Mapping eastern hemlock: Comparing classification techniques to evaluate susceptibility of a fragmented and valued resource to an exotic invader, the hemlock woolly adelgid

Joshua T. Clark a, Songlin Fei b,∗, Liang Liang c, Lynne K. Rieske a

a University of Kentucky, Department of Entomology, Lexington, KY 40546, USA
b Purdue University, Department of Forestry and Natural Resources, West Lafayette, IN 47907, USA
c University of Kentucky, Department of Geography, Lexington, KY 40506, USA

Abstract

Eastern hemlock (Tsuga canadensis Carrière), an ecologically important foundation species in forests of eastern North America, is currently threatened by the hemlock woolly adelgid (HWA, Adelges tsugae Annand, Hemiptera: Adelgidae), an aggressive invasive insect herbivore. HWA colonization of eastern hemlock results in rapid tree mortality. There is a pressing need to accurately determine eastern hemlock distribution in the face of expanding HWA populations to preserve this important forest species. However, efficient modeling of large geographic extents of eastern hemlock habitats to facilitate state-wide HWA management is lacking. We employ two modeling approaches, decision tree classification (based on presence-absence data) and maximum entropy (MaxEnt, based on presence-only data) method, to map eastern hemlock distribution in eastern Kentucky using a comprehensive suite of environmental parameters as predictor variables. Results demonstrate moderate model accuracies around 70%, supporting the practicality of mapping hemlock distribution over extensive regions. Comparison of the two modeling techniques suggests that decision tree classification has higher overall accuracies, while MaxEnt method was more efficient in model construction. In comparison to the decision tree method, MaxEnt suffered from possibly over-fitting as indicated by increased producer’s accuracies yet lower user’s accuracies. Our study provides useful references for selecting optimized approaches in accordance with study region characteristics and end user’s preferences.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Eastern hemlock (Tsuga canadensis Carrière) has suffered widespread mortality from an exotic invasive herbivore, the hemlock woolly adelgid (HWA, Adelges tsugae Annand, Hemiptera: Adelgidae) [McClure et al., 2001]. HWA was first reported in the eastern United States (US) in Virginia in the 1950s, and since the 1980s populations have expanded northward and westward (Havill et al., 2006), exploiting the large contiguous tracts of hemlock forest common in the northeast. More recently, HWA has expanded its geographic range southward along the southern Appalachian Mountains (Ward et al., 2004). HWA was first discovered in Kentucky in March 2006 (Kentucky Forest Health Task Force, 2006), and since then, infestations have been steadily expanding in the state. The continued spread of HWA through the region’s hemlock forests is imminent, but data concerning how that spread will develop is lacking, due in part to a poor understanding of the spatial distribution of the host tree species.

Eastern hemlock is an essential component of forests in eastern North America. It plays a vital role maintaining stream quality by regulating air and soil temperatures, soil moisture, hydrologic discharge, and amplitude of stream flow (Ford and Vose, 2007). Disruption of these processes can lead to loss of aquatic invertebrate and vertebrate biodiversity (Ross et al., 2003; Snyder et al., 2002). Loss of eastern hemlock will also cause changes in vegetation composition and structure; increased light penetration after hemlock death creates suitable habitat for less shade-tolerant tree species such as red maple (Acer rubrum) and black birch (Betula lenta) [Daley et al., 2007; Spaulding and Rieske, 2010; Yorks et al., 2003].

In spite of the ecological importance of eastern hemlock, resource managers have only limited information about the scope of its spatial distribution. The lack of detailed hemlock-specific maps is due in part to hemlock’s low economic importance, low density and basal area, and confined distribution to the study area (Little, 1971 and Fig. 1). Better mapping of eastern hemlock is...
critically needed to slow the spread of HWA by strategically monitoring and treating certain areas that can serve as HWA pathways or hotspots. To develop a detailed hemlock distribution map at the regional level, ground-based mapping of hemlock is impractical. Existing satellite image based maps such as the National Land Cover Dataset (NLCD, http://www.mrlc.gov/) provide general distribution information for evergreen forests, but do not provide any spatial information at the species level. The Kentucky Land Cover Dataset (KLCD, http://kygeonet.ky.gov/) does include species specific information for eastern hemlock, however their application to hemlock management is limited for the following two reasons. First, KLCD is intended for land cover and land type classification not for species classification, therefore, its accuracy at species level is limited. Second, due to its low density and sparse distribution, hemlock is often classified into other types/groups that do not have a defined hemlock component. Other products such as Forest Inventory and Analysis (FIA, Bechtold and Patterson, 2005) data may provide rich ecological information on sampled plots and unbiased statistical estimation of the overall abundance of hemlock distribution, but FIA data alone cannot produce continuous coverage that is necessary for hemlock management spatial prioritization and strategic planning, given their limited sampling density. Hence, we consider species distribution modeling as an effective way to bridge this gap and provide more physiologically coherent approximation of hemlock distribution with continuous coverage.

Species distribution models are frequently used to predict species distributions across landscapes or to gain insights into species ecological traits. Two types of species distribution models, presence–absence data based models and presence-only data based models, are widely used in various ecosystems (Elith and Leathwick, 2009). The decision tree classification method, a presence–absence data based method, is popular in species distribution prediction. Maingi and Luhn (2005) used decision tree classification to map conifer tree distribution with Landsat TM imagery and ancillary data in the Daniel Boone National Forest in Kentucky. Koch et al. (2005) used a decision tree classification aided with remotely sensed data to successfully map eastern hemlock in other regions in response to HWA-induced eastern hemlock decline. MaxEnt (Phillips et al., 2006), a presence-only data based method, is also a commonly used method in species distribution modeling and prediction. Elith et al. (2006, 2011) found that MaxEnt provides relatively consistent and reliable results compared to other presence-only data based methods.

In this study, both decision tree classification and MaxEnt were employed to map the distribution of eastern hemlock in Kentucky. The primary objective of our study was to explore effective ways to map eastern hemlock on a regional scale to facilitate state level management and conservation efforts. For comparison purposes, classification accuracy, processing complexity, and time to completion for the two techniques were evaluated.

2. Material and methods

2.1. Study area

Our study area covered 27,006 km² of eastern Kentucky (38.29–36.58°N, 81.96–84.83°W) and was partitioned into two physiographic regions: Upper Eastern Coal Field (hereafter, Coal Field) and Pine Mountain (Fig. 1). All models were constrained to the same geographic extent, shown as “Study Area” on the map. This mountainous region is composed of sandstone, shale, and siltstone and ranges in elevation from 154 to 1259 m (McDowell, 1986). Monthly average temperature ranges from 1.1 °C in January to 23.9 °C in July and monthly average precipitation ranges from 8.1 cm in October to 13.1 cm in May (NOAA, 2002). The forest type is mixed mesophytic, consisting primarily of pine-oak (Braun, 1950).

2.2. Modeling approach overview

We modeled eastern hemlock distribution using two modeling techniques in each physiographic region, for a total of four models. In the first approach we performed a decision tree analysis against environmental variables. A decision tree approach requires at least two classes of dependent variables, necessitating the collection of both presence and absence points. In a large study area this can be exceedingly time-consuming. For that reason, the second method we tested was maximum entropy species distribution modeling (MaxEnt), which requires only presence data (Phillips et al., 2006). Given the differences between the two physiographic regions with respect to elevation, slope, and geologic substrate, modeling was performed for each region separately.

2.3. Field data

Eastern hemlock presence and absence data were acquired from a variety of sources (Table 1); the majority of these data are a result of systematic sampling efforts. A total of 2624 points were used in this study. All points were georeferenced using a GPS unit in the field. Forest Inventory and Analysis data represent ground samples collected systematically at a rate of one plot per 2428 ha (6000 acre) hexagon. Actual FIA plot coordinates were used in this study. A plot is designated as a hemlock presence plot if any hemlock tree with >2.54 cm in diameter at breast height (d.b.h.) is present. The second source of systematic sampling was derived from an ongoing gypsy moth survey program. The traps for this monitoring effort were placed on a 2 × 2 km grid. The individuals setting and monitoring the traps were asked to record presence or absence of eastern hemlock within 30 m of the trap site. The remaining field data were derived from HWA monitoring by individuals representing various agencies, including the Kentucky Division of Forestry, the Kentucky Office of the State Entomologist, and the Kentucky State Nature Preserves Commission. Hemlock data from these sources were collected in an opportunistic manner. These field data were compiled into a database representing hemlock or non-hemlock categories; for all sampling efforts, any survey location with at least one hemlock tree was classified as hemlock. This liberal classification scheme reflects the overall goals of this research, to map hemlocks which may serve as a corridor for the spread of HWA. The data points were split by physiographic
and can utilize both continuous and categorical data (Huang et al., non-parametric, robust to noisy or missing data, easy to interpret, size of 30\text{m} using ArcGIS 9.3 (ESRI, Redlands, CA, US).}

### 2.4. Environmental data

In both modeling approaches, 11 environmental layers were used as independent variables representing physiographic and climatic characteristics. These environmental data were acquired or derived from various sources as detailed in Table 2. To maintain consistency across all models and for comparison to the NLCD (Homer et al., 2004) and KLCD (Commonwealth Office of Technology, 2004), each variable was converted to a raster layer with cell size of 30 × 30 m using ArcGIS 9.3 (ESRI, Redlands, CA, US).

#### 2.5. Classification tree

We chose a decision tree classification model because it is non-parametric, robust to noisy or missing data, easy to interpret, and can utilize both continuous and categorical data (Huang et al., 2003; Koch et al., 2005; Nelson et al., 2003). Similar methods have produced maps for eastern hemlock at a landscape scale (Koch et al., 2005) and for other evergreen trees at regional scales (Landenburger et al., 2008; McDermid and Smith, 2008). Using PASW Statistics software (SPSS Inc., Chicago, IL, US), decision tree analysis was performed to classify areas with presence or absence of eastern hemlock.

For the decision tree in each physiographic region a Chi-squared Automatic Interaction Detection (CHAID) growing method was applied with a Pearson chi-square statistic, maximum tree depth set to 10 levels, and a significance level of 0.1 for splitting nodes. The CHAID method was used in order to take advantage of multiple splitting pathways for each node, as opposed to binary splitting. This multi-splitting at each node in the CHAID method allows variables to be partitioned in a more biologically meaningful way. All environmental variables were designated as continuous variables except geologic formations and soil types, which were specified as categorical. The significance level was set to 0.1 in order to further divide nodes to minimize overestimation or underestimation. Node response was used as the threshold to assign the area represented by that node to either the presence or absence class; response values greater than 50% were designated as hemlock presence.

#### 2.6. MaxEnt modeling

We employed maximum entropy modeling using the MaxEnt species distribution program, version 3.2.19 (Phillips et al., 2006). The MaxEnt algorithm predicts the probability distribution of species occurrence based on the environmental constraints estimated from known locations.

All environmental variables were again designated as continuous variables except geologic formations and soil types (categorical) as for CHAID. Eighty percent of hemlock presence points were used to build the model, leaving one-fifth that were randomly selected by the program to be used in model validation procedures. The model was executed for 10 iterations with refreshed random partitions of training and validation data. We chose the best model to use based on highest area under the curve value (Boubli and de Lima, 2009; Pauchard and Alaback, 2004). The jackknife option was used to evaluate the importance of each environmental variable in the model. The raw model output was a continuous probability map with values between 0 and 1; therefore, a threshold was assigned in order to reclassify the output into dichotomous classes representing either suitable or unsuitable areas for hemlock habitat. The threshold was determined using a method based on the maximum difference of cumulative frequency of probabilities for observed hemlock locations and random points (Browning et al., 2005; Fei et al., 2007; Thompson et al., 2006). The geographic areas with probabilities above the threshold were designated as hemlock and those areas below the threshold were designated as non-hemlock.

#### 2.7. Accuracy assessments

For both modeling methods, accuracy assessments were performed for predicted eastern hemlock distribution using surveyed reference points. A total of 419 ground truthing reference points were used for Coal Field and 105 points for Pine Mountain. Overall accuracies with user’s and producer’s accuracies were calculated for each model.

### Table 1

<table>
<thead>
<tr>
<th>Source</th>
<th>Hemlock</th>
<th>Non-hemlock</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest Inventory and Analysis</td>
<td>Training</td>
<td>88</td>
<td>703</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>22</td>
<td>207</td>
</tr>
<tr>
<td>Gypsy Moth Survey</td>
<td>Training</td>
<td>7</td>
<td>349</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>HWA Monitoring</td>
<td>Training</td>
<td>749</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>202</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1068</td>
<td>1556</td>
</tr>
</tbody>
</table>

Table 1: Sources of hemlock and non-hemlock data points used in model building and accuracy assessments.

<table>
<thead>
<tr>
<th>Environmental Layer</th>
<th>Abbreviation</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEM</td>
<td>DEM</td>
<td>Elevation</td>
<td>USGS</td>
</tr>
<tr>
<td>Slope</td>
<td>Slope</td>
<td>Slope derived from DEM</td>
<td>Calculated using ArcGIS slope tool</td>
</tr>
<tr>
<td>Aspect</td>
<td>Aspect</td>
<td>Aspect derived from DEM</td>
<td>Calculated using ArcGIS aspect tool</td>
</tr>
<tr>
<td>Soil type</td>
<td>Soil</td>
<td>Soil Survey Geographic (SSURGO) data</td>
<td>USDA Geospatial Data Gateway</td>
</tr>
<tr>
<td>Moisture index</td>
<td>TRMIM</td>
<td>Topographic relative moisture index (modified)</td>
<td>TRMIM.aml script downloaded from <a href="http://earth.gis.usu.edu/swgap/landform.html">http://earth.gis.usu.edu/swgap/landform.html</a></td>
</tr>
<tr>
<td>Stream distance</td>
<td>Streams</td>
<td>Euclidean distance from nearest stream</td>
<td>Calculated using ArcGIS Euclidean distance tool</td>
</tr>
<tr>
<td>Topographic position</td>
<td>TPI</td>
<td>Topographic position index</td>
<td>Calculated using TPI extension (Jenness, 2006) for ArcView (ESRI, Redlands, CA, US)</td>
</tr>
<tr>
<td>Curvature</td>
<td>Curvature</td>
<td>Curvature of the terrain derived from DEM</td>
<td>Calculated using ArcGIS curvature tool</td>
</tr>
<tr>
<td>Geologic formation</td>
<td>Geo form</td>
<td>Geologic formations of Kentucky</td>
<td>Kentucky Geological Survey</td>
</tr>
</tbody>
</table>

Table 2: Environmental variables used in models with respective data sources.
3. Results

Both types of models generated maps of eastern hemlock distribution with moderate accuracies (Fig. 2, Table 3). Across models, overall accuracies were fairly close, ranging from 65.7% to 74.7%, but the total area classified as eastern hemlock differed between models by up to thousands of square kilometers (Fig. 3). Decision tree models appeared to yield slightly higher accuracies than the MaxEnt models in both regions (74.7% and 72.4%, respectively, for Coal Field and Pine Mountain). MaxEnt models had higher producer’s accuracies than the decision tree models in both regions.

The combined area (Coal Field and Pine Mountain) predicted to be eastern hemlock by decision tree models was 11.4% of the total area, approximately 2300 km² less than the MaxEnt models, which predicted 19.9% of the total area as eastern hemlock (Fig. 3). The environmental variable ranking (Table 4) for the decision trees revealed proximity to nearest stream to be a prominent classifier. The MaxEnt models, in particular, resulted in test AUC values (>0.80) that indicated good model performance (Fig. 4). The threshold values for an area to be designated as hemlock are 0.42 and 0.35 for Coal Field and Pine Mountain physiographic regions, respectively (i.e., areas with probability value higher than the threshold value will be classified as hemlock). The resulting binary map has a fairly high overall classification accuracy (72.8%) in the Coal Field region and slightly lower overall accuracy (65.7%) in the Pine Mountain region (Table 3). The MaxEnt models were associated with relatively large commission errors (37.6% and 39.0% for Coal Field and Pine Mountain, respectively) and small omission errors (16.4% and 11.3%). The predictor variables with the highest percent contributions were proximity to nearest stream, soil type, and geologic formation, which together made over 85% contributions to the models. The rank of these three variables varied between the two physiographic regions (Table 5).

We generated a synthesized intersection map (Fig. 2) which shows areas predicted as eastern hemlock by both modeling approaches. This intersection map was used to compare environmental data with the background because it likely indicates areas with a very high probability of finding hemlock. Hemlock coverage predicted by both methods is about 2300 km², approximately 56% of the predicted hemlock area by the decision tree models and 32% by MaxEnt models (Fig. 3). Considering that the study area lies on the edge of the distribution range for eastern hemlock, the intersection map reveals important environmental conditions that better support hemlock growth within the state (Table 6).
intersection areas predicted as eastern hemlock by both classification methods indicate that 95% of hemlock coverage is located less than 296 m from streams and between 189 and 433 m in elevation in the Coal Field region and less than 930 m from streams and between 238 and 779 m in elevation in the Pine Mountain region. Further, the majority of the eastern hemlock area is situated on northeast to southeast facing slopes and on geologic formations composed of shale, sandstone, or alluvium. The most abundant soils in the predicted hemlock areas are well drained, loamy, acidic ultisols and inceptisols including sheloca, helechawa, or grigsby complexes.

4. Discussion and conclusions

Both of our modeling approaches produced maps of eastern hemlock distribution in the Coal Field and Pine Mountain physiographic regions with moderate accuracy. The overall accuracy (~70%) was lower than that of similar studies (~80%) to map eastern hemlock in other regions (Koch et al., 2005; Maingi and Luhn, 2005). However, our study evaluated a much larger region with more complex topographic and physiographic characteristics, and greater vegetative variations. The reduced accuracies may be a consequence of the extensive spatial coverage which included greater landscape heterogeneity. We initially attempted a second decision tree classifier which included Landsat spectral bands as independent variables, but this method was abandoned. Reduced spectral resolution, cloud cover, and altered brightness values caused by the rugged terrain of eastern Kentucky, especially in the Pine Mountain area, affected the reliability of the satellite imagery. In addition, image processing time in areas with rugged terrain, such as our study area, can be extremely time consuming.

In general, MaxEnt modeling appeared to provide predictions with reasonable accuracy. Multiple model runs allowed for selecting optimum results. However, overall accuracies of MaxEnt models were a balance of relatively low user’s accuracy and high producer’s accuracy, which may imply over-fitting. The MaxEnt approach is known to be robust for presence-only models, but its generally high accuracy may be compromised by lacking absence data, as indicated by the low user’s accuracy in this study and suggested by related investigations (Elith et al., 2011). In addition, omission of absence data when building the model, a stipulation of the MaxEnt method, may also compromise the accuracy in comparison to the decision tree based method.

The difficulty and time required to complete each model varied considerably. Since the decision tree method required final maps to be built in GIS using the node-splitting conditions, the decision trees were significantly more time-intensive than the MaxEnt models. The existing MaxEnt program was straightforward with a simple user interface, and the processing was relatively efficient. While estimation of the thresholds added to the time committed, the MaxEnt model was the less complicated and quicker to execute.

The environmental characteristics associated with hemlock habitat prediction affirmed previous descriptions of eastern hemlock habitat in the southern Appalachian Mountains, which suggest that hemlock grows on north to east facing slopes in moist riparian zones and in neutral to acidic soils (Godman and Lancaster, 1990; Quimby, 1996). Soil was a significant contributor in the MaxEnt models in predicting the presence of hemlock, while it did not contribute in node-splitting in the decision trees. This could be partially due to the tight association between geologic formation and soil type and different statistical mechanisms used in the two methods in dealing with categorical variables. Eastern hemlock is commonly reported between 610 and 1520 m in this region (Southern Appalachians, Godman and Lancaster, 1990), much higher than predicted by our study (Table 6), which potentially reflects the regional difference of the river valley reliefs, as eastern Kentucky is located on the western edge of hemlock distribution with a relatively lower elevation range.

The total area predicted as eastern hemlock by each model is larger than national and state land cover datasets and FIA data (Fig. 3). The reason for these differences is twofold. First, areas predicted by the two models used in this study are potential niche/habitat, while the land cover datasets and FIA estimate the realized niche/habitat, which is always smaller than the former estimation. Second, classifications in the land cover datasets are often too broad and neglect the minor components in areas with mixed species. The total area predicted as hemlock by the decision tree model is slightly higher than the statistical estimation from FIA data. The intersected areas of the two models are more conservative than the FIA data based estimation.

Our results suggest that the most advantageous procedure for mapping hemlock resources in the central Appalachians may depend on the priorities set forth by the end users. This study demonstrates that a relatively accurate map can be produced with environmental data at the regional scale. In situations where absence data are unreliable or unavailable, a presence-only species distribution model such as MaxEnt is most promising. However, if absence data have been collected and resource managers desire to achieve a more accurate classification, the decision tree classifier may be a better option. Moreover, an ‘ensemble modeling’ approach could be considered to assist land management decision making (Jones-Farrand et al., 2011). A union of both predicted maps can be used if land managers want to be inclusive about all possible hemlock areas, or an intersection of predicted maps can be used if high certainty is needed.

---

**Table 5** Percent contribution of environmental variables used in MaxEnt models for the Coal Field and Pine Mountain physiographic regions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coal Field</th>
<th>Pine Mountain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Streams</td>
<td>41.0%</td>
<td>21.1%</td>
</tr>
<tr>
<td>Soil</td>
<td>37.1%</td>
<td>43.8%</td>
</tr>
<tr>
<td>Geo Form</td>
<td>12.0%</td>
<td>21.2%</td>
</tr>
<tr>
<td>DEM</td>
<td>2.4%</td>
<td>2.3%</td>
</tr>
<tr>
<td>TPI</td>
<td>2.0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Aspect</td>
<td>1.6%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Min Temp</td>
<td>1.5%</td>
<td>3.1%</td>
</tr>
<tr>
<td>Slope</td>
<td>1.0%</td>
<td>1.6%</td>
</tr>
<tr>
<td>TRMIM</td>
<td>0.5%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Curvature</td>
<td>0.5%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Max Temp</td>
<td>0.4%</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

*Detailed descriptions of variables can be found in Table 2.*
As non-native invaders continue to proliferate and threaten valued resources, land managers must utilize diverse approaches to develop increasingly aggressive tools for management and mitigation. The hemlock woolly adelgid is particularly threatening, since eastern hemlock is considered an essential foundation species in forests of eastern North America (Godman and Lancaster, 1990), and it is especially susceptible to adelgid feeding (Souto et al., 1996). The hemlock woolly adelgid threatens the persistence of the hemlock forest type (Spaulding and Rieske, 2010), which has tremendous consequences for ecosystem function (Orwig and Foster, 1998). Our results demonstrate that spatial habitat modeling using approaches such as the decision tree classification or the maximum entropy method is a feasible means of determining high-risk areas of invasion at the state level, and could aid in mitigating the effects of the hemlock woolly adelgid invasion.

Acknowledgments

We thank C. Harper, J.D. Loan and J. Obrycki for support and data collection and S.G. Lambert for the access of FIA plot data. This study is supported by the United States Department of Agriculture, Cooperative State Research, Education, and Extension Service (Proposal No. 2006-06176), and is published as Kentucky Agricultural Cooperative State Research, Education, and Extension Service (Prop-osal No. 2006-06176), and is published as Kentucky Agricultural Experiment Station Project 11-08-002.

References


