Landscape-Scale Prediction of Hemlock Woolly Adelgid, *Adelges tsugae* (Homoptera: Adelgidae), Infestation in the Southern Appalachian Mountains

F. H. KOCH,[1,2] H. M. CHESHIRE,[1] AND H. A. DEVINE[3]

Environ. Entomol. 35(5): 1313–1323 (2006)

ABSTRACT After causing substantial mortality in the northeastern and mid-Atlantic United States, the hemlock woolly adelgid, *Adelges tsugae* Annand (Homoptera: Adelgidae), has recently invaded the southern Appalachian region. Although general estimates of regional spread exist, the landscape-level dynamics of *A. tsugae* invasion are poorly understood—particularly factors predicting where the pest is likely to first infest a landscape. We examined first-year 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 variables using a geographic information system. After identifying significant variables, we applied four statistical techniques—discriminant analysis, *k*-nearest neighbor analysis, logistic regression, and decision trees—to derive classification functions separating the infested and uninfested groups. We used the resulting functions to generate maps of *A. tsugae* infestation risk in the Great Smoky Mountains. Three proximity variables (distance to the closest stream, trail, and road) appeared in all four classification functions, which performed well in terms of error rate. Discriminant analysis was the most accurate and efficient technique, but logistic regression best balanced accuracy, efficiency, and ease of use. Our results suggest that roads, major trails, and riparian corridors provide connectivity enabling long-distance dispersal of *A. tsugae*, probably by humans or birds. The derived classification functions can yield *A. tsugae* infestation risk maps for elsewhere in the southern Appalachian region, allowing forest managers to better target control efforts.

KEY WORDS hemlock woolly adelgid, southern Appalachians, dispersal, prediction, landscape connectivity

Forest product losses and control costs in the United States caused by nonindigenous insects have been estimated at $2.1 billion per year (Pimentel et al. 2000). In ecological terms, these invaders substantially change forest composition, structure, and microenvironment, alter critical ecosystem processes, and enable further invasion and disturbance (Liebhold et al. 1995, Orwig 2002). The hemlock woolly adelgid, *Adelges tsugae* Annand (Homoptera: Adelgidae), represents a significant threat to the long-term health of forests in the southern Appalachian Mountains of the United States. An Asian native accidentally introduced to Richmond, VA, in the 1950s, *A. tsugae* was at first considered a mere nuisance to ornamental hemlocks (McClure et al. 2001, Cheah et al. 2004). However, *A. tsugae* moved into natural forest stands during the 1980s, causing extensive hemlock mortality in the northeastern and mid-Atlantic United States before recently moving into the southern Appalachian region (Cheah et al. 2004).

A. tsugae attacks all age classes of eastern hemlock (*Tsuga canadensis* Carrière) and Carolina hemlock (*T. caroliniana* Engelmann), the region's native species. Neither species is resistant, although infested trees in more mesic sites or in deep ravines may experience slower mortality (Orwig and Foster 1998, McClure et al. 2001, Ward et al. 2004). Both *T. canadensis* and *T. caroliniana* are slow-growing, long-lived, and occupy ecological niches not easily filled by other tree species (Orwig and Foster 1998, Ward et al. 2004). *A. tsugae* is parthenogenetic, with two annual generations on hemlock: the sistens generation hatches in late spring and survives for about nine months, while the progrediens generation hatches in early spring and survives for about three months (Cheah et al. 2004). Typical egg production is high for both generations: 50–175 eggs for sistens, 25–125 eggs for progrediens (McClure et al. 2001). This supports explosive population growth that can quickly lead to severe needle loss and dieback, often killing trees within a few years

[1] Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695.

[2] Corresponding author: Department of Forestry and Environmental Resources, North Carolina State University, 3041 Cornwallis Rd., Research Triangle Park, NC 27709 (e-mail: fkoch@fs.fed.us).

[3] Department of Parks, Recreation, and Tourism Management, North Carolina State University, Campus Box 7106, Raleigh, NC 27695.

0046-225X/06/1313–1323$04.00/0 © 2006 Entomological Society of America

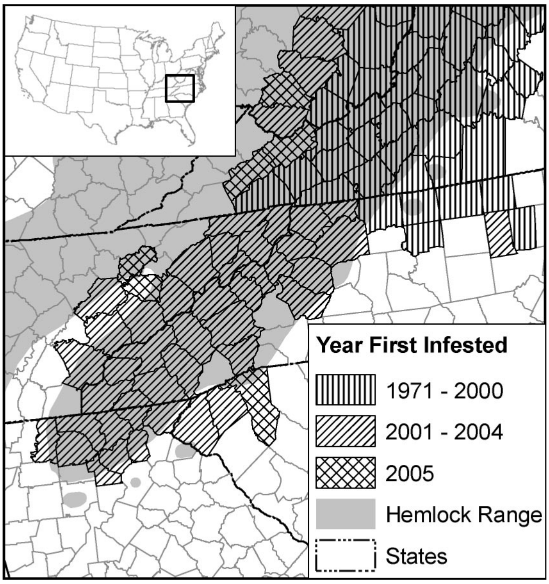

Fig. 1. Spatio-temporal pattern of spread of the hemlock woolly adelgid (by county) in the southern Appalachian region.

of infestation (McClure et al. 2001, Cheah et al. 2004). *A. tsugae* eggs and mobile first instar nymphs, or crawlers, are passively dispersed by wind, deer, birds, and humans, perhaps as much as 30 km per year (McClure 1990, 1996; Souto et al. 1996, Cheah et al. 2004).

Adelges tsugae has no natural enemies in the United States, and broad chemical control is unfeasible for natural hemlock stands because of cost or, in some cases, habitat sensitivity (Wallace and Hain 2000, Ward et al. 2004). Biological control has been embraced as the most effective approach for *A. tsugae* management. Imported predator species such as *Sasajiscymnus tsugae* Sasaji and McClure (Coleoptera: Coccinellidae) and *Laricobius nigrinus* Fender (Coleoptera: Derodontidae) have been released in forested areas, and there is ongoing research into insect-killing pathogens (Cheah and McClure 1998, Cheah et al. 2004, Flowers et al. 2005). Nevertheless, predators and other control agents take time to establish in a new setting, and establishment must happen before *A. tsugae* causes irreversible damage (Cheah and McClure 2002, Orwig et al. 2002, Cheah et al. 2004). Therefore, our chief research objective was to develop a method for prioritizing hemlock stands by infestation risk. This would, in turn, help forest managers narrow their area of focus and better target their *A. tsugae* control efforts.

Previously, the regional pattern of *A. tsugae* dispersal has been estimated at a coarse, county-level scale (Fig. 1). Finer-scale factors influencing spatial distribution—such as topography or connectivity of host species stands—have generally been neglected, in part because low-level *A. tsugae* populations can be difficult to detect (Orwig et al. 2002, Costa 2005). We set a goal to incorporate landscape-scale factors shaping *A. tsugae* spread into our prioritization method. Circumstances in the southern Appalachian region granted us a rare opportunity. Because we were able to procure data depicting specific geographic locations where *A. tsugae* was first detected in the region, we could apply multivariate statistical classification techniques to (1) identify key landscape variables distinguishing infested and noninfested sites, and (2) build functions based on these variables to predict the sites most likely to be infested. By implementing these classification functions in a geographic information system (GIS), we hoped to create useful maps of *A. tsugae* infestation risk.

Statistical classification methods that may be applied in such a manner include discriminant analysis (a parametric approach with linear and quadratic forms), *k*-nearest neighbor analysis (a type of nonparametric discriminant analysis), logistic regression, and decision trees. Generally, these techniques assign a given observation to one of two groups based on measurements for a suite of variables (Afifi et al. 2004). They require an initial sample where the group designation of each observation is known a priori, which serves as a training data set for calculating a function that can be used to predict the group membership of additional data points (McLachlan 1992, Afifi et al. 2004). While similar in application, the mechanics of these techniques are quite distinct, and each operates under different assumptions that may affect the implementation or the performance of the technique. For example, linear and quadratic discriminant analysis are restricted to continuous explanatory variables and assume the input data have a multivariate normal distribution, meaning data transformations are commonly necessary (Johnson and Wichern 2002, Afifi et al. 2004). Instead, a nonparametric technique such as *k*-nearest neighbor analysis may be used, but it is likely to be outperformed by linear/quadratic discriminant analysis if the data can be successfully transformed to multivariate normality (Khattree and Naik 2000). Logistic regression also does not assume normality and can handle both continuous and categorical input variables (Neter et al. 1996, Afifi et al. 2004). Nevertheless, it can be sensitive to outliers and multicollinearity and requires a large sample if many variables are used (Peduzzi et al. 1996, Allison 1999, Afifi et al. 2004). Decision tree approaches are nonparametric, can accommodate many variable types, and yield a set of interpretable rules for classifying additional data (Murthy 1998, De'Ath and Fabricius 2000). However, they may require large input samples for accurate classification, and the output decision trees may overfit the input data if not pruned using validation data or some other approach (Breiman et al. 1984, Jovanovic et al. 2002).

All of these techniques have been used with some success in predicting invasion patterns for a variety of landscapes (Beck et al. 1994, 1997, Gumpertz et al. 1999, Negrón et al. 2000, Kelly and Meentemeyer 2002, Liu et al. 2003, Wilson et al. 2003, Rouget et al. 2004, Jacquez et al. 2005). None consistently outperforms the others in all circumstances (Murthy 1998). For

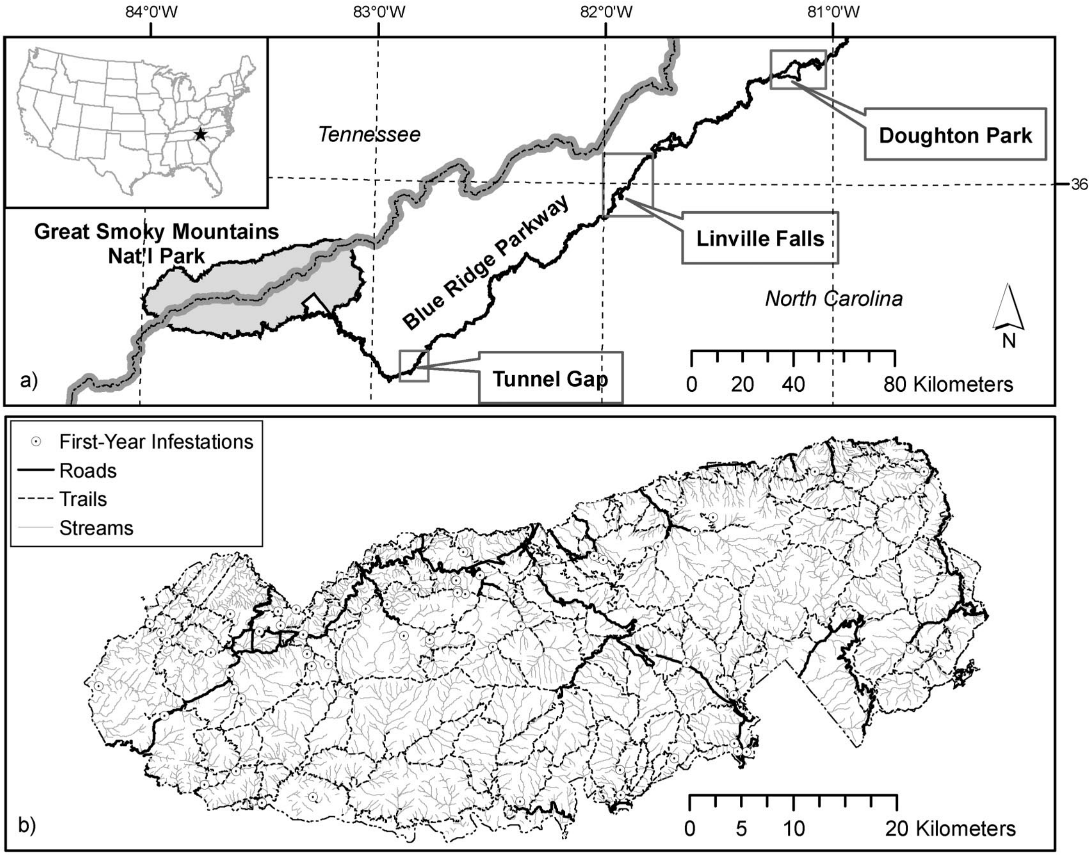

Fig. 2. General location of study areas. (a) Great Smoky Mountains National Park and the southern Blue Ridge Parkway region. (b) Close-up of Great Smoky Mountains National Park.

relatively small, heterogeneous environmental data sets such as ours, it can be difficult to predict which methods will work best. Therefore, to develop the best function for predicting areas most likely to be infested by *A. tsugae*, we compared the aforementioned classification approaches in terms of their accuracy, efficiency, and interpretability when used in a GIS context, as well as the landscape variables each identified as important for *A. tsugae* spread.

Materials and Methods

We performed all statistical analyses using SAS 9.1 software, including the SAS Enterprise Miner module for building decision trees (SAS Institute 2003). We used ArcGIS 8.3 software (Environmental Systems Research Institute 2002) for GIS operations.

Study Area. We used data from Great Smoky Mountains National Park (GSMNP) and the Blue Ridge Parkway (BRP), two U.S. National Park Service units in the southern Appalachian region (Fig. 2). A large park (>2,000 km^2), GSMNP has a significant *T. canadensis* presence in many areas (Whittaker 1956, Taylor 2002). *A. tsugae* was first detected in GSMNP in 2002. In contrast, BRP is a long, narrow road corridor connecting several small recreation areas. Both *T. canadensis* and *T. caroliniana* hemlocks occur within its borders, with sites such as Linville Falls exhibiting substantial hemlock presence. *A. tsugae* was first detected in the southern portion of BRP in 2003.

Data Compilation. After *A. tsugae* was discovered in GSMNP, park staff completed an extensive survey during the next few months to map the distribution of *A. tsugae* during the initial period of infestation. They recorded 67 distinct locations as points with global positioning system (GPS) units (Fig. 2b). The park also has a recent vegetation map, created from 1:12,000 scale aerial photographs, that depicts the distribution of hemlock stands (Welch et al. 2002). For comparison with the infested sites, we generated a random sample of 67 points from the mapped hemlock stands. These GIS points represented locations uninfested by *A. tsugae* during the initial period of invasion.

Park staff in BRP used GPS to record the locations of 17 distinct hemlock *A. tsugae* infestations in the months immediately after the pest was first detected in the park. Most of these locations were clustered in a few areas: Linville Falls, Doughton Park, and near Tunnel Gap (Fig. 2a). As with GSMNP, this layer is believed to be an accurate depiction of *A. tsugae* distribution for the southern portion of BRP during the initial period of infestation. However, BRP does not

Table 1. Variables tested for inclusion in the classification functions

Variable	Description	Mean (SD)	
		Infested	Uninfested
Aspect	Slope direction transformed to a measure of "northeast-ness" scaled 0–2; from Beers et al. (1966)	1.08 (0.69)	1.13 (0.71)
Curvature	Terrain convexity (+ values)/concavity (− values) measure	−0.37 (1.20)	−0.48 (1.38)
Elevation	Elevation in meters	793.05 (207.92)	1011.84 (342.81)
Landform index	Site exposure index scaled −1 (protected) to 1 (exposed); from McNab (1993)	−0.01 (0.02)	−0.01 (0.03)
Percent slope	Percent slope	29.93 (20.84)	44.94 (19.19)
Topographic relative moisture index	Site dryness-wetness index scaled 0 (xeric) to 60 (very mesic); from Parker (1982)	33.44 (10.03)	31.05 (9.49)
Distance to road	Distance in meters to closest road	1124.13 (1154.25)	696.69 (675.83)
Distance to stream	Distance in meters to closest stream	83.44 (110.25)	163.24 (148.16)
Distance to trail	Distance in meters to closest trail	254.55 (391.59)	696.69 (675.83)
Trail type	Composition of closest trail; categorical, three classes[a]	NA	NA
Disturbance history	Harvested or cleared land; categorical, five classes[a]	NA	NA
Geology	General bedrock formations; categorical, 26 different classes[a]	NA	NA
Vegetation	Dominant vegetation type in neighborhood around point of interest; categorical, 14 classes[a]	NA	NA

Mean and SD values are for the combined GSMNP and BRP data sets (infested $N = 84$, uninfested $N = 67$).
[a] Variables were only available for GSMNP data points.
NA, not available.

have a map of hemlock distribution, so we were unable to extract points representing uninfested hemlock stands.

For each of the points in GSMNP and BRP, we recorded values for a suite of landscape variables (Table 1) calculated from existing GIS layers. First, we chose topographic variables because of evidence that terrain characteristics may influence the rate of hemlock decline (Royle and Lathrop 1999, Orwig et al. 2002, Ward et al. 2004). We extracted these variables from 10-m digital elevation models of the two parks. Although each *A. tsugae* infestation was recorded as a point, it more realistically represented an area of infested hemlocks, so we derived mean values for the variables based on a 3-by-3 window of grid pixels around each point. For computation, we converted the mean aspect grid into a "northeast-ness" grid with rescaled aspect values based on their deviation from 45° (Beers et al. 1966).

We included road, stream, and trail proximity variables because all have been implicated as corridors of introduction for invasive pests (Zobel et al. 1985, McClure 1990, Forys et al. 2001, Maelzer et al. 2004, Ward et al. 2004). The GIS database for GSMNP included detailed (typically 1:24,000 scale) vector data on streams, roads, and trails, so we calculated the straight-line distance between each GSMNP sample point and the closest feature in each of these layers. We also labeled each point according to the composition (unimproved foot path, old roadbed, or paved path) of the closest trail. In addition, the GSMNP database had raster data layers for disturbance history, geology, and vegetation type. For disturbance history and geology, we assigned each sample point the value of the pixel (90-m resolution) in which it fell. For vegetation, we labeled each point according to the type exhibited by a majority of pixels (30-m resolution) in a 3-by-3 window around the point. Data layers for disturbance, geology, or vegetation were unavailable for BRP, nor did the park have extensive vector data. To generate proximity variables for BRP, we compiled stream, road, and trail vector data layers from U.S. Geological Survey 1:24,000 digital line graphs and calculated straight-line distances between each sample point and the closest feature in each category.

Exploratory Data Analysis. We generated a correlation matrix for all data points to detect collinear or redundant variables. We assessed univariate normality of the continuous variables based on Shapiro-Wilk tests and applied Box-Cox transformations as necessary for any non-normal variables (Johnson and Wichern 2002). We partitioned the GSMNP and BRP data sets to create separate training and validation data samples to, respectively, build the classification functions and evaluate their performance. For the GSMNP data set, we randomly held out one third of both the infested points and uninfested points (i.e., 22 of the 67 points in each group) as validation data, leaving the remainder as training data. Similarly, we randomly held out one third of the points in the BRP data set as validation data (i.e., 6 of the 17 points, all of which represented infested locations), leaving the remainder as training data. Merging the resulting subsets from GSMNP and BRP yielded a 101-point (56 infested, 45 uninfested) training sample and a 50-point (28 infested, 22 uninfested) validation sample. For variable selection and performance assessment, it is advantageous to use a repeated holdout approach where the original data are randomly partitioned several times in the above-described manner (Johnson and Wichern 2002). We repeated the random partitioning process four more times, yielding five differently partitioned training/validation data sets (i.e., each had a unique combination of 101 training and 50 validation observations). We tested each of the five training data sets for multivariate normality based on skewness and kurtosis measures (Mardia 1974).

Variable Selection. The inclusion of a large number of variables does not necessarily increase classification success and can be problematic with smaller samples, so we applied selection procedures to create a best subset of variables for each of our classification techniques (Khattree and Naik 2000, Johnson and Wichern 2002). Using the continuous variables, we performed stepwise discriminant analysis, which generates a sequence of models that adds or removes variables based on *F*-tests of significance (Khattree and Naik 2000). We selected variables with $P < 0.25$ for inclusion in our reduced discriminant function. While stepwise selection may also be used for logistic regression, the approach has been criticized for resulting in too few variables for successful prediction (Shtatland et al. 2001). Instead, we built a sequence of reduced logistic regression models and selected the model that minimized Akaike's Information Criterion (AIC), a measure that incorporates a penalty for model complexity (Shtatland et al. 2001). 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 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 built a second tree with the combined GSMNP and BRP data. We adopted 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, we used the validation data set to "prune" the tree, i.e., to identify the best subtree and thus the best subset of variables.

We applied these procedures to the five partitioned training data sets. For each procedure, if a variable was identified as significant for at least three of the five sets, we included it in a final variable subset. This yielded three variable subsets, each matched to one of the classification techniques. We also used the subset identified by stepwise discriminant analysis for our *k*-nearest neighbor classification procedures.

Classification. Using the selected variable subset, we applied each classification technique to the five partitioned data sets. For all but the decision tree approach, we derived a classification function from the training data and applied the resulting function to the validation data, recording the minimum, maximum, and mean classification error rate for both samples. For the decision tree, we used the validation data for tree pruning and classified both the training and validation data sets using the pruned tree, recording the associated error rates. We used Box-Cox-transformed variables for discriminant analysis as necessary, first testing the hypothesis of a pooled covariance to determine whether a linear or quadratic discriminant analysis approach was more appropriate (Johnson and Wichern 2002, Nelson et al. 2003). For the *k*-nearest neighbor approach, we used the eight closest neighbors based on Mahalanobis distance. For logistic regression, we classified an observation as infested if the equation yielded a predicted infestation probability ≥ 0.5.

One data set yielded the median or near-median error rate for all four classification techniques. For model verification, we used the results for this median data set to construct error matrices and compute Cohen's κ statistic. The κ statistic indicates how much of an improvement a classification effort represents over a completely random classification of the same data (Jensen 1996). Manel et al. (2001) suggested that κ is a simple, efficient statistic for comparing the success of predictive models, including those created with different algorithms.

Map Generation. We also used the classification functions generated from the median data set to create *A. tsugae* infestation maps for GSMNP. We prepared raster grids for the incorporated variables and applied each function to create binary maps of infestation risk in GSMNP. From each map, we recorded how much area was classified as likely to be infested. We expressed that value as a percentage of the park's total land area (2,093 km^2) and as a percentage of the park area with a mapped hemlock presence (294 km^2). These percentages served as relative measures of efficiency, i.e., how much total land area and hemlock area each classification function required in achieving its error rate.

After the first year of infestation, GSMNP staff members continued to regularly survey for *A. tsugae* and treated many sites by biological or chemical control. Between 2003 and 2005, 383 additional infestation points in the park were recorded with GPS. To test whether the infestation risk zones defined by our classification functions reasonably predicted the pattern of *A. tsugae* expansion after the first year, we recorded the percentage of survey points that fell in the areas delineated as likely to be infested on the binary maps generated with each classification function.

Results

Exploratory Data Analysis. We omitted curvature from all further analyses because it was redundant with landform index ($r = 0.99$). With the other continuous variables, none of the partitioned data sets exhibited significant multivariate skewness or kurtosis for either group, and χ^2 Q-Q plots suggested at least approximate multivariate normality for all five data sets. For discriminant analysis, we used a quadratic approach because testing indicated the group covariance matrices were unequal. Moreover, while a linear

Table 2. Variables selected for each classification function and resulting classification error rates (ERs) for the five partitioned data sets

Classification technique	Variables selected/incorporated in function	Training data ER			Validation data ER		
		Avg.	Min.	Max.	Avg.	Min.	Max.
Discriminant analysis	Distance to road, distance to stream, distance to trail, elevation	0.16	0.15	0.17	0.14	0.10	0.18
k-nearest neighbor	Distance to road, distance to stream, distance to trail, elevation	0.23	0.20	0.28	0.22	0.16	0.28
Logistic regression	Distance to road, distance to stream, distance to trail, elevation, percent slope	0.20	0.17	0.23	0.19	0.14	0.24
Decision tree	Distance to road, distance to stream, distance to trail	0.18	0.14	0.22	0.24	0.16	0.30

discriminant approach may still work in such cases, the quadratic approach yielded lower error rates in preliminary analyses.

Variable Selection. The three variable selection procedures identified similar variable sets, with the proximity variables most prominent (Table 2). Only logistic regression offered a straightforward numeric measure (generalized R^2) for model fit; values for the five training data sets ranged from 0.352 to 0.433. The fitted logistic equation (equation 1, for the median data set had distance to trail as the most significant variable and elevation the least significant based on χ^2 estimates).

$$\log\ (\text{odds of infestation}) = 5.3321 - 0.00497 \times \text{stream\_distance} - 0.00072 \times \text{road\_distance} - 0.00264 \times \text{trail\_distance} - 0.00150 \times \text{elevation} - 0.0242 \times \text{pct\_slope} \quad (1)$$

The decision tree for the median data set had six terminal nodes and a depth of three levels (Fig. 3). Neither the AIC analysis nor the pruned decision trees identified any categorical variables as significant.

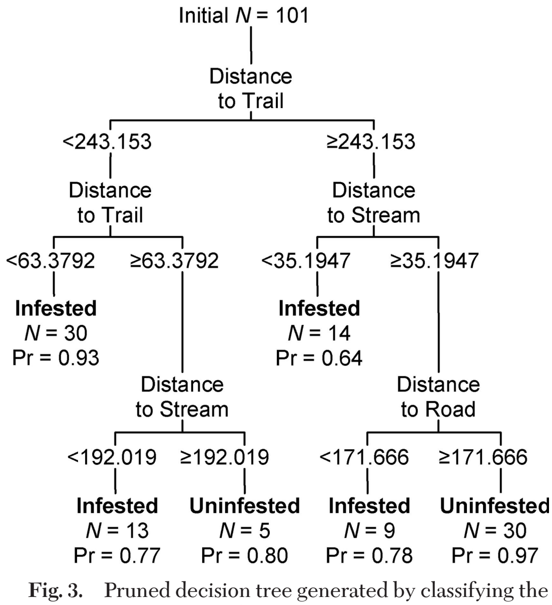

Fig. 3. Pruned decision tree generated by classifying the median data set. All variable values are in meters.

Classification Error Rates. Discriminant analysis had the lowest mean error rate for the training and validation data and exhibited the smallest range between minimum and maximum error rates for the five sets (Table 2). In contrast, *k*-nearest neighbor analysis performed worst for the training data sets and next to worst for the validation data sets. However, error rates for the training and validation data were similar, suggesting the *k*-nearest neighbor classification function remained consistent when applied to new data points. Logistic regression fell between discriminant analysis and *k*-nearest neighbor analysis. The decision tree approach performed almost as well as discriminant analysis for the training data sets, but it performed worst for the validation data sets, exhibiting the highest mean and maximum error rate. However, the error rate never exceeded 0.30 for the decision tree or any other approach, suggesting all four techniques performed moderately well.

Classification Error Matrices. Error matrices for the median data set (Tables 3 and 4) mirrored the general error rate results. For the training data subset, the decision tree had the highest overall accuracy, slightly ahead of discriminant analysis. Logistic regression performed better than *k*-nearest neighbor, but was still noticeably behind the other two approaches. κ values followed this same trend. Commission and omission errors—the percentage of observations assigned to the wrong group and excluded from the correct group, respectively—indicated that the discriminant analysis was well balanced (i.e., errors did not especially correspond with one group), as was logistic regression. The decision tree and *k*-nearest neighbor approaches were imbalanced, mostly attributable to "false positives" (i.e., uninfested observations mistakenly classified as infested). For the decision tree, the tendency toward false positives carried over to the validation data subset. It had the worst overall accuracy 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 overall, while logistic regression and *k*-nearest neighbor lagged slightly behind. κ values mirrored the overall accuracy estimates.

Prediction Maps. Binary maps generated for the median data set using the four classification functions

Table 3. Error matrices for the training subset of the median data set: (a) discriminant analysis, (b) *k*-nearest neighbor, (c) logistic regression, (d) decision tree

(a)	Reference group			(b)	Reference group		
Classified as . . .	Infested	Uninfested	CE	Classified as . . .	Infested	Uninfested	CE
Infested	*48*	7	12.7	Infested	*49*	15	23.4
Uninfested	8	*38*	17.4	Uninfested	7	*30*	18.9
OE	14.3	15.6		OE	12.5	33.3	
Overall accuracy = 86/100 = 86%				Overall accuracy = 79/100 = 79%			
Cohen's κ = 0.700				Cohen's κ = 0.551			
(c)	Reference group			(d)	Reference group		
Classified as . . .	Infested	Uninfested	CE	Classified as . . .	Infested	Uninfested	CE
Infested	*47*	11	19.0	Infested	*54*	12	18.2
Uninfested	9	*34*	20.9	Uninfested	2	*33*	5.7
OE	16.1	24.4		OE	3.6	26.7	
Overall accuracy = 81/100 = 81%				Overall accuracy = 88/100 = 88%			
Cohen's κ = 0.633				Cohen's κ = 0.713			

Cohen's κ statistic ranges from 0 to 1, representing the degree to which the model's prediction performance is better than chance: 0–0.4 = slight to fair model performance, 0.4–0.6 = moderate performance, 0.6–0.8 = substantial performance, 0.8–1 = near perfect performance (Landis and Koch 1977).

CE, percent commission error; OE, percent omission error.

(Fig. 4) differed in the amount of GSMNP area classified as likely to be infested (Table 5). Discriminant analysis seemed to be the most efficient, classifying the lowest percentage of GSMNP's total area and hemlock area as likely to be infested while exhibiting the lowest classification error. The logistic regression approach required a slightly higher percentage of total park area and hemlock area while exhibiting a higher mean error rate. Nevertheless, 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 likely to be infested.

The techniques also differed in the spatial pattern of the areas mapped as likely to be infested. They all emphasized areas near roads, trails, and streams. However, discriminant analysis defined a narrow buffer around almost every stream feature as part of its infestation zone, as did the decision tree. In contrast, logistic regression highlighted several large, discrete areas as likely to be infested by *A. tsugae*, as did the *k*-nearest neighbor analysis. With respect to capturing *A. tsugae* infestation sites recorded in GSMNP after the first year, the techniques all performed well despite their differences, capturing >85% of the surveyed points (Table 5).

Discussion

Selecting a Best Classification Method. Looking only at error rates, all four classification techniques performed reasonably well at distinguishing infested and uninfested sites, and it seems they would be at least moderately accurate if applied to new data sets. This is important given our goal of mapping *A. tsugae* infestation risk for the entire southern Appalachian region. However, error rate is a coarse measure, and further examination revealed aspects that make some of the techniques impractical for *A. tsugae* management purposes. For instance, the decision tree labeled a large percentage of GSMNP's hemlock area as likely to be infested; this large percentage probably explains why the approach resulted in so many false positives. Generally, our results recommend discriminant analysis as the best approach. It was efficient, delineating the least amount of area as high infestation risk while achieving the lowest error rates. Nonetheless, dis-

Table 4. Error matrices for the validation subset of the median data set: (a) discriminant analysis, (b) *k*-nearest neighbor, (c) logistic regression, (d) decision tree

(a)	Reference group			(b)	Reference group		
Classified as . . .	Infested	Uninfested	CE	Classified as . . .	Infested	Uninfested	CE
Infested	*25*	4	13.8	Infested	*24*	5	17.2
Uninfested	3	*18*	14.3	Uninfested	4	*17*	19.1
OE	10.7	18.2		OE	14.3	22.7	
Overall accuracy = 43/50 = 86%				Overall accuracy = 41/50 = 82%			
Cohen's κ = 0.715				Cohen's κ 0.633			
(c)	Reference group			(d)	Reference group		
Classified as . . .	Infested	Uninfested	CE	Classified as . . .	Infested	Uninfested	CE
Infested	*24*	5	17.2	Infested	*27*	10	27.1
Uninfested	4	*17*	19.1	Uninfested	1	*12*	7.7
OE	14.3	22.7		OE	3.6	45.5	
Overall accuracy = 41/50 = 82%				Overall accuracy = 39/50 = 78%			
Cohen's κ = 0.633				Cohen's κ = 0.533			

Cohen's κ statistic ranges from 0 to 1, representing the degree to which the model's prediction performance is better than chance: 0–0.4 = slight to fair model performance, 0.4–0.6 = moderate performance, 0.6–0.8 = substantial performance, 0.8–1 = near perfect performance (Landis and Koch 1977).

CE, percent commission error; OE, percent omission error.

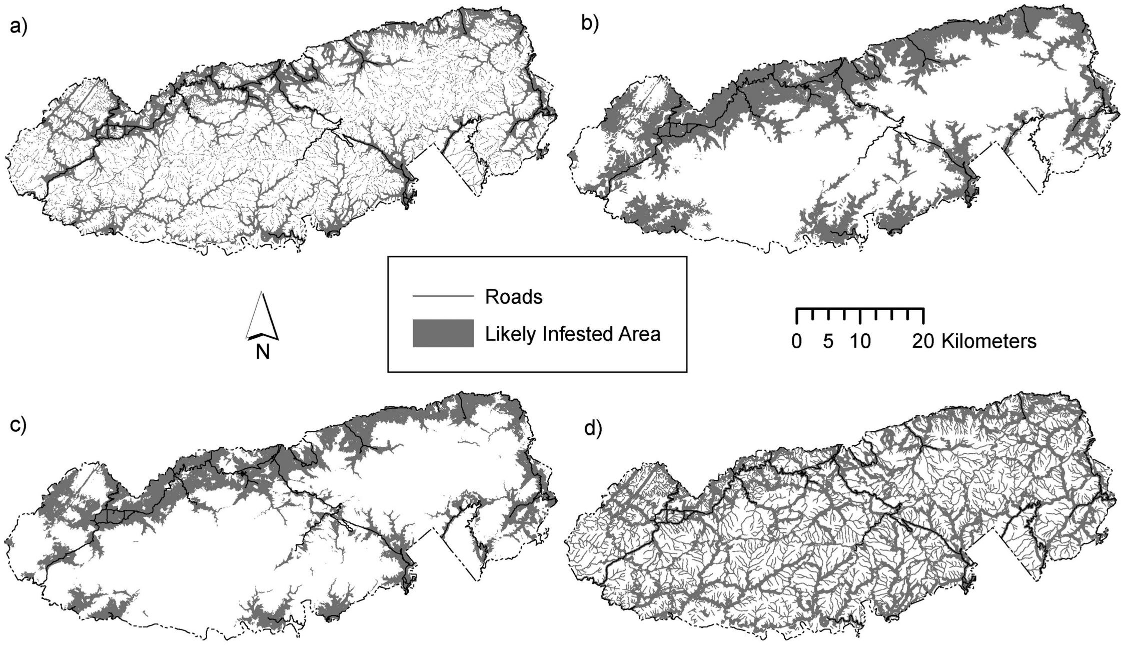

Fig. 4. Maps of Great Smoky Mountains National Park, showing the areas predicted by each classification function as likely to be infested by *A. tsugae*. (a) Discriminant analysis. (b) *k*-nearest neighbor. (c) Logistic regression. (d) Decision tree.

criminant analysis labeled almost all GSMNP areas adjacent to streams as likely to be infested by *A. tsugae* (Fig. 4). Although such areas might face somewhat elevated *A. tsugae* risk, the stream network is so diffuse across the park that it is difficult to isolate areas of focus for *A. tsugae* management. This suggests potential difficulty in applying the discriminant analysis function elsewhere in the southern Appalachians.

Logistic regression fell roughly in the middle of the four techniques in terms of classification error rate, was only slightly less efficient than discriminant analysis in terms of the area required to achieve that error rate, and captured a large percentage of *A. tsugae* points detected after the first year. In contrast with discriminant analysis, the resulting infestation map depicts discrete zones of *A. tsugae* infestation risk. These zones provide clear areas of focus for *A. tsugae* management efforts, supporting application elsewhere in the southern Appalachians, and also reflect the patchy pattern common in both host and pest species distributions. Furthermore, the logistic regression equation (either the one fitted to the median data set or perhaps an average across the five data sets) can be easily implemented in a GIS, and requires no prior variable transformation. Given its straightforwardness, we would recommend the logistic regression function over discriminant analysis or the other approaches, sacrificing some accuracy for ease of use.

Table 5. Prediction map efficiency in terms of the percentage of GSMNP total area and hemlock area mapped as likely to be infested, as well as the percentage of *A. tsugae* infestations surveyed after the first year that fell in these likely to be infested areas

Classification technique	Percent total area	Percent hemlock area	Percent surveyed points
Discriminant analysis	19.1	18.5	93.2
k-nearest neighbor	29.6	23.9	87.5
Logistic regression	22.8	20.1	85.4
Decision tree	30.7	34.6	93.5

Landscape Variables Affecting *A. tsugae* Spread. We completed our analyses at a limited spatial scale using a small sample of points. Because our primary objective was to find a tool for predicting *A. tsugae* landscape distribution throughout the southern Appalachian region, a question emerges as to whether the logistic regression function—as our chosen method—incorporated the appropriate variables and can be accurately generalized region-wide. All of the 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 *A. tsugae* is most likely to appear first in a landscape. This is consistent with research suggesting that, in general, the connectivity supplied by linear networks of landscape features exposes forested areas to various negative effects, including the spread of insect pests and disease (Simberloff and Cox 1987, Hess 1996, Trombulak and Frissell 2000). Corridors may be especially relevant for *A. tsugae* because they have the capability to amplify spread by three of the pest's major vectors: wind, birds, and humans. Forest edges along road corridors may experience increased wind exposure (Saunders et al. 1991). Major road corridors such as BRP serve as flyways for bird migration (Trani 2002). Human activity along trails and road corridors may facilitate pest dispersal by accidental transport of individuals or, for some species, altering microhabitats in a way that is favorable for a pest (Zobel et al. 1985, Forys et al. 2001, Jules et al. 2002). There is some relevant evidence of this for *A. tsugae:* surveys for the pest at Coweeta Hydrological Laboratory (near Franklin, NC) during the first year of infestation revealed that the heaviest infestations appeared along roadways, whereas forest interior areas remained uninfested or were only lightly infested (Lumpkin et al. 2003). Our analyses also seem to support the hypothesis that riparian corridors facilitate spread of *A. tsugae* by birds, although this is complicated by the fact that riparian corridors often include hiking trails with heavy human traffic (McWilliams and Schmidt 1999, Ward et al. 2004). Regardless, it is worth noting that crawlers of both the progrediens and sistens generations are active at times (early spring and early summer, respectively) when human recreational use and bird migration along corridors is likely to be highest.

In summary, the three proximity variables provide a reasonable, if simple, model for landscape-scale *A. tsugae* dispersal. The logistic regression function fitted to the median data set also includes elevation and percent slope as significant variables—basically, the odds of infestation are lower at high elevations and on steeper slopes. Because these factors could indeed influence accessibility of a site (especially by people), they seem like logical model components. Regarding the other topographic and categorical variables considered in our analyses, it may seem surprising that factors that shape hemlock distribution were not similarly significant for *A. tsugae* spread. However, while these variables may affect the hemlock abundance in a given forest stand—and thus the number of potential targets for the passively dispersed *A. tsugae*—the risk posed by greater hemlock presence pales in comparison to the risk associated with greater site accessibility. Likewise, although these variables may determine site quality and thus the persistence of infested hemlocks (Orwig and Foster 1998, Orwig et al. 2002, Ward et al. 2004), *A. tsugae* will eventually cause mortality in any location to which it spreads. Road, trail, and especially riparian corridors grant *A. tsugae* access to even fairly remote hemlock stands.

There are variables that we did not test in our analyses. We chose not to include limiting climate variables because the southern Appalachians are not consistently cold enough to curtail the regional expansion of *A. tsugae* (Ward et al. 2004). We did not consider patch characteristics because of too much patch-level heterogeneity in our data set. We also omitted traditional landscape metrics such as interpatch 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 dispersal. Incorporating such patch-level measures may be a fruitful direction for future research.

Because our approach is built on just a few easily derived spatial variables, it can be rapidly applied to generate *A. tsugae* infestation risk maps for the entire southern Appalachian region and at a finer spatial scale than existing portrayals of *A. tsugae* spread. These maps may be used to select monitoring sites before the pest's arrival in a particular locality, but our results also suggest that the maps can reliably predict *A. tsugae* distribution for at least the first few years after initial infestation. This could be advantageous for the targeted release and establishment of biological control agents (Cheah et al. 2004). Given that only an estimated 26% of hemlock habitat and 25% of hemlock basal area in the United States have already been invaded by *A. tsugae,* there are many areas for which this sort of information would be relevant (Morin et al. 2005). Ultimately, regional infestation probability maps might best be applied in conjunction with detailed maps of hemlock distribution. By overlaying infestation prediction maps with these distribution maps, forest managers could further reduce the area where they should target their *A. tsugae* monitoring or control efforts. This ability to prioritize would allow them to apply scarce management resources where they are most immediately needed.

Acknowledgments

We thank K. Johnson (GSMNP) and C. Ulrey (BRP) for providing the *A. tsugae* infestation data and many staff members from both parks for providing GIS data. This work was funded by the USDA Forest Service Forest Health Monitoring Program as part of an Evaluation Monitoring project. Comments from R. Jetton (NC State University) and two anonymous reviewers greatly improved this manuscript.

References Cited

Afifi, A., V. A. Clark, and S. May. 2004. Computer-aided multivariate analysis. Chapman & Hall, Boca Raton, FL.

Allison, P. D. 1999. Logistic regression using the SAS system. SAS Institute, Cary, NC.

Beck, L. R., M. H. Rodriguez, S. W. Dister, A. D. Rodriguez, E. Rejmankova, A. Ulloa, R. A. Meza, D. R. Roberts, J. F. Paris, M. A. Spanner, R. K. Washino, C. Hacker, and L. J. Legters. 1994. Remote-sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. Am. J. Trop. Med. Hyg. 51: 271–280.

Beck, L. R., M. H. Rodriguez, S. W. Dister, A. D. Rodriguez, D. R. Roberts, and M. A. Spanner. 1997. Assessment of a remote-sensing-based model for predicting malaria transmission risk in villages of Chiapas, Mexico. Am. J. Trop. Med. Hyg. 56: 99–106.

Beers, T. W., P. E. Dress, and L. C. Wensel. 1966. Aspect transformation in site productivity research. J. Forest. 64: 691–692.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. Stone. 1984. Classification and regression trees. Wadsworth, Belmont, CA.

Cheah, C.A.S.-J., and M. S. McClure. 1998. Life history and development of *Pseudoscymnus tsugae* (Coleoptera: Coccinellidae), a new predator of the hemlock woolly adelgid (Homoptera: Adelgidae). Environ. Entomol. 27: 1531–1536.

Cheah, C.A.S.-J., and M. S. McClure. 2002. *Pseudoscymnus tsugae* in Connecticut forests: the first five years, pp. 150–165. *In* B. Onken, R. Reardon, and J. Lashomb (eds.), Proceedings of the Hemlock Woolly Adelgid in Eastern North America Symposium, 5–7 February, East Brunswick, NJ.

Cheah, C., M. E. Montgomery, S. Salom, B. L. Parker, S. Costa, and M. Skinner. 2004. Biological control of hemlock woolly adelgid. U.S. Department of Agriculture, Forest Service Forest Health Technology Enterprise Team, Morgantown, WV.

Costa, S. D. 2005. Sampling for detection and monitoring of hemlock woolly adelgid within hemlock stands, pp. 57–62. *In* B. Onken and R. Reardon (eds.), Proceedings, Third Symposium on Hemlock Woolly Adelgid in the Eastern United States, 1–3 February 2005, Asheville, NC. U.S. Dep. Agric. Forest Service Forest Health Technology Enterprise Team, Morgantown, WV.

De'Ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological analysis. Ecology 81: 3178–3192.

Environmental Systems Research Institute. 2002. ArcGIS software user's manual, version 8.3. ESRI, Redlands, CA.

Flowers, R. W., S. M. Salom, and L. T. Kok. 2005. Competitive interactions among two specialist predators and a generalist predator of hemlock woolly adelgid, *Adelges tsugae* (Homoptera: Adelgidae), in the laboratory. Environ. Entomol. 34: 664–675.

Forys, E. A., A. Quistorff, and C. R. Allen. 2001. Potential fire ant impact (Hymenoptera: Formicidae) on the endangered Schaus swallowtail (Lepidoptera: Papilionidae). Fla. Entomol. 84: 254–258.

Gumpertz, M. L., C. Wu, and J. M. Pye. 1999. Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. Forest Sci. 46: 95–107.

Hess, G. R. 1996. Disease in metapopulation models: implications for conservation. Ecology 77: 1617–1632.

Jacquez, G. M., A. Kaufmann, J. Meliker, P. Goovaerts, G. AvRuskin, and J. Nriagu. 2005. Global, local and focused geographic clustering for case-control data with residential histories. Environmental Health 4: 4.

Jensen, J. R. 1996. Introductory digital image processing: a remote sensing perspective. Prentice Hall, Upper Saddle River, NJ.

Johnson, R. A., and D. W. Wichern. 2002. Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River, NJ.

Jovanovic, N., V. Milutinovic, and Z. Obradovic. 2002. Foundations of predictive data mining, pp. 53–58. *In* B. Reljin and S. Stankovic (eds.), Proceedings of the 6th Seminar on Neural Network Applications in Electrical Engineering. Faculty of Electrical Engineering, University of Belgrade, Yugoslavia.

Jules, E. S., M. J. Kauffman, W. D. Ritts, and A. L. Carroll. 2002. Spread of an invasive pathogen over a variable landscape: a nonnative root rot on Port Orford cedar. Ecology 83: 3167–3181.

Kelly, M., and R. K. Meentemeyer. 2002. Landscape dynamics of the spread of sudden oak death. Photogramm. Eng. Rem. S. 68: 1001–1009.

Khattree, R., and D. N. Naik. 2000. Multivariate data reduction and discrimination with SAS software. SAS Institute, Cary, NC.

Landis, J. R., and G. G. Koch. 1977. The measurements of observer agreement for categorical data. Biometrics 33: 159–174.

Liebhold, A. M., W. L. MacDonald, D. Bergdahl, and V. C. Maestro. 1995. Invasion by exotic forest pests—a threat to forest ecosystems. Forest Sci. 41: 1–49.

Liu, C., L. Zhang, C. J. Davis, D. S. Solomon, T. B. Brann, and L. E. Caldwell. 2003. Comparison of neural networks and statistical methods in classification of ecological habitats using FIA data. Forest Sci. 49: 619–631.

Lumpkin, S., P. Spaine, M. Hunter, and H. Keys. 2003. Responses of southern Appalachian ecosystems to hemlock woolly adelgid at the Coweeta watershed. Proceedings of the 14th Annual Southern Appalachian Man and the Biosphere Conference, 4–6 November 2003, Asheville, NC.

Maelzer, D. A., P. T. Bailey, and N. Perepelicia. 2004. Factors supporting the non-persistence of fruit fly populations in South Australia. Aust. J. Exp. Agr. 44: 109–126.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38: 921–931.

Mardia, K. V. 1974. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Indian J. Stat. 36(B): 115–128.

McClure, M. S. 1990. Role of wind, birds, deer, and humans in the dispersal of hemlock woolly adelgid (Homoptera, Adelgidae). Environ. Entomol. 19: 36–43.

McClure, M. S. 1996. Biology of *Adelges tsugae* and its potential for spread in the northeastern United States, pp. 16–25. *In* S. M. Salom, T. C. Tigner, and R. C. Reardon (eds.), Proceedings of the First Hemlock Woolly Adelgid Review. U.S. Department of Agriculture, Forest Service Forest Health Technology Enterprise Team, Morgantown, WV.

McClure, M. S., S. M. Salom, and K. S. Shields. 2001. Hemlock woolly adelgid. U.S. Department of Agriculture, Forest Service Forest Health Technology Enterprise Team, Morgantown, WV.

McLachlan, G. J. 1992. Discriminant analysis and statistical pattern recognition. Wiley, New York.

McNab, W. H. 1993. A topographic index to quantify the effect of mesoscale landform on site productivity. Can. J. Forest Res. 23: 1100-l107.

McWilliams, W. H., and T. L. Schmidt. 1999. Composition, structure, and sustainability of hemlock ecosystems in eastern North America, pp. 5–10. *In* K. A. McManus, K. S. Shields, and D. R. Souto (eds.), Proceedings of Symposium on Sustainable Management of Hemlock Ecosystems in Eastern North America. U.S. Department of Agriculture, Forest Service Northeastern Research Station, Newtown Square, PA.

Morin, R. S., A. M. Liebhold, E. R. Luzader, A. J. Lister, K. W. Gottschalk, and D. B. Twardus. 2005. Mapping host-species abundance of three major exotic forest pests. U.S. Department of Agriculture, Forest Service Northeastern Research Station, Newtown Square, PA.

Murthy, S. K. 1998. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min. Knowl. Disc. 2: 345–389.

Negrón, J. F., J. L. Wilson, and J. A. Ahnold. 2000. Stand conditions associated with roundheaded pine beetle (Coleoptera: Scolytidae) infestations in Arizona and Utah. Environ. Entomol. 29: 20–27.

Nelson, B. J., G. C. Runger, and J. Si. 2003. An error rate comparison of classification methods with continuous explanatory variables. IIE Trans. 35: 557–566.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. McGraw-Hill, Boston, MA.

Orwig, D. A. 2002. Ecosystem to regional impacts of introduced pests and pathogens: historical context, questions and issues. J. Biogeogr. 29: 1471–1474.

Orwig, D. A., and D. R. Foster. 1998. Forest response to the introduced hemlock woolly adelgid in southern New England, USA. J. Torrey Bot. Soc. 125: 60–73.

Orwig, D. A., D. R. Foster, and D. L. Mausel. 2002. Landscape patterns of hemlock decline due to the introduced hemlock woolly adelgid. J. Biogeogr. 29: 1475–1487.

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

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. Feinstein. 1996. A simulation of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 99: 1373–1379.

Pimentel, D., L. Lach, R. Zuniga, and D. Morrison. 2000. Environmental and economic costs of nonindigenous species in the United States. BioScience. 50: 53–65.

Rouget, M., D. M. Richardson, S. J. Milton, and D. Polakow. 2004. Predicting invasion dynamics of four alien *Pinus* species in a highly fragmented semi-arid shrubland in South Africa. Plant Ecol. 152: 79–92.

Royle, D. D., and R. G. Lathrop. 1999. The effect of site factors on the rate of hemlock decline: a case study in New Jersey, pp. 103. *In* K. A. McManus, K. S. Shields, and D. R. Souto (eds.), Proceedings of the Symposium on Sustainable Management of Hemlock Ecosystems in Eastern North America. U.S. Department of Agriculture, Forest Service Northeastern Research Station, Newtown Square, PA.

SAS Institute. 2003. SAS software user's manual, version 9. SAS Institute, Cary, NC.

Saunders, D. A., R. J. Hobbs, and C. R. Margules. 1991. Biological consequences of ecosystem fragmentation: a review. Conserv. Biol. 5: 18–32.

Shtatland, E. S., E. M. Cain, and M. B. Barton. 2001. The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system. Proceedings of the 26th Annual SAS Users Group International Conference, 22–25 April 2001, Long Beach, CA.

Simberloff, D., and J. Cox. 1987. Consequences and costs of conservation corridors. Conserv. Biol. 1: 63–71.

Souto, D., T. Luther, and B. Chianese. 1996. Past and current status of *A. tsugae* in eastern and Carolina hemlock stands, pp. 9–15. *In* S. M. Salom, T. C. Tigner, and R. C. Reardon (eds.), Proceedings of the First Hemlock Woolly Adelgid Review. U.S. Department of Agriculture, Forest Service Forest Health Technology Enterprise Team, Morgantown, WV.

Taylor, G. 2002. Hemlock resources in Great Smoky Mountains National Park. Proceedings of Hemlock Woolly Adelgid in Eastern North America Symposium, 5–7 February 2002, East Brunswick, NJ.

Trani, M. K. 2002. Chapter 1: terrestrial ecosystems, pp. 3–45. *In* D. N. Wear and J. G. Greis (eds.), Southern forest resource assessment. U.S. Department of Agriculture, Forest Service Southern Research Station, Asheville, NC.

Trombulak, S. C., and C. A. Frissell. 2000. Review of ecological effects of roads on terrestrial and aquatic communities. Conserv. Biol. 14: 18–30.

Wallace, S. M., and F. P. Hain. 2000. Field surveys and evaluation of native and established predators of the hemlock woolly adelgid and the balsam woolly adelgid in the southeastern United States. Environ. Entomol. 29: 638–644.

Ward, J. S., M. E. Montgomery, C.A.S.-J. Cheah, B. P. Onken, and R. S. Cowles. 2004. Eastern hemlock forests: guidelines to minimize the impacts of hemlock woolly adelgid. U.S. Department of Agriculture, Forest Service Northeastern Area State and Private Forestry, Morgantown, WV.

Welch, R., M. Madden, and T. Jordan. 2002. Photogrammetric and GIS techniques for the development of vegetation databases of mountainous areas: Great Smoky Mountains National Park. ISPRS J. Photogramm 57: 53–68.

Whittaker, R. H. 1956. Vegetation of the Great Smoky Mountains. Ecol. Monogr. 26: 1–80.

Wilson, B. A., A. Lewis, and J. Aberton. 2003. Spatial model for predicting the presence of cinnamon fungus (*Phytophthora cinnamomi*) in sclerophyll vegetation communities in south-eastern Australia. Austral Ecol. 28: 108–115.

Zobel, D. B., L. F. Roth, and G. M. Hawk. 1985. Ecology, pathology, and management of Port-Orford cedar (*Chamaecyparus lawsoniana*). U.S. Department of Agriculture, Forest Service Pacific Northwest Forest and Range Experiment Station, Portland, OR.

Received for publication 18 March 2006; accepted 23 June 2006.