are computed for each image. Furthermore, some researchers have recommended that cover type information be
included in the C-correction process, as with the Minnaert-based correction. McDonald et al. (2000) went so far as to suggest that an image should be partitioned by major vegetation type (e.g., forest/non-forest) and C-correction coefficients determined separately for each partitioned subset. The impracticality of such classification prior to normalization has already been noted, but it is a significant concern. Hale & Rock (2003) proposed partitioning an image into “sunlit” and “sunshade” subsets according to the solar azimuth as another way to improve classification accuracy.

Another important drawback of the C-correction or any other DEM-based normalization is propagation of error from the DEM (Gu et al. 1999). The correction is heavily dependent on the DEM’s accuracy and resolution, and this can lead to localized areas of over- or under-correction. The C-correction and similar approaches consider information only at the level of individual pixels, and drawing information from neighboring pixels may reduce such instances of error (Gu et al. 1999). Hale and Rock (2003) suggested that filtering a DEM via a neighborhood-based function (e.g., a low-pass statistical filter) might provide a workable solution. This will yield a smoother DEM as well as its corresponding aspect and slope surfaces, which may improve the C-correction in low illumination areas (Riano et al. 2003).

The C-correction and similar approaches employ a number of simplifying assumptions. Gu and Gillespie (1998) argued that, for dense forest, pixel brightness values are affected by sub-pixel characteristics of the forest canopy that cannot be modeled at the pixel scale. The C-correction and similar methods correct only at the pixel scale based on the geometric relationship between the sun, sensor, and terrain. As an alternative, Gu and Gillespie proposed a normalization method based on sun-canopy-sensor (SCS) geometry that
incorporates the forest canopy’s impact on radiance. They compared success of the cosine correction, the Minnaert constant, the C-correction, and their SCS method using Landsat TM images of a forested area, and found that the SCS performed comparably to the Minnaert and C-correction methods. They also argued that the SCS method was more consistent scene-to-scene than either of latter methods. On the other hand, Vincini and Frazzi (2003) compared the cosine correction, the SCS method, the Minnaert constant, the C-correction, and a normalization technique known as the B-correction, which is entirely empirical in nature in that it requires no photometric input. They applied these methods to Landsat TM imagery of a mixed deciduous forest region. The B-correction and C-correction methods outperformed all other methods, performing similarly in most cases, although the B-correction performed better than the C-correction in an image with a high solar zenith angle. In their study, the SCS method resulted in marked over-correction.

Newly developed algorithms such as the B-correction and SCS method may indeed be superior to the C-correction, but they have not been tested extensively. Meanwhile, the C-correction appears to be robust, standing up fairly well to these and other approaches in a number of comparative studies. It has a significant advantage in that it is easy to calculate based on readily available digital elevation data and simple regressions based on pixel values. Nonetheless, it is important to understand its limitations, and to realize that scene-to-scene consistency of any topographic normalization method is a concern.

2.2.6 Atmospheric Correction and Topographic Normalization

Related to this issue of consistency is the effect of the Earth’s atmosphere on remotely sensed imagery. Due to scattering and absorption by gases and aerosols,
electromagnetic radiation is changed as it passes through the atmosphere from the Earth to a satellite-based sensor (Song et al. 2001). This can substantially alter image pixel values. Atmospheric correction methods are commonplace in remote sensing, and range from simple approaches such as dark object subtraction to sophisticated models that are proprietary and expensive to purchase (Jensen 1996; Song et al. 2001). In cases where a single image is being classified (e.g., our project’s use of a Landsat image to generate an evergreen/non-evergreen map), atmospheric correction is unimportant, but for applications that use images from multiple dates, it is probably a mandatory step (Song et al. 2001). Some researchers have argued that atmospheric correction should be performed before topographic normalization, while others have suggested it should be performed concurrently—although the latter necessitates a complex, unwieldy processing model (Riano et al. 2003). Fortunately, for new sensors such as the EOS Terra satellite’s Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), atmospherically corrected products are readily available, and at spatial resolutions comparable to Landsat (LP-DAAC 2005a; 2005b). This offers a significant time saving for projects that will ultimately require regional image coverage.

2.2.7 Summary

Satellite imagery of mountainous areas exhibits highly variable illumination due to topographic relief, and this variability can be a serious impediment to successful image classification efforts. A number of methods have been proposed for normalizing this topographic effect. One method in particular, the C-correction, uses a DEM to model illumination and then regression to relate uncorrected pixel values to the modeled
illumination values. While it does have limitations and is built on simplifying assumptions, the C-correction has performed well compared to other topographic normalization methods. Because it is fairly easy to implement, the C-correction appears to be a logical first step in our hemlock mapping protocol. Moreover, imagery from new sensors such as ASTER are preferred for use in the protocol because atmospheric effects—which also contribute to image variability—have already been corrected by the data provider prior to distribution.

2.3 Approaches for Mapping Hemlocks in the Southern Appalachians

The general habitat descriptions for eastern hemlock (*Tsuga canadensis*) and Carolina hemlock (*T. caroliniana*) suggest a narrow window of suitable conditions, defined by a small suite of topographic and environmental factors: moist, rocky soils; slopes, hillsides, ravines, and riparian zones; and northern or eastern aspects with cool, shaded conditions (Delcourt & Delcourt 2000; Godman & Lancaster 1990; Quimby 1996). This is something of a simplification; in reality, hemlocks can also be important in well-drained upland sites—including ridgetops—and at western aspects (Boyce 1999; McWilliams & Schmidt 1999). The aforementioned factors are certainly suitable as general rules of thumb for mapping hemlock distribution. Nevertheless, it should be noted that these general rules are likely based on studies from the mid-Atlantic and northeastern U.S., where *T. canadensis* is regularly a dominant or co-dominant forest species, and which has been the setting for most hemlock and hemlock woolly adelgid research to date (Shriner 2001). The southern Appalachians, by comparison, hold only 8% of the U.S. hemlock inventory, distributed across 2% of the U.S. hemlock habitat (McWilliams & Schmidt 1999). While some hemlock-dominated stands do exist in the southern Appalachian region, the species is
commonly a secondary associate of a variety of hardwood species, particularly species from
the oak group (Godman & Lancaster 1990; McWilliams & Schmidt 1999). This is not meant
to marginalize the importance of hemlocks in the region. On the contrary, an analysis of the
relative importance of vegetation species in Great Smoky Mountains National Park ranked
eastern hemlock as the second most common and important species in the park, if all
vegetation strata are considered (Shriner 2001). However, when only tree data were
considered, *T. canadensis* dropped to sixth in importance in the park. This echoes the
seminal work of Whittaker (1956), who noted the existence of “hemlock forest”—where
hemlocks occupied 70-80% of the canopy—in some high elevation areas of the park, but also
asserted that hemlock was just one of several dominant species in low-elevation cove forests,
though perhaps the most important. In other words, while eastern hemlock is extremely
important and widespread in the Great Smoky Mountains, it is not always dominant or even a
secondary component of the canopy. This suggests that simple rules of thumb may not be
adequate for complete mapping of southern Appalachian hemlock distribution.

2.3.1 Remote Sensing and Hemlock

Air- or space-borne remotely sensed imagery can provide useful, spatially explicit
information about an area’s vegetation. In their guidebook for minimizing *hemlock woolly
adelgid* impacts in the eastern U.S., Ward et al. (2004) suggested that hemlock stands could
be mapped successfully with aerial photography. Although this is true, air photo
interpretation is an expensive and time-consuming proposition for mapping a fairly large
region such as the southern Appalachians. For instance, large-scale (1:12,000) air photo
coverage of Great Smoky Mountains National Park for a recent vegetation mapping effort
required more than 1000 photos and many months of delineation work by interpreters (Welch et al. 2002). Airborne hyperspectral imagery might be able to capture demonstrated spectral differences between species with overlapping spatial distributions—for example, *T. canadensis* and red spruce, *Picea rubens* (Rock et al. 1994)—but it is also quite expensive and difficult to apply regionally; indeed, each hyperspectral image may have hundreds of bands that must be both atmospherically corrected and topographically normalized prior to use. An automated mapping approach using moderate-scale satellite imagery seems more practical and efficient, given the amount of hemlock habitat in the southern Appalachian region (and elsewhere) that has not been mapped but is at imminent risk of hemlock woolly adelgid infestation.

A small group of researchers have already used moderate-scale imagery for hemlock studies, but have tended to focus on mapping change in hemlock health due to the adelgid. Royle and Lathrop (1997) used Landsat TM imagery to identify areas in New Jersey’s Highlands region that had changed between 1984 and 1994 due to adelgid-induced hemlock mortality or decline. For each pixel in their study region, they calculated the change in a simple vegetation index (near infrared/red band ratio) between the two anniversary dates. They then performed a linear regression of hemlock damage class (healthy/light, moderate, severe, dead) recorded from ground survey plots on the vegetation index difference, and used this equation to map estimated hemlock damage across the region. Their four-class map was 64% accurate. Royle and Lathrop (2002a) then expanded this pilot study, exploring the applicability of other vegetation indices: the Normalized Difference Vegetation Index (NDVI), the Atmospherically Resistant Vegetation Index (ARVI), and the Modified Soil Adjusted Vegetation Index (MSAVI$_2$). A discriminant function combining all of these
indices performed fairly well (73% accurate) at identifying hemlock health to five levels of damage, and even better with a four-level damage scale (82% accuracy). These researchers also used an extended time series (1984, 1992, 1996, 1998) of Landsat images to better quantify the temporal pattern of hemlock decline, as well as to examine landscape-level factors affecting that decline, the latter with modest results (Royle & Lathrop 2000a, 2002b). Bonneau et al. (1999a; 1999b) performed similar change detection analyses for a study area in southern Connecticut using a Landsat time series spanning eleven years (1985-1995). Calculating change using the MSAVI2, they classified hemlock stands to four levels of health with 82% overall accuracy.

Of course, these studies required mapping of the hemlock distribution in order to perform meaningful analysis. Bonneau et al. (1999b) performed a straightforward supervised classification of their 1985 Landsat image to create a 10-class land cover map of their Connecticut study area, with a resulting overall accuracy of 86%, and better for hemlock classification (88.5% producer’s accuracy, 94.7% user’s accuracy). Unfortunately, their approach does not serve as a particularly good example for the southern Appalachians. First, the gentler topography in Connecticut required no explicit correction prior to classification and exhibited no major shadowing. Second, relatively pure areas of hemlock were narrowly restricted to steep riparian zones that had escaped earlier clearing for agriculture use, rather than being diffused throughout a forested landscape, and were thus more easily discernible. Moreover, the authors were able to complete their classification with relatively few training signatures (27 signatures, including 4 for hemlocks at different topographic positions), suggesting little likelihood of confusion between land cover classes. Indeed, pine was the
only other evergreen land cover class besides hemlock, and occupied a much smaller percentage of the landscape (0.88% versus 2.03 % for hemlock).

In contrast, Royle and Lathrop (1997) noted substantial difficulty in separating hemlock from pine and spruce in either coarse-scale aerial photography or satellite data of their New Jersey study area. Their solution was to create a base map of evergreen/coniferous forest cover from one of their Landsat images using spectral modeling techniques. They performed a regression analysis relating percent conifer cover values estimated from aerial photos to corresponding Landsat near-infrared/red band ratio brightness values ($R^2=0.96$). After applying the resulting model, they labeled pixels determined to be 30-100% conifer as hemlock forest, both because hemlock was by far the dominant evergreen tree species in the area, and because the 30% threshold likely masked out deciduous forest pixels with a broadleaf evergreen shrub layer. By their admission, this map was highly accurate for evergreen presence, but less accurate regarding hemlock specifically. In later work, Royle and Lathrop (2002a) acknowledged that their limited ability to distinguish hemlock from other evergreens created a ceiling for the accuracy of their change analyses, and asserted that improvement was likely impossible without ancillary data beyond those provided by Landsat imagery.

Royle and Lathrop’s observation on ancillary data highlights a possible tactic for mapping hemlock distribution in the southern Appalachians. Ancillary data sources can help to distinguish classes that are difficult to identify from remotely sensed imagery alone, particularly when mapping a fairly large area and when the classes of interest exhibit high within-class variability (Jensen 1996; McIver & Friedl 2002). It can be a challenge, however, to identify and procure the most helpful ancillary data. Ground surveys are
extremely useful for ecological analyses, but often unfeasible due to time and cost, leaving GIS-based data sources as an alternative (He et al. 1998). With such data sets, scale can be a concern. For example, the USDA Forest Service Forest Inventory and Analysis (FIA) plot data offer a wealth of information on the tree species they contain, and have even been used to map hemlock tree distribution, but they are of limited sampling density: 1-2 plots/10 km$^2$ (He et al. 1998; Morin et al. 2005). This limits their usefulness for analyses that require a high degree of spatial explicitness. Even at an appropriate resolution, ancillary data can be difficult to integrate before key variables and their threshold values have been determined. Such decisions can be further complicated by the huge amounts of data that may be stored in a GIS or other database structure and that must be filtered or reduced in some manner prior to use (Murthy 1998).

2.3.2 Decision Trees

Fortunately, there is a growing body of literature on automated data exploration techniques and their application in both remote sensing and the spatial modeling of species distributions. The decision tree concept broadly includes a number of algorithms for the automated analysis and reduction of complex multi-factor data, as well as the subsequent application of those data in classification and prediction. Decision trees have roots in statistics, engineering, and decision theory, and are also commonly used in disciplines such as data mining (Jovanovic et al. 2002; Murthy 1998). For example, during the 1970s, efforts to classify Landsat and other remotely sensed imagery drove the examination of decision trees for pattern recognition purposes (Murthy 1998). Regrettably, the numerous disciplines currently employing decision trees have developed slightly different terminology and
nuances of approach, creating some confusion (Murthy 1998). For example, some researchers sub-categorize decision trees according to whether the response variable is categorical (classification trees) or numeric (regression trees) (De’Ath & Fabricius 2000). Nonetheless, the basic mechanism is nearly universal: Tree building (or induction) occurs in a top-down fashion, starting with an empty tree and a training data set. The training data are split into mutually exclusive groups that are as homogeneous as possible, and then (if possible) each of these groups is split independently, and so on. The process is repeated until the terminal groups exhibit maximum possible homogeneity (De’Ath & Fabricius 2000). Each split is a decision rule employing a threshold value for one (or more, in some software packages) of the input variables (De’Ath & Fabricius 2000; Murthy 1998). Decision trees are often represented graphically with the root node—the unsplit training data—at the top, followed by the branches, internal nodes, and then the terminal nodes or leaves (De’Ath & Fabricius 2000; Murthy 1998).

Decision tree algorithms use a variety of splitting criteria, often based on distance measures. Breiman et al. (1984) provided a detailed treatise on the classification and regression tree (CART) algorithm, which was strongly endorsed for ecological data analysis by De’Ath & Fabricius (2000). The CART algorithm employs the Gini index, a measure of the relative purity of partitioned subgroups relative to their parent group (De’Ath & Fabricius 2000; Gehrke 2000; Murthy 1998). The Chi-Squared Automatic Interaction Detection (CHAID) algorithm uses the $\chi^2$ statistic for distance measurement. There are other distance measures (e.g., Bhattacharya distance) as well as information-theory-based splitting criteria, and hybrids of various approaches. There are other noteworthy distinctions. For example, CART is limited to binary (i.e., two-way) splits, while CHAID allows for an $n$-way number
of partitions (Jovanovic et al. 2002; Nelson et al. 2003). CART typically requires smaller training data sets than CHAID (Lee & Siau 2001). How decision tree algorithms limit tree size is quite variable, and together with the choice of distance measure, this has motivated the creation of the myriad of decision tree algorithms that are available (Murthy 1998).

In a general sense, there are a number of advantages to decision trees over other classification tools such as linear discriminant analysis or neural networks. Most importantly, they employ very few assumptions, which makes them quite flexible (McIver & Friedl 2002). They are non-parametric, so normality of variables is not critical; in fact, decision trees can handle a wide variety of input data distributions as well as missing data values (De’Ath & Fabricius 2000; Murthy 1998). Decision trees can accommodate categorical and/or numeric explanatory variables, and response variables can take many forms: numeric, categorical, ratings, and survival data (De’Ath & Fabricius 2000). The ability to accept categorical data is a major advantage over neural networks, which require numeric inputs and, as a result, often yield a large number of predictor variables (Jovanovic et al. 2002). The algorithm chooses the best splitting variables to maximize homogeneity, so this process effectively eliminates variables that prove extraneous or redundant. Finally, the clearly defined splitting rules of a decision tree are easy to interpret, and can be readily transported to GIS or other analytical settings. Neural networks, by comparison have some hidden parameters that are difficult to translate into explicit rationale, and can have unpredictable results (Jovanovic et al. 2002; Lawrence & Wright 2001).

There are also criticisms of decision tree algorithms. Training data samples generally must be quite large for good decision rule development and to avoid overly optimistic estimates of the classifier’s success (Murthy 1998). Since decisions at lower levels of a tree
are based on smaller and smaller fragments of the training data, some leaves generated from these fragments may be unimportant, probabilistically. Furthermore, since many leaves may be necessary to capture different instances of the class of interest, large decision trees may result (Murthy 1998). Such large trees can be difficult to interpret, but more significantly, they can over-fit the data, making them less generalizable to other data sets (Jovanovic et al. 2002). Nelson et al. (2003) noted that over-fitting can be particularly pronounced when working with a large number of continuous variables, and provided evidence that discriminant analysis or neural networks may be more efficient in such cases. At times, intricate decision trees can result even when fundamental relationships among the variables are quite simple (Jovanovic et al. 2002). Moreover, several different decision trees can produce similar or even identical degrees of accuracy, making the selection of the best candidate a difficult proposition (Murthy 1998).

Tree size is a major concern, so different methods have been proposed for limiting the size of trees. Setting a minimum number of observations per node is one possibility, though this has been demonstrated to not be a robust measure (Murthy 1988). Another approach is to limit the maximum depth to which the tree may grow, though this may eliminate splits that are justified statistically (Jovanovic et al. 2002). In terms of the latter, there are several statistical approaches for limiting tree size. Size can be limited beforehand; for example, a purity threshold (e.g., a Gini index or $\chi^2$ value) can be set, such that if a particular split does not meet the threshold, then tree growth is stopped (Li et al. 2001; Murthy 1998). This is the approach applied by older tree-based algorithms such as CHAID (Li et al. 2001). As an alternative, Breiman et al. (1984) suggested growing over-sized trees and then pruning them back by removing branches that are not contributing significantly to the tree’s accuracy.
Pruning is generally accomplished by dividing the input data into a training set to grow an over-fitted decision tree and a validation set to remove branches judged insignificant or not beneficial (Murthy 1998; Nelson et al. 2003). Breiman et al. (1984) introduced the cost-complexity pruning approach, which is probably the most widely accepted pruning method (Li et al. 2001). This method proceeds by building nested, increasingly smaller trees from the training data based on assessment of the per-leaf accuracy cost. Then, one of these smaller trees is chosen based on its accuracy at classifying the validation data set (Li et al. 2001; Murthy 1998). Breiman et al. (1984) also introduced a cross-validation technique that eliminates the need for a separate validation data set, but it has been criticized in several regards, particularly that it has a large variance and may not yield the most optimal tree (Murthy 1998). Ultimately, no pruning method associated with a particular decision tree method has been demonstrated to consistently outperform other methods (Murthy 1998).

2.3.3 Examples

Decision tree algorithms and their operational parameters continue to evolve, and in the interim, have engendered ongoing debate on the “best” approach. Still, during the past several years a number of remote sensing studies have been published that used decision tree approaches for land cover or vegetation classification. Friedl and Brodley (1997) demonstrated, with three different remotely sensed data sets, that decision trees outperformed the commonly used maximum likelihood algorithm as well as linear discriminant functions in accurately classifying land cover. In a more specific example, Lawrence and Wright (2001) applied the CART algorithm for land cover classification in the Greater Yellowstone Ecosystem. They employed 26 explanatory variables including Landsat TM data (raw pixel
and difference values between images transformed by the Tasseled Cap algorithm),
elevation, slope, and aspect. They built trees to classify the area’s land cover to three
hierarchical levels, with good results: 96% overall accuracy for the broadest and 65% for the
most detailed classification scheme, the latter of which included classes distinguishing
between dominant tree species (e.g., aspen, cottonwood, willow, Douglas-fir). Similarly,
Rogan et al. (2003) used the CART algorithm and multitemporal Landsat TM data to
quantify land cover change for a six-year period (1990-1996) in the San Diego area. Input
variables included image data treated to the Multitemporal Kauth Thomas transformation,
which yields values for stable brightness, greenness, and wetness, as well the degree of
change in those three measures. The authors also incorporated environmental variables such
as elevation, slope, aspect, fire history, and existing land cover. They derived a three-level
hierarchical land cover classification, with the most detailed level (Level III) including nine
different classes for percent forest canopy change as well as shrub/canopy and developed
area change. The Level III classifier employed a tree with 20 terminal nodes, with an overall
accuracy of 72%. The authors noted that most of their classification errors occurred between
classes depicting forest canopy decreases, suggesting that subtle changes in forest pixel
values may not always be discernible using moderate-resolution satellite data, even with a
fairly sophisticated decision tree classifier.

Brown de Coulston et al. (2003) used C5.0, a commercially available decision tree
classifier, to delineate the vegetation of the Delaware Water Gap National Recreation Area to
the formation level of the U.S. National Vegetation Classification. They built their decision
tree using Landsat Enhanced Thematic Mapper (ETM+) image band values as well as NDVI
values, abandoning terrain data because of highly variable elevation, slope, and aspect within
their eleven fairly broad formation classes. Their final land cover map had an overall accuracy of 82% based on field data, although this figure improved to 99.5% when comparing only forest versus non-forest.

Joy et al. (2003) modeled forest vegetation types at 10-m spatial resolution using field data, applying topographic information (elevation, slope, aspect, and landform) and Landsat TM data as ancillary variables. Elevation was their primary discriminating variable, but several Landsat bands were also important in the final decision tree. Overall accuracy was 74.5%. Other variables that the authors included in their analyses, such as canopy closure, understory vegetation species, and proportion of ground covered, proved to be unimportant for vegetation type delineation. The authors theorized that these variables reflected differences in forest stocking levels rather than more important differences in species composition.

In addition to these studies, a number of others have attempted to predict vegetation type based simply on environmental data, i.e., without remotely sensed image data. As Guisan and Zimmermann (2000) noted, modeling the potential distribution of plant species or communities is equivalent to modeling their potential habitat, so variables employed in these studies commonly came from broad categories generally considered important in plant geography, such as climate and soils. Franklin (2002) used decision tree models to predict the distribution of eight dominant shrub species across ~3900 km² in California, for the purpose of augmenting an existing land cover map. She employed climatic (minimum and maximum temperatures, annual precipitation) and terrain (topographic moisture index and southwestness) variables, and the resulting trees predicted shrub presence/absence (based on field plot data) with 75% accuracy. Cairns (2001) used general linear models, neural
networks, and classification tree methods to predict vegetation type from alpine treeline ecotone data in a portion of Glacier National Park. He applied a suite of environmental indicator variables including elevation, moisture potential, solar radiation potential, and disturbance history, and then compared the predicted vegetation type to the actual vegetation. Prediction accuracy was highly variable, which Cairns attributed to variability between the watersheds used in the analysis. His conclusion, ultimately, was that no single classification method performed best, but that all could be used to address this sort of problem, and could be generalized fairly well.

The notion that—at least for mountainous environments—a digital elevation model DEM can provide highly diagnostic variables for vegetation prediction is perhaps unremarkable (Franklin & Wilson 1992). The widely cited mosaic diagrams of Whittaker (1956) distributed the vegetation types of the Great Smoky Mountains, including hemlock forests, along principal axes of elevation and landform. Numerous studies in the southern Appalachian region have confirmed the importance and, in some cases, the predictive power of topography. Callaway et al. (1987) applied multivariate analysis in a study limited to just the lower-elevation western Smokies, and determined the most significant variables in vegetation distribution to be elevation and a terrain-derived protection index. McNab (1991) performed a canonical discriminant analysis to identify important variables distinguishing four hardwood forest types in Bent Creek Experimental Forest, near Asheville, NC. He found that five topographic variables—elevation, aspect, slope, landform index, and a surface shape index—accounted for 97% of the variation between the forest types. McNab argued that his analysis suggested forest types in the southern Appalachians could be predicted fairly accurately based just on variables determined from a DEM, although he noted in a later study
that extrapolating relationships from one landscape-scale study area to another is problematic without incorporating additional data (McNab 1996).

In another project in the southern Appalachians, Bolstad et al. (1998) found strong relationships between several canopy species (including *T. canadensis*) and DEM-derived measurements of elevation and the terrain shape index (McNab 1989). However, at least by themselves, terrain shape and elevation had limited power to predict study area vegetation in four broad classes (northern hardwood, mixed deciduous, cove, xeric oak/pine). None of their prediction approaches exceeded 57% overall accuracy. Treitz and Howarth (2000) echoed this finding, noting the weak discriminatory power of terrain variables by themselves, compared to their strong contribution to forest ecosystem classification when used in conjunction with spectral variables at a landscape scale.

This last point suggests that for mapping hemlocks in the southern Appalachians, digital elevation data would be most effective when used in conjunction with satellite imagery. In addition, even if topography is the primary overall factor in vegetation distribution, environmental factors such as proximity to streams can be extremely important for individual species, as is the case with hemlock (Franklin et al. 2001). Variables representing soil characteristics can also add some distinguishing power (Elliott et al. 1999), but may not be available at the appropriate scale. Site moisture indices, calculated in from spatial data and other sources, may serve as adequate surrogates for certain soil characteristics (Parker 1982). Adopting a decision tree approach to combine all of these factors offers us flexibility and, perhaps more importantly, will yield threshold values for the application of these variables in regional mapping of hemlocks.
Admittedly, a given environmental factor might strongly influence vegetation
distribution at a particular location with little or no impact on a nearby site (Abella et al.
2003). Furthermore, forest species can be distributed along “subtle and nonlinear gradients
of environmental complexes” (Abella et al. 2003, p.1943). To accommodate such variety,
the area chosen to train a decision tree should be topographically and environmentally
diverse; otherwise, it will not yield a good set of generalized rules. Of course, the data
representing the variable that control species distribution often do not truly capture on-the-
ground complexity, due to issues of scale, accuracy, and precision (Abella et al. 2003). Our
best options for hemlock mapping are to procure the finest available spatial data and to be
aware of the assumptions underlying our analysis.

2.3.4 Summary

There are a number of obstacles to accurate hemlock mapping in the southern
Appalachians using satellite imagery. While hemlocks are widespread in the region, they are
rarely in large homogeneous stands. Instead, they tend to be interspersed with other
evergreen and deciduous species, with their distribution broadly influenced by topographic
and environmental factors. Traditional image classification techniques using only spectral
information have limited ability to distinguish hemlocks. In contrast, a decision tree
classification approach incorporates ancillary spatial data to generate rules defining hemlocks
and other categories of interest. Decision trees have become increasingly popular in remote-
sensing-based land cover and vegetation classification due to the flexibility of the approach:
They are non-parametric, can accommodate a wide range of variable types, and yield clearly
defined splitting rules highlighting the most important variables. Decision trees are not
without criticism; in particular, they can be over-fitted to the training data set used to generate the rules. Nevertheless, the combined use of satellite image data, DEM-derived topographic variables, and environmental variables in a decision tree classifier appears to be a logical component of our hemlock mapping protocol.

2.4 Predicting Hemlock Woolly Adelgid Invasion at the Landscape Level

A review of the existing hemlock woolly adelgid literature yields a rather wide estimate for the pest’s general rate of spread—8 to 30 km per year—and suggests that a main front is expanding rapidly into the southeastern U.S., with isolated infestations ahead of the main front (Cheah et al. 2004; Evans 2004; McClure 1996; Souto et al. 1996; Ward et al. 2004). The only existing field study on hemlock woolly adelgid dispersal behavior (McClure 1990) asserted that wind, deer, and especially birds can all spread eggs and crawlers some distance from an infested site. Indeed, birds may carry the pest a few kilometers or more in a single season, and may in fact be most responsible for the expanding front (McClure 1987; Ward et al. 2004). Humans, of course, have similarly been implicated in rapid, long-distance transport of invasive pests including the hemlock woolly adelgid (Pimentel et al. 2000; SAMAB 2004).

Because the hemlock woolly adelgid may be spread by several different mechanisms, the appearance of new infestations in the southern Appalachians may seem unpredictable when viewed at a broad scale. In other words, while it is relatively easy to explain the spread of adelgids from one infested hemlock stand into nearby stands, it is harder to explain why new infestations might appear in sites that are geographically distant from existing infestations, and in some cases, separated by natural boundaries such as mountains. Long-
distance dispersal events are simply more difficult to model adequately (Hulme 2003). In the case of the hemlock woolly adelgid, landscape-level factors are almost certainly at work, though their importance may just amount to influencing the behavior of vector organisms (e.g., birds following riparian corridors). The challenge is to find a way to identify key landscape-level variables and represent them in a spatially explicit manner. Our ultimate aim is to identify areas at the highest risk of imminent hemlock woolly adelgid infestation, and thus allow forest managers to target their resources on those areas rather than the entire forest. This goal echoes Andersen et al. (2004), who asserted that a better understanding of landscape ecology and structure are important facets of good invasive species management because they permit a habitat-oriented view of the problem and may ultimately enhance the possible management options.

A number of different statistical and spatial analysis techniques can be employed to examine landscape-scale data sets, identify key variables, and generate functions to map infestation risk. Discriminant analysis, logistic regression, and decision tree approaches have all been used for spatially structured problems with some success. At least superficially, the three multivariate techniques resemble each other. Their basic purpose is to classify a given observation into one of two or more alternative groups (or populations) based on a set of measurements (Afifi et al. 2004). All require an initial “training” sample where the group designation of each observation is known a priori. For example, we have a sample of the first locations where the hemlock woolly adelgid was detected in the Great Smoky Mountains and along the Blue Ridge Parkway (the “positive” group), which we can pair with a sample of hemlock stands that were not infested during the first year (the “negative” group). The sample serves as the training data set that determines the parameters of the
discriminant function, logistic equation, or decision tree, which can then be used for predicting group membership of additional data points (Afifi et al. 2004; McLachlan 1992). In addition, all of the techniques enable determination of the variables that contribute best to the classification process (Afifi et al. 2004).

Of course, each technique has particular advantages and disadvantages, and these are discussed in subsequent paragraphs. Typically, no classification technique offers an error-free solution, since the attributes of the groups in question may not provide clear distinction (Johnson & Wichern 2002). The idea is to find the “optimal” classification that offers the lowest misclassification rate and the best efficiency. The notion of efficiency is context-specific; for example, with respect to hemlock woolly adelgid, efficiency likely means a high accuracy rate but a relative small forest area designated as high-risk—if the area is too large, it does not adequately reduce the territory that must be covered by forest managers.

2.4.1 Discriminant Analysis

Discriminant analysis is widely used in medical diagnosis and pattern recognition, and examples of its application can be found in chemistry, biology, genetics, education, sociology, and many other disciplines (Khattree & Naik 2000; McLachlan 1992). Discriminant analysis can be divided into two separate goals, generally termed discrimination and allocation (Johnson & Wichern 2002). The first goal is to identify “discriminants”: variables whose numerical values separate input observations as much as possible into distinct groups (Johnson & Wichern 2002). The second goal is to derive a function for optimally classifying new observations into the defined groups (Johnson & Wichern 2002). Essentially, allocation follows from discrimination of a training sample.
Though discriminant analysis can be used to assign observations to more than two groups, it is best suited for two-group classification (Afifi et al. 2004).

Fisher (1936) depicted the discriminant function as

\[ Z = a_1 X_1 + a_2 X_2 \ldots a_p X_p \]

where \( X_1 \ldots X_p \) are the variables used in classification. The coefficients \( a_1 \ldots a_p \) are selected to maximize the distance between groups. Mahalanobis distance has become the common distance measure when the employed variables are continuous, and squared Mahalanobis distance, \( D^2 \), is calculated as

\[ D^2 = \frac{(\overline{Z}_1 - \overline{Z}_2)^2}{S^2_Z} \]

where \( \overline{Z}_1 \) and \( \overline{Z}_2 \) represent the mean values of \( Z \) for the two groups and \( S^2_Z \) is the pooled sample variance (Afifi et al. 2004; McLachlan 1992). The larger the value of \( D^2 \), the easier it is to distinguish between groups (Afifi et al. 2004).

The discriminant rule can also be depicted in matrix terms. An observation with the attribute vector \( x_0 \) will be allocated to Group 1 (of two) if

\[ (\overline{x}_1 - \overline{x}_2)' S^{-1}_{pooled} x_0 - \frac{1}{2} (\overline{x}_1 - \overline{x}_2)' S^{-1}_{pooled} (\overline{x}_1 + \overline{x}_2) \geq \ln \left[ \begin{pmatrix} c(1|2) \\ c(2|1) \end{pmatrix} \begin{pmatrix} p_2 \\ p_1 \end{pmatrix} \right] \]

where \( \overline{x}_1 \) and \( \overline{x}_2 \) are the mean vectors of the two groups, \( S^{-1}_{pooled} \) is the pooled covariance matrix, \( p_1 \) and \( p_2 \) represent prior probabilities, and \( c(1|2) \) and \( c(2|1) \) represent the per-class costs of misclassifying an observation (Johnson & Wichern 2002). Often, misclassification costs and prior probabilities are unknown and thus assumed equal between groups. Since this results in \( \ln(1) = 0 \), the discriminant rule amounts to a comparison of the test statistic against zero. If the test statistic value is less than zero, the observation will be allocated to...
Group 2 (Johnson & Wichern 2002). Discriminant analysis also generates posterior probabilities of class membership for each observation, which can be used to assign observations instead of applying the discriminant rule (Johnson & Wichern 2002; Khattree & Naik 2000).

Discriminant analysis assumes the input data have a multivariate normal distribution when performing certain statistical tests, e.g., testing whether the means of the populations associated with the two groups are significantly different (Afifi et al. 2004; Johnson & Wichern 2002). By extension, each explanatory variable is typically univariate normal as well. Though approximate normality may suffice for many analyses, very skewed or long-tailed distributions for some variables can increase the total error rate and the results can be misleading (Afifi et al. 2004; Khattree & Naik 2000). In such cases, transformation to near normality is advised, and there are a number of methods for assessing normality and identifying the best variable transformations (Johnson & Wichern 2002; Khattree & Naik 1999; Khattree & Naik 2000). The overall quality of the data is also important to the error rate. Outliers should be removed and the data should be checked for independence, and the number of missing data values should be similar for the two groups (Afifi et al. 2004).

Success of a discriminant analysis function is typically judged by calculating its error rate. The true error rate generally cannot be calculated because it requires information about the parameters of the parent populations that is generally unknown (Johnson & Wichern 2002). Nevertheless, the apparent error rate can be easily calculated by

\[ \frac{n_{1M} + n_{2M}}{n_1 + n_2} \]

where \( n_{1M} \) is the number of Group 1 observations misclassified as Group 2, \( n_{2M} \) is the number of Group 2 observations misclassified as Group 1, \( n_1 \) is the actual number of Group 1
observations in the training sample, and $n_2$ is the actual number of Group 2 observations (Johnson & Wichern 2002). Unfortunately, the apparent error rate tends to underestimate the actual error rate except for very large samples, primarily because the function is tested using the same data with which it was constructed (Johnson & Wichern 2002). An alternative is cross-validation, in which each observation is individually omitted from the sample. A discriminant function is generated with the remaining observations, and the function is used to classify the omitted observation. By tabulating the number of misclassified holdout observations in the same manner as the apparent error rate, a nearly unbiased estimate of the actual error rate is possible (Johnson & Wichern 2002).

The discriminant rule described above is linear and applies only in cases where the covariance matrixes of the two groups are equal. If their covariance matrices are unequal, a pooled variance and linear discriminant function are no longer appropriate (Johnson & Wichern 2002). Instead a quadratic rule is used to allocate $x_0$ to group 1:

$$-\frac{1}{2}x_0' (S_1^{-1} - S_2^{-1})x_0 + (\bar{x}_1'S_1^{-1} + \bar{x}_2'S_2^{-1})x_0 - k \geq \ln \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right)$$

where $S_1$ and $S_2$ are the covariance matrices of the two groups, and

$$k = \frac{1}{2} \ln \left( \frac{|S_1|}{|S_2|} \right) + \frac{1}{2} \left( \bar{x}_1'S_1^{-1} - \bar{x}_2'S_2^{-1} \right)$$

(Johnson & Wichern 2002).

As with the linear discriminant rule, this amounts to a comparison versus zero if the misclassification costs or prior probabilities are all equal. A major limitation of the quadratic discriminant function is that it is particularly sensitive to even minor departures from multivariate normality. In addition, a single observation can be extremely influential on the membership probabilities of the other observations in a data set, necessitating close attention to possible outliers when using a quadratic rule (Fung 1996). Furthermore, classification by
this quadratic rule is awkward in more than two dimensions and can yield unusual results (Johnson & Wichern 2002). Though quadratic discriminant analysis is more effective than linear discriminant analysis in cases where the group variances are unequal, it generally will not be if the variances are in fact equal, so testing for this condition prior to analysis is advisable (Mason 1998).

There are additional caveats for discriminant analysis, whether linear or quadratic. First, users must be cautious about applying discriminant analysis on small samples with many variables (Afifi et al. 2004). Second, it is a common practice to divide the input sample data into two groups, where two-thirds of the data are used to generate a function for classifying the remaining one-third, and see if the misclassification rates are similar. However, when the function is applied to yet another sample, the misclassification rate can change substantially, perhaps through removal of key information (Johnson & Wichern 2002). This is a difficult problem to avoid—the only guideline is to pick a sample that is very representative of the larger population to which the function will ultimately be applied, which may be quite difficult to judge (Afifi et al. 2004). Third, discriminant analysis is generally reserved for continuous, normally distributed variables. Theory for the use of both continuous and categorical variables in discriminant analysis is limited, and performance is likely linked to the correlations between the continuous and categorical variables (Johnson & Wichern 2002). Finally, discriminant analysis may be ineffective if the population mean vectors are not significantly different. Assuming the data are approximately multivariate normal, the difference between the populations associated with the groups of interest can be tested via multivariate analysis of variance (MANOVA) prior to discriminant analysis (Johnson & Wichern 2002).
In cases where multivariate normality cannot be established, non-parametric discriminant analysis is an alternative. The primary methods are the nearest neighbor and kernel methods. Both methods substitute non-parametric estimates of population density functions rather than assuming the populations are multivariate normal (Khattree & Naik 2000; McLachlan 1992). A \( k \)-nearest neighbor approach assigns observations to a class based on its closest neighbors, determined from attribute vector distances. After computing the distance between an observation and all other data points, the \( k \) closest observations can be chosen, and based on these the posterior probability of belonging in either group 1 or 2 can be determined (Khattree & Naik 2000). The \( k \)-nearest neighbor approach can be an excellent choice when the statistical surface separating classes is complex, but it can be sensitive to data scaling (Mazzatorta et al. 2004). The value for \( k \) may be somewhat artificial: Experiments have demonstrated that increasing this value appears to have little or no effect on performance (Bressan & Vitria 2003). Kernel methods, which can employ a variety of different kernel shapes, estimate the group-specific probability densities for each observation. These density functions are shaped by the other observations in the training data and a radius-based smoothing parameter provided by the user (Khattree & Naik 2000). Notably, a non-parametric discriminant analysis approach offers no improvement over linear or quadratic methods if the input sample is multivariate normal (Khattree & Naik 2000).

2.4.2 Logistic Regression

Logistic regression is an extremely flexible method for handling both continuous and categorical data. Historically, discriminant analysis was the primary method used for statistical classification, but logistic regression has become common in the health sciences
and many other disciplines (Afifi et al. 2004). Logistic regression is commonly used to perform analysis on a response variable with two qualitative outcomes, e.g., infected or not infected, though it can be extended to three or more outcomes as well (Neter et al. 1996). Unlike discriminant analysis, it does not require multivariate normality or equal covariances (Afifi et al. 2004). It does require knowledge about the explanatory and response variables. This allows the results to be used in future classification efforts when the response variable is unknown (Afifi et al. 2004).

A data set where each observation is labeled, for example, “infected” and “not infected” can be thought of as (0,1) binary. When a response variable is binary, the shape of the response function is typically curvilinear (Neter et al. 1996). In addition, the mean response is probabilistic (Neter et al. 1996). The logistic response function—which is sigmoidal in shape—has the form

\[
P_z = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}
\]

where \(Z\) is the function \(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\), a linear combination of the explanatory variables (Afifi et al. 2004). This logistic regression equation can be linearized as the logit transformation:

\[
\ln(\text{odds}) = \ln \left( \frac{P_z}{1 - P_z} \right)
\]

where the ratio \(P_z/(1-P_z)\) is termed the odds ratio (Neter et al. 1996). In words, the odds ratio is the ratio of the probability of an event occurring to the probability of the event not occurring (Jovanovic et al. 2002). In any case, the formula above leads to the logit response function:
\[
\ln(\text{odds}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p
\]

which assumes that the logarithm of the odds is linear in the explanatory variables, and can
be expressed in a form essentially the same as a multiple linear regression equation (Afifi et
al. 2004; Neter et al. 1996). In summary, because the response variable is discrete, it cannot
be modeled directly by linear regression. To accomplish this, instead of predicting whether
an event itself will occur, the model is built to predict the logarithm of the odds of its
occurrence (Jovanovic et al. 2002).

Logistic regression models the probability that the dependent variable is 1, which
corresponds to one of the classes of interest (Afifi et al. 2004). (Instead, some software
packages model the probability that the dependent variable is 0, the other class of interest.)
Once the coefficients of the equation have been fitted, it is possible to calculate the class
probabilities for any observation. A threshold value (e.g., 0.5) can be used to divide
observations into one class or the other. By comparing to the known class, it is possible to
determine misclassification rates in a manner essentially the same as discriminant analysis.

The logistic regression model can be fitted (and coefficients determined) through a
few common methods, including ordinary and weighted least squares (Allison 1999).
Maximum likelihood is quite popular, for a number of reasons. First, it is the only method in
general use for individual-level data, i.e., ordinary and weighted least squares require
grouping of data (Allison 1999). Second, maximum likelihood estimators are approximately
unbiased and asymptotically normal for large samples, meaning that confidence intervals and
p-values can be computed. Finally, maximum likelihood handles categorical variables quite
good (Allison 1999).
Maximum likelihood proceeds by approximating coefficients through repeated iterations. When the change in coefficients between approximations becomes very small, maximum likelihood is said to have converged on a solution (Allison 1999). This is rarely a problem for large samples. For relatively small samples including categorical variables with many classes, maximum likelihood may fail due to a condition of quasi-complete separation (Allison 1999). Typically, this means that, at one level of a categorical variable, either every observation is classified as 0 or every observation is classified as 1. This can easily happen if the observations of a small input sample are sparsely distributed across the many levels of the categorical variables. Possible solutions are to increase the sample size (not always a realistic option), convert the problematic variable to an ordinal or quantitative variable, reduce the number of levels for the variable by combining classes, or remove it from the analysis (Allison 1999).

As this suggests, logistic regression is not perfectly flexible with small samples. Sample size also bears on the number of variables that can be realistically included: Peduzzi et al. (1996) suggested no fewer than ten observations per model parameter. Logistic regression can be sensitive to sample bias, outliers, and collinearity between included variables (Afifi et al. 2004). As with linear regression, there can be meaningful interactions between variables. The importance of an interaction term to model fit can be evaluated by methods such as the Wald test (Afifi et al. 2004). Interactions make odds ratios and model coefficients significantly more difficult to interpret (Afifi et al. 2004).

The overall assumption that the response is linear in the coefficients of the explanatory variables should be checked using goodness-of-fit measures (Afifi et al. 2004; Jovanovic et al. 2002). There are a number of goodness-of-fit statistics (e.g., the Hosmer-
Lemeshow test) that basically amount to chi-square tests of observed versus expected outcomes for the observations, though they are most reliable with large samples (Afifi et al. 2004). In addition, several different methods have been put forward for computing the coefficient of determination, $R^2$, any of which can be used to compare the information content of fitted models (Afifi et al. 2004; Shtatland et al. 2000).

2.4.3 Variable Selection in Discriminant Analysis and Logistic Regression

Variable selection can be a difficult process. Doing so in a way that properly denotes the relationship between explanatory and response variables requires skill and experience (Jovanovic et al. 2002). A preliminary reduction of variables may be relatively straightforward. For example, initial screening can be performed using univariate tests of equal group means for continuous variables, allowing for a large significance level so that useful predictors are not missed (Afifi et al. 2004). However, while this usually reduces the number of variables to less than ten (Afifi et al. 2004), additional screening methods may be necessary to further simplify the model.

SAS and other statistical software packages contain processes for stepwise discriminant analysis, enabling variable selection. While stepwise discriminant analysis can be useful when many variables are employed in the classification, it has a number of potential problems. In particular, there is no guarantee that the process will yield the best subset, and the subset may perform poorly with future samples. Selection is especially problematic when there are large correlations among the variables (Johnson & Wichern 2002). On the other hand, the addition of many variables does not necessarily increase, and may in fact decrease, the classification success (Khattree & Naik 1999). Johnson & Wichern
(2002) recommended at least one validation sample to test a reduced variable set, but further suggested it might be wiser to split the sample into a number of batches and then determine the “best” variable subset for each batch, such that the number of times a particular variable appears in a subset is a good indication of its worth.

Most of the above points also apply for logistic regression. Stepwise selection methods for logistic regression have been criticized because they often result in models that have too few variables for successful prediction (Shtatland et al. 2001, 2003). On the other hand, Pearce and Ferrier (2000) used logistic regression models to predict the distributions of several species in New South Wales, Australia, and found that employing stepwise selection with strict confidence limits—to eliminate any extraneous variables—maximized the models’ predictive power. There are alternative ways of selecting the best logistic regression models, including the Akaike and Schwarz information criteria, although all should be used with caution (Afifi et al. 2004, Shtatland et al. 2001, 2003).

2.4.4 Decision Trees

The theoretical aspects of decision trees were presented in Section 2.3, but a few points are worth repeating here, at least for the sake of comparison to discriminant analysis and logistic regression. In short, decision trees may be preferred for some classification and prediction efforts due to their low number of assumptions. Decision trees are non-parametric, and so are not tied to a normal or other population distribution, nor are they tied to a particular optimization criterion (Johnson & Wichern 2002; Murthy 1998). Like logistic regression, decision trees can accommodate many variable types, including both categorical and continuous variables. The clearly defined rules of a decision tree are generally more
straightforward to interpret than the coefficients of a discriminant function or logistic
equation (i.e., the cases where a selected variable is important in classification is spelled out
explicitly). However, decision trees often require large training samples—hundreds of
observations and many variables—for accurate classification (Johnson & Wichern 2002).
While small samples can be successfully classified, there is often substantial variance, and as
a result trees with different variable sets and threshold values can perform equally well
(Murthy 1998). This, in turn, makes it difficult to identify the tree most conducive to
generalization. On other hand, large training samples can lead to large, intricate decision
trees even when fundamental relationships among the variables are quite simple (Jovanovic
et al. 2002). To avoid over-fitting, the resulting trees must be pruned based on some
subjective judgment (Nelson et al. 2003).

2.4.5 Choice of Method

Studies comparing the above-described classification techniques have yielded
equivocal results in terms of performance. Murthy (1998) detailed several studies from
medical disciplines comparing logistic regression, discriminant analysis, decision trees, and
other classification methods, and suggested no one method appeared to be consistently better
than the others in terms of misclassification rates. Mazzatorta et al. (2004) predicted the
ecotoxicity of 235 pesticides based on a large set of physical, chemical, and structural
attributes. The authors built their models using several different classification methods,
including linear and quadratic discriminant analysis, decision trees, and k-nearest neighbor
classification (essentially non-parametric discriminant analysis). All of the methods
performed well, but each had distinct disadvantages (e.g., the decision trees tended to over-
fit; the $k$-nearest neighbor approach was sensitive to data scaling). Similarly, Worth and Cronin (2003) used models developed via logistic regression, discriminant analysis, and a classification tree approach to predict whether 139 chemicals are eye irritants based on a single variable (molecular weight). The classification tree model was the most accurate (97% correctly classified as irritants, versus 75% for the discriminant model and 27% for the logistic model), but with a much higher percentage of false positives (51%, versus 38% for the discriminant model and 7% for the logistic model). Reichard and Hamilton (1997) used both discriminant analysis and decision trees to classify woody plant species introduced to North America prior to 1930 as either invaders or non-invaders based on structural, biogeographical, and life history attributes. Discriminant analysis predicted 97% of invaders correctly, but only 70% of non-invaders. While the decision tree correctly predicted only 73% of invaders, it captured more than 83% of non-invaders.

Certainly, the choice of classification method may depend on the nature of the input data—in particular, the distribution or size of the input sample. Press and Wilson (1978) tested logistic regression and discriminant analysis using empirical data, and found that logistic regression estimators outperformed discriminant analysis in cases of non-normality, but not by a large amount. They further suggested the two methods would rarely yield substantially different results. In cases of normal populations with identical covariance matrices, discriminant analysis is preferred, particularly due to its greater efficiency. However, noting the difficulties in achieving multivariate normality, Press and Wilson suggested logistic regression is often more practical. Moreover, if the assumptions of binary logistic regression or discriminant analysis cannot be justified from the data, a decision tree or non-parametric discriminant analysis approach can be successful (Afifi et al. 2004). As
already noted, logistic regression and decision trees can be sensitive to sample size, although there is no strict threshold on what qualifies as too small a sample (Afifi et al. 2004; Murthy 1998; Pearce & Ferrier 2000).

The purpose of the analysis may also be a factor. While decision trees may provide the highest classification accuracy for the class of interest, they may also result in high rates of false positives (e.g., Worth & Cronin 2003), and this inefficiency may be problematic. Logistic regression and discriminant analysis provide group-membership probabilities for individual observations, but decision trees do not yield such individualized information (Worth & Cronin 2003). Perhaps the best way to assess whether a model developed with a particular classification technique is the best for a particular data set is to try all of the techniques and compare their success. This includes a full analysis of the assumptions (i.e., whether any were violated) and a comparison of classification rates. Manel et al. (2001) suggested the use of Cohen’s kappa statistic as an easy-to-compute method for comparing the classification accuracies of different models.

2.4.6 Examples

Studies from a diverse assortment of disciplines have relevance for our landscape-scale analysis of the hemlock woolly adelgid. Gumpertz et al. (1999) used a marginal logistic regression model, adjusted for temporal and spatial autocorrelation, to predict the likelihood of southern pine beetle (*Dendroctonus frontalis* Zimm.) outbreaks in North Carolina, South Carolina, and Georgia counties. They developed their model using 31 years of outbreak presence or absence. Probability of outbreak was described via several explanatory variables: elevation, longitude, host availability, climate (seasonal precipitation
and temperatures), and other physiographic variables. The amount of fall precipitation was
the single best predictor in the fitted model, with outbreak probability increasing with higher
fall precipitation. Areas with dry summers also tended to have higher outbreak probability.

Insects can serve as vectors for human diseases as well. Beck et al. (1994) applied a
landscape approach to identify villages at high risk of malaria transmission in southern
Chiapas, Mexico. The researchers processed Landsat TM images to delineate different land
cover classes in the areas surrounding 40 villages in the region where they had also recorded
mosquito (Anopheles albimanus) abundance. They used stepwise discriminant analysis and
stepwise linear regression to investigate the relationship between land cover and mosquito
abundance. The proportions of transitional swamp and unmanaged pasture were the most
important variables. Discriminant functions generated for these two variables distinguished
between high-abundance and low-abundance villages with 90% accuracy. In a follow-up
study, Beck et al. (1997) applied these models in another part of southern Chiapas. The
discriminant model predicted 79% of the high-risk and 50% of the low risk villages, for an
overall accuracy of 70%, suggesting that a simple landscape model can be generalized to
another region successfully, and that GIS and remotely sensed data can be employed to map
risk explicitly.

Like insect pests, forest pathogens require a landscape perspective, because pathogens
propagate at regional scales and “according to heterogeneous spatial patterns of flow and
isolation” (Holdenreider et al. 2004, p.446). For instance, Wilson et al. (2003) used a GIS to
predict infection sites for cinnamon fungus (Phytophthora cinnamomi) in southeastern
Australia. At 50 sites, they performed infection tests on plants and soils and recorded
numerous variables, including elevation, aspect, slope, proximity to roads, and soil and
vegetation characteristics. They used logistic regression models to predict the probability of each site being infected by cinnamon fungus, and compared their prediction results with the test results from the site assays. The only variables identified as significant by backwards stepwise selection were elevation (negatively associated) and solar index (positively associated with infection). Their logistic regression model was 97% accurate at predicting sites with cinnamon fungus but far less accurate at predicting sites where it was not present (58%).

Similarly, Kelly and Meentemeyer (2002) generated decision tree models for assessing oak tree mortality due to sudden oak death (*Phytophthora ramorum*) in Marin County, California. A number of spatial variables emerged as important in the spread of the pathogen: density of dead tree crowns, density of host shrub species, elevation, topographic moisture index, summer solar radiation, distance to forest edge, and distance to trails. The best decision tree explained 63% of the variability in the location of oak mortality for the training data, and 58% of the variability for a separate test data set.

Several invasive plant studies suggest general techniques and variables that may also be meaningful in assessing the spread risk of insect pests at the landscape level. Dirnbock et al. (2003) examined invasive plants on the Juan Fernandez Archipelago, a national park and biosphere preserve in Chile. Using data spanning several decades, the authors constructed a history of vegetation change in the park and identified the most significant invasive plant species. They used logistic regression models with a number of landscape variables (elevation, slope position, topographic similarity index, solar radiation) to predict the likely distributions of the worst invaders, as well as the distributions of some native species. All of the modeled species except one were strongly controlled by the landscape variables, although
models for the natives predicted more reliably, suggesting these species had narrower environmental constraints.

Rouget et al. (2004) explored the ability of models to predict the spread of four invasive *Pinus* species in southern Africa. They looked at a variety of explanatory variables, including altitude, slope, soil pH, vegetation type, height, and density, proximity to watercourses, and vegetation age post-fire. They used two different statistical techniques to generate models: logistic regression and decision trees. They found that the decision trees made more accurate predictions: 65% of the current pine distribution was correctly predicted by the distribution of the first *Pinus* trees that invaded.

Deckers et al. (2005) examined landscape scale factors affecting the spread of black cherry (*Prunus serotina*) in agricultural landscapes in the Flanders region of Belgium, where the tree species is considered invasive. Their 251-ha study area contained 2962 black cherry trees, largely distributed in a hedgerow network. Through spatial point pattern analysis and logistic regression, they discovered that the trees were clustered near hedgerow intersections, seed sources, and roosting trees, suggesting that long-distance dispersal by birds was a substantial factor in black cherry invasion, in addition to local seed dispersal.

As is true of invasives, wildland fire risk is a regional-scale issue that can involve a large number of spatially structured factors affecting the rate and pattern of spread. Rollins et al. (2004) set out to design a system permitting repeatable generation of spatial data for assessing fire risk. They recorded 38 physiographic, climatic, and environmental variables in a GIS from remote sensing and other sources. Important variables included topographic factors such as elevation, slope, and curvature. Using logistic regression, discriminant analysis, and decision trees, they developed models by which they could generate GIS maps
of fuel load, fuel model type, fire interval, and fire severity. Accuracies of maps ranged from 51 to 80%, with the fire interval map being the most accurate and fuel load the least. The authors noted that their system’s performance depended heavily on the input data, but noted that could be adapted if better inputs become available.

Non-invasive species can also be relevant examples. Maehr and Cox (1995) used GIS to analyze twenty geographic variables potentially influencing the distribution of Florida panthers (Felis concolor coryi) in the southern part of the state. They compared radio telemetry points marking the movements of 23 panthers with a large number of randomly generated points in the region. Variables included proximity to different land cover types and roads, as well as the area of preferred cover types. The authors used regression analysis and discriminant analysis to classify the panther locations and random points and to generate GIS maps predicting panther distribution. Four variables (preferred patch size, proximity to preferred habitat type, diversity of preferred habitats, and the background matrix type) were effective in predicting > 80% of the points, and the GIS maps identified large areas of suitable land cover on private ranches in south Florida.

2.4.7 Summary

While spread of the hemlock woolly adelgid may be difficult to simulate, it is clear that landscape-level factors influence how the pest is distributed across a region. A number of statistical classification methods—parametric and non-parametric discriminant analysis, logistic regression, and decision trees—can be used to distinguish areas at high risk from those at low risk of hemlock woolly adelgid infestation. All of these techniques use a multivariate training sample to construct a classification function that can in turn be applied
to new data sets. However, each has a different set of simplifying assumptions. Comparative examples from the literature appear to confirm the assertion that none of the outlined techniques is definitively better than the others. For our analysis, therefore, trying all of them and comparing their results seems like a logical solution. With respect to the variables important for classification, topographic attributes appear to be significant quite often in examples from the literature, perhaps because they delineate ideal and/or suitable habitat for pest or host. Proximity variables (to streams, roads, trails, etc.) can also be relevant, depending on the vectors of spread for a pest and the importance of connectivity. A number of other factors (e.g., site disturbance history) could be important, but unfortunately such information may not be available at the appropriate scale, or at all. For our analysis of hemlock woolly adelgid, we will focus on readily available or easily calculable data sets. This will make our approach more likely to be generalized and used by the forest managers who are currently managing the threat on the ground.

2.5 References


Abstract: A problematic feature of remotely sensed imagery of mountainous environments is variable illumination due to topographic relief, which can result in image information loss and diminished accuracies of derived products. The C-correction belongs to a family of statistical-empirical equations for topographic normalization that employ solar properties and digital elevation data to model the effect of surface orientation on illumination. We applied the C-correction to a Landsat Enhanced Thematic Mapper (ETM+) image and an Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) image of Great Smoky Mountains National Park. After geometric correction, we partitioned each image by aspect into sunlit (less corrected) and sunshade (more corrected) subsets, and applied the C-correction separately to each subset. We employed a 10-m digital elevation model of the park and solar parameters from the date of image capture, and determined C-correction equation coefficients for each subset via linear regressions based on 10,000-pixel samples. To test normalization success, we examined changes in actual pixel values and per-band standard deviation values. We assessed the relative thematic accuracy of the image subsets based on derived vegetation maps: a two-class (evergreen, non-evergreen) map for the Landsat image, and a three-class map (hemlock, non-hemlock evergreen, non-evergreen) for the ASTER image. We used orthophotos as reference for the Landsat-derived map and a combination of field data and aerial photos as reference for the ASTER-derived map. Generally, the C-correction performed well, with both images displaying overall decreases in standard deviation between 3 and 33%, and similar reductions within different vegetation classes. The sunshade subsets, specifically, exhibited increases in standard deviation for certain bands, as well as substantial increases in the actual values of isolated pixels; this is likely because poorly illuminated areas in these subsets required major correction. For each image and its associated map, the overall thematic accuracy of the sunlit subset was higher than for the sunshade subset: 90% vs. 81% for Landsat, 93% vs. 88% for ASTER. Although this suggests the topographic effect was not completely removed by normalization, all accuracies met reasonable standards for remote-sensing-derived maps. Notably, aspect partitioning is recommended prior to normalization to enable targeted correction of the differently illuminated portions of an image.

Key Words: C-correction; topographic normalization; remote sensing; ASTER; Landsat

3.1 Introduction

A major challenge when using remotely sensed data of mountainous regions is the highly variable level of ground illumination due to topographic relief. Shadows caused by topographic slope and aspect result in loss and alteration of image information, such that pixels representing objects with nearly identical spectral reflectance properties—e.g., two
homogeneous stands of the same forest type—may end up with very different values simply due to the orientation of the terrain underlying those objects (Colby & Keating 1998; Jensen 1996). Naturally, the situation may also be reversed: Two pixels of quite different phenomena may have very similar spectral values primarily due to topographic effects. Numerous researchers have demonstrated that topographic effects on satellite imagery greatly diminish classification accuracy (e.g., Colby 1991; Ekstrand 1996; Hale & Rock 2003; Justice et al. 1981). Not surprisingly, during the last several decades, a number of theoretical approaches have been proposed for minimizing or removing these topographic effects (Richter 1997). The most successful methods typically involve direct modeling of the Earth’s surface via elevation data (Riano et al. 2003; Richter 1997). In this paper, we present the application of an elevation-based normalization method, the C-correction, to imagery from two different moderate-resolution platforms: the Landsat Enhanced Thematic Mapper (ETM+) and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), one of several sensors on the NASA Earth Observing System Terra satellite (ERSDAC 2003a). The images cover the North Carolina portion of Great Smoky Mountains National Park. We selected this study area as the setting for development of a protocol for mapping hemlock (Tsuga spp.) forest distribution. This mapping protocol is intended for general application throughout the southern Appalachian region and elsewhere, and we wanted to evaluate whether the C-correction would be a useful step in this protocol. Notably, the literature on C-correction and similar methods has focused largely on Landsat or SPOT (Satellites Pour l’Observation de la Terre) imagery, so this appears to be one of few, if any, tests of the response of ASTER imagery to the C-correction (Jensen 1996).
Previous analyses (e.g., Hale & Rock 2003; McDonald et al. 2000; Riano et al. 2003) have employed field and image measurements to compare the properties of various topographic normalization models, noting particular situations when models tended to over-correct or under-correct pixel values. While these studies informed our approach, we were ultimately more concerned with the map products derived from the normalized images than the properties of the actual images. Previous research has suggested that topographically normalized images will result in better forest classification maps than uncorrected images (Civco 1989; Jensen 1996; Meyer et al. 1993; and others). Subsequently, while we examined changes in image spectral heterogeneity and pixel values to assess the effectiveness of the C-correction, we also focused on the thematic accuracy of derived maps.

Topography does not exert a uniform influence across a satellite image (Jensen 1996). Acknowledging this, we performed the our analyses using subsets of the images: “sunlit” subsets, covering areas where topographic normalization had a minimal effect on the input image, as well as more heavily corrected “sunshade” subsets, corresponding to the areas of deepest shadow (Hale & Rock 2003). The “sunshade” subsets were of particular interest, because some of these heavily shadowed areas are likely to exhibit hemlock dominance (Delcourt & Delcourt 2000; Godman & Lancaster 1990; Quimby 1996). We believe that use of the subsets enhanced our correction efforts and—because we were able to examine their results separately—yielded a more comprehensive overall picture of how the C-correction performs in some of the most rugged topography in the southern Appalachian region. In turn, we were able to achieve our primary objective of predicting the efficacy of including the method in our hemlock mapping protocol.
3.2 Background

The concept of *illumination* is fundamental to the discussion of image topographic correction. In terms of sun-terrain-sensor geometry, illumination can be basically represented as the cosine of the incidence angle, representing the proportion of direct solar radiation that strikes the ground area captured by an image pixel. The incidence angle, $i$, is the angle between the surface normal for a particular pixel and the solar zenith direction (Figure 3-1), where the surface normal is a vector perpendicular to the pixel (Jensen 1996; Smith et al. 1980). One can also distinguish between *direct* and *diffuse* irradiance. The amount of irradiance that directly reaches a pixel on the ground is (directly) proportional to the cosine of the incidence angle (Ekstrand 1996). Most topographic normalization techniques focus only on direct irradiance in terms of the incidence angle, ignoring the impact of diffuse sources, such as light reflected from surrounding mountainsides.

![Figure 3-1. Solar incidence angle ($i$) and surface normal for a location on inclined terrain (based on Dymond & Shepherd 1999).](image)

Topographic normalization techniques can be divided into two broad categories: those based on band ratios and those that employ digital elevation data (Riano et al. 2003). Band ratio approaches assume that image reflectance will increase or decrease proportionally
between two ratio bands, thus compensating for topographic effects. A significant problem with this assumption is that the relative contribution of diffuse irradiance to pixel values varies between spectral bands, such that direct irradiance is not truly proportional. Band ratios also result in a loss of spectral resolution that might be important in classification (Colby 1991; Justice et al. 1981; Riano et al. 2003).

Another family of corrections uses a digital elevation model (DEM) to calculate the cosine of the incidence angle $i$ (i.e., direct irradiance) according to the formula:

$$ \cos i = \cos \theta_z \cos \theta_n + \sin \theta_z \sin \theta_n \cos (\phi_s - \phi_n) $$

where $\theta_z$ is the solar zenith angle, $\phi_s$ is the solar azimuth, $\theta_n$ is the surface slope, and $\phi_n$ is the surface aspect (Hale & Rock 2003). DEM-based correction methods are further divided into two categories: Lambertian and non-Lambertian. Lambertian methods assume that reflectance is independent of observation and reflectance angles, and thus equally bright from all directions. In other words, such methods ignore the typically variable illumination geometry of rugged, real-world surfaces in favor of the simplifying assumption of an isotropic reflectance law (Riano et al. 2003; Richter 1997).

Most currently used topographic correction methods accept that surfaces exhibit non-Lambertian behavior, especially vegetative surfaces (Li et al. 1998). A common non-Lambertian approach is based on the Minnaert constant (Minnaert 1941). Teillet et al. (1982) introduced the Minnaert constant to modify the Lambertian cosine correction:

$$ \rho_H = \rho_T \left( \frac{\cos \theta_z}{\cos i} \right)^k $$

where $\rho_H$ is the horizontal surface radiance (i.e., the terrain-corrected image data), $\rho_T$ is the inclined surface radiance (i.e., the uncorrected data), and $k$ is the Minnaert constant. Minnaert constant values must be calculated for each image band separately and,
furthermore, may be dependent on cover type (Ekstrand 1996; Teillet et al. 1982).

Estimating cover-type-specific Minnaert constants requires *a priori* knowledge of the vegetation that covers a study area; unfortunately, this is usually the information that is being sought by the classification that will occur after normalization, and so is not known beforehand without field sampling or other prior source of vegetation information (Hale & Rock 2003). Furthermore, extraction of the necessary values from field data can be problematic unless the sample is large, unbiased, and free of outliers (Tokola et al. 2001).

Statistical-empirical correction methods offer alternatives to the Minnaert constant. In particular, Teillet et al. (1982) introduced a modified statistical-empirical approach named the C-correction, computed according to the formula:

\[
\rho_H = \rho_T \left( \frac{\cos \theta_c + c_k}{\cos i + c_k} \right)
\]

where \( c_k = b_k / m_k \), for \( \rho_T = b_k + m_k \cos i \). This model introduces a parameter, \( c_k \), that is the quotient between the slope (\( b_k \)) and intercept (\( m_k \)) of the regression equation \( \rho_T \) versus \( \cos i \), i.e., the regression of pixel spectral values versus DEM-derived illumination values (Riano et al. 2003). The C-correction is relatively easy to implement and has performed well when compared to other normalization techniques. Riano et al. (2003) found that C-correction of a Landsat Thematic Mapper (TM) image from southwestern Spain performed substantially better than the Minnaert constant and Lambertian cosine correction methods at reducing pixel value heterogeneity caused by topography. Similarly, McDonald et al. (2000) compared several different normalization methods—including the cosine correction, the Minnaert correction, and the C-correction—using Landsat TM imagery of New South Wales, Australia. The Minnaert and C-correction approaches both performed well statistically and visually, while the others resulted in substantial under- or over-correction. The authors
recommended the C-correction over the Minnaert approach primarily because model parameters are more easily obtained from the data.

Admittedly, the C-correction employs simplifying assumptions. Alternative methods have been proposed that model fine-scale (i.e., forest canopy) geometry instead of coarse terrain geometry (Gu & Gillespie 1998; Borel et al. 1991), use a multi-scale approach (Li et al. 1998), or that do not depend on a photometric function at all (Vincini & Frazzi 2003). These methods have not been tested extensively, although the C-correction performed well versus alternative methods in at least one study (Vincini & Frazzi 2003). Furthermore, while the C-correction is scene-dependent (i.e., separate correction coefficients must be determined for each image that is processed), it is easy to calculate based on readily available digital elevation data and simple regressions using image pixel values. This makes it a strong choice for application in mapping efforts—such as ours—that will ultimately employ numerous images over a study region, and thus must prioritize efficiency and ease of use.

3.3 Study Area

We performed our analysis using satellite imagery from the eastern portion of Great Smoky Mountains National Park (Figure 3-1). This area contains some of the most rugged topography in the southern Appalachians, although rounded, forested summits are more typical than rock outcrops or similar examples of extreme relief (Wiser et al. 1996). Elevation in the study area ranges from ~650 to ~2000 m and slopes from 0 to 56 degrees, with many minor peaks and valleys bisecting the region. A number of forest types are found in the study area, included high-elevation spruce-fir (*Picea rubens*-*Abies fraseri*) forest, cove
hardwoods, mixed mesophytic forest, and forests dominated by eastern hemlock (*Tsuga canadensis*).

![Figure 3-2. Great Smoky Mountains National Park, with a close-up of the study area (outlined in red).](image)

### 3.4 Methods

We used ERDAS Imagine 8.7 software for image processing and SAS Version 9 software for statistical analyses (Leica Geosystems 2001; SAS Institute 2003).

#### 3.4.1 Image Acquisition

We acquired a geometrically corrected October 2001 leaf-off Landsat ETM+ image (Path 18, Row 35) from the Global Land Cover Facility at the University of Maryland ([http://glcf.umiacs.umd.edu/index.shtml](http://glcf.umiacs.umd.edu/index.shtml)). The Global Land Cover Facility is an archive of satellite imagery from a variety of platforms, often used in previous projects but now freely downloadable through the World Wide Web (GLCF 2005). We also acquired a leaf-on, early
September 2000 ASTER scene that covered the study area. We chose to use an ASTER image because of the sensor’s improved spatial (15-m in the visible and near infrared versus 28.5 or 30-m for Landsat) resolution, as well as its negligible cost: Archived reflectance and radiance ASTER scenes can be freely acquired through the NASA Earth Observing System Data Gateway (http://edcimswww.cr.usgs.gov/pub/imswelcome/), while raw scenes may be acquired for a nominal fee (ERSDAC 2003a; 2003b). ASTER is actually a combination of three different radiometers that yield output image data with differing spatial resolutions: a 15-m resolution, 3-band radiance image in the visible and near infrared, a 30-m, 6-band radiance image in the short-wave infrared, and a 90-m, 5-band image in the thermal infrared range (ERSDAC 2003a). During import from their native hierarchical data format (HDF) to ERDAS Imagine format, the visible and near infrared and short-wave infrared images were automatically subjected to a preliminary geometric correction using a third-order polynomial equation based on parameter values from an associated metadata file. Ignoring the thermal infrared image, we merged the other two images into a single 15-m image, sidestepping any resampling of the short-wave infrared values by simply dividing the image’s pixels into 4 sub-pixels of equal value.

3.4.2 Image Pre-processing

To facilitate their combined use in our later mapping effort, we applied data fusion to the Landsat ETM+ image in order to bring its spatial resolution close to that of the ASTER image. We fused six bands of the 28.5-m multispectral Landsat image (excluding the thermal band) with the matching 14.25-m panchromatic image via an algorithm that utilizes correspondence analysis to fuse data with minimal alteration of pixel values (Cakir 2003).
To correct a moderate level of noise in the panchromatic image—noise that would otherwise be carried through to the fused image—we processed the image with a 3×3 low pass filter (Figure 3-2) prior to fusion (Jensen 1996).

![Figure 3-3. Detail of the Landsat panchromatic image prior to (left) and after (right) low pass filtering.](image)

Although both images had received prior geometric correction, we performed additional correction to better register them to each other and improve their joint application in our mapping effort (Jensen 1996). We corrected the fused Landsat image using a third-order polynomial equation and 92 ground control points (GCPs) collected from a color-infrared digital orthophoto quarter quad (DOQQ) mosaic of the study area. Root mean squared error (RMSE) for the Landsat image was less than half a pixel (4.1420 m). We refined the ASTER image using a fourth-order polynomial equation and 80 DOQQ-derived GCPs. The resulting RMSE was less than half a pixel (6.4694 m).

### 3.4.3 Topographic Normalization

Typically, topographic normalization equations are designed to use radiance rather than raw digital number values (Teillet et al. 1982; Jensen 1996). The ASTER data were already converted to atmospherically corrected radiance prior to downloading from the
NASA Earth Observing System Data Gateway (ERSDAC 2003a). For the Landsat image, we converted the initial digital number (DN) values to radiance via the formula:

\[ L = \left[ \frac{(LMAX - LMIN)}{255} \right] DN + LMIN \]

where LMIN and LMAX represent the dynamic range of each Landsat channel/band (Lillesand & Kiefer 2000). A calibration parameter file uniquely associated with the Landsat image provided LMIN and LMAX values for each band.

Staff from Great Smoky Mountains National Park supplied a 10-meter DEM of the study area in grid format. We applied two iterations of a 3×3 low pass filter to the DEM, a technique that has been demonstrated to minimize systematic errors in DEMs and their resulting effects on the topographic normalization process (Hale & Rock 2003). We resampled the DEM to 15 m so that it was similar spatial resolution to our images and because inadequate DEM resolution can influence the correction process (Jensen 1996; Richter 1998). We clipped the DEM to the extent of our study region, and then generated slope and aspect images for use in calculating the C-correction parameters.

Prior to normalization, we also clipped the Landsat and ASTER images to the study region extent, and then divided each into two subset images via aspect partitioning. Aspect partitioning stratifies an image into one subset with aspect values within or equal to 180 degrees of the solar azimuth—the “sunlit” image—and a second subset with aspect values greater than 180 degrees from the solar azimuth—the “sunshade” image (Hale & Rock 2003). Aspect partitioning has some potential for classification improvement because normalization functions are developed and applied separately for these very distinct subsets (Hale & Rock 2003). For each image, we partitioned the aspect image generated from the DEM according to the corresponding solar azimuth value: 155.156 degrees for the Landsat
image and 164.676 degrees for the ASTER image, derived from parameter files distributed with the raw imagery. This yielded “sunshade” and “sunlit” masks that we applied to the two images to generate appropriate subsets. The two subsets of the ASTER image are shown in Figure 3-3 as an illustration of partitioning.

![Figure 3-4. Aspect partitioned subsets for the ASTER image.](image)

### Table 3-1. Parameter estimates and C-correction coefficients for the Landsat image subsets.

<table>
<thead>
<tr>
<th>Band</th>
<th>Sunlit Subset</th>
<th>Sunshade Subset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b_k$ (intercept)</td>
<td>$m_k$ (slope)</td>
</tr>
<tr>
<td>1</td>
<td>30.31296</td>
<td>6.42533</td>
</tr>
<tr>
<td>2</td>
<td>17.23457</td>
<td>11.66936</td>
</tr>
<tr>
<td>3</td>
<td>8.57892</td>
<td>15.65514</td>
</tr>
<tr>
<td>4</td>
<td>9.43453</td>
<td>48.0018</td>
</tr>
<tr>
<td>5</td>
<td>0.69694</td>
<td>9.44858</td>
</tr>
<tr>
<td>7</td>
<td>0.24613</td>
<td>1.62931</td>
</tr>
</tbody>
</table>
To determine the necessary set of $c_k$ coefficients for the two images, we generated random samples of $(x, y)$ coordinate points to extract radiance values from each band of the images—six values for Landsat, nine values for ASTER—as well as corresponding $\theta_n$ and $\phi_n$ values from the DEM-derived slope and aspect images. We extracted separate 10,000-point samples (using random-point-generation and pixel-to-ASCII utilities in ERDAS Imagine) from the sunlit and sunshade subsets. No pixel was sampled more than once. We removed a
small number of points (~100) from each of these samples because the slope and/or aspect values were zero; these points were outliers with a disproportionate influence on the regression parameter estimates. We executed a regression of \( \cos i \) on the sampled radiance values. We calculated the \( \cos i \) terms for the sample points using the equation:

\[
\cos i = \cos (90 - \theta_s) \cos \theta_n + \sin (90 - \theta_s) \sin \theta_n \cos (\varphi_s - \varphi_n).
\]

In this case, \((90 - \theta_s)\), where \(\theta_s\) is the solar elevation value, is substituted for the solar zenith angle \(\theta_z\) in the original \(\cos i\) equation (Hale & Rock 2003). As with the solar azimuth values \((\varphi_s)\), solar elevation values for each image—37.846 degrees for Landsat, 50.986 degrees for the ASTER image—were available from its corresponding parameter file. We performed separate regressions of \( \cos i \) versus radiance for each band, yielding slope \((b_k)\) and intercept \((m_k)\) estimates that we used to calculate \(c_k\) for each (Tables 3-1 and 3-2).

With the \(c_k\) values determined, we constructed an ERDAS Imagine model that applies the \(C\)-correction to each pixel in a band-by-band fashion (Figure 3-4). The model features are straightforward: (1) the illumination angle is determined based on the solar elevation and azimuth, as well as the slope and aspect values derived from a DEM of the image area; (2) the illumination angle is used to determine \( \cos i \); and (3) the solar elevation, \( \cos i \), and per-band coefficients are applied in the \(C\)-correction equation outlined by Teillet et al. (1982).

After performing the \(C\)-correction, we merged the separate sunlit and sunshade subsets for each image into a single file. The normalization process failed to correct a small number of pixels in the Landsat image—13,608 out of 2,386,304 pixels (0.57%)—because the illumination angle was particularly low in some areas. We ignored the non-normalized pixels until after performing image classification (described below), when we used the color-infrared DOQQ mosaic to label these remaining pixels as either evergreen or non-evergreen.
For the ASTER image, fewer than 20 pixels failed to be normalized, so for these pixels we simply copied the original values into the normalized image.

### 3.4.4 Assessment of the C-correction

We performed a number of analyses to diagnose the success of the C-correction and observe how the technique affected different portions of the Landsat and ASTER images. First, we recorded coefficient of determination ($r^2$) values for the regression equations used to determine $b_k$, $m_k$ and $c_k$ for each image band. We used the $r^2$ values to compare the proportions of variability in radiance explained by the regressions for both the sunlit and sunshade subsets, as well as to assess the relative impact of the C-correction on the subsets (Neter et al. 1996). To examine the actual alteration of image data resulting from the correction process, we created difference images for each subset by subtracting uncorrected pixel radiance values from corresponding normalized pixel radiance values and then recording the absolute value of that difference. For each band, we recorded maximum per-pixel difference (i.e., extreme changes in the data) and mean per-pixel difference (i.e., typical changes) as percentages of the range in radiance values exhibited prior to correction.

Across an image, topographic normalization should result in a reduction in the per-band standard deviation (SD) of pixel values due to decreased spectral heterogeneity (Riano et al. 2003). Furthermore, if an image has been classified into a number of cover types, the SD within each cover type should be similarly reduced (Civco 1989; Colby 1991; Riano et al. 2003). We computed per-band percentage changes in SD for the full Landsat and ASTER images as well as the sunlit and sunshade subsets. In addition, we divided the Landsat image into two cover classes (evergreen and non-evergreen) and the ASTER image into three
classes (hemlock, non-hemlock evergreen, non-evergreen) by overlaying vegetation maps generated from the images. We calculated the per-band percentage change in SD separately for each cover class.

We also used the derived vegetation maps to examine differences in classification accuracy between the sunlit and sunshade portions of each image. We created the two-class (evergreen vs. non-evergreen) map directly from the normalized Landsat image via cluster busting (Jensen 1996). This process required several iterations of unsupervised classification, where the iterations after the first focused only on pixels that could not be clearly distinguished during the previous iteration. The three-class “hemlock” map, on the other hand, was the product of an expert classifier built on several different input data layers, including the ASTER image. Furthermore, we used the Landsat-derived binary map to mask out non-evergreen areas prior to employing the classifier. As a result, the accuracy of the three-class map could not be entirely attributed to the ASTER image. However, we were still able to evaluate any differences in accuracy corresponding to the image’s sunlit and sunshade subsets.

We partitioned the binary map in the same way as the Landsat image (i.e., according to the same aspect range) and extracted stratified random samples of pixels from the resulting subsets: 50 pixels in evergreen and 50 in non-evergreen classified areas, based on the sample size recommendations of Congalton (1991). We judged accuracy based on a 3-by-3 majority window around the sampled pixel, i.e., if the pixel’s class as determined from reference data (the CIR DOQQ mosaic) was the same as the majority of pixels in the window, the sample pixel was labeled as correct (Congalton & Green 1999).
We had previously completed an assessment of the hemlock map based on 206 reference points gathered largely from field visits, but supplemented by viewing aerial photos where appropriate. We partitioned these reference points according to the aspect value used to subsets the ASTER image. This yielded reference sample sets of 92 and 114 corresponding to the ASTER sunlit and sunshade subsets, respectively. We determined accuracy differently than for the binary map: To accommodate limitations in the positional accuracy of the reference data, and because we were interested in capturing very small stands of vegetation, we looked at pixel values within a 22.5-meter radius of a sample point. If the reference point’s class appeared within that radius, then the classification was judged to be correct.

We created error matrices corresponding to each subset, allowing us to calculate overall accuracy, producer’s and user’s accuracies for each vegetation class, as well as a value for Cohen’s Kappa statistic. Percent correctly classified can be a misleading statistic, because a certain number of correctly identified classes is expected to occur by chance (Goodchild 1994). In contrast, Cohen’s Kappa statistic indicates how much of an improvement a classification effort is over a completely random classification of the same area (Congalton 1991). We calculated Kappa, or $K_{hat}$, according to the formula:

$$K_{hat} = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} \times x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} \times x_{+i})}$$

where $r$ is the number of classes in the matrix, $x_{ii}$ is the number of observations in the diagonal (i.e., row $i$ and column $i$), $x_{i+}$ and $x_{+i}$ are the marginal totals for row $i$ and column $i$ respectively, and $N$ is the total number of survey points (Jensen 1996). $K_{hat}$ can range from 0 to 1, with 0 being the least possible improvement and 1 being the most (Jensen 1996).
3.5 Results

A visual comparison of the original and normalized images (Figures 3-5 and 3-6) suggests that the C-correction successfully removed much of the obvious shadowing caused by terrain. The pixels in the Landsat image that went uncorrected by normalization are visible as black spots of “no data”. In addition, there are dark areas, particularly in the Landsat image, that still have low pixel values. Ultimately, it is difficult to completely remove the topographic effect from deeply shadowed valleys or similar features, regardless of approach (Kawata et al. 1988). Overall visual consistency and interpretability, nevertheless, were improved by the normalization process.

Tables 3-3a and 3-3b list \( r^2 \) values for the linear regression equations used to determine \( b_k, m_k \) and \( c_k \) for each image spectral band. For the sunlit subsets, the regressions explained little of the variation in spectral values (typically, \( r^2 \leq 0.09 \)), although \( r^2 \) values for bands 4 and 5 of the Landsat image were somewhat larger (\( r^2 = 0.13 \) and \( r^2 = 0.25 \), respectively). For the sunshade subsets, the regressions explained a substantially greater proportion of the variation. This was particularly true for bands 4, 5, and 7 of the Landsat sunshade subset, which all had \( r^2 \) values of 0.30 or greater.
Figure 3-6. Uncorrected (left) and topographically normalized (right) Landsat image.

Figure 3-7. Uncorrected (left) and topographically normalized (right) ASTER image.
Tables 3-3a and 3-3b. Per-band $r^2$ values for the Landsat (l) and ASTER image subsets (r).

<table>
<thead>
<tr>
<th>Band</th>
<th>Sunlit $r^2$</th>
<th>Sunshade $r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.04</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>0.39</td>
</tr>
<tr>
<td>5</td>
<td>0.13</td>
<td>0.34</td>
</tr>
<tr>
<td>7</td>
<td>0.09</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Tables 3-4a and 3-4b. Maximum and mean percent per-pixel change (absolute value) between the uncorrected and normalized Landsat (l) and ASTER (r) image subsets.

<table>
<thead>
<tr>
<th>Band</th>
<th>Sunlit Max (%)</th>
<th>Sunlit Mean (%)</th>
<th>Sunshade Max (%)</th>
<th>Sunshade Mean (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.6</td>
<td>0.8</td>
<td>5.7</td>
<td>1.7</td>
</tr>
<tr>
<td>2</td>
<td>6.4</td>
<td>1.7</td>
<td>15.2</td>
<td>3.1</td>
</tr>
<tr>
<td>3</td>
<td>9.9</td>
<td>2.4</td>
<td>26.7</td>
<td>3.6</td>
</tr>
<tr>
<td>4</td>
<td>25.9</td>
<td>6.9</td>
<td>98.3</td>
<td>10.4</td>
</tr>
<tr>
<td>5</td>
<td>27.1</td>
<td>6.4</td>
<td>110.5</td>
<td>9.0</td>
</tr>
<tr>
<td>7</td>
<td>21.3</td>
<td>4.0</td>
<td>82.9</td>
<td>6.3</td>
</tr>
</tbody>
</table>

A similar pattern appeared with respect to the actual change of pixel values due to normalization (Tables 3-4a and 3-4b). Mean percent per-pixel changes were quite low for all bands of the ASTER sunlit subset (≤ 3.5%), but somewhat higher for bands 4, 5, and 7 of the Landsat sunlit subset (as high as 6.9% for band 4). Nevertheless, even for the sunshade subsets, mean percent per-pixel change never exceeded 8.3% for the ASTER image or 10.4% for the Landsat image. For both images, the greatest mean percent per-pixel changes were associated with spectral bands spanning the near-infrared and lower mid-infrared (MIR) portions of the electromagnetic spectrum: bands 4 and 5 of the Landsat image and bands 3 and 4 of the ASTER image (Jensen 1996). This is unsurprising, since infrared, particularly
MIR, bands are generally most sensitive to topographic effects (Jensen 1996; Kawata et al. 1988). Notably, as indicated by the maximum percent change values, some individual pixels underwent substantial alteration, especially in bands 4 and 5 of the Landsat sunshade subset. In a few cases, individual pixel values were doubled or nearly doubled.

Tables 3-5a and 3-5b. Change in standard deviation (SD) due to normalization for the Landsat (l) and ASTER (r) images.

<table>
<thead>
<tr>
<th>Band</th>
<th>Pre-Norm. SD</th>
<th>Post-Norm. SD</th>
<th>% Change</th>
<th>Band</th>
<th>Pre-Norm. SD</th>
<th>Post-Norm. SD</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4.717</td>
<td>3.677</td>
<td>-22.1</td>
<td>2</td>
<td>2.182</td>
<td>2.116</td>
<td>-3.0</td>
</tr>
<tr>
<td>3</td>
<td>6.223</td>
<td>5.002</td>
<td>-19.6</td>
<td>3</td>
<td>18.452</td>
<td>15.846</td>
<td>-14.1</td>
</tr>
<tr>
<td>4</td>
<td>13.945</td>
<td>9.367</td>
<td>-32.8</td>
<td>4</td>
<td>1.455</td>
<td>1.151</td>
<td>-20.9</td>
</tr>
<tr>
<td>5</td>
<td>3.073</td>
<td>2.309</td>
<td>-24.9</td>
<td>5</td>
<td>0.185</td>
<td>0.153</td>
<td>-17.3</td>
</tr>
<tr>
<td>7</td>
<td>0.601</td>
<td>0.483</td>
<td>-19.6</td>
<td>7</td>
<td>0.159</td>
<td>0.131</td>
<td>-17.6</td>
</tr>
</tbody>
</table>

Table 3-6. Change in Landsat image SD within the classes of the binary vegetation map.

<table>
<thead>
<tr>
<th>Band</th>
<th>Pre-Norm. SD</th>
<th>Post-Norm. SD</th>
<th>% Change</th>
<th>Pre-Norm. SD</th>
<th>Post-Norm. SD</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.193</td>
<td>2.646</td>
<td>-17.1</td>
<td>3.282</td>
<td>2.970</td>
<td>-9.5</td>
</tr>
<tr>
<td>2</td>
<td>1.895</td>
<td>1.873</td>
<td>-1.2</td>
<td>4.140</td>
<td>3.362</td>
<td>-18.8</td>
</tr>
<tr>
<td>4</td>
<td>1.257</td>
<td>0.973</td>
<td>-22.6</td>
<td>14.23</td>
<td>10.192</td>
<td>-28.4</td>
</tr>
<tr>
<td>5</td>
<td>0.152</td>
<td>0.121</td>
<td>-20.4</td>
<td>2.355</td>
<td>1.958</td>
<td>-16.9</td>
</tr>
<tr>
<td>7</td>
<td>0.08</td>
<td>0.067</td>
<td>-16.3</td>
<td>0.442</td>
<td>0.396</td>
<td>-10.4</td>
</tr>
</tbody>
</table>

Table 3-7. Change in ASTER image SD within the classes of the three-class vegetation map.

<table>
<thead>
<tr>
<th>Band</th>
<th>Pre-Norm. SD</th>
<th>Post-Norm. SD</th>
<th>% Change</th>
<th>Pre-Norm. SD</th>
<th>Post-Norm. SD</th>
<th>% Change</th>
<th>Pre-Norm. SD</th>
<th>Post-Norm. SD</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.93</td>
<td>3.496</td>
<td>-11.0</td>
<td>3.137</td>
<td>2.595</td>
<td>-17.3</td>
<td>3.137</td>
<td>2.595</td>
<td>-17.3</td>
</tr>
<tr>
<td>2</td>
<td>2.35</td>
<td>2.267</td>
<td>-3.5</td>
<td>2.098</td>
<td>2.028</td>
<td>-3.3</td>
<td>2.098</td>
<td>2.028</td>
<td>-3.3</td>
</tr>
<tr>
<td>4</td>
<td>1.548</td>
<td>1.317</td>
<td>-14.9</td>
<td>1.378</td>
<td>0.938</td>
<td>-31.9</td>
<td>1.378</td>
<td>0.938</td>
<td>-31.9</td>
</tr>
<tr>
<td>5</td>
<td>0.185</td>
<td>0.158</td>
<td>-14.6</td>
<td>0.091</td>
<td>0.084</td>
<td>-7.7</td>
<td>0.091</td>
<td>0.084</td>
<td>-7.7</td>
</tr>
<tr>
<td>6</td>
<td>0.215</td>
<td>0.18</td>
<td>-16.3</td>
<td>0.1</td>
<td>0.089</td>
<td>-11.0</td>
<td>0.1</td>
<td>0.089</td>
<td>-11.0</td>
</tr>
<tr>
<td>7</td>
<td>0.163</td>
<td>0.138</td>
<td>-15.3</td>
<td>0.055</td>
<td>0.052</td>
<td>-5.5</td>
<td>0.055</td>
<td>0.052</td>
<td>-5.5</td>
</tr>
<tr>
<td>8</td>
<td>0.118</td>
<td>0.1</td>
<td>-15.3</td>
<td>0.097</td>
<td>0.089</td>
<td>-8.2</td>
<td>0.097</td>
<td>0.089</td>
<td>-8.2</td>
</tr>
<tr>
<td>9</td>
<td>0.1</td>
<td>0.091</td>
<td>-9.0</td>
<td>0.07</td>
<td>0.065</td>
<td>-7.1</td>
<td>0.07</td>
<td>0.065</td>
<td>-7.1</td>
</tr>
</tbody>
</table>
Both images exhibited substantial reduction in per-band standard deviation (SD), with the Landsat image displaying greater reduction (Tables 3-5a and 3-5b). The decrease in spectral heterogeneity this represents is a primary aim of the normalization process (Riano et al. 2003), and it appears the decrease in heterogeneity was consistent across vegetation classes for both images (Tables 3-6 and 3-7). The actual degree of SD reduction varied between vegetation classes, but this might be attributable to basic differences in their spatial distribution (e.g., a vegetation class could be more strongly associated with shadowed areas and thus tends to occur in pixels receiving greater topographic correction).

Table 3-8. Change in SD due to normalization for the Landsat image subsets.

| Band | Sunlit | | | Sunshade | | |
|------|-------|----|---|--------|----|
| | Pre-Norm. | Post-Norm. | % Change | Pre-Norm. | Post-Norm. | % Change |
| 1 | 3.448 | 3.287 | -4.7 | 3.037 | 2.993 | -1.5 |
| 2 | 4.141 | 3.7 | -10.7 | 3.666 | 3.644 | -0.6 |
| 3 | 6.172 | 5.267 | -14.7 | 4.319 | 4.695 | 8.7 |
| 5 | 2.816 | 2.18 | -22.6 | 1.957 | 2.414 | 23.4 |
| 7 | 0.576 | 0.462 | -19.8 | 0.397 | 0.501 | 26.2 |

Table 3-9. Change in SD due to normalization for the ASTER image subsets.

| Band | Sunlit | | | Sunshade | | |
|------|-------|----|---|--------|----|
| | Pre-Norm. | Post-Norm. | % Change | Pre-Norm. | Post-Norm. | % Change |
| 1 | 3.233 | 3.081 | -4.7 | 2.903 | 2.893 | -0.3 |
| 2 | 2.197 | 2.041 | -7.1 | 1.772 | 2.187 | 23.4 |
| 3 | 16.784 | 15.456 | -7.9 | 15.186 | 16.149 | 6.3 |
| 4 | 1.236 | 1.146 | -7.3 | 1.154 | 1.140 | -1.2 |
| 5 | 0.168 | 0.160 | -4.8 | 0.148 | 0.142 | -4.0 |
| 6 | 0.186 | 0.177 | -4.8 | 0.172 | 0.166 | -3.5 |
| 7 | 0.138 | 0.132 | -4.4 | 0.132 | 0.128 | -3.0 |
| 8 | 0.105 | 0.101 | -3.8 | 0.096 | 0.093 | -3.1 |
| 9 | 0.093 | 0.091 | -2.2 | 0.084 | 0.082 | -2.4 |

There were notable differences between the images’ aspect-partitioned subsets. In particular, their sunlit subsets exhibited SD reduction across all bands. Bands 3, 4, 5, and 7 of the Landsat sunshade subset and bands 2 and 3 of the ASTER sunshade subset actually
displayed increases in SD (Tables 3-8 and 3-9). Intuitively, this may seem like a negative result, but these increases appear to be associated with the shadowed pixels of the sunshade subsets. These pixels generally exhibited low (i.e., near zero) radiance values in certain bands prior to correction, and thus could only become more variable with the introduction of terrain information.

Error matrices for the image subsets are shown in Tables 3-10 through 3-13. The sunlit subsets for both images exhibited higher overall accuracy values than their corresponding sunshade subsets. The point is further emphasized by the gap in Kappa values for the images: Landsat $K_{hat}$ values were 0.80 for the sunlit and 0.62 for the sunshade subset, while ASTER $K_{hat}$ values were 0.90 for the sunlit and 0.78 for the sunshade subset. Still, overall classification accuracy exceeded 80% for all four subsets. Similarly, while per-class producer’s and user’s accuracies were lower for the sunshade than the sunlit subsets, they still exceeded 75% in all cases. The gap in the evergreen class user’s accuracy between the Landsat sunlit and sunshade subsets is the most noteworthy difference between matrices.

Table 3-10. Error matrix for Landsat sunlit subset.

<table>
<thead>
<tr>
<th>Class</th>
<th>Reference Totals</th>
<th>User's Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Evergreen</td>
<td>Non-Evergreen</td>
</tr>
<tr>
<td>Evergreen</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Producer's Accuracy (%)</td>
<td>90.0</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Table 3-11. Error matrix for Landsat sunshade subset.

<table>
<thead>
<tr>
<th>Class</th>
<th>Reference Totals</th>
<th>User's Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Evergreen</td>
<td>Non-Evergreen</td>
</tr>
<tr>
<td>Evergreen</td>
<td>38</td>
<td>12</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>Producer's Accuracy (%)</td>
<td>84.44</td>
<td>78.18</td>
</tr>
</tbody>
</table>
### Table 3-12. Error matrix for ASTER sunlit subset.

<table>
<thead>
<tr>
<th>Class</th>
<th>Hemlock</th>
<th>Non-Evergreen</th>
<th>Non-Hemlock Evergreen</th>
<th>Total</th>
<th>User’s Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemlock</td>
<td>42</td>
<td></td>
<td></td>
<td>42</td>
<td>100.0</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>5</td>
<td>24</td>
<td></td>
<td>29</td>
<td>82.76</td>
</tr>
<tr>
<td>Non-Hemlock Evergreen</td>
<td>1</td>
<td>20</td>
<td></td>
<td>21</td>
<td>95.24</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>25</td>
<td>20</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

**Classification Totals**

<table>
<thead>
<tr>
<th>Producer’s Accuracy (%)</th>
<th>89.36</th>
<th>96.0</th>
<th>100.0</th>
</tr>
</thead>
</table>

### Table 3-13. Error matrix for ASTER sunshade subset.

<table>
<thead>
<tr>
<th>Class</th>
<th>Hemlock</th>
<th>Non-Evergreen</th>
<th>Non-Hemlock Evergreen</th>
<th>Total</th>
<th>User’s Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemlock</td>
<td>62</td>
<td>3</td>
<td>1</td>
<td>66</td>
<td>93.94</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>6</td>
<td>24</td>
<td>1</td>
<td>30</td>
<td>80.0</td>
</tr>
<tr>
<td>Non-Hemlock Evergreen</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>18</td>
<td>77.78</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>29</td>
<td>15</td>
<td>114</td>
<td></td>
</tr>
</tbody>
</table>

**Classification Totals**

<table>
<thead>
<tr>
<th>Producer’s Accuracy (%)</th>
<th>88.57</th>
<th>82.76</th>
<th>93.33</th>
</tr>
</thead>
</table>

### 3.6 Discussion

In mountainous environments, topographically normalized satellite images can greatly improve forest stand classification success as compared to uncorrected imagery (Ekstrand 1996; Hale & Rock 2003; Meyer et al. 1993). For our goal of mapping hemlock distribution in the southern Appalachians, topographic normalization is basically essential, since hemlock stands are prevalent in the ravines and valleys typically covered in shadow in uncorrected montane imagery ((Delcourt & Delcourt 2000; Godman & Lancaster 1990; Quimby 1996). Our results indicate that, at least for this study, the C-correction performed well. In particular, the normalized Landsat and ASTER images displayed a strong overall reduction in spectral heterogeneity across all image bands, which is a fundamental aim of the normalization process. Furthermore, the reduction in heterogeneity carried over well to individual vegetation classes determined *a posteriori*, suggesting that *a priori* information
about the cover types present in an image, while possibly beneficial, is not absolutely necessary for successful normalization.

Nevertheless, the C-correction did not perform flawlessly. Differences in overall classification accuracy and $K_{\text{hat}}$ values between the sunlit and sunshade subsets of each image revealed that the C-correction did not completely counteract the topographic effect, and as a result classification errors remained more likely in the sunshade portions of both images after correction. Moreover, while mean per-band changes in pixel radiance values were generally low, individual pixels saw substantial increases for specific bands, which could be a sign of over-correction due to the model (Riano et al. 2003). This possibility is echoed by the difference in the evergreen class user’s accuracy between the Landsat sunlit and sunshade subsets. In this case, over-corrected values, particular in infrared bands, might have suggested the presence of living green vegetation in areas of the sunshade subset where, in fact, there was none (Jensen 1996).

The issue of over-correction is far less important than adequate accuracy of derived map products. Although the sunshade subsets of both images exhibited lower accuracies than their sunlit subsets, the lowest overall accuracy value—81% for the Landsat sunshade subset—is still a good result for classifications derived from remotely sensed data (Congalton & Green 1999). While it would be ideal for per-class user’s and producer’s accuracies to similarly exceed 80%, the results we received are generally adequate for remotely sensed data, and sufficiently consistent between subsets. Similarly, the lower $K_{\text{hat}}$ values of the sunshade subsets still represent substantial improvements on random classifications (Jensen 1996).
All of these observations suggest that, despite its limitations, the C-correction is appropriate for inclusion in our hemlock mapping protocol. The large differences in $r^2$ values between the sunlit and sunshade subsets appear to recommend the use of aspect partitioning prior to applying the correction. Admittedly, the regression equations explained just a small proportion of the variation in the sunlit subsets, translating to less actual impact on the radiance values of the sunlit areas of each image. Yet, this also meant that we were able to build regression models specific to the sunshade portions of each image, resulting in higher $r^2$ values and better model fit for these areas than if we had simply applied a blanket set of models across the entire image (Hale & Rock 2003). Given the ease with which aspect partitioning can be performed, and that applying the C-correction to two subsets requires only a handful of additional processing steps, there is little reason not to include it in our mapping protocol.

Although the topographic normalization literature has historically focused on Landsat and SPOT imagery, our findings suggest that the C-correction generalizes to ASTER imagery as well. In terms of changes in actual pixel values and in SD, results for the ASTER image were typically not as dramatic as for the Landsat image. This can be ascribed to two factors. First, the ASTER’s lower solar elevation angle meant fewer poorly illuminated areas requiring dramatic correction. Second, the ASTER image was captured in the summer, and summer images typically have less topographic shadowing than winter images (Jensen 1996). The solar angle and time of image capture may be factors worth considering when selecting images for use in remote sensing projects.

There are several potential sources of error that must be considered in assessing this methodology. First, the images were geometrically corrected using a polynomial equation
method. This approach is standard for satellite imagery, but it ignores the effects of
topographic relief. Although the geometric correction for both images resulted in RMSE
values less than half a pixel, this is simply an overall measure. In areas of extreme relief, the
localized positional accuracy of the corrected images is likely to be worse than the RMSE
value indicates (Tucker et al. 2004). Orthorectification (i.e., the incorporation of a digital
elevation model into the geometric correction process) of satellite images might result in
fewer errors due to misregistration, but the process requires a sufficient number of high-
quality ground control points, which are not always available (Tucker et al. 2004). We
decided that the degree of positional error was acceptable for our analysis, but would
recommend the use of orthorectified satellite imagery if it were available for future
applications.

Another important drawback of the C-correction or any other elevation-model-based
normalization is propagation of error from the DEM (Gu et al. 1999). The correction is
heavily dependent on the DEM’s accuracy and resolution, and this can lead to localized areas
of over- or under-correction. The C-correction and similar approaches consider information
only at the level of individual pixels, making them susceptible to error from flawed input
data, especially the DEM (Gu et al. 1999). Drawing information from neighboring pixels in
some fashion may reduce such instances of error. Hale and Rock (2003) suggested that
filtering a DEM via a neighborhood-based function (e.g., a low-pass statistical filter) might
provide a workable solution, while Albani and Klinkenberg (2003) outlined an algorithm for
clearing up systematic DEM artifacts, some of which can affect aspect and slope. Smoother
surfaces may also improve the C-correction in low illumination areas (Riano et al. 2003).
While higher-quality DEMs from other sources, such as airborne laser altimetry (i.e.,
LIDAR), might improve the topographic normalization results, such resources are not currently available for large regions in the southern Appalachians. If such data become available, their impact on topographic normalization may be worth testing.

Finally, to determine the C-correction equation parameters for the models applied here, we used 10,000-point samples from each subset. It is possible that a larger sample—i.e., a sample including every pixel—might change the C-correction coefficients and therefore the normalization result. Nonetheless, a 10,000-point random sample is reasonably large (roughly 1-2% of pixels) for a subset of an already-reduced study image, and also should be unbiased (Neter et al. 1996). Given the generally high level of variation in pixel values due to environmental factors other than topographic relief, it seems unlikely that equations can be developed that have a substantially better fit than the ones created.

We have already noted some of the drawbacks to the C-correction and similar topographic normalization methods. Spectral reflectance is often strongly anisotropic and wavelength dependent, so solar, surface, and sensor geometry must all be considered and modeled well (Jensen 1996); of course, DEM-based normalization methods have been criticized for representing these geometries too simplistically (Gu & Gillespie 1998). This is in addition to the fact that the C-correction and similar methods ignore the potential effect of diffuse irradiance. Certainly, a simple model is being used to explain complex phenomena, so there are bound to be inadequacies and potentially some errors.

Another possible issue is the effect of the Earth’s atmosphere on remotely sensed imagery. Due to scattering and absorption by gases and aerosols, electromagnetic radiation is changed as it passes through the atmosphere from the Earth to a satellite-based sensor (Song et al. 2001). This can substantially alter image pixel values. In cases where a single
image is being classified (e.g., our project’s use of a Landsat image to generate an evergreen/non-evergreen map), atmospheric correction is unimportant. However, for applications that are likely to employ images from multiple dates—such as our regional hemlock mapping project—it is a far more important step (Song et al. 2001). Fortunately, image products that have already been atmospherically corrected are readily available from the ASTER sensor (LP-DAAC 2005a; 2005b). This offers a significant time saving for projects that will ultimately require regional image coverage, and is one of the reasons why we chose to employ ASTER imagery for our hemlock classification effort.

3.7 Conclusions

We applied a C-correction topographic normalization to two images—a leaf-off Landsat ETM+ image and a leaf-on ASTER image—covering the North Carolina portion of Great Smoky Mountains National Park. We hoped to minimize the effect of topography on image pixel values, because these images were subsequently used in deriving vegetation maps of the region. Prior to normalization, we partitioned the images according to topographic aspect into “sunlit” and “sunshade” subsets. This allowed us to develop correction coefficients specific to the poorly and well illuminated portions of the images. The correction process resulted in a reduction of per-band standard deviation for each image, indicating an overall decrease in spectral heterogeneity. While there may have been some instances of over-correction, accuracies of derived maps were high enough for all subsets to suggest that over-correction had a negligible effect on the practical results.

The C-correction depends on an adequate representation of topography for modeling sun-terrain-sensor geometry. To this end, DEMs used should be as free of errors as possible.
In addition, images must be adequately geometrically corrected beforehand, to ensure the best possible correspondence with the DEM. Samples of pixel values are required to determine the C-correction parameters—via linear regression—so a representative sample of adequate size and distribution must be collected from the image. Perhaps most importantly, separate samples should be extracted from both “sunlit” and “sunshade” partitions of the image, so that C-correction coefficients are better targeted at these distinct image segments.

The C-correction appears to be a workable solution for areas of moderate elevation and rugged topography such as in our study region. Users should consider that the C-correction—or any other topographic normalization method—does alter the original data, and this may introduce error into classifications or yield patterns that really do not exist except as an artifact of the normalization process. Nonetheless, potential gains in information and consistency across an image may be an acceptable tradeoff.

3.8 References


Leica Geosystems. 2001. ERDAS Imagine 8.7 software. Atlanta, GA: Leica Geosystems GIS & Mapping, LLC.


4. MAPPING HEMLOCKS VIA TREE-BASED CLASSIFICATION OF SATELLITE IMAGERY AND ENVIRONMENTAL DATA

Abstract: Within the last few years, the hemlock woolly adelgid has made significant inroads into the southern Appalachians. Since the region’s native hemlock species are not resistant to the pest, timely application of control measures is critical to minimizing mortality. Unfortunately, hemlock stands in the region are poorly mapped and general characteristics of their distribution present serious mapping challenges. One approach for improving map success is to augment medium-resolution satellite imagery with ancillary environmental data in a decision-tree-structured classification. We applied such an approach using Terra ASTER and Landsat ETM+ images of eastern and western study areas in Great Smoky Mountains National Park. First, we performed multiple unsupervised classifications (cluster busting) of leaf-off images to distinguish and remove non-evergreen pixels from co-registered leaf-on images. We then used the masked leaf-on images to separate hemlocks from other evergreen species. We extracted large random samples of points for each study area, stratifying the samples according to an air-photo-derived vegetation map of the park (>14,000 points total). At each point, we recorded vegetation class, image spectral data, and values for a suite of environmental variables recorded in a geographic information system. We generated decision trees for mapping vegetation: an initial tree using only points from the eastern area as training data, and an enhanced tree using points from both areas. Thematic accuracy assessment of the resulting maps indicated the enhanced tree performed better, yielding 90% overall accuracy in the eastern study area and 75% success at capturing hemlocks in a partial assessment of the western study area. With nine variables and 27 terminal nodes, the enhanced tree is compact and general enough for application elsewhere in the southern Appalachians for hemlock and hemlock woolly adelgid management purposes. Moreover, additional use might offer a chance to further refine the tree’s decision rules.

Key Words: hemlock woolly adelgid, Landsat, ASTER, classification, decision trees, southern Appalachians

4.1 Introduction

Accidentally introduced from Asia, the hemlock woolly adelgid (Adelges tsugae Annand) is now a significant insect pest in the eastern United States. First observed in Richmond, Virginia around 1951, the hemlock woolly adelgid was considered little more than a nuisance pest of ornamental hemlocks until the 1980s, when it invaded natural forest stands and began causing substantial mortality (Cheah et al. 2004; McClure et al. 2001). Unfortunately, both hemlock species native to the eastern U.S.—eastern (Tsuga canadensis (L.) Carr.) and Carolina hemlock (T. caroliniana Engelm.)—have proven to be extremely
susceptible hosts for the pest, exhibiting no significant resistance (McClure et al. 2001). Indeed, spurred by exponential population growth, adelgid infestations can kill hemlocks in as few as four years (Cheah et al. 2004; McClure et al. 2001; Ward et al. 2004).

The hemlock woolly adelgid is readily dispersed by wind, deer, birds, and people (McClure 1990). During the last couple of decades, it has expanded its range, primarily to the northeast, by approximately 17 km per year (Cheah et al. 2004; McClure 1990). More recently, the hemlock woolly adelgid has made significant advances into the southern Appalachian region, with isolated infestations now appearing as far south as Georgia (Ward et al. 2004). There has been a concerted effort to respond to this invasion with biological control via predators introduced from Asia (Asaro et al. 2005; Cheah et al. 2004, 2005; Johnson et al. 2005). This approach holds some promise for combating the pest in natural stands. Unfortunately, application of this or any other countermeasure faces a significant obstacle: The distribution of hemlock stands in the southern Appalachians has not been well delineated, except in isolated locations such as Great Smoky Mountains National Park (Johnson et al. 2005; Welch et al. 2002). This, in turn, makes it difficult to identify hemlock stands for predator release, or even to efficiently survey forests regarding current hemlock woolly adelgid presence. The lack of knowledge may be particularly costly because there is only a small window of opportunity for successfully establishing predators in infested stands (Cheah & McClure 2000). Therefore, a means to map hemlock distribution and prioritize stands is a key component of a successful effort to combat the hemlock woolly adelgid in the southern Appalachians or elsewhere.

Satellite data is a logical foundation for a hemlock mapping protocol: It offers region-wide coverage, can be processed quickly using automated classification techniques, and is
relatively inexpensive for moderate spatial resolutions (Jensen 1996). However, previous hemlock mapping efforts in the northeastern U.S. (Bonneau et al. 1999; Royle & Lathrop 1997, 2002) have used Landsat imagery and reported varying levels of success. For example, Royle and Lathrop (1997) noted major difficulties in separating eastern hemlock from other evergreen species in their study area in New Jersey based on image data alone. A number of factors make mapping hemlocks using medium-resolution satellite data a challenging prospect, especially for the southern Appalachians. First, both eastern and Carolina hemlock are commonly found in moist valleys, riparian areas, coves, steep ravines, or on shaded, north-facing bluffs (Delcourt & Delcourt 2000; Godman & Lancaster 1990). Such areas may be in complete shadow in mountainous satellite imagery (Jensen 1996).

More importantly, while small hemlock-dominated stands do exist in the southern Appalachians, hemlocks are more widespread as secondary associates of various hardwood species, such that they may or may not be a major presence in upper forest strata (Godman & Lancaster 1990; McWilliams & Schmidt 1999; Shriner 2001; Whittaker 1956).

Ancillary data sources can help to distinguish classes, like hemlock, that are difficult to identify from image spectral data alone, particularly if those classes exhibit high variability (Jensen 1996; McIver & Friedl 2002). Vegetation pattern is shaped by environmental gradients, many of which can be reasonably depicted with geospatial data sets. For example, many studies (e.g., Bolstad et al. 1998; Cairns 2001; Franklin 2002; Treitz & Howarth 2000) have employed terrain-based variables derived from digital elevation models to predict vegetation distribution. However, environmental gradients in southern Appalachian forests can be highly complex, suggesting that key variables and threshold values may be difficult to discern regardless of their source of input (Abella et al. 2003).
Fortunately, there is a growing body of literature on automated data exploration techniques applicable to remote sensing. The decision tree concept broadly includes a number of algorithms for the automated analysis and reduction of complex multi-factor data, as well as the subsequent application of those data in classification and prediction (Jovanovic et al. 2002; Murthy 1998). The basic mechanism of these algorithms is nearly universal: Tree building (or induction) occurs in a top-down fashion, starting with a “branchless” tree and a training data set, e.g., a set of records containing image spectral data and other geospatial data (Lawrence & Wright 2001). First, the training data are split into mutually exclusive groups that are as homogeneous as possible, and then (if possible) each of these groups is split again, and so on; the process is repeated until the terminal groups achieve maximum homogeneity (De’Ath & Fabricius 2000). Each split of the data is a decision rule employing a threshold value for one (or more, in some software packages) of the input variables (De’Ath & Fabricius 2000; Murthy 1998). Decision trees are often represented graphically with the root node—the unsplit training data—at the top, followed by the branches, the intermediate splitting nodes, and then the terminal nodes or “leaves” (De’Ath & Fabricius 2000; Murthy 1998).

Decision tree algorithms use a variety of splitting criteria, often based on statistical distance measures. The popular classification and regression tree (CART) algorithm employs the Gini index, a measure of the relative purity of partitioned subgroups relative to their parent group, while the Chi-Squared Automatic Interaction Detection (CHAID) algorithm uses the $\chi^2$ statistic for distance measurement (Breiman et al. 1984; De’Ath & Fabricius 2000; Murthy 1998). There are other noteworthy distinctions. For example,
CART is limited to binary (i.e., two-way) splits, while CHAID allows for an \( n \)-way number of partitions (Jovanovic et al. 2002; Nelson et al. 2003).

Decision trees employ few assumptions, making them quite flexible (McIver & Friedl 2002). They are non-parametric, so normality of variables is not critical (De’Ath & Fabricius 2000; Murthy 1998). Decision trees can accommodate categorical and/or numeric explanatory variables, and response variables can take many forms: numeric, categorical, ratings, and survival data (De’Ath & Fabricius 2000). The tree-building process is essentially a process of variable selection, though the inclusion of large numbers of highly correlated variables may be problematic (Murthy 1998). Finally, the clearly defined splitting rules of a decision tree are easy to interpret, and can be readily transported to geographic information systems (GIS) or other analytical settings.

On the other hand, since many leaves may be required to capture different instances of the classes of interest, large decision trees may result (Murthy 1998). Large trees can be difficult to interpret and may over-fit the training data, making it difficult to generalize them to other data sets (Jovanovic et al. 2002; Nelson et al. 2003). However, tree size can be controlled by automatic pruning techniques. While some algorithms simply limit how large a tree can grow according to some significance threshold, pruning generally involves dividing the input data into (1) a training set to grow the decision tree and (2) a validation set to remove branches judged insignificant or not beneficial (Murthy 1998; Nelson et al. 2003). The cost-complexity pruning approach—one of the most widely utilized pruning methods—proceeds by building nested, increasingly smaller trees from the training data based on assessment of the per-leaf accuracy cost. Then, one of the smaller trees is chosen based on its accuracy at classifying the validation data set (Breiman et al. 1984; Li et al. 2001).
Decision trees have been used quite successfully for a number of land cover and vegetation classification efforts based on satellite imagery and other geospatial data. Friedl and Brodley (1997) demonstrated, with three different remotely sensed data sets, that decision trees outperformed the conventional maximum likelihood algorithm in accurately classifying land cover. Lawrence and Wright (2001) classified the vegetation of the Greater Yellowstone ecosystem with decision trees built on Landsat Thematic Mapper (TM) data as well as terrain variables. They achieved 96% overall accuracy for their broadest classification scheme and 65% for a scheme distinguishing the dominant tree species. Joy et al. (2003) modeled vegetation types in the Kaibab National Forest based on field data, applying topographic information and Landsat TM data as ancillary variables. Overall accuracy of their output map was 74.5%. Brown de Coulston et al. (2003) delineated the vegetation of the Delaware Water Gap National Recreation Area to the formation level of the U.S. National Vegetation Classification. They built their decision tree using raw Landsat Enhanced Thematic Mapper (ETM+) imagery as well as normalized difference vegetation index (NDVI) values. Their final land cover map had an overall accuracy of 82%, and was 99.5% accurate at classifying forest versus non-forest. Rogan et al. (2003) used the CART algorithm and multitemporal Landsat TM imagery to quantify land cover change for a six-year period (1990-1996) in the San Diego area. They employed transformed spectral data and variables such as elevation, slope, aspect, fire history, and existing land cover to derive a three-level hierarchical land cover classification. The most detailed level of their scheme included nine different classes for percent forest canopy change as well as shrub/grass and developed area change. The decision tree created for this level performed with an overall accuracy of 72%.
Drawing on these and other studies as examples, we adopted a decision tree approach as the key component of a hemlock mapping classifier. Starting with a relatively small study area, our primary objective was to create rules general enough that they could be applied for mapping hemlock throughout the southern Appalachian region. To enable rapid application of the resulting rules, we employed image spectral data and variables that were readily available or calculable in a GIS as inputs. Moreover, we kept the inputs simple, so it would be easy to refine the rules if necessary. In any case, the hemlock maps resulting from application of our approach—especially when combined with maps predicting hemlock woolly adelgid occurrence—are intended to help forest managers prioritize areas for control measures or specifically targeted analysis.

4.2 Study Area

We selected portions of Great Smoky Mountains National Park (GSMNP) in North Carolina and Tennessee for our decision tree analyses. GSMNP has a significant eastern hemlock presence in many different areas and in association with a variety of evergreen and non-evergreen species (Taylor 2002; Whittaker 1956). The park’s range of habitats and conditions is a fairly representative sample of the entire southern Appalachians. Furthermore, the park has one of the most comprehensive GIS databases in the region. We first developed a hemlock classifier (i.e., a decision tree) using an area of approximately 482 km² from the eastern side of the park (Figure 4-1). To create a more comprehensive tree (described in Section 4.3.4), we added a second study area of approximately 108 km² on the western side of the park. The western study area is characterized by gentler topography and less elevation change than the eastern study area.
4.3 Methods

Except where specified, we used ERDAS Imagine 8.7 for image processing, ESRI ArcGIS 8.3 for GIS analysis, and SAS Enterprise Miner Release 4.3 for building our decision trees (ESRI 2002; Leica Geosystems 2001; SAS Institute 2003).

4.3.1 Image Acquisition and Pre-Processing

We acquired a late October 2001 leaf-off Landsat ETM+ image (Path 18, Row 35) from the Global Land Cover Facility, an archive of satellite imagery freely downloadable through the World Wide Web (GLCF 2005). We downloaded both multispectral and panchromatic bands and combined them into a 14.25-m multispectral image using a correspondence-analysis-based data fusion algorithm (Cakir 2003). We geometrically corrected the fused Landsat image using a third-order polynomial equation and 92 ground control points collected from color infrared digital orthophoto quarter quads (DOQQs) of the area, with a resulting root mean square error (RMSE) of 4.142 m. We did not perform
atmospheric correction because the Landsat image was used in a single-image classification where atmospheric effects are likely trivial (Song et al. 2001).

We also acquired a leaf-on, early September 2000 ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) image for the study area. We chose to use ASTER imagery because of its comparability to Landsat and low cost: Archived and atmospherically corrected reflectance and radiance scenes can be freely downloaded through the NASA Earth Observing System Data Gateway (ERSDAC 2003a; 2003b). While ASTER imagery has regularly been used to study thermal phenomena such as surface emissivity (e.g., Schmugge et al. 2002; Dash et al. 2002), it has also been applied successfully for land cover classification (Sharma et al. 2004). The ASTER sensor is actually a combination of radiometers that yield simultaneous and co-registered image data of differing spatial resolutions (ERSDAC 2003a). For our analysis, we merged a 15-m resolution, 3-band radiance image in the visible and near infrared and a corresponding 30-m, 6-band radiance image in the short-wave infrared, sidestepping any resampling of the short-wave infrared image’s values by simply dividing its pixels into four 15-m sub-pixels of equal value. Though the ASTER data were already geometrically corrected, we further corrected the merged image using a fourth-order polynomial and 80 ground control points from the color infrared DOQQs (RMSE = 6.469 m). This second correction better aligned the ASTER image with the Landsat image. We then clipped each image to fit the eastern study area.

As is typical for satellite images of mountainous regions, the ASTER and Landsat images contained heavily shadowed areas due to topographic relief. A number of equations have been proposed that correct images using digital elevation models (DEM}s) to account for terrain-induced variation. We topographically normalized both images using the C-
correction method introduced by Teillet et al. (1982). This statistical-empirical approach requires calculation of the cosine of illumination (cos $i$) for each pixel of an image:

$$\cos i = \cos (90 - \theta_s) \cos \theta_n + \sin (90 - \theta_s) \sin \theta_n \cos (\varphi_s - \varphi_n),$$

where $\theta_s$ is the solar elevation value, $\varphi_s$ is the solar azimuth, $\theta_n$ is the surface slope, and $\varphi_n$ is the surface aspect (Hale & Rock 2003). The slope and aspect values are derived from a DEM, while the solar elevation and azimuth are available from a parameter file associated with the image. The C-correction is implemented separately for each band of an image according to the formula:

$$\rho_T = \rho_T \left( \frac{\cos \theta_s + c_k}{\cos i + c_k} \right),$$

where $\rho_T$ represents the original pixel values, $\rho_H$ the normalized pixel values, and $c_k$ is based on the coefficients of a regression of the original pixel values versus cos $i$—i.e., $c_k = b_k / m_k$ for $\rho_T = b_k + m_k \cos i$ (Riano et al. 2003; Teillet et al. 1982). Prior to normalization, we subdivided the Landsat and ASTER images according to their solar azimuth values into sunlit and sunshade subsets (Hale & Rock 2003). This “aspect partitioning” process allowed us to develop C-corrections specific to these quite differently illuminated subsets. We calculated $c_k$ and applied C-corrections separately for each subset before reassembling the images.

4.3.2 Masking of Non-Evergreen Areas

The leaf-off Landsat image served as a means to mask out all non-evergreen pixels in our study area. We separated the image into evergreen and non-evergreen vegetation classes via cluster busting. Cluster busting is a process of applying repeated runs of an unsupervised classification algorithm to an image (Jensen 1996). Each subsequent run after the first is an
attempt to repartition the image feature space for improved classification of the areas deemed unclassifiable during the previous iteration (Jensen 1996). Initially, we applied the ISODATA algorithm to divide the image into 100 different spectral clusters. Based on visual comparison to the color infrared DOQQs for the area, 59 of these clusters (and roughly 60% of the image pixels) were easily distinguishable as evergreen or non-evergreen. We masked out all pixels that were classified into one of these clusters during the first ISODATA run, and then classified the remaining pixels through six additional iterations of ISODATA classification, employing 25 clusters in the first three iterations, 10 clusters in the fourth and fifth iterations, and 5 clusters in the sixth. Upon completion, we combined the results of all iterations into a single binary map. When assessed using the DOQQs, the map had an overall accuracy greater than 85%, and producer’s and user’s accuracies for each class exceeded 83%. After resampling the map from 14.25 m to 15 m resolution, we used it to remove all non-evergreen pixels from the ASTER image.

4.3.3 Initial Decision Tree Generation

Summer (or leaf-on) images like the ASTER image offer good spectral conditions for species separation, and have less topographic shadowing than leaf-off images (Jensen 1996). After masking out non-evergreen pixels, we used the ASTER image to separate hemlock from non-hemlock evergreen classes. To do this, we created a data set suitable for training and pruning a decision tree. Our guiding source for the data set was a GIS-based vegetation map for GSMNP 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. Using a macro developed for ArcGIS (Sawada 2002), we generated a large random sample of points for each of the 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). We made sure no image pixel was sampled more than once, and we scaled the sample sizes to roughly 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 “environmental” variables derived from raster data layers in the GIS database for GSMNP (Table 4-1). We also recorded pixel data from the masked ASTER image. We used normalized band ratios rather than individual per-band pixel values (Table 4-1). This approach reduced the chance that image-related threshold values in the decision tree were scene-dependent and thus could not be generalized well. Because a decision tree fit to five different classes can be unwieldy and inaccurate (Murthy 1998), we simplified the training sample by combining the dominant with the co-dominant class and the secondary component with the inclusion class.

SAS Enterprise Miner is a flexible software package, allowing a user to approximate the parameters of CHAID, CART, or other decision tree algorithms (SAS Institute 2003). We employed a CART-like approach; in particular, all data splits were binary and evaluated based on minimization of the Gini impurity index (Breiman et al. 1984). To avoid over-fitting of the tree, we also employed CART-style automatic pruning. We divided the training data set into two equal parts, the first to generate a sequence of nested sub-trees, and the second (the validation set) to select the smallest sub-tree with the best classification accuracy.
We imported the pruned tree into the Expert Classifier module of ERDAS Imagine, where we converted it into decision rules for assigning pixels to an output class based on the raster data layer for each input variable. We created a final four-class map—hemlock dominant/co-dominant, hemlock secondary/inclusion, non-hemlock evergreen, non-evergreen—for the study area by applying the expert classifier to the masked ASTER image and associated GIS data, and then merging this result with the pixels identified as non-evergreen in the cluster-busted Landsat image.

Table 4-1. Variables tested for incorporation in the initial and enhanced decision trees.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Resolution (m)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect</td>
<td>10</td>
<td>Slope direction based on DEM, transformed to a measure of “northeast-ness”; scaled 0 to 2 (most NE) (Beers et al. 1966)</td>
</tr>
<tr>
<td>Curvature</td>
<td>10</td>
<td>Convexity (+ values) / concavity (- values) measure based on DEM; units=1/100 meters</td>
</tr>
<tr>
<td>Elevation</td>
<td>10</td>
<td>Elevation from DEM; units=meters</td>
</tr>
<tr>
<td>Landform Index</td>
<td>10</td>
<td>Exposure index calculated from DEM; scaled –1 (protected) to 1 (exposed) (McNab 1993)</td>
</tr>
<tr>
<td>Percent Slope</td>
<td>10</td>
<td>Percent slope based on DEM</td>
</tr>
<tr>
<td>Topographic Relative</td>
<td>10</td>
<td>Dryness-wetness index calculated from DEM; scaled 0 (xeric) to 60 (very mesic) (Parker 1982)</td>
</tr>
<tr>
<td>Moisture Index</td>
<td>10</td>
<td>Proximity grid developed from GSMNP stream vector layer; distance to closest stream from each cell’s centroid (meters)</td>
</tr>
<tr>
<td>Distance to Stream</td>
<td>10</td>
<td>Harvested or cleared land (GSMNP data)</td>
</tr>
<tr>
<td>Disturbance History</td>
<td>90</td>
<td>Reoccurring burns, 1920s-80s (GSMNP)</td>
</tr>
<tr>
<td>Fire Frequency</td>
<td>90</td>
<td>Decades of fires, 1920s-80s (GSMNP)</td>
</tr>
<tr>
<td>Fire History</td>
<td>90</td>
<td>General bedrock formations (GSMNP)</td>
</tr>
<tr>
<td>Geology</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>ASTER Ratios:</td>
<td></td>
<td>Normalized difference indices to allow generalization of any image-based rules in the output trees. Ratios were chosen based on band-to-band correlations. Indices calculated as (band a – band b) / (band a + band b)</td>
</tr>
<tr>
<td>Band 3/Band 1</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Band 1/Band 2</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Band 4/Band 5</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Band 4/Band 6</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Band 4/Band 7</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Band 4/Band 8</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Band 4/Band 9</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

4.3.4 Enhanced Decision Tree Generation

Although we employed automatic pruning to control over-fitting and keep the decision tree general in nature, we also realized the eastern study area might not be a true
representation of the variation of conditions in the southern Appalachians, thus limiting its universality as a source for training data. To improve region-wide applicability of our hemlock classifier, we extracted supplementary training data from a study area in the western portion of GSMNP. Eastern hemlock is far less dominant in the coves of the western Smokies than elsewhere in the park, although it is still an important component of topographically protected areas such as stream valleys (Callaway et al. 1987).

We incorporated the western study area into our analysis by essentially duplicating our earlier process steps. We procured atmospherically corrected ASTER radiance data, captured in June 2000 (leaf-on) and November 2003 (leaf-off), through the NASA Earth Observing System Data Gateway. For each date, we merged separate visible and near infrared and short-wave infrared images into a single 15-m resolution, 9-band image. Though subjected to preliminary geometric correction, we further corrected the images with polynomial equations (39 ground control points and RMSE = 6.155 for June 2000; 24 ground control points and RMSE = 4.644 for November 2003) and clipped them to the western study area extent. We then partitioned the images into sunlit and sunshade subsets based on the solar azimuth and topographically normalized them via the C-correction method (Hale & Rock 2003; Teillet et al. 1982). We generated an evergreen/non-evergreen map from the November 2003 image via several iterations of cluster busting, and used this to mask out non-evergreen areas in the June 2000 image.

From the remaining evergreen portion of the masked June 2000 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 recorded image band ratio values as well as values for additional environmental variables from raster layers in the GSMNP GIS database (Table 4-1). We then combined these sample points with our eastern area training data, yielding a substantially larger set of more than 14,000 sample points across five classes.

We used SAS Enterprise Miner to generate an enhanced decision tree based on this larger data set. We again merged the dominant and co-dominant classes and the secondary component and inclusion classes to simplify the output tree. As with the initial decision tree analysis, we divided the training data into two equal sets and employed a CART-like approach to build and prune our decision tree. We imported the leaves from the tree into the Expert Classifier module of Imagine, and used the resulting rules to classify the evergreen pixels from both study areas. We then created final four-class output maps by merging these results with the non-evergreen areas identified in the cluster-busted October 2001 Landsat (for the eastern study area) and November 2003 ASTER (for the western area) images. For the sake of comparison, we also used our initial decision tree (which was built only with eastern area data) to construct a hemlock map of the western area.

4.3.3 Accuracy Assessment

We performed accuracy assessment of the eastern area maps generated with the initial and enhanced decision trees using 206 reference points, gathered largely from field surveys or by viewing the color infrared DOQQs where appropriate. Based on the limited information collected for many of the points, 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 judged accuracy (i.e., whether a reference point and its
corresponding map point matched) in two ways. First, we examined only the pixel containing the reference point and noted whether their classes were in agreement. Second, 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 class of the reference point corresponded to the class of any pixel falling within the window, then the map and reference point were considered in agreement. 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 recreation-grade global positioning system (GPS) units (± 15 m). Such use of windows is not unusual; for remote-sensing-derived products, accuracy is commonly judged based on the majority of pixels in a window around a reference point (Congalton & Green 1999). However, hemlock stands in the southern Appalachians can be quite small and isolated from one another (McWilliams & Schmidt 1999), making the majority rule inappropriate.

For each map, we created error matrices and calculated 95% confidence intervals for overall accuracy based on a binomial distribution (Snedecor & Cochran 1967). We also calculated an overall value for Cohen’s kappa statistic. The 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=61) of hemlock survey points that provided some indication of how the initial and enhanced trees might perform in this sort of
region. As with the eastern maps, we judged accuracy for both the immediate pixel and within a 22.5-m buffer.

These survey data were the best we could compile given the budget and timeframe for the project. Certainly, we could have utilized more data on the non-evergreen and non-hemlock evergreen classes, especially for the western study area. Nevertheless, we believe we have the necessary information to judge whether either decision tree performed well enough to consider its application to other parts of the southern Appalachian region.

4.4 Results

4.4.1 Decision Trees: General Characteristics

The initial decision tree (Figure 4-2) includes 11 variables, 33 terminal nodes, and extends to a depth of nine levels. Table 4-2 shows a ranking of the variables included in the initial tree. The ranking is based on number of appearances in the tree and the depth of those appearances. Elevation is the most important splitting variable, appearing eight times in the top seven levels of the initial tree. It is followed in importance by three ratios involving short-wave infrared bands of the ASTER image: Band 4/Band 9, Band 4/Band 5, and Band 4/Band 6. Topographic relative moisture index is the only other environmental (i.e., not image-based) variable besides elevation to appear in the top five levels of the initial decision tree. Also noteworthy, distance to closest stream is of negligible importance, appearing only once—in the eighth level of the tree—as part of a cluster of decision rules separating the hemlock secondary component/inclusion class from non-hemlock evergreen.

The enhanced decision tree (Figure 4-3) includes nine variables, 27 terminal nodes, and extends to a depth of ten levels. As in the initial tree, elevation is the most important
splitting variable, appearing a total of nine times—including seven in the top six levels of the tree. Elevation is followed in importance by the Band 4/Band 9 ratio, topographic relative moisture index, and distance to closest stream (Table 4-3). In contrast with the initial tree, distance to closest stream appears twice in the enhanced tree, and at the second and fourth levels, arguably making it second only to elevation in importance as a splitting variable. While the Band 1/Band 2 and Band 4/Band 5 ratios are incorporated, the Band 1/Band 3 ratio is the only image ratio besides Band 4/Band 9 to be included in the tree’s upper levels.

Table 4-2. Ranking of variables in the initial decision tree. Right-hand columns indicate the number of nodes associated with a variable at a given tree level. TRMI= topographic relative moisture index.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variable</th>
<th># of Nodes</th>
<th>Level of Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1  2  3  4  5  6  7  8  9</td>
</tr>
<tr>
<td>1</td>
<td>Elevation</td>
<td>8</td>
<td>1  1  1  1  1  2  1</td>
</tr>
<tr>
<td>2</td>
<td>B4/B9 Ratio</td>
<td>5</td>
<td>1  2  1</td>
</tr>
<tr>
<td>3</td>
<td>B4/B5 Ratio</td>
<td>4</td>
<td>1  1  1  1  1</td>
</tr>
<tr>
<td>4</td>
<td>B4/B6 Ratio</td>
<td>3</td>
<td>1  1  1  1</td>
</tr>
<tr>
<td>5</td>
<td>TRMI</td>
<td>3</td>
<td>1  2</td>
</tr>
<tr>
<td>6</td>
<td>Percent Slope</td>
<td>3</td>
<td>2  1</td>
</tr>
<tr>
<td>7</td>
<td>B1/B2 Ratio</td>
<td>2</td>
<td>1  1</td>
</tr>
<tr>
<td>8</td>
<td>B4/B8 Ratio</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Landform Index</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>Curvature</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Distance to Stream</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4-3. Ranking of variables in the enhanced decision tree. Right-hand columns indicate the number of nodes associated with a variable at a given tree level. TRMI= topographic relative moisture index.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variable</th>
<th># of Nodes</th>
<th>Level of Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1  2  3  4  5  6  7  8  9  10</td>
</tr>
<tr>
<td>1</td>
<td>Elevation</td>
<td>9</td>
<td>1  1  3  1  1  1  1  1</td>
</tr>
<tr>
<td>2</td>
<td>B4/B9 Ratio</td>
<td>4</td>
<td>1  1  1  1  1</td>
</tr>
<tr>
<td>3</td>
<td>TRMI</td>
<td>3</td>
<td>1  2</td>
</tr>
<tr>
<td>4</td>
<td>Distance to Stream</td>
<td>2</td>
<td>1  1</td>
</tr>
<tr>
<td>5</td>
<td>B3/B1 Ratio</td>
<td>2</td>
<td>1  1</td>
</tr>
<tr>
<td>6</td>
<td>Percent Slope</td>
<td>2</td>
<td>1  1</td>
</tr>
<tr>
<td>7</td>
<td>B1/B2 Ratio</td>
<td>2</td>
<td>1  1  1  1</td>
</tr>
<tr>
<td>8</td>
<td>B4/B5 Ratio</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Curvature</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 4-2. Initial decision tree for hemlock classification. Pathways start from top center and terminate when reaching one of the three map classes in question, designated by shape (ellipse, rectangle, hexagon—see legend). Note that this tree only applies to evergreen areas of the image. TRMI = topographic relative moisture index.
Figure 4-3. Enhanced decision tree for hemlock classification. Pathways start from top center and terminate when reaching one of the three map classes in question, designated by shape (ellipse, rectangle, hexagon—see legend). Note that this tree only applies to evergreen areas of the image. TRMI = topographic relative moisture index.
Overall, the enhanced decision tree has a more compact, vertical structure, with maximum branching at the fifth level (five nodes) and a noticeable bottleneck (down to two nodes) at the sixth level. By comparison, the initial tree exhibits significant branching as deep as the sixth (seven nodes) and seventh (six nodes) levels. The two trees have different upper elevation limits for hemlock presence: The initial tree allows for the presence of hemlock secondary component/inclusion up to 1538.65 m under certain conditions, while the enhanced decision tree classifies all areas greater than 1489.9 m as non-hemlock evergreen. Both values are reasonably close to the estimated historical limit (~1500 m) for eastern hemlock in the southern Appalachians (Godman & Lancaster 1990; NatureServe 2005).

Notably, the first split in the enhanced decision tree is at 976.6 m (~3200 ft) elevation—roughly in the middle of the range for hemlocks (Godman & Lancaster 1990). This split may have occurred due to the inclusion of training data from the western study area. In any case, the lower elevation (i.e., less than 976.6 m) sub-tree delineates narrow riparian zones in which hemlock is largely dominant/co-dominant, except on the steepest slopes. In addition, areas further away from streams may be hemlock dominant/co-dominant, but distribution is restricted by environmental factors. Meanwhile, the higher elevation (> 976.6 m) sub-tree is devoted to rules separating the hemlock secondary component/inclusion class from non-hemlock evergreen.

While the branching structure of the initial decision tree is not as clear-cut, many of the same trends are similar to those in the enhanced tree. In particular, hemlock prevalence is generally associated with higher topographic relative moisture index values, lower values of the Band 4/Band 9 ratio, and lower slope values. How the variables interact to create the final classification is quite different; for instance, because distance to closest stream is only a
minor variable in the initial tree, it relies more heavily on variables such as slope, topographic relative moisture index, and even band ratio values to define the riparian (and similar) areas where hemlocks are commonplace.

4.4.2 Comparison of Output Maps

The initial and enhanced decision trees yielded visually similar maps of hemlock distribution for the eastern study area (Figure 4-4). The initial tree classified 14% of the eastern study area as hemlock dominant/co-dominant and 24% as hemlock secondary component/inclusion, while the enhanced tree classified 15% as hemlock dominant/co-dominant and 23% as hemlock secondary component/inclusion. (Both classified 29% as non-hemlock evergreen, with the remaining 33% in non-evergreen based on the cluster-
busted Landsat image.) The most significant difference between the maps was in their spatial representation of the hemlock dominant/co-dominant class. Because the enhanced tree placed greater importance on the distance to closest stream variable, its map exhibited discrete riparian corridors of hemlock presence in portions of the eastern area where the initial tree portrayed a more diffuse distribution. Overall, the two maps’ depictions of the hemlock secondary component/inclusion class are comparable, although the initial tree assigned slightly more territory to this class in western portions of the eastern study area.

Figure 4-5. Hemlock distribution maps for the western study area: a) initial decision tree; b) enhanced decision tree; c) GSMNP map based on photo interpretation.
An examination of the distribution maps for the western study area reveals marked differences in the two classifiers (Figure 4-5a and 4-5b). Most obviously, the initial decision tree classified 71% of the western area as hemlock dominant/co-dominant, while the enhanced tree classified only 31% of the area into the class. In fact, the initial tree classified essentially all but the most upland areas as hemlock dominant/co-dominant. Simple visual comparison of the trees’ maps to the photo-interpreted map for GSMNP (Figure 4-5c) reveals that the enhanced tree generated a far more reasonable depiction of hemlock distribution. Nonetheless, while the enhanced decision tree closely followed the photo-interpreted map with respect to capturing the overall hemlock pattern, it also over-predicted the width of some hemlock riparian zones and omitted the hemlock presence along some minor tributaries. Furthermore, it failed to distinguish some areas delineated as hemlock secondary component/inclusion in the photo-interpreted map, instead lumping all hemlock presence in the western area into the hemlock dominant/co-dominant class.

4.4.3 Accuracy Assessment Results

Given the visual similarity between the maps of the eastern study area generated by the two trees, it is unsurprising that the corresponding error matrices were also similar. The two matrices (Tables 4-4 and 4-5) for “at-point” accuracy—i.e., based only on the pixel in which each survey point actually fell—had similar overall accuracy numbers: 67.5% (95% CI = 60.8% to 74.1%) for the initial tree and 68.9% (95% CI = 62.4% to 75.5%) for the enhanced tree. Their Kappa ($K_{hat}$) values were also similar: 0.484 for the initial tree and 0.501 for the enhanced tree. While these broad measures suggest that the enhanced tree performed slightly better overall, the low producer’s accuracy values for hemlock in both
matrices demonstrate that each tree under-predicted hemlock presence in some instances. The matrices’ identical producer’s accuracies for the non-evergreen class resulted from the cluster-busted Landsat mask, which, because it was applied to each map, dictated how many non-evergreen reference points would be correctly and incorrectly classified. However, the initial decision tree labeled more of the misclassified non-evergreen points as hemlock, which might be a more preferable error pattern than labeling them as non-hemlock evergreen (i.e., we would prefer hemlock over-prediction over any other type of error).

Table 4-4. Error matrix for initial decision tree, accuracy based on immediate pixel.

<table>
<thead>
<tr>
<th>Reference Totals</th>
<th>Hemlock</th>
<th>Non-Evergreen</th>
<th>Non-Hemlock Evergreen</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemlock</td>
<td>75</td>
<td>14</td>
<td>6</td>
<td>95</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>23</td>
<td>35</td>
<td>5</td>
<td>63</td>
</tr>
<tr>
<td>Non-hemlock Evergreen</td>
<td>19</td>
<td>5</td>
<td>29</td>
<td>53</td>
</tr>
<tr>
<td>Total</td>
<td>117</td>
<td>54</td>
<td>35</td>
<td>206</td>
</tr>
</tbody>
</table>

Producer's Accuracy: 64.10 64.81 82.86 67.48 Overall Accuracy

Table 4-5. Error matrix for enhanced decision tree, accuracy based on immediate pixel.

<table>
<thead>
<tr>
<th>Reference Totals</th>
<th>Hemlock</th>
<th>Non-Evergreen</th>
<th>Non-Hemlock Evergreen</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemlock</td>
<td>78</td>
<td>10</td>
<td>6</td>
<td>94</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>23</td>
<td>35</td>
<td>5</td>
<td>58</td>
</tr>
<tr>
<td>Non-hemlock Evergreen</td>
<td>16</td>
<td>9</td>
<td>29</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td>117</td>
<td>54</td>
<td>35</td>
<td>206</td>
</tr>
</tbody>
</table>

Producer's Accuracy: 66.67 64.81 82.86 68.93 Overall Accuracy

Table 4-6. Error matrix for initial decision tree, accuracy based on 22.5-m window.

<table>
<thead>
<tr>
<th>Reference Totals</th>
<th>Hemlock</th>
<th>Non-Evergreen</th>
<th>Non-Hemlock Evergreen</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemlock</td>
<td>104</td>
<td>5</td>
<td>2</td>
<td>111</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>12</td>
<td>48</td>
<td>33</td>
<td>60</td>
</tr>
<tr>
<td>Non-hemlock Evergreen</td>
<td>1</td>
<td>1</td>
<td>33</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>117</td>
<td>54</td>
<td>35</td>
<td>206</td>
</tr>
</tbody>
</table>

Producer's Accuracy: 88.89 88.89 94.29 89.21 Overall Accuracy
Table 4-7. Error matrix for enhanced decision tree, accuracy based on 22.5-m window.

<table>
<thead>
<tr>
<th>Classification Totals</th>
<th>Reference Totals</th>
<th>User's Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemlock</td>
<td>Non-Evergreen</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>Hemlock</td>
<td>104</td>
<td>3</td>
</tr>
<tr>
<td>Non-Evergreen</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>Non-hemlock Evergreen</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>117</td>
<td>54</td>
</tr>
</tbody>
</table>

When looking at accuracy in terms of the 22.5-m window, both trees performed extremely well in the eastern study area (Tables 4-6 and 4-7). Overall accuracy was 89.2% (95% CI = 85.4% to 94.2%) for the initial tree and 90.3% (95% CI = 86.0% to 94.6%) for the enhanced tree. Kappa values were similarly high: 0.827 for the initial tree and 0.837 for the enhanced tree. Perhaps most notably, no producer’s or user’s accuracy value fell below 80% in either matrix. Indeed, both decision trees met or exceeded most reasonable thematic accuracy standards for remote-sensing-derived products (Congalton & Green 1999).

Our limited reference data for the western study area misleadingly suggest a high level of accuracy for the initial decision tree. In terms of at-point accuracy, the initial tree captured 52 of 61 (85.3%) hemlock points. When evaluated within a 22.5-m window, the number of points captured jumps to 100%. Of course, this result is artificially high, because comparison to the photo-interpreted GSMNP map clearly indicates substantial over-estimation of hemlock presence by the initial decision tree. If the western area reference data included non-hemlock evergreen and non-evergreen points, chances are that many of these points would have been misclassified by the initial tree, and as a result, the overall accuracy of the tree would have been quite low.

In contrast, the enhanced tree captured just 19 of 61 (31.2%) hemlock points in the western area when judged in terms of at-point accuracy. Nine (14.8%) of the hemlock points
were misclassified as non-evergreen and 33 (54.1%) as non-hemlock evergreen. After incorporating the 22.5-m window, the number of accurately captured points jumped to 46 of 61 (75.4%). Four (6.6%) points were misclassified as non-evergreen and 11 (18.0%) points as non-hemlock evergreen. The large difference between at-point and within-window accuracies suggests that the riparian areas outlined by the enhanced tree, while emulating the source of the training data (i.e., the photo-interpreted GSMNP map), might be too narrowly defined. Many of the survey points fell just outside the edges of the hemlock riparian zones defined by the enhanced tree, so that they were considered correctly classified when we used the 22.5-m window. Still, 75.4% accuracy is a reasonable result for a remote-sensing-based classification effort (Congalton & Green 1999). Furthermore, this limited analysis of the western area basically amounts to an assessment of producer’s accuracy for the hemlock class. Overall accuracy for the western areas would likely be higher than this single-class value if a full suite of reference data were available.

4.5 Discussion

The hemlock woolly adelgid is estimated to have infested 26% of the hemlock habitat and 25% of the total hemlock basal area in the eastern U.S. (Morin et al. 2005). The most severe impacts have been felt in the Northeast, where the pest has been present in natural stands for roughly two decades (Orwig et al. 2002). Since the adelgid is a relatively recent arrival in the southern Appalachians, there is still a chance to preserve the region’s hemlock resource with appropriate management, and a better picture of the host species’ distribution is an integral part of the solution. Even if the hemlock woolly adelgid cannot be successfully controlled in the short term, better mapping of hemlock stands in the region may help
eventual restoration efforts. For example, researchers may be able to identify remote stands that can be sampled for *ex situ* conservation or signs of host resistance (Esham et al. 2005; Tighe et al. 2005).

Thus, there is a clear need for a hemlock mapping protocol that can be quickly pressed into service in the southern Appalachians. We believe that our enhanced decision tree, when combined with careful masking of non-evergreen areas, is a good candidate for this protocol. Like any approach, it has limitations that must be considered critically. In particular, the primary concerns are whether the approach is accurate and general enough to justify its application elsewhere in the southern Appalachian region.

### 4.5.1 Accuracy

Our enhanced tree displayed a lot of promise, yielding accurate maps for two very distinct parts of GSMNP: 90% overall accuracy for the eastern area map, and 75% accuracy in capturing hemlocks for the western area map, based on occurrence with a 22.5-m window. While our use of the 22.5-m standard may seem lenient, we believe the values are more realistic than the at-point accuracy values for the maps, which did not account for positional errors in the input and reference data or the small, isolated nature of hemlock stands (Godman & Lancaster 1990; Jensen 1996). Judging accuracy based on a majority rather than simple occurrence within the window is similarly impractical because of the nature of hemlock stands. Admittedly, all of our accuracy numbers pertained to a single hemlock class, rather than the separate dominant/co-dominant and secondary component/inclusion classes delineated by the tree. If we had reference data to assess these classes separately, overall accuracy might have been lower due to confusion between the two. Nevertheless,
any decline in accuracy due to confusion between hemlock classes is less significant to our objectives than misclassifying hemlocks as non-evergreen or non-hemlock evergreen.

While the numbers were generally good, there was a disparity in accuracy between the two study areas exceeding any possible effect of our limited reference data for the western area. The disparity may be partially explained by differences in hemlock spatial pattern between the study areas and subsequent limitations of the training data. For instance, hemlock is less dominant in the western part of GSMNP than elsewhere in the park, so it is more likely to be found in small inclusions rather than large, continuous stands (Callaway et al. 1987). The photo-derived GSMNP vegetation map, however, was developed with a nominal minimum mapping unit of 0.5 ha (Welch et al. 2002), meaning that these small clusters of hemlocks were not delineated and thus did not serve as training data for the enhanced decision tree. Similarly, the photo-derived map mostly restricts hemlocks to narrow riparian zones in the western study area. Though represented as distinct lines in the photo-derived map, the borders of these hemlock riparian areas are likely to be fuzzy on the ground; moreover, when delineating the boundaries of such areas from aerial photographs, there is a tendency to underestimate the boundaries in a downslope direction (Barrette et al. 2000; Goodchild 1994). This may have translated to the training data, affecting threshold values for distance to closest stream and other variables.

Other factors may have affected accuracy for the western area map. The elevation range of the area is small (267 to 952 m), thus limiting the utility of a variable that—in the eastern study area—greatly aided distinction of certain species (e.g., red spruce from eastern hemlock). Of particular note, the western area falls below the higher-elevation “belt” where hemlock is dominant elsewhere in GSMNP (Whittaker 1956). Western area hemlocks,
rather, are distributed throughout a heterogeneous forest containing a variety of evergreen species, including white pine (*Pinus strobus*) and various yellow pines (Callaway et al. 1987). While these species are spatially segregated to a certain degree by moisture status (Whittaker 1956), moisture differences between sites are subtle in the relatively gentle topography of the western area, making separation difficult.

Some of our classification errors may have been due to limitations in the input data—although better data may not be worth the potential tradeoffs. For example, the 15-m spatial resolution of the ASTER imagery is not ideal for distinguishing hemlocks in cases where they are commingled with other trees, especially evergreens. While we used ancillary data to help make sub-pixel-level distinctions, they could not prevent all of our classification errors. Better resolution imagery might reduce such errors by increasing within-pixel homogeneity, but would cost considerably more and present a new set of classification problems (Jensen 1996). In another example, the application of cluster busting to mask out non-evergreen areas was an important component in the analysis. By first removing all non-evergreen areas from consideration, we were able to develop our enhanced decision tree by focusing only on the distinction between a small set of evergreen classes. Cluster busting of the leaf-off Landsat image yielded a non-evergreen/evergreen map that was accurate by remote sensing standards, i.e., greater than 85% overall accuracy (Congalton & Green 1999). However, this also meant that, when we used the map to mask non-evergreen areas, ~15% of pixels were erroneously included in or excluded from the decision tree step based on this mask. This factor largely accounts for the hemlock reference points mistakenly classified as non-evergreen—the most common classification error in the eastern study area. Unfortunately, it
may be difficult to develop a better non-evergreen mask without employing data of a higher spatial resolution, which could substantially raise data costs and/or processing time.

Ultimately, the enhanced decision tree is probably as accurate as possible given our input data. Certainly, we could have added other environmental variables (e.g., soils) that may have been relevant. We omitted a number of these due to a lack of consistent regional availability and concerns about their accuracy, as well as a sense that the included variables already incorporated the most critical attributes affecting hemlock distribution.

4.5.2 Generalization

We have ignored our initial decision tree in this discussion, basically because it over-estimated hemlock presence in the western study area due to a lack of training data. Anticipating this problem, we used two study areas to train and prune our enhanced tree, but these training data still may not adequately represent the southern Appalachian region. Thus, it is possible that the enhanced tree will not generalize well to locations outside GSMNP. On the other hand, the tree’s rules appear to logically follow region-wide trends for hemlocks (Delcourt & Delcourt 2000; Godman & Lancaster 1990; McWilliams & Schmidt 1999). Essentially, the rules describe a shift of hemlock from riparian areas and valley flats at low elevations to protected slopes and ridges at high elevations (Whittaker 1956). At lower elevations, distance to the closest stream is the major distinguishing variable, although hemlock may be dominant/co-dominant outside these zones under the right conditions of slope, topographic relative moisture index, and image ratio values. The tree allows hemlock dominance/co-dominance up to 1236 m in elevation given appropriate Band 4/Band 9 ratio values. Above 1236 m, hemlock may be present as a secondary component/inclusion in a
variety of sites, dictated mostly by topographic relative moisture index and image ratio values.

The exact contributions of the image band ratios to the enhanced tree are harder to interpret. Some pixel values reflect the presence of canopy-level hardwoods, because the masking process did not eliminate pixels with a hardwood-dominated overstory but dense evergreen understory. Indeed, one of the reasons we used leaf-on imagery was to exploit associations between hemlocks and certain hardwood tree species, while reducing confusion between hemlocks and two widespread shrubs, mountain laurel (*Kalmia latifolia*) and rhododendron (*Rhododendron maximum*). Looking at the tree’s two most important image ratios—and regardless of exactly what species defined the pixel values—hemlock presence correlates with higher values for the Band 3/Band 1 ratio and lower values for the Band 4/Band 9 ratio. Overall, it makes sense for these bands to figure prominently in vegetation discrimination: Band 1 corresponds to the green reflectance of vegetation, Band 3 covers wavelengths responsive to vegetative biomass, and Bands 4 and 9 span wavelengths associated with leaf moisture content (ERSDAC 2003a; Jensen 1996). The threshold values for these ratios may represent subtle distinctions between vegetation types, meaning that any application of the tree in other parts of the region depends on consistency between the ASTER images used to train the tree and all future ASTER imagery. This should not be an issue if the images are acquired with atmospheric correction and topographically normalized prior to classification.

We should note that Carolina hemlock does not occur in GSMNP (Taylor 2002), so it could not be incorporated in the tree construction process. Because Carolina hemlock occupies a slightly different niche than eastern hemlock, it is possible the enhanced tree fails
to address conditions of particular relevance for Carolina rather than eastern hemlock. On the other hand, the range of Carolina hemlock is completely overlapped by eastern hemlock’s range. We suspect this will have no more than a minor effect on accuracy for region-wide application.

Despite our best efforts, the enhanced decision tree may be over-fitted. This can be corrected by manual pruning; indeed, the bottleneck at the sixth level of the enhanced tree may be a logical place to prune the tree and generalize it even further. Certainly, we could continue testing the tree using additional GSMNP data, changing the threshold values and adding or deleting variables before releasing a final version. However, given the immediacy of the hemlock woolly adelgid threat, it seems like the best course of action is to apply our hemlock mapping approach in other parts of the southern Appalachian region. Feedback from these applications might suggest further refinement. Any progress towards a regional map of hemlock distribution would certainly be advantageous in the current circumstances.

We would like to emphasize that our hemlock mapping approach is just one component of a system for managing the hemlock woolly adelgid in the southern Appalachians. In a separate analysis, we have been working on GIS-based models to predict the locations in the southern Appalachian region that are at the greatest risk of imminent hemlock woolly adelgid infestation. These models yield probability maps that can be used to rank areas at the highest risk of infestation. By overlaying these risk probability maps with decision-tree-based maps of hemlock distribution, forest managers can target specific areas for their control efforts. This will substantially reduce the territory they must cover, allowing them to allocate scarce management resources more quickly and efficiently.
4.6 Conclusions

To enable better management of the hemlock woolly adelgid, we developed an approach for mapping hemlock distribution in the southern Appalachians. The approach has two components: (1) masking of non-evergreen areas using leaf-off satellite imagery; and (2) separation of hemlock from other evergreens using a decision tree classifier. The decision tree is built on available or easily calculable spatial variables and moderate-resolution satellite imagery. We trained the tree and assessed the success of our approach using data from two different study areas in Great Smoky Mountains National Park. Overall accuracy of the classifier appears to be quite good given the relatively coarse nature of the input data. Furthermore, the approach appears to be general enough that it should be applicable throughout the southern Appalachian region. Given the current peril that the southern Appalachian region faces from continued hemlock woolly adelgid infestation, we recommend that our approach be rapidly applied region-wide to help forest managers prioritize hemlock stands for hemlock woolly adelgid control measures.

4.7 References


148


Leica Geosystems. 2001. ERDAS Imagine 8.7 software. Atlanta, GA: Leica Geosystems GIS & Mapping, LLC.


152

5. LANDSCAPE-LEVEL PREDICTION OF HEMLOCK WOOLLY ADELGID INFESTATION IN THE SOUTHERN APPALACHIANS

Abstract: After causing substantial mortality in the northeastern U.S. during the last two decades, the insect pest hemlock woolly adelgid \textit{(Adelges tsugae)} has recently made inroads into the southern Appalachian region. Neither of the region’s native hemlocks \textit{(Tsuga canadensis, T. caroliniana)} is resistant, and infestations can kill trees within four years. Although general estimates of regional hemlock woolly adelgid spread do exist, landscape-level dynamics of invasion are not well understood—in particular, factors predicting where the pest is likely to first infest a landscape. We examined first-year hemlock woolly adelgid infestation locations from Great Smoky Mountains National Park and the Blue Ridge Parkway to identify possible factors. For 84 infested and 67 uninfested sites, we calculated values for a suite of topographic, environmental, and proximity variables using a geographic information system. We randomly sorted these sites into five 101-point training and 50-point validation sets. After identifying the most significant variables, we applied four classification techniques—discriminant analysis, $k$-nearest neighbor analysis, logistic regression, and decision trees—to derive functions separating the infested and uninfested groups. We then used the resulting classification functions to generate maps of hemlock woolly adelgid infestation risk in the Great Smokies. Three proximity variables (distance to the closest stream, trail, and road) were incorporated in all four classification functions, which performed well in terms of mean error rate (0.16-0.23 for training data sets, 0.14-0.24 for validation data sets). Depending on approach, the derived infestation probability maps placed between 18.5\% and 34.6\% of the park’s hemlock areas at high risk (prob. $\geq 0.50$) of imminent infestation. Discriminant analysis was the most accurate and classified the least area as high infestation risk, but logistic regression best balanced accuracy, efficiency, and interpretability. In any case, our results suggest that roads, major trails, and riparian corridors provide an important degree of connectivity enabling long-distance dispersal of the hemlock woolly adelgid, probably by humans or birds. Notably, given adequate spatial data, any of the derived classification functions can be used to create hemlock woolly adelgid infestation risk maps for elsewhere in the southern Appalachian region. This would allow forest managers to better target monitoring and control efforts.

Key Words: hemlock woolly adelgid, invasive, southern Appalachians, prediction, landscape connectivity

5.1 Introduction

Non-indigenous invasive insects are considered a critical threat to North American forest ecosystems. Nearly 400 non-indigenous insect species have invaded American forests since the arrival of Europeans, of which roughly 30\% are now significant pests (Liebhold et al. 1995; Pimentel et al. 2000). Invasive insects change forest composition, structure, and
microenvironment, alter critical ecosystem processes, and increase susceptibility of areas to further invasion and disturbance (Orwig 2002). Forest product losses and control costs due to non-indigenous insects equal roughly $2.1 billion per year (Campbell 2002). While this is a small percentage of the $137 billion lost annually due to all non-indigenous species activity in the U.S., it is not a trivial amount. Moreover, it does not include certain impacts that are difficult to measure economically, such as loss of habitat for already-imperiled native species (Pimentel et al. 2000).

Our current framework for dealing with non-indigenous pests tends to focus on points early in the cycle of invasion: quarantine efforts to prevent species from successfully invading, and eradication if a species manages to invade a few locations but has not become well established (Elton 1958). If an invasive species does become established, the only option is control: attempting to manage the pest so that its impact is minimized (Elton 1958). This is a much more difficult proposition, because control measures can quickly become expensive and laborious as a pest expands its range. Therefore, effective control of a non-indigenous pest requires some understanding of its spread in order that control measures can be efficiently targeted. Such understanding is particularly important in cases where an invasive pest, if unchecked, may cause substantial damage either immediately or within a few generations of its arrival (Pedgley 1993).

The hemlock woolly adelgid (Adelges tsugae Annand) represents this type of threat to southern Appalachian forests. An Asian insect that was first introduced to Richmond, Virginia in the 1950s, the adelgid was considered little more than a nuisance to ornamental hemlock trees in home landscapes for a couple of decades (Cheah et al. 2004; McClure et al. 2001). Unfortunately, when the hemlock woolly adelgid moved into natural stands during
the 1980s, it spread rapidly, and has since resulted in extensive hemlock mortality in the northeastern and mid-Atlantic U.S. (Cheah et al. 2004).

The hemlock woolly adelgid attacks both of the southern Appalachian region’s native hemlock species, eastern (Tsuga canadensis) and Carolina hemlock (T. caroliniana), and neither is resistant (McClure et al. 2001). Both species occupy unusual ecological niches that will not be easily filled by other (mostly hardwood) tree species (Orwig & Foster 1998; Ward et al. 2004). The hemlock woolly adelgid quickly causes severe needle loss and then dieback, often resulting in mortality within a few years of infestation (McClure et al. 2001). It attacks all age classes, including the youngest juveniles (Orwig & Foster 1998; Ward et al. 2004). It has two reproductive generations per year on hemlock, supporting explosive population growth after arriving at a site (Cheah et al. 2004). It is also parthenogenetic (only females occur), such that the introduction of a single individual can theoretically lead to an infestation (McClure & Cheah 2002). The pest has no natural enemies in the U.S., and chemical control is unfeasible for natural hemlock stands due to the high cost (Wallace & Hain 2000; Ward et al. 2004).

The hemlock woolly adelgid is passively dispersed by a wide range of possible vectors: wind, deer, birds, and humans (McClure 1990). Estimates for long-distance dispersal range as high as 30 km per year (Cheah et al. 2004; Evans 2004; McClure 1996; Souto et al. 1996; Ward et al. 2004). Typically, these dispersal estimates have been based only on coarse scale data—essentially, the first year the pest was detected and recorded in a county—and the spatial pattern of hemlock woolly adelgid dispersal is represented with the same coarseness (e.g., Evans 2004; USDA-FS 2002). The lack of resolution is perhaps unsurprising: Low-level adelgid populations can be difficult to detect, and conversely, it is
particularly difficult to prove that a stand has *not* been infested (Costa 2005). As a result, forest managers often assume a region is saturated by the adelgid once the pest has been detected, although distribution is unlikely to be uniform (Evans 2005). In any case, finer-scale factors influencing distribution have generally been ignored (Orwig et al. 2002).

![Figure 5-1. Hemlock woolly adelgid range expansion in the southern Appalachian region.](image)

An estimated 26% of hemlock habitat and 25% of hemlock basal area in the U.S. has already been invaded by the hemlock woolly adelgid (Morin et al. 2005), meaning a large percentage of hemlock stands have not been invaded yet (Figure 5-1). Furthermore, the pest is a fairly recent arrival in the southern Appalachian region. Although Figure 5-1 indicates many of the region’s counties have already been invaded by the hemlock woolly adelgid, infestation of individual hemlock stands is far from absolute. Thus, there are still many areas for which management measures are relevant. Biological control has been embraced as the most effective approach for on-the-ground hemlock woolly adelgid management (Cheah et
al. 2004; Orwig et al. 2002). However, predators and other control agents take time and money to establish in a new setting, and this must happen before a hemlock stand is irreversibly damaged by the pest (Cheah & McClure 2002). For these reasons, any method to prioritize sites for management would be extremely advantageous.

Therefore, our chief research objective was to develop a site prioritization method that incorporates the landscape-level factors influencing hemlock woolly adelgid spread. Circumstances in the southern Appalachian region granted us a unique opportunity. Because we were able to procure data depicting specific geographic locations where the pest was first detected in the region, we could apply statistical classification techniques to (1) identify key landscape variables distinguishing infested and non-infested sites, and (2) build functions based on these variables to predict the sites most likely to be infested. By implementing these functions in a geographic information system (GIS), we hoped to create accurate maps of infestation risk. These risk maps, when combined with detailed maps of host species range, could help forest managers target their hemlock woolly adelgid control efforts by significantly narrowing their area of focus.

5.2 Statistical Classification Techniques

Discriminant analysis, logistic regression, and decision tree approaches may all be applied to spatially structured classification problems. The basic aim of these multivariate techniques is to classify a given observation into one of two (or more) alternative groups based on measurements for a suite of variables (Afifi et al. 2004). All three techniques require an initial sample where the group designation of each observation is known \textit{a priori}. The initial sample then serves as a training data set for determining the parameters of a function which can be used to predict the group membership of additional data points (Afifi
et al. 2004; McLachlan 1992). While superficially similar, the techniques differ in approach, particularly in the assumptions under which they operate.

5.2.1 Discriminant Analysis

Discriminant analysis is widely used in medical diagnosis and pattern recognition (Khattree & Naik 2000; McLachlan 1992). The approach identifies “discriminants”—variables with values that best segregate the input observations into distinct groups—and then constructs a function based on these discriminants for optimally allocating new observations into the labeled groups (Johnson & Wichern 2002). Essentially, discriminant analysis works by calculating the statistical distance between an observation—represented as a vector of variable values—and the mean vector of each group, and then assigning the observation to the closest group (Nelson et al. 2003). The most commonly used distance measure is (squared) Mahalanobis distance, which incorporates the covariance of the observations (Johnson & Wichern 2002; Nelson et al. 2003). The appropriate form of discriminant function depends on the covariance matrices of the groups. Linear discriminant analysis assumes that the matrices are equal and constructs a function using the pooled covariance matrix. If the groups’ covariance matrices are unequal, quadratic discriminant analysis is the more appropriate choice (Johnson & Wichern 2002; Nelson et al. 2003).

Discriminant analysis is limited to continuous explanatory variables, and assumes the input data have a multivariate normal distribution (Afifi et al. 2004; Johnson & Wichern 2002; Nelson et al. 2003). Though approximate normality may suffice for linear discriminant analysis, the quadratic form is particularly sensitive to departures from multivariate normality (Afifi et al. 2004; Fung 1996; Johnson & Wichern 2002). There are
numerous methods to test for normality and to transform variables to near-normality if necessary (Johnson & Wichern 2002; Khattree & Naik 1999).

In cases where multivariate normality cannot be established, non-parametric discriminant analysis is an alternative. To cite one example, $k$-nearest neighbor analysis looks at the distance (based on variable values) from an observation to its $k$ closest neighbors, and assigns the observation to the group containing the majority of its neighbors (Jovanovic et al. 2002; Khattree & Naik 2000). It is a good choice when the statistical surface separating classes is complex; however, the approach can be sensitive to data scaling (Mazzatorta et al. 2004). Notably, a non-parametric approach offers no improvement over linear or quadratic methods if the training sample turns out to be multivariate normal (Khattree & Naik 2000).

5.2.2. Logistic Regression

Logistic regression is a flexible method for handling both continuous and categorical data that does not assume normality (Afifi et al. 2004). The approach is a generalization of linear regression for fitting a model to a response variable with two (or occasionally more) qualitative outcomes—e.g., membership in either an infested or uninfested group (Allison 1999; Jovanovic et al. 2002; Neter et al. 1996). Logistic regression relies on the concept of the odds ratio, which is

$$\frac{\text{probability of an outcome occurring}}{\text{probability of an outcome not occurring}} = \frac{p}{1-p} \quad \text{(Allison 1999)}.$$

Instead of fitting a linear model to predict a response variable based on explanatory variables, logistic regression fits a model to predict the logarithm of the odds that one of the qualitative outcomes will occur (Jovanovic et al. 2002). The logistic regression (or logit) model for $n$ explanatory variables is
\[
\ln \left[ \frac{p}{1-p} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n,
\]
where \( p \) is the probability of one of the qualitative outcomes being true (Afifi et al. 2004; Allison 1999). The equation returns a value between 0 and 1, and a threshold value (e.g., 0.5) can be used to assign an observation to one particular outcome or another.

While logistic regression has few assumptions, sample size bears on the number of variables that can be realistically included: Peduzzi et al. (1996) suggested no fewer than ten observations per model parameter. Logistic regression can be sensitive to sample bias, outliers, and multicollinearity (Afifi et al. 2004; Allison 1999). Furthermore, the overall assumption that the response is linear in the coefficients of the explanatory variables should be checked using goodness-of-fit measures (Afifi et al. 2004; Jovanovic et al. 2002).

5.2.3 Decision Trees

The decision tree concept includes a number of algorithms for automated classification of multivariate data (Murthy 1998). The basic mechanism is nearly universal: Tree building occurs in a top-down fashion, starting with an empty tree and a training data set (Murthy 1998). The training data are split into mutually exclusive groups that are as homogeneous as possible, and then (if possible) each of these groups is split independently, and so on until the terminal groups, or “nodes”, exhibit maximum possible homogeneity (De’Ath & Fabricius 2000). Typically, each split is a decision rule employing a threshold value for one of the input variables (De’Ath & Fabricius 2000; Murthy 1998). Decision trees are often represented graphically with the root node—the unsplit training data—at the top, followed by the branches and then the terminal nodes (De’Ath & Fabricius 2000; Murthy 1998).
Decision tree algorithms use a variety of splitting criteria. The classification and regression tree (CART) algorithm employs the Gini index, a measure of the relative purity of partitioned subgroups relative to their parent group (De’Ath & Fabricius 2000; Gehrke 2000; Murthy 1998). The Chi-Squared Automatic Interaction Detection (CHAID) algorithm uses the $\chi^2$ statistic for distance measurement. There are other noteworthy distinctions between algorithms. For example, CART is limited to binary (i.e., two-way) splits, while CHAID allows for an $n$-way number of partitions (Jovanovic et al. 2002; Nelson et al. 2003).

Decision tree approaches are non-parametric, and so are not tied to a normal or other population distribution (Johnson & Wichern 2002; Murthy 1998). They can accommodate both categorical and continuous variables, and the rules of a decision tree are generally more straightforward to interpret than the coefficients of a discriminant function or logistic equation. However, they often require large training samples for accurate classification, and large samples can lead to intricate trees even when fundamental relationships among variables are quite simple (Jovanovic et al. 2002). To avoid over-fitting, the resulting trees must be pruned carefully. Breiman et al. (1984) introduced the cost-complexity pruning approach, which is probably the most widely accepted method (Li et al. 2001). This method proceeds by building nested, increasingly smaller trees from the training data based on assessment of the per-leaf accuracy cost. Then, one of these smaller trees is chosen based on its accuracy at classifying the validation data set (Li et al. 2001; Murthy 1998).

5.2.4 Choice of Techniques

A second objective of our analysis was to identify the best classification technique for predicting hemlock woolly adelgid risk. Each of the techniques has been used with some
success in similar cases. For example, discriminant analysis has been applied in the Chiapas region of Mexico for predicting the landscape pattern of malaria (*Plasmodium* sp.) spread due to mosquito (*Anopheles albimanus*) presence (Beck et al. 1994, 1997). Logistic regression has been employed to model the landscape-scale movement of cinnamon fungus (*Phytophthora cinnamomi*) in Australia (Wilson et al. 2003) and to predict southern pine beetle (*Dendroctonus frontalis*) outbreaks in the southeastern U.S. (Gumpertz et al. 1999). Decision tree approaches have been used to model the landscape-scale spread of the sudden oak death pathogen (*Phytophthora ramorum*) in California (Kelly & Meentemeyer 2002) and invasive *Pinus* species in southern Africa (Rouget et al. 2004). The $k$-nearest neighbor approach has been used to predict habitat distribution based on environmental variables (Liu et al. 2003) and has been adapted for identifying geographic clusters of disease occurrence (Jacquez et al. 2005).

None of the methods consistently outperforms the others in all circumstances (Murthy 1998). For instance, logistic regression may outperform discriminant analysis in cases of non-normality, although not by much; indeed, Press and Wilson (1978) suggested their accuracies for a given data set will rarely differ substantially. The $k$-nearest neighbor approach may compete with discriminant analysis or logistic regression, but careful attention must be paid to the number of included neighbors (Raudys & Jain 1991). A decision tree approach might be more flexible than either discriminant analysis or logistic regression, but it can provide high classification accuracy at the cost of a high rate of false positives (e.g., Worth & Cronin 2003). Similarly, any of the methods might accurately capture one group of interest but poorly capture another (e.g., Reichard & Hamilton 1997). Given that none is definitively better, we decided to test all four approaches—discriminant analysis, $k$-nearest
neighbor, logistic regression, and decision tree—using the hemlock woolly adelgid data and compare their results in terms of accuracy, efficiency, and interpretability.

![Map of Great Smoky Mountains National Park (GSMNP) and southern Blue Ridge Parkway (BRP).](image)

Figure 5-2. Great Smoky Mountains National Park (GSMNP) and southern Blue Ridge Parkway (BRP).

### 5.3 Study Area

We performed our analyses using data from Great Smoky Mountains National Park (GSMNP) and the Blue Ridge Parkway (BRP), two national park units in the southern Appalachian region (Figure 5-2). GSMNP is large (> 2000 km²), with a significant eastern hemlock presence in many areas (Taylor 2002; Whittaker 1956). The hemlock woolly adelgid was first detected in GSMNP in 2002. BRP is basically a narrow road corridor connecting a number of small recreation areas. Both eastern and Carolina hemlocks occur within its borders (notably, the latter species has never been documented in GSMNP). Key
BRP sites such as Linville Falls have a substantial hemlock presence. The hemlock woolly adelgid was first detected in the southern portion of BRP (i.e., in North Carolina) in 2003.

5.4 Methods

Except where noted, we performed all statistical analyses using SAS 9.1 statistical software, including the SAS Enterprise Miner Release 4.3 module for building decision trees (SAS 2003). We used ArcGIS 8.3 software (ESRI 2002) for GIS operations, most notably for the generation of infestation risk maps.

![Figure 5-3. First-year hemlock woolly adelgid infestation sites in GSMNP.](image)

5.4.1 Data Preparation

After the hemlock woolly adelgid was discovered in GSMNP, park staff undertook an unsystematic but extensive survey during the next few months, identifying 67 distinct infestation locations (Figure 5-3). They recorded these locations with recreational-grade (or
occasionally mapping-grade) global positioning system (GPS) units and converted the locations into a GIS point layer. We believe this layer to be a reasonable representation of the park-wide distribution of the hemlock woolly adelgid during the first “wave” of infestation.

GSMNP also has a recent vegetation map, created from 1:12,000 aerial photographs, that depicts the distribution of hemlock stands in the park (Welch et al. 2002). For comparison to the infested sites, we used an ArcGIS macro (Sawada 2002) to randomly select 67 points from GSMNP hemlock stands. These points represented locations considered uninfested by the adelgid during the first wave. While it is admittedly difficult to detect hemlock woolly adelgid presence at low densities, we believe our assumption that these sites were uninfested is reasonable given the extent of the staff’s surveys in the months after the pest was first detected.

Park staff in BRP similarly used GPS to record the location of 17 distinct hemlock woolly adelgid infestations in the months immediately after first detection. Most of these locations were clustered in a few areas, such as Linville Falls, Doughton Park, and near Tunnel Gap (Figure 5-2). As with GSMNP, the GPS-derived point layer is believed to be an accurate representation of adelgid distribution for the southern portion of BRP during the first wave of infestation. BRP does not have a map of hemlock stand distribution, so we were unable to extract any points representing uninfested hemlock stands.

For each of the 134 points in GSMNP and 17 points in BRP, we recorded values for a suite of landscape-level variables (Table 5-1). We chose variables that were readily available or easily calculable from existing GIS data layers, yet potentially important to hemlock woolly adelgid distribution. For example, we chose topographic variables because of
evidence that terrain characteristics may influence the rate of hemlock decline (Orwig et al. 2002; Royle & Lathrop 1999; Ward et al. 2004). We hypothesized that this was because topography predisposed certain sites to earlier and/or more intense infestations, or rather, acted as a barrier temporarily preventing the infestation of certain sites. We also included variables depicting proximity to streams, roads, and trails because all have been implicated as potential corridors of introduction for the hemlock woolly adelgid or other invasive pests (Forys et al. 2001; Maelzer et al. 2004; McClure 1990; Ward et al. 2004; Zobel et al. 1985). We did not include patch-level characteristics because the size and hemlock percentage of patches was extremely variable: Some infestations occurred in mixed forest stands where hemlock is only an inclusion, while others occurred in hemlock-dominated stands.

<table>
<thead>
<tr>
<th>Table 5-1. Variables tested for inclusion in classification efforts.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>Aspect</td>
</tr>
<tr>
<td>Curvature</td>
</tr>
<tr>
<td>Elevation</td>
</tr>
<tr>
<td>Landform Index</td>
</tr>
<tr>
<td>Percent Slope</td>
</tr>
<tr>
<td>Topographic</td>
</tr>
<tr>
<td>Relative Moisture Index</td>
</tr>
<tr>
<td>Distance to Stream</td>
</tr>
<tr>
<td>Distance to Road</td>
</tr>
<tr>
<td>Distance to Trail</td>
</tr>
<tr>
<td>Trail Type</td>
</tr>
<tr>
<td>Disturbance History</td>
</tr>
<tr>
<td>Geology</td>
</tr>
<tr>
<td>Vegetation</td>
</tr>
</tbody>
</table>
We extracted all topographic variables using 10-m resolution digital elevation models (DEM) of the two parks, edited and refined by park staff from National Elevation Dataset files. Although each infestation was recorded as a point, it more realistically represented an area of infested hemlock trees. To address this—and also to accommodate the limited positional accuracy of the GPS units (± 15 m) used to record the points—we derived mean topographic values based on a 3×3 window of pixels around each point. Because aspect is a circular measure, calculation of a neighborhood mean is not a straightforward operation in GIS. So, we derived sine and cosine grids from the aspect grid, processed both using a 3×3 neighborhood mean filter, and constructed a mean aspect surface via a conditional operation on the mean sine and cosine grids. For computational purposes, we converted the mean aspect surface into a “northeast-ness” surface that re-scaled aspect values (range 0-2) based on their deviation from 45 degrees (Beers et al. 1966). We then assigned each sample point the northeast-ness value for the pixel in which it fell.

The GIS database for GSMNP includes detailed (typically 1:24,000 nominal scale) vector data on streams, roads, and trails, as well as raster data layers for disturbance history, geology, and vegetation. We calculated the straight-line distance between each GSMNP sample point and the closest stream, road, and trail feature. We also labeled each point with the type classification (unimproved foot path, old roadbed, or paved path) of the closest trail. For the disturbance history and geology layers, we assigned each sample point the value of the pixel (90-m resolution) in which it fell. We also labeled each sample point with a vegetation class based on the class exhibited by a majority of the vegetation map’s pixels (30 m) in a 3×3 window around the point. We included this variable to test whether the
surrounding vegetation type might predispose a site to early hemlock woolly adelgid infestation.

While BRP does have a 10-m DEM, it does not have data layers for any of the categorical variables included in the GSMNP sample, nor does it have extensive vector data. To generate proximity variables, we compiled stream, road, and trail vector data from U.S. Geological Survey 1:24000 digital line graphs for the BRP region, supplemented them with the limited data available from the park, and calculated straight-line distances between each sample point and the closest stream, road, and trail feature.

5.4.2 Initial Analyses

We assessed univariate normality of the continuous variables based on Shapiro-Wilk tests (Johnson & Wichern 2002). All except topographic relative moisture index were non-normal, though elevation and landform index were approximately normal based on examination of normal Q-Q plots. For the other continuous variables, we used the BOXCOX macro in SAS to identify and apply appropriate Box-Cox transformations (SAS 2003). After transformation, we used another SAS macro to test multivariate normality of each group (i.e., infested and uninfested) based on skewness and kurtosis (Mardia 1974).

To identify possible collinearity or redundancy of variables, we generated a correlation matrix for all data points. We immediately eliminated curvature from further analysis because it strongly correlated ($r=0.99$) with landform index. While topographic relative moisture index was negatively correlated with the landform index ($r=-0.60$) and positively correlated with northeast-ness/aspect ($r=0.49$), we included all three variables in our initial analyses, anticipating that variable selection procedures would identify which, if
any, should be included in the final classification functions. No other variables exhibited strong correlations.

A classification function derived from a single training sample may perform well on initial data yet poorly classify subsequent data sets. At a minimum, a validation sample is typically held out from the original data set as an additional test of the function’s accuracy (Johnson & Wichern 2002). We partitioned the GSMNP data set by randomly selecting 45 (67%) of the infested points and 45 of the uninfested points as training data, setting aside the remaining 22 points from each group as validation data. We similarly partitioned the BRP data set by randomly choosing 11 points as training data and setting aside the remaining six points as validation data. We merged these samples to create a 101-point (56 infested, 45 uninfested) training sample and 50-point (28 infested, 22 uninfested) validation sample. Johnson & Wichern (2002) suggested that, especially for variable selection, it is advantageous to generate multiple batches from the original data and compare the output of each batch. Following this line of thought, we repeated the random partitioning process four more times, yielding five differently partitioned training/validation data sets (i.e., each data set had a random, unique combination of 101 training and 50 validation observations). We tested each of the training data sets for multivariate normality to see if they substantially differed from the original data set.

The inclusion of a large number of variables does not necessarily increase classification success, and in fact can be problematic with small sample sizes (Johnson & Wichern 2002; Khattree & Naik 2000). Therefore, it is a common practice to select a variable subset that contains nearly as much information as the full set (Johnson & Wichern 2002). We applied three different variable selection procedures, each chosen to create a
subset of variables for one of our classification techniques. First, we performed stepwise discriminant analysis using the continuous variables. Stepwise discriminant analysis generates a sequence of models that adds or removes variables based on F-tests of significance, ultimately resulting in a “best” subset of variables (Khattree & Naik 2000). We chose p-value < 0.25 as our significance threshold—a fairly permissive standard—to ensure that no important variables were omitted from future analyses (Khattree & Naik 2000).

While stepwise selection is also used for logistic regression, this process has been criticized for resulting in too few variables for successful prediction (Shtatland et al. 2001, 2003). Instead, we followed the procedures of Shtatland et al. (2001): We built a sequence of reduced logistic regression models and selected the model (i.e., variable subset) that minimized Akaike’s Information Criterion (AIC), a measure that incorporates a penalty for model complexity. To keep the analysis simple, we did not test for any variable interactions. Because the BRP data did not include values for the categorical variables, we first performed the AIC analysis using only the GSMNP data to see if any of the categorical variables were potentially significant. We then performed a second AIC analysis for the combined GSMNP and BRP data.

Similarly, we built an initial decision tree using just the GSMNP data to identify any important categorical variables, and then built a second tree with the combined data. We employed SAS Enterprise Miner for tree construction, adopting an approach similar to the classification and regression tree (CART) algorithm; in particular, we used the Gini impurity index as our splitting criterion (Breiman et al. 1984). Instead of a formal variable selection procedure as with discriminant analysis and logistic regression, we used the validation data
set to “prune” the tree, i.e., we used the validation data to identify the best sub-tree and, thus, the best subset of variables.

We applied the three variable selection procedures to our five partitioned training data sets. For each procedure, if a variable was identified as significant for at least three of the five partitioned data sets, we highlighted it for inclusion in a final variable subset. The variable selection procedures yielded three variable subsets, each specifically matched to one of our classification techniques. We employed the variable subset identified by stepwise discriminant analysis for our discriminant analysis and $k$-nearest neighbor classification procedures.

5.4.3 Classification Procedures

Using the appropriate variable subset, we applied each of our chosen classification techniques to the five partitioned data sets. For all but the decision tree approach, we utilized the technique to derive a classification function from the training data and then applied the resulting function to the validation data, recording the classification error rate for both samples. For the decision tree approach, we used the validation data to prune an initial tree defined by the training data. We then classified both the training and validation data sets using the pruned tree and recorded their classification error rates.

We used Box-Cox-transformed variables for discriminant analysis as necessary; for the other three techniques, we employed untransformed variables. For discriminant analysis, we first tested the hypothesis of a pooled covariance to determine whether a linear or quadratic approach was more appropriate. Moreover, we assessed the classification error rate for the training data based on cross-validation, since the apparent error rate—calculated
directly from the data—tends to be optimistically biased (Johnson & Wichern 2002; SAS 2003). For the $k$-nearest neighbor approach, we employed the 8 closest neighbors, and again judged the training data error rate by cross-validation. Similarly, we used cross-validation to calculate predicted probabilities for the logistic regression. We classified observations as infested if the logistic regression equation yielded a predicted probability $\geq 0.5$ for that class; otherwise, we classified the observation as uninfested. Finally, we adopted CART-like settings for decision tree construction: the Gini impurity index as our splitting criterion and pruning using a best sub-tree approach.

For each of the classification techniques, we calculated minimum, maximum, and mean classification error rates for the five training samples and for the five validation samples. One of these partitioned data sets yielded the median or near-median error rate for all four classification techniques. So, we used the classification results for this “median” data set to build error matrices. The matrices allowed us to compare the error patterns of each classification technique and compute Cohen’s kappa statistic. The kappa statistic indicates how much of an improvement a classification effort is over a completely random classification of the same data (Jensen 1996). It can range from 0 to 1, with 0 being the least possible improvement and 1 being the most. Manel et al. (2001) suggested that kappa is a simple, efficient statistic for comparing the success of predictive models, including those created with different algorithms.

5.4.4 Risk Map Generation

We also used the classification functions generated for the “median” data set to create hemlock woolly adelgid infestation maps for GSMNP. If they did not already exist, we
prepared 10-m resolution grids for each variable incorporated in the functions. We treated the stacked grids as a new multivariate data set to be classified, i.e., we treated each pixel as a single, unique observation with values for each of the selected variables. For discriminant analysis and \(k\)-nearest neighbor, we imported the new data set into SAS and processed it using the original procedure code. We then derived grids of the predicted infestation probability from the resulting SAS output. From the discriminant analysis and \(k\)-nearest neighbor probability grids, we also created binary (high infestation risk/low infestation risk) grid layers based on a threshold of infestation probability \(\geq 0.5\).

To generate an infestation probability grid for the logistic regression approach, we converted the fitted logistic regression equation into a GIS map algebra statement for calculating each pixel’s probability:

\[
\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) / (1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p))
\]

where \(\alpha\) was the intercept of the fitted equation, \(\beta_1 \ldots \beta_p\) were the equation coefficients, and \(x_1 \ldots x_p\) were values for the appropriate variable grids. From this probability grid, we also created a binary map based on an infestation probability \(\geq 0.5\).

Our simple decision tree approach did not offer predicted probabilities as output. However, we did convert the tree’s decision rules into a conditional map algebra statement that we applied to the appropriate variable grids to generate a binary map.

We used the binary maps to determine the percentage of GSMNP’s total area that was classified as high infestation risk by the four functions. Moreover, we calculated the percentage of GSMNP’s delineated hemlock areas that were classified as high risk. These percentages served as relative measures of efficiency, i.e., how much area each classification technique required to achieve its level of accuracy. Such comparisons are important in a
practical sense, because if the identified risk area is too large, then it will be of little use to
forest managers for prioritizing hemlock stands and reducing their area of focus. In addition,
we calculated the percent overlap in area designated as high infestation risk by each
classification technique with the areas labeled high risk by the other three techniques.

Figure 5-4. Hemlock woolly adelgid survey and treatment sites recorded after the first year of infestation.

After the first year of infestation, GSMNP staff members continued to regularly
survey for hemlock woolly adelgid, and treated many sites via biological or chemical control.
During the last three years, 224 treatment sites and 159 additional infestations in the park
were recorded with GPS. We used these GPS locations (Figure 5-4) to test whether the
infestation risk zones defined by our classification functions reasonably predicted the pattern
of hemlock woolly adelgid expansion after the initial invasion. We recorded the percentages
of treated and surveyed points that fell in the infested areas delineated on each of the four
binary maps. In addition, for the discriminant analysis, $k$-nearest neighbor, and logistic
regression approaches, we noted the percentages of treated and surveyed points that fell in areas with an infestation probability $\geq 0.75$.

5.5 Results

5.5.1 Multivariate Normality

For the continuous variables of the full, unpartitioned data set, results of multivariate normality testing were mixed. The infested group exhibited neither significant skewness (p-value=0.1269) or kurtosis (p=0.4447). On the other hand, testing suggested the uninfested group was skewed (p=0.021), although it did not demonstrate significant kurtosis (p=0.6662). Nonetheless, chi-square Q-Q plots (Figure 5-5) suggested at least approximate multivariate normality for the uninfested group. Notably, the five partitioned training data sets performed slightly better than the full data set from which they were extracted (Table 5-2): None exhibited a skewness p-value below 0.05 or kurtosis p-value less than 0.27 for either group. Overall, multivariate normality seemed like a reasonable assumption for discriminant analysis, especially since we only used a subset of the variables that omitted some (e.g., landform index) that were the least univariate normal.

![Figure 5-5. Chi-square Q-Q plots for multivariate normality: infested (left) and uninfested (right) of the full data set.](image)
Table 5-2. Multivariate normality test results for the five training data sets.

<table>
<thead>
<tr>
<th>Training Data Set Number</th>
<th>Infested Group Skewness</th>
<th>Infested Group Kurtosis</th>
<th>Uninfested Group Skewness</th>
<th>Uninfested Group Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0528</td>
<td>0.2741</td>
<td>0.5875</td>
<td>0.4155</td>
</tr>
<tr>
<td></td>
<td>0.1175</td>
<td>0.4028</td>
<td>0.0652</td>
<td>0.4198</td>
</tr>
<tr>
<td></td>
<td>0.4785</td>
<td>0.2926</td>
<td>0.2168</td>
<td>0.5111</td>
</tr>
<tr>
<td></td>
<td>0.6147</td>
<td>0.4491</td>
<td>0.1018</td>
<td>0.6051</td>
</tr>
<tr>
<td></td>
<td>0.5059</td>
<td>0.698</td>
<td>0.0696</td>
<td>0.2878</td>
</tr>
</tbody>
</table>

5.5.2 Variable Selection and Error Rate Performance

Testing indicated the group covariance matrices were unequal, so we used a quadratic discriminant analysis approach. Stepwise selection identified four significant variables: distance to stream, distance to road, distance to trail, and elevation. Using these four variables, discriminant analysis was the most successful classification technique when judged by error rate (Table 5-2). Discriminant analysis had the lowest mean error rate for the training and validation data sets, and performed the most consistently, exhibiting the smallest difference between minimum and maximum error rates for the five sets. In contrast, \( k \)-nearest neighbor analysis (built with the same four variables) performed the worst of the four classification techniques for the training data, and next to worst for the validation data. However, error rates for the training and validation data were similar, indicating the \( k \)-nearest neighbor classification function remained relatively accurate when applied to new data points.

Table 5-3. Error rate summaries across the five partitioned data sets.

<table>
<thead>
<tr>
<th>Classification Technique</th>
<th>Training Data</th>
<th>Validation Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Error Rate</td>
<td>Min.</td>
</tr>
<tr>
<td>Discriminant Analysis</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>( k )-Nearest Neighbor</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.18</td>
<td>0.14</td>
</tr>
</tbody>
</table>
For logistic regression, the AIC analysis identified a five-variable subset similar to that noted by stepwise discriminant analysis: the three proximity variables, elevation, and percent slope. Notably, no categorical variables were identified as significant. Generalized $R^2$ values for the logistic regression equations fitted to the five training data sets ranged from 0.352 to 0.433. The fitted equation for the “median” data set,

$$5.3321 - 0.00497 \times \text{stream\_distance} - 0.00072 \times \text{road\_distance} - 0.00264 \times \text{trail\_distance} - 0.00150 \times \text{elevation} - 0.0242 \times \text{pct\_slope},$$

was typical of the other data sets, with distance to trails as the most significant variable and elevation the least significant variable based on chi-square estimates. In terms of error rates, the logistic regression performed between discriminant analysis and $k$-nearest neighbor analysis. As with these two approaches, the logistic regression error rates were fairly consistent between the training and validation data sets.

![Decision tree diagram](link_to_diagram)

Figure 5-6. Decision tree for the median data set; road\_dist=distance to closest road, strm\_dist=distance to closest stream, and trl\_dist=distance to closest trail.

After pruning, the decision tree for the five sets generally included only the three proximity variables. The decision tree constructed for the median data set (Figure 5-6) was typical, with six terminal nodes and a depth of three levels. Based on mean error rate for the
training data set, the decision tree approach performed almost as well as discriminant analysis. However, it performed worst of all techniques for the validation data sets, exhibiting the highest mean and maximum error rate, as well as the largest range between minimum and maximum. Still, it is worth noting the error rate never exceeded 0.30 for the decision tree, or any other, approach. This suggests that all of the approaches performed at least moderately well for basic classification.

5.5.3 Error Matrices

The error matrices for the median data set (Tables 5-4 and 5-5) set largely echo the general error rate results. For the training data subset of the median data set, the decision tree had the highest overall accuracy (86.14%), slightly ahead of discriminant analysis (85.15%). Logistic regression (80.20%) performed better than $k$-nearest neighbor (78.22%), but was still noticeably behind the other two approaches. Kappa values followed this same trend: 0.713 for the decision tree, 0.700 for discriminant analysis, 0.633 for logistic regression, and 0.551 for $k$-nearest neighbor. Commission and omission errors—the percentage of observations included in the wrong group and excluded from the correct group, respectively—indicate that the discriminant analysis was well balanced (i.e., errors did not especially correspond with one class). Logistic regression, while not as accurate, was also fairly well balanced. The decision tree, while quite accurate as capturing infested locations, was imbalanced, with much of the error attributable to “false positives” (i.e., uninfested observations mistakenly classified as infested). The $k$-nearest neighbor approach similarly resulted in many false positives.
For the decision tree, the tendency towards false positives carried over to the validation data subset. It had the worst overall accuracy (78%) and a large difference in omission error between the two groups. In contrast, the other three techniques were well balanced with respect to error. Discriminant analysis was the most accurate technique overall (86%), while logistic regression and $k$-nearest neighbor lagged slightly behind (82%). Kappa values reflected the overall accuracy results: 0.715 for discriminant analysis, 0.633 for logistic regression and $k$-nearest neighbor, and 0.533 for the decision tree approach.

Table 5-4. Error matrices for the training subset of the median data set: a) discriminant analysis, b) $k$-nearest neighbor, c) logistic regression, d) decision tree; CE=% commission error, OE=% omission error.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Reference Group</th>
<th>Infested</th>
<th>Uninfested</th>
<th>Total</th>
<th>CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infested</td>
<td></td>
<td>48</td>
<td>7</td>
<td>55</td>
<td>12.73</td>
</tr>
<tr>
<td>Uninfested</td>
<td></td>
<td>8</td>
<td>38</td>
<td>46</td>
<td>17.39</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>56</td>
<td>45</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>OE</td>
<td></td>
<td>14.29</td>
<td>15.56</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classification</th>
<th>Reference Group</th>
<th>Infested</th>
<th>Uninfested</th>
<th>Total</th>
<th>CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infested</td>
<td></td>
<td>47</td>
<td>11</td>
<td>58</td>
<td>18.97</td>
</tr>
<tr>
<td>Uninfested</td>
<td></td>
<td>9</td>
<td>34</td>
<td>43</td>
<td>20.93</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>56</td>
<td>45</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>OE</td>
<td></td>
<td>16.07</td>
<td>24.44</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5-5. Error matrices for the validation subset of the median subset: a) discriminant analysis, b) $k$-nearest neighbor, c) logistic regression, d) decision tree; CE=% commission error, OE=% omission error.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Reference Group</th>
<th>Infested</th>
<th>Uninfested</th>
<th>Total</th>
<th>CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infested</td>
<td></td>
<td>25</td>
<td>4</td>
<td>29</td>
<td>86.21</td>
</tr>
<tr>
<td>Uninfested</td>
<td></td>
<td>3</td>
<td>18</td>
<td>21</td>
<td>85.71</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>28</td>
<td>22</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>OE</td>
<td></td>
<td>10.71</td>
<td>18.18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classification</th>
<th>Reference Group</th>
<th>Infested</th>
<th>Uninfested</th>
<th>Total</th>
<th>CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infested</td>
<td></td>
<td>24</td>
<td>5</td>
<td>29</td>
<td>17.24</td>
</tr>
<tr>
<td>Uninfested</td>
<td></td>
<td>4</td>
<td>17</td>
<td>21</td>
<td>19.05</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>28</td>
<td>22</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>OE</td>
<td></td>
<td>14.29</td>
<td>22.73</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 5-7. Areas at high risk of imminent hemlock woolly adelgid infestation, as defined by classification functions: a) discriminant analysis, b) k-nearest neighbor, c) logistic regression; d) decision tree.
5.5.4 Prediction Map Performance

Binary maps generated for the median data set using the four classification functions (Figure 5-7) differed substantially in the amount of GSMNP area classified as at risk for hemlock woolly adelgid infestation (Table 5-6). Discriminant analysis appeared to perform the most efficiently: It classified the lowest percentage of GSMNP’s total area and hemlock area as infested, while at the same time it was the most accurate based on mean error rates and the error matrices. The logistic regression approach was not quite as efficient as discriminant analysis, requiring a slightly higher percentage of total park area and hemlock area while exhibiting a higher mean error rate. However, it was considerably more efficient than the $k$-nearest neighbor and decision tree approaches which labeled a large percentage (~30%) of the park’s area as high infestation risk.

Table 5-6. GSMNP total area and hemlock area designated as high risk by each classification technique based on the median data set.

<table>
<thead>
<tr>
<th>Classification Technique</th>
<th>% GSMNP Total Area</th>
<th>% GSMNP Hemlock Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discriminant Analysis</td>
<td>19.1</td>
<td>18.5</td>
</tr>
<tr>
<td>$k$-Nearest Neighbor</td>
<td>29.6</td>
<td>23.9</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>22.8</td>
<td>20.1</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>30.7</td>
<td>34.6</td>
</tr>
</tbody>
</table>

The four techniques also differed in pattern of areas they mapped as high infestation risk (Figure 5-7). Naturally, with their similar variable subsets, they all emphasized areas near roads, trails, and streams to some degree. However, discriminant analysis defined a narrow buffer around almost every stream feature as part of its high-risk zone. The decision tree approach behaved similarly, although it segmented GSMNP even further with wide buffers around streams as well as roads and trails. In contrast to these dendritic representations, logistic regression highlighted several large, discrete areas as high-risk for hemlock woolly adelgid infestation. The $k$-nearest neighbor analysis highlighted
geographically similar areas to the logistic regression, but defined them with wider boundaries.

Table 5-7. Percentage of the area mapped as high infestation risk by each technique that was also designated high risk by the other three techniques.

<table>
<thead>
<tr>
<th>Discriminant Analysis</th>
<th>% Overlapped</th>
<th>k-Nearest Neighbor</th>
<th>% Overlapped</th>
<th>Logistic Regression</th>
<th>% Overlapped</th>
<th>Decision Tree</th>
<th>% Overlapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>By k-Nearest Neighbor</td>
<td>80.2</td>
<td>By Discriminant Analysis</td>
<td>51.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>By Logistic Regression</td>
<td>73.4</td>
<td>By Logistic Regression</td>
<td>68.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>By Decision Tree</td>
<td>87.5</td>
<td>By Decision Tree</td>
<td>58.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Examination of the overlap percentages between the four techniques (Table 5-7) further highlights their differences and commonalities. Notably, the $k$-nearest neighbor, logistic regression, and decision tree approaches all designated as high risk greater than 73% of the area labeled as such by discriminant analysis. The $k$-nearest neighbor approach displayed substantially more overlap with logistic regression than the other two techniques, capturing ~89% of the area designated high risk by logistic regression—further emphasizing the geographic similarity of the areas designated high risk by $k$-nearest neighbor and logistic regression. The decision tree captured ~88% of the area labeled high risk by discriminant analysis, and therefore appears to agree more with this technique than $k$-nearest neighbor or logistic regression, but no technique captured a large percentage of the extensive area designated high risk by the decision tree.

Infestation probability maps for discriminant analysis and logistic regression (Figure 5-8) confirm much overlap in the areas delineated as high risk by the two techniques, but also emphasize a few noteworthy differences in pattern. In particular, certain trails in the southwest part of GSMNP were highlighted by discriminant analysis and not logistic
regression. Conversely, sections of road in the park’s central area were highlighted by logistic regression and not discriminant analysis.

Figure 5-8. Risk probability surfaces developed with discriminant analysis (top) and logistic regression (bottom).

With respect to the surveyed and treated infestation sites recorded after the first year (Table 5-8), the risk maps generated with the four techniques performed well, capturing ~80% or more of the surveyed points and ~89% or more of the treated points. In other words, despite exhibiting different spatial patterns, all techniques appeared to correctly predict a high percentage of second-, third- and fourth-year infestations. When the probability threshold was increased to 0.75 (from the default 0.50), discriminant analysis performed nearly as well. Logistic regression exhibited some drop-off at this higher threshold, while \(k\)-nearest neighbor analysis performed rather poorly, suggesting that the \(k\)-
nearest neighbor function was far less definitive in its classifications than discriminant analysis—though this may only be a minor concern.

Table 5-8. Percentage of surveyed and treated infestation points captured by each classification technique.

<table>
<thead>
<tr>
<th>Classification Technique</th>
<th>Surveyed Points</th>
<th>Treated Points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Prob. ≥ 0.5</td>
<td>% Prob. ≥ 0.75</td>
</tr>
<tr>
<td>Discriminant Analysis</td>
<td>89.9</td>
<td>86.2</td>
</tr>
<tr>
<td>k-Nearest Neighbor</td>
<td>79.9</td>
<td>37.7</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>80.5</td>
<td>58.5</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>93.1</td>
<td>n/a</td>
</tr>
</tbody>
</table>

5.6 Discussion

5.6.1 “Best” Classification Method

Looking only at error rates, the four classification techniques performed reasonably well at the basic task of distinguishing infested and uninfested sites. Although the decision tree was the worst performer with respect to the validation data sets, it still had a mean error rate of 0.24 and a maximum error rate of 0.30 for those sets. This implies that the derived classification functions—even the decision tree—would be at least moderately accurate if applied to new data sets. Since our main goal is to apply one of these functions to map risk for the entire southern Appalachian region, this is especially meaningful. Still, error rate is only a single coarse measure. Our other tests revealed classification imbalances or problematic characteristics that make certain of the techniques less than practical for hemlock woolly adelgid management purposes. For instance, the decision tree approach, while highly successful at capturing infested sites, also misclassified a large number of uninfested sites as infested. In addition, the decision tree approach designated more than one-third of GSMNP’s hemlock area as high risk; in fact, this large proportion may explain why the approach
resulted in so many false positives. Of course, we chose certain parameters when implementing the decision tree (e.g., univariate vs. linear-combination splitting rules, discrete vs. probabilistic output) that may have affected its success. Nevertheless, as implemented, the lack of efficiency and lopsided error pattern suggest that the decision tree approach is not a practical choice for further application.

Most of our test results recommended discriminant analysis as the best classification approach. It exhibited low error rates that were well balanced between the two classes, delineated the least amount of area as high risk while achieving those low error rates, and captured a large percentage of subsequent hemlock woolly adelgid infestations in a high risk probability (>=0.75) zone. Nonetheless, visual examination of the generated risk map indicates at least one potential problem with using the approach: Discriminant analysis labeled almost all areas adjacent to streams with a high infestation risk. Although those areas might actually face elevated infestation risk, the stream network is so densely distributed across the GSMNP area that this representation makes it difficult to identify areas of focus for hemlock woolly adelgid management efforts. Certainly, the narrowest riparian areas could simply be ignored when the discriminant analysis approach is used to map other parts of the southern Appalachian region, but the appropriate threshold for choosing which areas to ignore is unclear. To do so may also result in the loss of important classification information.

Arguably, the logistic regression approach performed second best when considering all of our results. In terms of the error rates and error matrices, it fell roughly in the middle of the four techniques, and it designated only slightly more area as high risk than discriminant analysis (indeed, there was substantial overlap between the two techniques) while achieving that degree of accuracy. However, a noteworthy feature of the logistic
Regression approach is that, unlike discriminant analysis, it defined discrete zones of high infestation risk. While not necessarily a more accurate portrayal of regional risk, these zones provide clear areas of focus for hemlock woolly adelgid management efforts. The \( k \)-nearest neighbor analysis also generated discrete zones, but with less accuracy and efficiency than logistic regression—making it a less preferred choice.

Perhaps most importantly, the fitted logistic regression equation is easily implemented as a raster calculator function in a GIS. In contrast, because discriminant analysis was executed using a quadratic approach, there is no easily interpretable output function. Application of discriminant analysis would also require transformation of some variables prior to use. Given the straightforwardness and transparency of the logistic regression function, we would recommend the use of this approach over discriminant analysis (as well as the \( k \)-nearest neighbor or decision tree approaches), thus sacrificing some accuracy for interpretability and ease of use.

5.6.2 Landscape Variables Affecting Hemlock Woolly Adelgid Spread

We completed our analyses at a limited scale using a small sample of points from GSMNP and the southern BRP. Since our primary objective was to find a tool for predicting hemlock woolly adelgid landscape distribution throughout the southern Appalachian region, a question emerges as to whether the logistic regression approach—as our chosen method—incorporated the appropriate variables, as well as whether the function built on those variables can be accurately generalized region-wide. It is noteworthy that all of the classification techniques (or rather, their variable selection procedures) incorporated the three proximity variables: distance to closest road, distance to closest trail, and distance to closest
stream. This suggests that potential corridors of spread are by far the most important factors for predicting where hemlock woolly adelgid infestations are likely to occur in a landscape. This conclusion is not surprising. In general, the connectivity supplied by linear networks of landscape features—even habitat corridors—may expose forested areas to various negative effects, including the spread of insect pests and disease (Hess 1996; Saunders et al. 1991; Simberloff & Cox 1987; Trombulak & Frissell 2000). Roads and trails are especially relevant for the hemlock woolly adelgid because they may amplify spread by three of the pest’s known vectors: wind, birds, and humans. Forest edges along road corridors may experience increased wind exposure (Saunders et al. 1991). Major roads such as the Blue Ridge Parkway serve as flyways for bird migration (Wear & Greis 2002). Of course, human traffic along trails and road corridors can greatly increase pest dispersal distance (Forys et al. 2001; Hulme 2003; Jules et al. 2002; Zobel et al. 1985). Similarly, riparian corridors have been implicated for the spread of the hemlock woolly adelgid by birds in the northeastern U.S. (Ward et al. 2004). This is complicated by the fact that riparian corridors often include hiking trails with heavy human traffic (McWilliams & Schmidt 1999), but this does not diminish the importance of stream distance as a discriminatory variable.

In summary, the three proximity variables provide a reasonable, if simple, model for landscape-scale hemlock woolly adelgid dispersal. Notably, the fitted logistic regression equation also includes elevation and percent slope as significant variables—basically, the odds of infestation are lower at high elevations and on steep slopes. Since these factors could influence accessibility of a site (especially by people), they seem like logical components of a model predicting hemlock woolly adelgid infestations. But why were the remaining topographic variables not considered important? It is likely because these variables are
strongly tied to hemlock distribution, but not necessarily adelgid presence. For example, in an earlier analysis we found that hemlock occurrence corresponded to higher topographic relative moisture index values. However, the hemlock woolly adelgid is fairly indiscriminate about hemlock age or the degree of hemlock presence—as long as a site has hemlocks, it is a candidate for infestation—so a factor such as topographic relative moisture index does not directly determine the level of infestation risk (Orwig & Foster 1998; Orwig et al. 2002; Ward et al. 2004). The categorical variables may have been omitted from all of the final functions for similar reasons, although they may also have been too low in resolution—either spatial or attribute resolution—to be effective as discriminatory variables.

Certainly, there are variables that we did not include in our analyses. We have already noted that we did not consider patch characteristics because of too much patch-level heterogeneity in our data set. Furthermore, we did not have reliable hemlock patch information for the BRP points. A more comprehensive data set may have revealed significant differences between infested and uninfested patches; indeed, this is an area that may deserve further analysis. We also omitted some traditional landscape metrics such as inter-patch distance. Although regions with high average distances between hemlock patches might be invaded more slowly, we reasoned that this was secondary to the density of roads and other corridors of spread. Finally, we chose not to include climate variables. While winter temperatures may suppress hemlock woolly adelgid populations and limit the northward expansion of the pest, the southern Appalachian climate is probably not consistently cold enough to have the same effect (Ward et al. 2004). At a landscape scale, there may be microclimatic characteristics that predispose a site to hemlock woolly adelgid infestation, but microclimatic data are sparsely available, if at all.
We built our classification functions using a small amount of GSMNP and BRP data. We hope that these data, while limited, offer a reasonable representation of conditions throughout the southern Appalachian region. However, we have previously noted our most important assumption, that the recorded points are an accurate reflection of the hemlock woolly adelgid presence during the first year of detected infestation. Despite our best efforts, there is some possibility that the sample is biased. There may have been some backcountry infestations during the first year that simply eluded survey. However, evidence from another southern Appalachian site bolsters our assumption: Surveys for hemlock woolly adelgid at Coweeta Hydrological Laboratory (near Franklin, North Carolina) during the first year of infestation revealed that the heaviest infestations appeared along roadways, while forest interior areas remained uninfested or were only lightly infested (Lumpkin et al. 2003).

Beyond these concerns, our approach is limited in scope. While it does predict where infestations are most likely to occur in a landscape, it does not explicitly incorporate regional spread. In other words, it does not indicate which previously uninfested portions of the southern Appalachians have the most immediate risk of hemlock woolly adelgid infestation. Instead, it assumes that long-distance dispersal puts the entire region at risk. Fortunately, because our approach is built on just a few readily available or easily derived spatial variables, it can be rapidly applied to generate infestation risk maps for the entire southern Appalachian region. Undoubtedly, these maps may be used to select monitoring sites prior to the adelgid’s arrival in a particular locality. Nonetheless, our results also suggest that the maps can predict the pest’s likely distribution for at least the first few years after infestation. This could be particularly advantageous for the targeted release of biological control agents, which can take some time to establish in infested sites (Cheah et al. 2004).
Ultimately, regional infestation probability maps would be best applied in conjunction with a second tool we have developed for managing the hemlock woolly adelgid problem. In a separate analysis, we have developed a decision tree classifier for mapping hemlock distribution in the southern Appalachians using remotely sensed imagery and GIS data for a suite of environmental variables. By overlaying the infestation probability maps with hemlock distribution maps, forest managers can further reduce the area where they should target their monitoring or control efforts (see Figure 5-9 for an example). This ability to prioritize should allow them to apply scarce management resources where they are most immediately needed.

![Figure 5-9. Overlay of logistic regression risk map and hemlock distribution in GSMNP.](image)

5.7 Conclusions

To enable better management of the hemlock woolly adelgid in the southern Appalachians, we examined landscape-level factors influencing the pest’s distribution across
the region. Using spatial data on first-year locations in Great Smoky Mountains National Park and along the southern Blue Ridge Parkway, we tested four different classification techniques to see which could most effectively predict locations that are likely to be infested by the pest. We considered a suite of topographic, environmental, and proximity variables as possible factors in the landscape pattern of infestation. Of the four techniques tested, a logistic regression equation using five variables—proximity to roads, trails, and streams, as well as slope and elevation—represented the best combination of classification accuracy and interpretability. A resulting infestation probability map successfully predicted > 80% of new infestations appearing within three years of the first detection of the adelgid. Overall, the approach appears to be general enough for application throughout the southern Appalachian region. Given the level of threat to the southern Appalachian hemlock resource due to hemlock woolly adelgid infestation, we recommend that our approach be rapidly applied region-wide to help forest managers prioritize areas and allocate scarce resources for hemlock woolly adelgid control measures.

5.8 References


6. APPENDICES
6.1 SAS Code: Regressions for Generating C-correction Coefficients

/* C_correct.sas */
/* SAS code for generating C-correction coefficients */
/* Regressions yield Bk and Mk values for calculating Ck */
/* Per-band Ck coefficients can then be loaded into Imagine 8.7 model */

/**************************************************************/
/* Part I: for Landsat imagery (six bands) */
/* Separate sunlit and sunshade subsets created via aspect partitioning */
/**************************************************************/

/* load comma-delimited text file with image radiance values */
/* as well as aspect and slope values from corresponding DEM */
/* aspect and slope values are used in the calculation of cos i */

data landsat_sunlit ;
   infile 'c:/temp/landsat_sunlit.dat' dsd ;
   input x y b1 b2 b3 b4 b5 b7 aspect slope ;    /* B1-B7=band #s */
   pi = constant('pi') ;
   phi_s = 155.1560165 ;    /* solar azimuth from Landsat parameter file */
   theta_s = 37.8461488 ;   /* solar elevation from Landsat parameter file */
   cos_i = cos((90-theta_s)*pi/180)*cos(slope*pi/180) + sin((90-theta_s)*pi/180)*sin(slope*pi/180)*cos((phi_s-aspect)*pi/180) ;
run ;

/* optional printout of the data – can be commented out */
proc print data=landsat_sunlit ;
   var b1 b2 b3 b4 b5 b7 aspect slope phi_s theta_s cos_i ;
   title 'Landsat sunlit data';
run ;

/* per-band regressions of radiance values on cos i */
proc reg data=landsat_sunlit ;
   model b1 = cos_i ;
   title 'Landsat sunlit regressions';
proc reg data=landsat_sunlit ;
   model b2 = cos_i ;
proc reg data=landsat_sunlit ;
   model b3 = cos_i ;
proc reg data=landsat_sunlit ;
   model b4 = cos_i ;
proc reg data=landsat_sunlit ;
   model b5 = cos_i ;
proc reg data=landsat_sunlit ;
   model b7 = cos_i ;
run ;
/* load comma-delimited text file with image radiance values */
/* as well as aspect and slope values from corresponding DEM */
data landsat_sunshade ;
  infile 'c:/temp/landsat_sunshade.dat' dsd ;
  input x y b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope ; /* b1-b7=band #s */
/* NOTE: No B6 because dropped the thermal IR band */
  pi = constant('pi') ;
  phi_s = 155.1560165 ; /* solar azimuth */
  theta_s = 37.8461488 ; /* solar elevation */
/* cos i term: cos(90-theta_s)cos(theta_n) + sin(90-theta_s)sin(theta_n)cos(phi_s-phi_n) */
/* phi_n is the surface aspect */
/* theta_n is the surface slope */
/* pi/180 = degree to radians conversion */
cos_i = cos((90-theta_s)*pi/180)*cos(slope*pi/180) + sin((90-theta_s)*pi/180)*sin(slope*pi/180)*cos((phi_s-aspect)*pi/180) ;
run ;

/* optional printout of data */
proc print data=landsat_sunshade ;
  var b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope phi_s theta_s cos_i ;
  title 'Landsat sunshade data' ;
run ;
/* per-band regressions of radiance values on cos i */
proc reg data=landsat_sunshade ;
  model b1 = cos_i ;
  title 'Landsat sunshade regressions' ;
proc reg data=landsat_sunshade ;
  model b2 = cos_i ;
proc reg data=landsat_sunshade ;
  model b3 = cos_i ;
proc reg data=landsat_sunshade ;
  model b4 = cos_i ;
proc reg data=landsat_sunshade ;
  model b5 = cos_i ;
proc reg data=landsat_sunshade ;
  model b7 = cos_i ;
run ;

************************************************************************/
/* Part II: for ASTER VNIR and SWIR imagery (nine bands) */
************************************************************************/
/* load comma-delimited text file with image radiance values */
/* as well as aspect and slope values from corresponding DEM */
data aster_sunlit ;
infile 'c:/temp/aster_sunlit.dat' dsd ;
input x y b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope ; /* b1-B9=band #s */
  pi = constant('pi') ;
  phi_s = 164.67564 ; /* solar azimuth from ASTER parameter file */
  theta_s = 50.986448 ; /* solar elevation from ASTER parameter file */
/* cos i = cos((90-theta_s)*pi/180)*cos(slope*pi/180) + sin((90-theta_s)*pi/180)*sin(slope*pi/180)*cos((phi_s-aspect)*pi/180) ;
run ;

/* optional printout of data */
proc print data=aster_sunlit ;
  var b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope phi_s theta_s cos_i ;
  title 'ASTER sunshade data' ;
run ;
/* per-band regressions of radiance values on cos i */
proc reg data=aster_sunlit ;
  model b1 = cos_i ;
  title 'ASTER sunshade regressions' ;
proc reg data=aster_sunlit ;
  model b2 = cos_i ;
proc reg data=aster_sunlit ;
  model b3 = cos_i ;
proc reg data=aster_sunlit ;
  model b4 = cos_i ;
proc reg data=aster_sunlit ;
  model b5 = cos_i ;
proc reg data=aster_sunlit ;
  model b7 = cos_i ;
run ;

************************************************************************/
/* Part II: for ASTER VNIR and SWIR imagery (nine bands) */
************************************************************************/
/* Separate sunlit and sunshade subsets created via aspect partitioning */
************************************************************************/
/* load comma-delimited text file with image radiance values */
/* as well as aspect and slope values from corresponding DEM */
data aster_sunlit ;
infile 'c:/temp/aster_sunlit.dat' dsd ;
input x y b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope ; /* b1-B9=band #s */
  pi = constant('pi') ;
  phi_s = 164.67564 ; /* solar azimuth from ASTER parameter file */
  theta_s = 50.986448 ; /* solar elevation from ASTER parameter file */
/* cos i = cos((90-theta_s)*pi/180)*cos(slope*pi/180) + sin((90-theta_s)*pi/180)*sin(slope*pi/180)*cos((phi_s-aspect)*pi/180) ;
run ;

/* optional printout of data */
proc print data=aster_sunlit ;
  var b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope phi_s theta_s cos_i ;
  title 'ASTER sunshade data' ;
run ;
/* per-band regressions of radiance values on cos i */
proc reg data=aster_sunlit ;
  model b1 = cos_i ;
  title 'ASTER sunshade regressions' ;
proc reg data=aster_sunlit ;
  model b2 = cos_i ;
proc reg data=aster_sunlit ;
  model b3 = cos_i ;
proc reg data=aster_sunlit ;
  model b4 = cos_i ;
proc reg data=aster_sunlit ;
  model b5 = cos_i ;
proc reg data=aster_sunlit ;
  model b7 = cos_i ;
run ;

************************************************************************/
/* Part II: for ASTER VNIR and SWIR imagery (nine bands) */
************************************************************************/
/* Separate sunlit and sunshade subsets created via aspect partitioning */
************************************************************************/
/* load comma-delimited text file with image radiance values */
/* as well as aspect and slope values from corresponding DEM */
data aster_sunlit ;
infile 'c:/temp/aster_sunlit.dat' dsd ;
input x y b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope ; /* b1-B9=band #s */
  pi = constant('pi') ;
  phi_s = 164.67564 ; /* solar azimuth from ASTER parameter file */
  theta_s = 50.986448 ; /* solar elevation from ASTER parameter file */
/* cos i = cos((90-theta_s)*pi/180)*cos(slope*pi/180) + sin((90-theta_s)*pi/180)*sin(slope*pi/180)*cos((phi_s-aspect)*pi/180) ;
run ;

/* optional printout of data */
proc print data=aster_sunlit ;
  var b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope phi_s theta_s cos_i ;
  title 'ASTER sunshade data' ;
run ;
/* per-band regressions of radiance values on cos i */
proc reg data=aster_sunlit ;
  model b1 = cos_i ;
  title 'ASTER sunshade regressions' ;
proc reg data=aster_sunlit ;
  model b2 = cos_i ;
proc reg data=aster_sunlit ;
  model b3 = cos_i ;
proc reg data=aster_sunlit ;
  model b4 = cos_i ;
proc reg data=aster_sunlit ;
  model b5 = cos_i ;
proc reg data=aster_sunlit ;
  model b7 = cos_i ;
run ;
\[
\text{theta}_s \times (\pi/180) \times \sin(\text{slope} \times \pi/180) \times \cos((\text{phi}_s - \text{aspect}) \times \pi/180);
\]

run;

/* optional printout of the data */
proc print data=aster_sunlit;
  var b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope phi_s theta_s cos_i;
  title 'ASTER sunlit data';
run;

/* per-band regressions of radiance values on \(\cos i\) */
proc reg data=aster_sunlit;
  model b1 = cos_i;
  title 'ASTER sunlit regressions';
proc reg data=aster_sunlit;
  model b2 = cos_i;
proc reg data=aster_sunlit;
  model b3 = cos_i;
proc reg data=aster_sunlit;
  model b4 = cos_i;
proc reg data=aster_sunlit;
  model b5 = cos_i;
proc reg data=aster_sunlit;
  model b6 = cos_i;
proc reg data=aster_sunlit;
  model b7 = cos_i;
proc reg data=aster_sunlit;
  model b8 = cos_i;
proc reg data=aster_sunlit;
  model b9 = cos_i;
run;

/* load comma-delimited text file with image radiance values */
/* as well as aspect and slope values from corresponding DEM */
data ASTER_sunshade;
  infile 'c:/temp/aster_sunshade.dat' dsd;
  input x y b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope; /* b1-b9=band #s */
  pi = constant('pi');
  phi_s = 164.67564; /* solar azimuth */
  theta_s = 50.986448; /* solar elevation */
  cos_i = \[\cos((90-\text{theta}_s) \times \pi/180) \times \cos(\text{slope} \times \pi/180) + \sin((90-\text{theta}_s) \times \pi/180) \times \sin(\text{slope} \times \pi/180) \times \cos((\text{phi}_s - \text{aspect}) \times \pi/180);\]
run;

/* optional printout of data */
proc print data=third_order;
  var b1 b2 b3 b4 b5 b6 b7 b8 b9 aspect slope phi_s theta_s cos_i;
  title 'aster sunshade data';
run;

/* per-band regressions of radiance values on \(\cos i\) */
proc reg data=aster_sunshade;
  model b1 = cos_i;
  title 'ASTER sunshade regressions';

proc reg data=aster_sunshade ;
  model b2 = cos_i ;
proc reg data=aster_sunshade ;
  model b3 = cos_i ;
proc reg data=aster_sunshade ;
  model b4 = cos_i ;
proc reg data=aster_sunshade ;
  model b5 = cos_i ;
proc reg data=aster_sunshade ;
  model b6 = cos_i ;
proc reg data=aster_sunshade ;
  model b7 = cos_i ;
proc reg data=aster_sunshade ;
  model b8 = cos_i ;
proc reg data=aster_sunshade ;
  model b9 = cos_i ;
run ;
6.2 ERDAS Macro Language: Topographic Normalization via the C-correction

The following script is the ERDAS Macro Language (EML) equivalent of the modeling diagram depicted in Figure 3-4. The code can be saved as a *.mdl file and run through the Model Librarian in ERDAS Imagine.

```emd
# c-correction.mdl
#
# TOPOGRAPHIC NORMALIZATION VIA C-CORRECTION
#
# Notes: (1) surface slope and aspect calculated from a digital elevation model
# (2) coefficient table (n32) has nine rows for ASTER, may have to be altered for other image types
#
# set cell size for the model
# SET CELLSIZE MIN;
#
# set window for the model
# SET WINDOW UNION;
#
# set area of interest for the model
# SET AOI NONE;
#
# declarations
#
# Raster Inputs: Surface slope (n1), Surface aspect (n4), Non-normalized image (n16)
# Raster Output: Normalized image (n18)
# Table Input: C-correction coefficients (n32)
# Scalar Inputs: Solar azimuth (n8), Solar elevation (n9), pi/180 radians conversion (n12)
#
Integer RASTER n1_PROMPT_USER FILE OLD NEAREST NEIGHBOR AOI NONE "c:/temp/slope15m.img";
Integer RASTER n4_PROMPT_USER FILE OLD NEAREST NEIGHBOR AOI NONE "c:/temp/aspect15m.img";
Float RASTER n16_PROMPT_USER FILE OLD NEAREST NEIGHBOR AOI NONE "c:/temp/aster_sunlit.img";
Float RASTER n18_PROMPT_USER FILE DELETE_IF_EXISTING USEALL ATHEMATIC FLOAT SINGLE "c:/temp/aster_sunlit_ccorr.img";
FLOAT TABLE n32_Custom_Float [9];
FLOAT SCALAR n8_Float;
FLOAT SCALAR n9_Float;
FLOAT SCALAR n12_Float;
#
# set output projection for the model
# SET PROJECTION USING n16_PROMPT_USER;
```
# load scalar n8_Float == solar azimuth from image parameter file
n8_Float = 164.67564;

# load scalar n9_Float == solar elevation from image parameter file
n9_Float = 50.986448;

# load scalar n12_Float == pi/180 to do degrees to radians conversions
n12_Float = 0.01745;

# load table n32_Custom_Float == per-band C-correction coefficients
# (from Band 1 thru...9 for an ASTER image, 6 for a Landsat image)
n32_Custom_Float = TABLE(1.944451101, 0.253207509, 0.864129119, 1.285429462, 2.410090057, 1.883850859, 2.427955786, 2.222367173, 4.455151823);

# function definitions
# (1) create cos i surface (n6_memory)
# (2) Execute C-correction
#define n6_memory Float(EITHER COS($n12_Float * (90-$n9_Float)) IF ($n1_PROMPT_USER==0.0) OR COS($n12_Float * (90-$n9_Float)) * COS($n12_Float*$n1_PROMPT_USER) + SIN($n12_Float * (90-$n9_Float)) * SIN($n12_Float*$n1_PROMPT_USER) * COS($n12_Float * ($n8_Float-$n4_PROMPT_USER)) OTHERWISE)

n18_PROMPT_USER = EITHER 0 IF ($n6_memory <= 0.0) OR $n16_PROMPT_USER*((COS((90-$n9_Float)*$n12_Float) + $n32_Custom_Float)/($n6_memory + $n32_Custom_Float)) OTHERWISE ;
6.3 SAS Enterprise Miner: Process Diagram for Decision Tree Construction

We developed our initial and enhanced decision trees using SAS Enterprise Miner. Enterprise Miner allows a user to construct multi-stage data analysis models using a graphical interface. Each node in the graphic represents a module with several user-defined options. Figure 7-1 displays our process diagram for constructing three-class decision trees to map hemlocks. We have also listed the specific module settings we used in our analysis.

![Figure 6-1. SAS Enterprise Miner process diagram for tree construction.](image)

**Input Data Source**
- After loading data set, automatically labels variables as class or interval
- Model role of main class variable (hemlock_class) set as target, all others set as input

**Transform Variables**
- Converted aspect to a measure of “northeastness” using the formula $\cos((45-\text{aspect})*3.14159/180)+1$

**Filter Outliers**
- Used to filter out any erroneous or missing data points
- Set to remove any points with “no data” values (i.e., -9999) for any variables; only lost one point

**Data Set Attributes**
- Used to drop any extraneous variables
- Removed x-coordinate and y-coordinate variables not required in tree construction
Data Partition

- Partitions data into training, validation, and/or testing sets
- Split the hemlock data into 50% training and 50% validation via simple random sampling

Tree

- Emulated a classification and regression tree (CART) approach:
  - Gini reduction as the splitting criterion
  - Maximum number of branches from a node = 2
  - Surrogate rules saved in each node = 5
  - Unchecked option to treat missing as an acceptable value
  - Subtree method set to best assessment value (uses validation data)
  - Observations sufficient for split search = 1000
  - Maximum tries in an exhaustive split search = 5000
- Minimum number of observations in a leaf = 10
- Maximum depth of tree = 12
6.4 Rules List for the Initial Decision Tree for Mapping Hemlocks

The following rules list (Table 7-1) is equivalent to the initial decision tree diagram in Figure 4-2. These rules can be imported into the Expert Classifier in ERDAS Imagine 8.7 or another GIS/image processing software package. Nodes are numbered according to tree position: top-down and then from left to right at each depth level (Figure 7-2). Only terminal nodes are included in the list.

<table>
<thead>
<tr>
<th>Node</th>
<th>Variable</th>
<th>Values</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.698657</td>
<td>Hemlock Dominant/Co-dominant</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1236.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&lt; 0.814015</td>
<td></td>
</tr>
<tr>
<td>Node</td>
<td>Variable</td>
<td>Values</td>
<td>Class</td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------------</td>
<td>-----------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>52</td>
<td>PCT_SLOPE</td>
<td>&lt; 74.479</td>
<td>Hemlock Dominant/Co-dominant</td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.708949</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.698657</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1236.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&lt; 0.814015</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>PCT_SLOPE</td>
<td>&gt;= 74.479</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.708949</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.698657</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1236.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&lt; 0.814015</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>ELEVATION</td>
<td>&lt; 817.9</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.708949</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.698657</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&lt; 0.814015</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>TRMI</td>
<td>&lt; 20.5</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[ 817.9 , 1236.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.708949</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.698657</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&lt; 0.814015</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>TRMI</td>
<td>&gt;= 20.5</td>
<td>Hemlock Dominant/Co-dominant</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[ 817.9 , 1236.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.708949</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.698657</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&lt; 0.814015</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ELEVATION</td>
<td>[1236.45 , 1497.45)</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&lt; 0.814015</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>B1_B2_RATIO</td>
<td>&lt; 0.538413</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1368.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.814015</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>ELEVATION</td>
<td>&lt; 1047.7</td>
<td>Hemlock Dominant/Co-dominant</td>
</tr>
<tr>
<td></td>
<td>LANDFORM</td>
<td>&lt; 0.0155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1_B2_RATIO</td>
<td>&gt;= 0.538413</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.814015</td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>ELEVATION</td>
<td>[ 1047.7 , 1368.85)</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>LANDFORM</td>
<td>&lt; 0.0155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1_B2_RATIO</td>
<td>&gt;= 0.538413</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.814015</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>LANDFORM</td>
<td>&gt;= -0.0155</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>B1_B2_RATIO</td>
<td>&gt;= 0.538413</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1368.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.814015</td>
<td></td>
</tr>
<tr>
<td>Node</td>
<td>Variable</td>
<td>Values</td>
<td>Class</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>64</td>
<td>B4_B8_RATIO</td>
<td>&lt; 0.81349</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.827666)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1368.85, 1497.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.713024</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>B4_B8_RATIO</td>
<td>&gt;= 0.81349</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.827666)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1368.85, 1497.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.713024</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.827666</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1368.85, 1497.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&lt; 0.713024</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>TRMI</td>
<td>&lt; 13.5</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.814015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt;= 1497.45</td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>TRMI</td>
<td>[13.5, 27.5)</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1219.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.826824)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>TRMI</td>
<td>&gt;= 27.5</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1219.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.826824)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>CURVATURE</td>
<td>&lt; -1.15</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1219.8, 1497.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.826824)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td>162</td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.72267</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>DISTANCE</td>
<td>&lt; 283.655</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CURVATURE</td>
<td>&gt;= -1.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1219.8, 1497.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.826824)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td>163</td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.72267</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>DISTANCE</td>
<td>&lt; 283.655</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CURVATURE</td>
<td>&gt;= -1.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1219.8, 1497.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.826824)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td>139</td>
<td>DISTANCE</td>
<td>&gt;= 283.655</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>CURVATURE</td>
<td>&gt;= -1.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1219.8, 1497.45)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>[0.814015, 0.826824)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td>Node</td>
<td>Variable</td>
<td>Values</td>
<td>Class</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------</td>
<td>--------------------------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>140</td>
<td>B4_B9_RATIO</td>
<td>[0.826824, 0.827329)</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.726166</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1497.45</td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>B4_B6_RATIO</td>
<td>[0.713024, 0.715789)</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>(0.827329, 0.83183)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.726166</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1497.45</td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.715789</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>(0.827329, 0.83183)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.726166</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1497.45</td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.833183</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.726166</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1497.45</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.726166</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B9_RATIO</td>
<td>&gt;= 0.826824</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRMI</td>
<td>&gt;= 13.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B6_RATIO</td>
<td>&gt;= 0.713024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 1497.45</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>PCT_SLOPE</td>
<td>&lt; 26.8065</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1497.45, 1538.65)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1_B2_RATIO</td>
<td>&lt; 0.837155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.714865</td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>PCT_SLOPE</td>
<td>[26.8065, 63.704)</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1497.45, 1538.65)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1_B2_RATIO</td>
<td>&lt; 0.837155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.714865</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>PCT_SLOPE</td>
<td>&gt;= 63.704</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1497.45, 1538.65)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1_B2_RATIO</td>
<td>&lt; 0.837155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.714865</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>ELEVATION</td>
<td>[1538.65, 1624.35)</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B1_B2_RATIO</td>
<td>&lt; 0.837155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.714865</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>B1_B2_RATIO</td>
<td>&gt;= 0.837155</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>B4_B5_RATIO</td>
<td>&lt; 0.714865</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1497.45, 1624.35)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>B4_B5_RATIO</td>
<td>&gt;= 0.714865</td>
<td>Non-hemlock Evergreen</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>[1497.45, 1624.35)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ELEVATION</td>
<td>&gt;= 1624.35</td>
<td>Non-hemlock Evergreen</td>
</tr>
</tbody>
</table>
6.5 Rules List for the Enhanced Decision Tree for Mapping Hemlocks

The following rules list (Table 7-2) is equivalent to the enhanced decision tree diagram in Figure 4-3. These rules can be imported into the Expert Classifier in ERDAS Imagine 8.7 or another GIS/image processing software package. Nodes are numbered according to tree position: top-down and then from left to right at each depth level (Figure 7-3). Only terminal nodes are included in the list.

![Figure 6-3. Node index for the enhanced decision tree](image)

### Table 6-2. Decision rules for the enhanced tree.

<table>
<thead>
<tr>
<th>Node</th>
<th>Variable</th>
<th>Values</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>PCT_SLOPE</td>
<td>&lt; 49.0247</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td></td>
<td>ELEVATION</td>
<td>&lt; 412.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B3_B1_RATIO</td>
<td>&lt; 0.554116</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DISTANCE</td>
<td>&lt; 49.4345</td>
<td></td>
</tr>
<tr>
<td>Node</td>
<td>Variable</td>
<td>Values</td>
<td>Class</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------</td>
<td>-------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>33</td>
<td>PCT_SLOPE, ELEVATION, B3_B1_RATIO, DISTANCE</td>
<td>( \geq 49.0247 ), (&lt; 412.25 ), (&lt; 0.554116 ), (&lt; 49.4345 )</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>17</td>
<td>ELEVATION, B3_B1_RATIO, DISTANCE</td>
<td>([ 412.25, 976.6 ) , (&lt; 0.554116 ), (&lt; 49.4345 )</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>9</td>
<td>B3_B1_RATIO, DISTANCE, ELEVATION</td>
<td>( \geq 0.554116 ), (&lt; 49.4345 ), (&lt; 976.6 )</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>20</td>
<td>ELEVATION, TRMI, DISTANCE</td>
<td>(&lt; 801.9 ), (&lt; 27.5 ), ( \geq 49.4345 )</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>40</td>
<td>B4_B9_RATIO, ELEVATION, TRMI, DISTANCE</td>
<td>(&lt; 0.79721 ), ([ 801.9, 976.6 ) , (&lt; 27.5 ), ( \geq 49.4345 )</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>41</td>
<td>B4_B9_RATIO, ELEVATION, TRMI, DISTANCE</td>
<td>( \geq 0.79721 ), ([ 801.9, 976.6 ) , (&lt; 27.5 ), ( \geq 49.4345 )</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>22</td>
<td>DISTANCE, TRMI, ELEVATION</td>
<td>([49.4345, 94.6041 ) , ( \geq 27.5 ), (&lt; 976.6 )</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>130</td>
<td>B1_B2_RATIO, B4_B5_RATIO, ELEVATION, DISTANCE, TRMI</td>
<td>(&lt; 0.577497 ), (&lt; 0.684353 ), (&lt; 594.8 ), ( \geq 94.6041 ), ( \geq 27.5 )</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>131</td>
<td>B1_B2_RATIO, B4_B5_RATIO, ELEVATION, DISTANCE, TRMI</td>
<td>( \geq 0.577497 ), (&lt; 0.684353 ), (&lt; 594.8 ), ( \geq 94.6041 ), ( \geq 27.5 )</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>132</td>
<td>PCT_SLOPE, B4_B5_RATIO, ELEVATION, DISTANCE, TRMI</td>
<td>(&lt; 23.5896 ), ( \geq 0.684353 ), (&lt; 594.8 ), ( \geq 94.6041 ), ( \geq 27.5 )</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>Node</td>
<td>Variable</td>
<td>Values</td>
<td>Class</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>--------</td>
<td>----------------------</td>
</tr>
<tr>
<td>166</td>
<td>B1_B2_RATIO, PCT_SLOPE, B4_B5_RATIO, ELEVATION, DISTANCE, TRMI</td>
<td>&lt;= 0.536382, &gt;= 23.5896, &gt;= 0.684353, &lt; 594.8, &gt;= 94.6041, &gt;= 27.5</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>167</td>
<td>B1_B2_RATIO, PCT_SLOPE, B4_B5_RATIO, ELEVATION, DISTANCE, TRMI</td>
<td>&gt;= 0.536382, &gt;= 23.5896, &gt;= 0.684353, &lt; 594.8, &gt;= 94.6041, &gt;= 27.5</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>67</td>
<td>ELEVATION, DISTANCE, TRMI</td>
<td>[594.8, 893.3), &gt;= 94.6041, &gt;= 27.5</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>45</td>
<td>ELEVATION, DISTANCE, TRMI</td>
<td>[893.3, 976.6), &gt;= 94.6041, &gt;= 27.5</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>24</td>
<td>ELEVATION, B4_B9_RATIO</td>
<td>[976.6, 1236.45), &lt; 0.813516</td>
<td>Hemlock Dominant/Co-Dominant</td>
</tr>
<tr>
<td>25</td>
<td>ELEVATION, B4_B9_RATIO</td>
<td>[1236.45, 1489.9), &lt; 0.813516</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td>50</td>
<td>TRMI, B4_B9_RATIO, ELEVATION</td>
<td>&lt; 12.5, [0.813516, 0.820715), [976.6, 1489.9)</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>51</td>
<td>TRMI, B4_B9_RATIO, ELEVATION</td>
<td>&gt;= 12.5, [0.813516, 0.820715), [976.6, 1489.9)</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td>52</td>
<td>TRMI, B4_B9_RATIO, ELEVATION</td>
<td>&lt; 21.5, &gt;= 0.820715, [976.6, 1489.9)</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>114</td>
<td>B3_B1_RATIO, B4_B9_RATIO, TRMI, ELEVATION</td>
<td>&lt; 0.52983, [0.820715, 0.836017), &gt;= 21.5, [976.6, 1489.9)</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>154</td>
<td>ELEVATION, B3_B1_RATIO, B4_B9_RATIO, TRMI</td>
<td>[976.6, 1096.55), &gt;= 0.52983, [0.820715, 0.836017), &gt;= 21.5</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>Node</td>
<td>Variable</td>
<td>Values</td>
<td>Class</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>--------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>194</td>
<td>ELEVATION CURVATURE B3_B1_RATIO B4_B9_RATIO TRMI</td>
<td>[1096.55, 1414.05) &lt; 0.25 &gt;= 0.52983 [0.820715, 0.836017] &gt;= 21.5</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td>195</td>
<td>ELEVATION CURVATURE B3_B1_RATIO B4_B9_RATIO TRMI</td>
<td>[1414.05, 1489.9) &lt; 0.25 &gt;= 0.52983 [0.820715, 0.836017] &gt;= 21.5</td>
<td>Non-Hemlock Evergreen</td>
</tr>
<tr>
<td>181</td>
<td>CURVATURE ELEVATION B3_B1_RATIO B4_B9_RATIO TRMI</td>
<td>&gt;= 0.25 [1096.55, 1489.9) &gt;= 0.52983 [0.820715, 0.836017] &gt;= 21.5</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td>81</td>
<td>B4_B9_RATIO TRMI ELEVATION</td>
<td>&gt;= 0.836017 &gt;= 21.5</td>
<td>Hemlock Secondary Component/Inclusion</td>
</tr>
<tr>
<td>7</td>
<td>ELEVATION</td>
<td>&gt;= 1489.9</td>
<td>Non-Hemlock Evergreen</td>
</tr>
</tbody>
</table>
6.6 SAS Code: Discriminant Analysis; k-Nearest Neighbor; Logistic Regression

/* import infested locations */
proc import out= WORK.infest
datafile= "C:\temp\gsm_first_year_infestations.csv"
dbms=csv replace ;
   getnames=yes ;
dastrarow=2 ;
   guessingrows=20 ;
data infest1 ;
   set infest ;
   neness = cos((45-mean_as)*3.14159/180) + 1 ; /* convert aspect to
   'northeast-ness' */
   drop trl_type mean_as mean_ds distcode geolcode vegcode disturb geol
   veg ; /* don't need these, based on GSMNP findings */
   rename mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   _label mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   strm_bc = (strm_dist ** 0.2 - 1) / 0.2 ; /* "best" box-cox
   transformation */
    road_bc = (road_dist ** 0.3 - 1) / 0.3 ; /* "best" box-cox
   transformation */
    trl_bc = (trl_dist ** 0.1 - 1) / 0.1 ; /* "best" box-cox
   transformation */
    elev_bc = (mean_el ** 0.8 - 1)/0.8 ; /* "best" box-cox
   transformation */
    pslope_bc = (mean_ps ** 0.6 - 1)/0.6 ; /* "best" box-cox
   transformation */
    neness_bc = (neness ** 0.6 - 1) / 0.6 ; /* "best" box-cox
   transformation */
/* import BRP locations */
proc import out= WORK.blri
datafile= "C:\temp\blri_sample_pts.csv"
dbms=csv replace ;
   getnames=yes ;
dastrarow=2 ;
   guessingrows=20 ;
data blriinf ;
   set blri ;
   neness = cos((45-mean_as)*3.14159/180) + 1 ; /* convert aspect to
   'northeast-ness' */
   drop x y mean_ds ; /* don't need degree slope */
   rename mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   _label mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   strm_bc = (strm_dist ** 0.2 - 1) / 0.2 ; /* "best" box-cox
   transformation */
    road_bc = (road_dist ** 0.3 - 1) / 0.3 ; /* "best" box-cox
   transformation */
    trl_bc = (trl_dist ** 0.1 - 1) / 0.1 ; /* "best" box-cox
   transformation */
/* import infested locations */
proc import out= WORK.infest
datafile= "C:\temp\gsm_first_year_infestations.csv"
dbms=csv replace ;
   getnames=yes ;
dastrarow=2 ;
   guessingrows=20 ;
data infest1 ;
   set infest ;
   neness = cos((45-mean_as)*3.14159/180) + 1 ; /* convert aspect to
   'northeast-ness' */
   drop trl_type mean_as mean_ds distcode geolcode vegcode disturb geol
   veg ; /* don't need these, based on GSMNP findings */
   rename mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   _label mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   strm_bc = (strm_dist ** 0.2 - 1) / 0.2 ; /* "best" box-cox
   transformation */
    road_bc = (road_dist ** 0.3 - 1) / 0.3 ; /* "best" box-cox
   transformation */
    trl_bc = (trl_dist ** 0.1 - 1) / 0.1 ; /* "best" box-cox
   transformation */
/* import BRP locations */
proc import out= WORK.blri
datafile= "C:\temp\blri_sample_pts.csv"
dbms=csv replace ;
   getnames=yes ;
dastrarow=2 ;
   guessingrows=20 ;
data blriinf ;
   set blri ;
   neness = cos((45-mean_as)*3.14159/180) + 1 ; /* convert aspect to
   'northeast-ness' */
   drop x y mean_ds ; /* don't need degree slope */
   rename mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   _label mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
   mean_1f=landform ;
   strm_bc = (strm_dist ** 0.2 - 1) / 0.2 ; /* "best" box-cox
   transformation */
    road_bc = (road_dist ** 0.3 - 1) / 0.3 ; /* "best" box-cox
   transformation */
    trl_bc = (trl_dist ** 0.1 - 1) / 0.1 ; /* "best" box-cox
   transformation */
elev_bc = (mean_el ** 0.8 - 1)/0.8 ; /* "best" box-cox transformation */
pslope_bc = (mean_ps ** 0.6 - 1)/0.6 ; /* "best" box-cox transformation */
neness_bc = (neness ** 0.6 - 1) / 0.6 ; /* "best" box-cox transformation */

/* import random, uninfested locations */
proc import out= WORK.noninf
datafile= "C:\temp\gsm_random_pts.csv"
dbms=dbf replace;
getdeleted=no;

/* convert variable trl_type to integer, set unspecified trl type to 1 */
data noninf1;
set noninf;
  neness = cos((45-mean_as)*3.14159/180) + 1 ; /* convert aspect to 'northeast-ness' */
  drop trl_type mean_as mean_ds distcode geolcode vegcode disturb geol veg ; /* don't need these, based on GSMNP findings */
  rename mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
  mean_lf=landform;
  label mean_el=elev mean_ps=pslope mean_cr=curvat mean_tr=trmi
  mean_lf=landform;
  strm_bc = (strm_dist ** 0.2 - 1) / 0.2 ; /* "best" box-cox transformation */
  road_bc = (road_dist ** 0.3 - 1) / 0.3 ; /* "best" box-cox transformation */
  trl_bc = (trl_dist ** 0.1 - 1) / 0.1 ; /* "best" box-cox transformation */
  elev_bc = (mean_el ** 0.8 - 1)/0.8 ; /* "best" box-cox transformation */
  pslope_bc = (mean_ps ** 0.6 - 1)/0.6 ; /* "best" box-cox transformation */
  neness_bc = (neness ** 0.6 - 1) / 0.6 ; /* "best" box-cox transformation */
run;

/* all infested points - for normality testing */
data allinfpts;
set infest1 blriinf;
run;

/* all uninfested points - for normality testing */
data allnoninfpts;
set noninf1;
run;

/* correlations betw variables */
title 'correlation matrix';
proc corr data=allpts outp=correlations1;
  var strm_dist road_dist trl_dist elev pslope curvat trmi landform neness;
run;

title 'correlation matrix w/ transformed';
proc corr data=allpts outp=correlations2;
proc import datafile="infest" out=infest1 dbms=dlm replace;
    delimiter="	"
run;

proc import datafile="blriinf" out=blriinf dbms=dlm replace;
    delimiter="	"
run;

proc import datafile="noninf" out=noninf1 dbms=dlm replace;
    delimiter="	"
run;

proc import datafile="infest" out=infest2 dbms=dlm replace;
    delimiter="	"
run;

proc import datafile="blriinf" out=blriinf2 dbms=dlm replace;
    delimiter="	"
run;

/* 1st random partition of the infestation data */
data infa1 infh1;
    set infest1;
    retain k 45 n 67;
    if ranuni(358798) < = k/n then do;
        k = k-1;
        output infa1;
    end;
    else output infh1;
    n = n-1;
    drop k n;
/* 1st random partition of the BRP data */
data blra1 blrh1;
    set blriinf;
    retain k 11 n 17;
    if ranuni(358798) < = k/n then do;
        k = k-1;
        output blra1;
    end;
    else output blrh1;
    n = n-1;
    drop k n;
/* 1st random partition of the non-infestation data */
data noninfa1 noninfh1;
    set noninf1;
    retain k 45 n 67;
    if ranuni(358798) < = k/n then do;
        k = k-1;
        output noninfa1;
    end;
    else output noninfh1;
    n = n-1;
    drop k n;
/* 2nd random partition of the infestation data */
data infa2 infh2;
    set infest2;
    retain k 45 n 67;
    if ranuni(432634) < = k/n then do;
        k = k-1;
        output infa2;
    end;
    else output infh2;
    n = n-1;
    drop k n;
/* 2nd random partition of the BRP data */
data blra2 blrh2;
    set blriinf;
    retain k 11 n 17;
    if ranuni(432634) < = k/n then do;
k = k-1 ;
output blra2 ;
end ;
else output blrh2 ;
n = n-1 ;
drop k n ;

/* 2nd random partition of the non-infestation data */
data noninfa2 noninfh2 ;
set noninf1 ;
retain k 45 n 67 ;
if ranuni(432634) <= k/n then do ;
k = k-1 ;
output noninfa2 ;
end ;
else output noninfh2 ;
n = n-1 ;
drop k n ;

/* 3rd random partition of the infestation data */
data infa3 infh3 ;
set infest1 ;
retain k 45 n 67 ;
if ranuni(547856) <= k/n then do ;
k = k-1 ;
output infa3 ;
end ;
else output infh3 ;
n = n-1 ;
drop k n ;

/* 3rd random partition of the BRP data */
data blra3 blrh3 ;
set blriinf ;
retain k 11 n 17 ;
if ranuni(547856) <= k/n then do ;
k = k-1 ;
output blra3 ;
end ;
else output blrh3 ;
n = n-1 ;
drop k n ;

/* 3rd random partition of the non-infestation data */
data noninfa3 noninfh3 ;
set noninf1 ;
retain k 45 n 67 ;
if ranuni(547856) <= k/n then do ;
k = k-1 ;
output noninfa3 ;
end ;
else output noninfh3 ;
n = n-1 ;
drop k n ;

/* 4th random partition of the infestation data */
data infa4 infh4 ;
set infest1;
retain k 45 n 67;
if ranuni(721324) <= k/n then do;
   k = k-1;
   output infa4;
end;
else output infh4;
n = n-1;
drop k n;

/* 4th random partition of the BRP data */
data blra4 blrh4;
   set blriinf;
   retain k 11 n 17;
   if ranuni(721324) <= k/n then do;
      k = k-1;
      output blra4;
   end;
   else output blrh4;
n = n-1;
drop k n;

/* 4th random partition of the non-infestation data */
data noninfa4 noninfh4;
   set noninf1;
   retain k 45 n 67;
   if ranuni(721324) <= k/n then do;
      k = k-1;
      output noninfa4;
   end;
   else output noninfh4;
n = n-1;
drop k n;

/* 5th random partition of the infestation data */
data infa5 infh5;
   set infest1;
   retain k 45 n 67;
   if ranuni(698921) <= k/n then do;
      k = k-1;
      output infa5;
   end;
   else output infh5;
n = n-1;
drop k n;

/* 5th random partition of the BRP data */
data blra5 blrh5;
   set blriinf;
   retain k 11 n 17;
   if ranuni(698921) <= k/n then do;
      k = k-1;
      output blra5;
   end;
   else output blrh5;
n = n-1;
drop k n;
/* 5th random partition of the non-infestation data */
data noninfa5 noninfh5 ;
set noninf1 ;
retain k 45 n 67 ;
if ranuni(698921) <= k/n then do ;
  k = k-1 ;
  output noninfa5 ;
end ;
else output noninfh5 ;
n = n-1 ;
drop k n ;
run ;

/* create first analysis and holdout sets */
data analysis1 ;
  set infa1 blra1 noninfal ;
data holdout1 ;
  set infh1 blrh1 noninfh1 ;

/* create second analysis and holdout sets */
data analysis2 ;
  set infa2 blra2 noninf2 ;
data holdout2 ;
  set infh2 blrh2 noninfh2 ;

/* create third analysis and holdout sets */
data analysis3 ;
  set infa3 blra3 noninf3 ;
data holdout3 ;
  set infh3 blrh3 noninfh3 ;

/* create fourth analysis and holdout set */
data analysis4 ;
  set infa4 blra4 noninf4 ;
data holdout4 ;
  set infh4 blrh4 noninfh4 ;

/* create fifth analysis and holdout set */
data analysis5 ;
  set infa5 blra5 noninf5 ;
data holdout5 ;
  set infh5 blrh5 noninfh5 ;
run ;

/* for multivariate normality testing of each partitioned set's infested group */
data infpt1 ;
  set infa1 blra1 ;
data infpt2 ;
  set infa2 blra2 ;
data infpt3 ;
  set infa3 blra3 ;
data infpt4 ;
  set infa4 blra4 ;
data infpt5 ;
  set infa5 blra5 ;
/* tests for univariate normality */
symbol v=+;
title 'normal QQ plots - all';
proc univariate data = allpts noprint;
    qqplot strm_dist road_dist trl_dist elev pslope curvat trmi landform
    neness/normal(mu=est sigma=est);
title 'histograms - all';
proc univariate data=allpts noprint;
    histogram strm_dist road_dist trl_dist elev pslope curvat trmi
    landform neness;
title 'tests for normality - all';
ods select TestsForNormality;
proc univariate data=allpts normaltest;
    var strm_dist road_dist trl_dist elev pslope curvat trmi
    landform neness;
run;
/* tests for univariate normality with transformed variables */
symbol v=+;
title 'normal QQ plots - transformed';
proc univariate data = allpts noprint;
    qqplot strm_bc road_bc trl_bc elev pslope_bc curvat trmi landform
    neness_bc/normal(mu=est sigma=est);
run;
title 'histograms - transformed';
proc univariate data=allpts noprint;
    histogram strm_bc road_bc trl_bc trl_type elev pslope_bc curvat trmi
    landform neness_bc distcode;
run;
title 'tests for normality - transformed';
ods select TestsForNormality;
proc univariate data=allpts normaltest;
    var strm_bc road_bc trl_bc elev pslope_bc curvat trmi
    landform neness_bc;
run;
/*
title 'Box-Cox for strm_dist';
%BOXCOXAR(allpts, strm_dist, LAMBDAHI=5.0, LAMBDALO=-2.0, NLAMBDA=71,
OUT=bcstrmd);
title 'Box-Cox for road_dist';
%BOXCOXAR(allpts, road_dist, LAMBDAHI=5.0, LAMBDALO=-2.0, NLAMBDA=71,
OUT=bcroadd);
title 'Box-Cox for trl_dist';
%BOXCOXAR(allpts, trl_dist, LAMBDAHI=5.0, LAMBDALO=-2.0, NLAMBDA=71,
OUT=bcotr1d);
title 'Box-Cox for pslope';
%BOXCOXAR(allpts, pslope, LAMBDAHI=5.0, LAMBDALO=-2.0, NLAMBDA=71,
OUT=bcpslope);
title 'Box-Cox for elev';
%BOXCOXAR(allpts, elev, LAMBDAHI=5.0, LAMBDALO=-2.0, NLAMBDA=71,
OUT=bcelev);

/* Box-Cox for neness; */
%BOXCOXAR(allpts, neness, LAMBDAHI=5.0, LAMBDALO=-2.0, NLAMBDA=71,
OUT=bcneness);

*****************************************************************************
********* multivariate normality macro ***********/
/* macro multnorm ( */
data=_last_, /* input data set */
var= , /* REQUIRED: variables for test */
/* May NOT be a list e.g. var1-var10 */
plot=yes , /* create chi-square plot? */
hires=yes /* high resolution plot? */
);
options nonotes;
%let lastds=&syslast;
/* Verify that VAR= option is specified */
%if &var= %then %do;
  %put ERROR: Specify test variables in the VAR= argument;
  %goto exit;
%end;
/* Parse VAR= list */
%let _i=1;
%do %while (%scan(&var,&_i) ne %str() );
  %let arg&_i=%scan(&var,&_i);
  %let _i=%eval(&_i+1);
%end;
%let nvar=%eval(&_i-1);
/* Remove observations with missing values */
%put MULTNORM: Removing observations with missing values...;
data _nomiss;
  set &data;
  if nmiss(of &var )=0;
run;
/* Quit if covariance matrix is singular */
%let singular=nonsingular;
%put MULTNORM: Checking for singularity of covariance matrix...;
proc princomp data=_nomiss outstat=_evals noprint;
  var &var ;
run;
%if &syserr=3000 %then %do;
  %put MULTNORM: PROC PRINCOMP required for singularity check.;
  %put %str( Covariance matrix not checked for singularity.);
  %goto findproc;
%end;
data _null_;
set _evals;
where _TYPE_='EIGENVAL';
if round(min(of &var ),1e-8)<=0 then do;
   put 'ERROR: Covariance matrix is singular.';
   call symput('singular','singular');
end;
run;
%if &singular=singular %then %goto exit;

%findproc:
/* Is IML or MODEL available for analysis? */
%let mult=yes; %let multtext=%str( and Multivariate);
%put MULTNORM: Checking for necessary procedures...;
proc iml; quit;
%if &syserr=0 %then %goto iml;
proc model; quit;
%if &syserr=0 and (%substr(&sysvlong,1,9)>=6.09.0450 and %substr(&sysvlong,3,2) ne 10) %then %goto model;
%put MULTNORM: SAS/ETS PROC MODEL with NORMAL option or SAS/IML is required;
%put %str( to perform tests of multivariate normality.
Univariate);
%let mult=no; %let multtext=;
%goto univar;

%iml:
proc iml;
reset;
use _nomiss; read all var {&var} into _x; /* input data */
/* compute mahalanobis distances */
_n=nrow(_x); _p=ncol(_x);
_c= _x-j(_n,1)*_x[:,]; /* centered variables */
_s=( _c`* _c)/_n; /* covariance matrix */
_rij= _c*inv(_s)* _c`; /* mahalanobis angles */
/* get values for probability plot and output to data set */
%if &plot=yes %then %do;
_d=vecdiag( _rij#(_n-1)/_n); /* squared mahalanobis distances */
_rank=ranktie(_d); /* ranks of distances */
_chisq=cinv(( _rank-.5)/_n,_p);/* chi-square quantiles */
_chiplot= d||_chisq;
create _chiplot from _chiplot [colname={'MAHDIST' 'CHISQ'}];
append from _chiplot;
%end;
/* Mardia tests based on multivariate skewness and kurtosis */
_b1p= ( _rij##3)[+,]+/(_n##2); /* skewness */
_b2p=trace( _rij##2)/_n; /* kurtosis */
_k=(_p+1)#(_n+1)#(_n+3)/(_n#(_n+1)#(_p+1)-6)); /* small sample correction */
_b1pchi= _b1p/(_n#_k/6); /* skewness test statistic */
_b1pdf=(_p+1)(_p+2)/6; /* and df */
_b2pnorm=(b2p-_p(_p+2))/sqrt(8(_p+2)/n); /* kurtosis test statistic */
_probblp=1-probchi(_b1pchi,_b1pdf); /* skewness p-value */
_probblp=2*(1-probnorm(abs(_b2pnorm))); /* kurtosis p-value */

/* output results to data sets */
_names="Mardia Skewness", "Mardia Kurtosis";
create _names from _names [colname='TEST'];
append from _names;
_probs=(n b1p b1pchi probblp) // (n b2p b2pnorm probblp);
create _values from _probs [colname=('N' 'VALUE' 'STAT' 'PROB')];
append from _probs;
quit;

data _mult;
merge _names _values;
run;

%univar:
/* get univariate test results */
proc univariate data=_nomiss noprint;
var &var;
output out=_stat normal=&var;
output out=_prob probn=&var;
output out=_n n=&var;
run;
data _univ;
set _stat _prob _n;
run;
proc transpose data=_univ name=variable
out=_tuniv(rename=(col1=stat col2=prob col3=n));
var &var;
run;
data _both;
length test $15.;
set _tuniv
  %if &mult=yes %then _mult;;
  if test='' then if n<=2000 then test='Shapiro-Wilk';
  else test='Kolmogorov';
run;
proc print data=_both noobs split='/';
var variable n test %if &mult=yes %then value;
stat prob;
format prob pvalue.;
*title "MULTNORM macro: Univariate&multtext Normality Tests";
label variable="Variable" test="Test" %if &mult=yes %then value="Multivariate/Skewness &/Kurtosis";
stat="Test/Statistic/Value"
prob="p-value";
run;
%if &plot=yes %then
  %if &mult=yes %then %goto plotstep;
  %else %goto plot;
%else %goto exit;

%model:
/* Multivariate and Univariate tests with MODEL */
proc model data=_nomiss;
  %do _i=1 %to &nvar;
    &&arg&_i = _a;
  %end;
  fit &var / normal;
  title "MULTNORM macro: Univariate\multtext Normality Tests";
run;
%if &plot ne yes %then %goto exit;

%plot:
/* compute values for chi-square Q-Q plot */
proc princomp data=_nomiss std out=_chiplot noprint;
  var &var ;
  run;
%if &syserr=3000 %then %do;
  %put ERROR: PROC PRINCOMP in SAS/STAT needed to do plot.;
  %goto exit;
%end;
data _chiplot;
  set _chiplot;
  mahdist=uss(of prin1-prin&nvar );
  keep mahdist;
  run;
proc rank data=_chiplot out=_chiplot;
  var mahdist;
  ranks rdist;
  run;
data _chiplot;
  set _chiplot nobs=_n;
  chisq=cinv((rdist-.5)/_n,&nvar);
  keep mahdist chisq;
  run;

%plotstep:
/* Create a chi-square Q-Q plot
   NOTE: Very large sample size is required for chi-square asymptotics
   unless the number of variables is very small. */
%if &hires=yes %then proc gplot data=_chiplot;
%else                proc  plot data=_chiplot;
  plot mahdist*chisq;
  label mahdist="Squared Distance"
       chisq="Chi-square quantile";
  title "MULTNORM macro: Chi-square Q-Q plot";
run;
quit;
%if &syserr=3000 %then %do;
   %put MULTNORM: PROC PLOT will be used instead.;
   %let hires=no;
   %goto plotstep;
%end;

%exit:
options notes _last_=&lastds;
title;
%mend;
/**************************************************************************
***************************/

/* tests of multivariate normality - all infested data */
/* original variables */
title 'Multivariate normality, infested group' ;
%multnorm(data=allinfpts, var=strm_dist road_dist trl_dist elev pslope curvat trmi landform neness) ;
/* original variables plus transformations */
title 'Multivariate normality, infested group, with transformations' ;
%multnorm(data=allinfpts, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc) ;

/* tests of multivariate normality - all uninfested data */
/* original variables */
title 'Multivariate normality, infested group' ;
%multnorm(data=allnoninfpts, var=strm_dist road_dist trl_dist elev pslope curvat trmi landform neness) ;
/* original variables plus transformations */
title 'Multivariate normality, uninfested group, with transformations' ;
%multnorm(data=allnoninfpts, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc) ;

/* multivariate analysis of variance (MANOVA) - all data */
/* will also yield univariate tests of significance */
proc glm data = allpts ;
class class ;
model strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc = class ;
manova h=class /printh printe ;
title 'MANOVA' ;
run ;

/* multivariate normality of partitioned analysis sets */
/* purpose: see if they maintain approximate normality, similar to the full data set */
title 'Multivariate normality, analysis1, infested group' ;
%multnorm(data=infpt1, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc) ;
title 'Multivariate normality, analysis2, infested group' ;
%multnorm(data=infpt2, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc) ;
title 'Multivariate normality, analysis3, infested group' ;
%multnorm(data=infpt3, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc) ;
title 'Multivariate normality, analysis4, infested group' ;
%multnorm(data=infpt4, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc);
title 'Multivariate normality, analysis5, infested group';
%multnorm(data=infpt5, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc);
title 'Multivariate normality, analysis1, uninfested group';
%multnorm(data=noninfa1, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc);
title 'Multivariate normality, analysis2, uninfested group';
%multnorm(data=noninfa2, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc);
title 'Multivariate normality, analysis3, uninfested group';
%multnorm(data=noninfa3, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc);
title 'Multivariate normality, analysis4, uninfested group';
%multnorm(data=noninfa4, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc);
title 'Multivariate normality, analysis5, uninfested group';
%multnorm(data=noninfa5, var=strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc);

/* stepwise discriminant analyses with partitioned analysis sets */
/* purpose: to see which variables are consistently selected */
proc stepdisc data=analysis1 method=stepwise all slentry=0.25 slstay=0.25;
class class;
var strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc;
title 'Stepwise Discriminant Analysis1';
proc stepdisc data=analysis2 method=stepwise all slentry=0.25 slstay=0.25;
class class;
var strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc;
title 'Stepwise Discriminant Analysis2';
proc stepdisc data=analysis3 method=stepwise all slentry=0.25 slstay=0.25;
class class;
var strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc;
title 'Stepwise Discriminant Analysis3';
proc stepdisc data=analysis4 method=stepwise all slentry=0.25 slstay=0.25;
class class;
var strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc;
title 'Stepwise Discriminant Analysis4';
proc stepdisc data=analysis5 method=stepwise all slentry=0.25 slstay=0.25;
class class;
var strm_bc road_bc trl_bc elev pslope_bc trmi landform neness_bc;
title 'Stepwise Discriminant Analysis5';
run;

/* logistic regression - variable selection via AIC */
/* purpose: to see which variables are consistently selected */
/* logistic regression */
proc logistic data=analysis1 outmodel=sasuser.logregout1aic;
class class;
model class = strm_dist road_dist trl_dist elev pslope trmi landform neness/ rsquare details selection=stepwise sle=1 sls=1 lackfit;
output out=outlr1aic prob=prob1 predprobs=crossvalidate ;
title 'Logistic Regression analysis1 aic' ;
ods output ModelBuildingSummary=SUM1 ;
ods output FitStatistics=FIT1 ;
proc logistic data = analysis2 outmodel = sasuser.logregout2aic ;
class class ;
model class = strm_dist road_dist trl_dist elev pslope trmi landform
neness/ rsquare details selection=stepwise sle=1 sls=1 lackfit;
output out=outlr2aic prob=prob2 predprobs=crossvalidate ;
title 'Logistic Regression analysis2 aic' ;
ods output ModelBuildingSummary=SUM2 ;
ods output FitStatistics=FIT2 ;
proc logistic data = analysis3 outmodel = sasuser.logregout3aic ;
class class ;
model class = strm_dist road_dist trl_dist elev pslope trmi landform
neness/ rsquare details selection=stepwise sle=1 sls=1 lackfit;
output out=outlr3aic prob=prob3 predprobs=crossvalidate ;
title 'Logistic Regression analysis3 aic' ;
ods output ModelBuildingSummary=SUM3 ;
ods output FitStatistics=FIT3 ;
proc logistic data = analysis4 outmodel = sasuser.logregout4aic ;
class class ;
model class = strm_dist road_dist trl_dist elev pslope trmi landform
neness/ rsquare details selection=stepwise sle=1 sls=1 lackfit;
output out=outlr4aic prob=prob4 predprobs=crossvalidate ;
title 'Logistic Regression analysis4 aic' ;
ods output ModelBuildingSummary=SUM4 ;
ods output FitStatistics=FIT4 ;
proc logistic data = analysis5 outmodel = sasuser.logregout5aic ;
class class ;
model class = strm_dist road_dist trl_dist elev pslope trmi landform
neness/ rsquare details selection=stepwise sle=1 sls=1 lackfit;
output out=outlr5aic prob=prob5 predprobs=crossvalidate ;
title 'Logistic Regression analysis5 aic' ;
ods output ModelBuildingSummary=SUM5 ;
ods output FitStatistics=FIT5 ;
run ;
/*************************************************************************/
/****************//* ANALYSIS SECTION */
/*************************************************************************/

discriminant analysis for the five subsets */
/* decided to try just the four variables that were highlighted by
stepwise process */
/* also tests whether should be linear or quadratic DA, turns out to be
quadratic */
proc discrim data=analysis1 method=normal outstat=analysis1da
ptestdata=holdout1 testout=holdout1da pool=test testlist list crossvalidate
manova anova ;
class class ;
testclass class ;
var strm_bc road_bc trl_bc elev ;
title 'Discrim Analysis 1' ;
proc discrim data=analysis2 method=normal outstat=analysis2da
ptestdata=holdout2 testout=holdout2da pool=test testlist list crossvalidate
manova anova ;
class class;
testclass class;
var strm_bc road_bc trl_bc elev;
title 'Discrim Analysis 2';
proc discrim data=analysis3 method=normal outstat=analysis3da
testdata=holdout3 testout=holdout3da pool=test testlist list crossvalidate
manova anova;
class class;
testclass class;
var strm_bc road_bc trl_bc elev;
title 'Discrim Analysis 3';
proc discrim data=analysis4 method=normal outstat=analysis4da
testdata=holdout4 testout=holdout4da pool=test testlist list crossvalidate
manova anova;
class class;
testclass class;
var strm_bc road_bc trl_bc elev;
title 'Discrim Analysis 4';
proc discrim data=analysis5 method=normal outstat=analysis5da
testdata=holdout5 testout=holdout5da pool=test testlist list crossvalidate
manova anova;
class class;
testclass class;
var strm_bc road_bc trl_bc elev;
title 'Discrim Analysis 5';
run;

/* non-parametric discriminant analysis for the five subsets */
/* since not working under assumption of normality, didn't bother with
transformations of variables */
proc discrim data=analysis1 method=npar k=8 outstat=analysis1np
testdata=holdout1 testout=holdout1np testlist list crossvalidate;
class class;
testclass class;
var strm_dist road_dist trl_dist elev;
title 'Non-Parametric Discrim Analysis 1, no transforms';
proc discrim data=analysis2 method=npar k=8 outstat=analysis2np
testdata=holdout2 testout=holdout2np testlist list crossvalidate;
class class;
testclass class;
var strm_dist road_dist trl_dist elev;
title 'Non-Parametric Discrim Analysis 2, no transforms';
proc discrim data=analysis3 method=npar k=8 outstat=analysis3np
testdata=holdout3 testout=holdout3np testlist list crossvalidate;
class class;
testclass class;
var strm_dist road_dist trl_dist elev;
title 'Non-Parametric Discrim Analysis 3, no transforms';
proc discrim data=analysis4 method=npar k=8 outstat=analysis4np
testdata=holdout4 testout=holdout4np testlist list crossvalidate;
class class;
testclass class;
var strm_dist road_dist trl_dist elev;
title 'Non-Parametric Discrim Analysis 4, no transforms';
proc discrim data=analysis5 method=npar k=8 outstat=analysis5np
testdata=holdout5 testout=holdout5np testlist list crossvalidate;
class class;
testclass class;
var strm_dist road_dist trl_dist elev;
title 'Non-Parametric Discrim Analysis 5, no transforms';
run;

/*logistic regression for each of the subsets*/
proc logistic data = analysis1 outmodel = sasuser.analysis1lr;
class class;
model class = strm_dist road_dist trl_dist elev pslope/ rsquare selection=none ctable;
output out=allr prob=prob predprobs=crossvalidate;
title 'Logistic Regression Analysis 1';
/* logistic regression classification tables */
proc print data = allr;
var _FROM_ _INTO_ XP_Infested XP_Non_Infested _LEVEL_ prob;
title 'LR classif tables Analysis 1';
proc logistic inmodel = sasuser.analysis1lr;
score data = analysis1 out=allscore fitstat;
title 'log reg scores Analysis 1';
proc print data=allscore;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Analysis 1';
proc logistic inmodel = sasuser.analysis1lr;
score data = holdout1 out=h1score fitstat;
title 'log reg scores Holdout 1';
proc print data=h1score;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Holdout 1';
run;

proc logistic data = analysis2 outmodel = sasuser.analysis2lr;
class class;
model class = strm_dist road_dist trl_dist elev pslope/ rsquare selection=none ctable;
output out=a2lr prob=prob predprobs=crossvalidate;
title 'Logistic Regression Analysis 2';
/* logistic regression classification tables */
proc print data = a2lr;
var _FROM_ _INTO_ XP_Infested XP_Non_Infested _LEVEL_ prob;
title 'LR classif tables Analysis 2';
proc logistic inmodel = sasuser.analysis2lr;
score data = analysis2 out=a2score fitstat;
title 'log reg scores Analysis 2';
proc print data=a2score;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Analysis 2';
proc logistic inmodel = sasuser.analysis2lr;
score data = holdout2 out=h2score fitstat;
title 'log reg scores Holdout 2';
proc print data=h2score;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Holdout 2';
run;

proc logistic data = analysis3 outmodel = sasuser.analysis3lr;
class class;
model class = strm_dist road_dist trl_dist elev pslope/ rsquare
selection=none ctable;
output out=a3lr prob=prob predprobs=crossvalidate;
title 'Logistic Regression Analysis 3';
/* logistic regression classification tables */
proc print data = a3lr;
var _FROM_ _INTO_ XP_Infested XP_Non_Infested _LEVEL_ prob;
title 'LR classif tables Analysis 3';
proc logistic inmodel = sasuser.analysis3lr;
score data = analysis3 out=a3score fitstat;
title 'log reg scores Analysis 3';
proc print data=a3score;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Analysis 3';
proc logistic inmodel = sasuser.analysis3lr;
score data = holdout3 out=h3score fitstat;
title 'log reg scores Holdout 3';
proc print data=h3score;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Holdout 3';
run;

proc logistic data = analysis4 outmodel = sasuser.analysis4lr;
class class;
model class = strm_dist road_dist trl_dist elev pslope/ rsquare
selection=none ctable;
output out=a4lr prob=prob predprobs=crossvalidate;
title 'Logistic Regression Analysis 4';
/* logistic regression classification tables */
proc print data = a4lr;
var _FROM_ _INTO_ XP_Infested XP_Non_Infested _LEVEL_ prob;
title 'LR classif tables Analysis 4';
proc logistic inmodel = sasuser.analysis4lr;
score data = analysis4 out=a4score fitstat;
title 'log reg scores Analysis 4';
proc print data=a4score;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Analysis 4';
proc logistic inmodel = sasuser.analysis4lr;
score data = holdout4 out=h4score fitstat;
title 'log reg scores Holdout 4';
proc print data=h4score;
var F_class I_class P_infested P_non_infested;
title 'log reg score table Holdout 4';
run;

proc logistic data = analysis5 outmodel = sasuser.analysis5lr;
class class;
model class = strm_dist road_dist trl_dist elev pslope/ rsquare
selection=none ctable;
output out=a5lr prob=prob predprobs=crossvalidate;
title 'Logistic Regression Analysis 5';
/* logistic regression classification tables */
proc print data = a5lr;
var _FROM_ _INTO_ XP_Infested XP_Non_Infested _LEVEL_ prob;
title 'LR classif tables Analysis 5';
proc logistic inmodel = sasuser.analysis5lr;
score data = analysis5 out=a5score fitstat ;
title 'log reg scores Analysis 5' ;
proc print data=a5score ;
var F_class I_class P_infested P_non_infested ;
title 'log reg score table Analysis 5' ;
proc logistic inmodel = sasuser.analysis5lr ;
score data = holdout5 out=h5score fitstat ;
title 'log reg scores Holdout 5' ;
proc print data=h5score ;
var F_class I_class P_infested P_non_infested ;
title 'log reg score Table Holdout 5' ;
run ;