Developing a Methodology to Predict Oak Wilt Distribution Using Classification Tree Analysis

Marla C. Downing, USDA Forest Service, Forest Health Technology Enterprise Team, Fort Collins, CO
Vernon L. Thomas, INTECS International, Inc., Fort Collins, CO
Robin M. Reich, Department of Forest Sciences, Colorado State University, Fort Collins, CO

Abstract—*Oak wilt (*Ceratocystis fagacearum*), a fungal disease that causes some species of oak trees to wilt and die rapidly, is a threat to oak forested resources in 22 states in the United States. We developed a methodology for predicting the Potential Distribution of Oak Wilt (PDOW) using Anoka County, Minnesota as our study area. The PDOW utilizes GIS; the classification tree statistical test; field sample data; commonly acquired, coarse-resolution auxiliary data; and a unique application of data from the Landsat Thematic Mapper (TM) satellite to predict the spatial distribution of oak wilt. Two accuracy assessments, one being a ten-fold cross validation, the other using verified oak wilt data from a later date, indicate that, at the landscape scale, PDOW correctly models the presence of oak wilt, and accurately predicts oak wilt distribution in Anoka County. Important variables in predicting oak wilt distribution in Anoka County included: Landsat TM Bands 3, 4, 5, and 7; distances between sample sites and lakes (Distance to Lakes); density of streams within a 400 x 400 meter grid surface (Stream Density); slope; aspect; and density of roads within a 400 x 400 meter grid surface (Road Density).*

Introduction

Oak wilt (*Ceratocystis fagacearum*) is a serious tree disease that kills many species of oaks. The disease spreads in two ways, via insect vectors and root grafts. Untreated infection centers kill thousands of oak trees annually throughout 22 middle and eastern states in the US, and as far south as Texas (Appel and Maggio, 1984; Juzwik, 2000; USDA Forest Service, 2001; O'Brien et. al., 2003), (fig. 1), creating ever widening pockets of disturbance in forests, woodlots, and home landscapes.Oak trees dominate the Minneapolis/St. Paul urban landscape and are both a valued element of the forest ecosystem (International Society of Arboriculture, 1992; Nowak et. al., 2001; USDA, 2001) and an important food source for many animals (Juzwik, 2000; Sander, 2003). But the oaks are threatened: in 1998, the Minnesota Department of Natural Resources (MNDNR) identified and treated 3,182 acres of infected oak wilt trees in Anoka County. The MNDNR projects that, at current infection rates, there will be a two-fold increase in oak wilt by 2008 (MNDNR, 2001).The purpose of our study was to test the feasibility of using a classification tree on commonly acquired datasets and location data collected in Anoka County, Minnesota, to create a surface of potential oak wilt distribution. Tree-based models (decision trees) are useful for both categorical classification and regression problems. The classification or regression tree is a collection of many rules displayed in the form of a binary tree. The rules are determined by a recursive partitioning procedure (MathSoft, 1999). Such an approach offers a way to describe the spatial continuity that is an essential feature of many natural phenomena (Isaacs and Srivastasa, 1989), and have been used to predict spatial patterns and develop indicators of sudden oak death in California, USA (Kelly and Meentemeyer, 2002), classify remote sensing imagery (Michaelson and others 1994; Friedl and Brody, 1997; Joy and others 2002; Ruefenacht and others 2003), and estimate fuel loads in the Black Hills, USA (Reich and others 2004).

Advantages of using decision trees include the nonparametric nature of the model, ease of interpretation, and the robustness of the test (De'Ath and Fabricius, 2000). Decision trees are appropriately used with datasets that are a mix of both numeric and factor information, nonlinear and/or non-additive data. A potential disadvantage to using decision trees is the need for large numbers of samples (Joy, 2003).

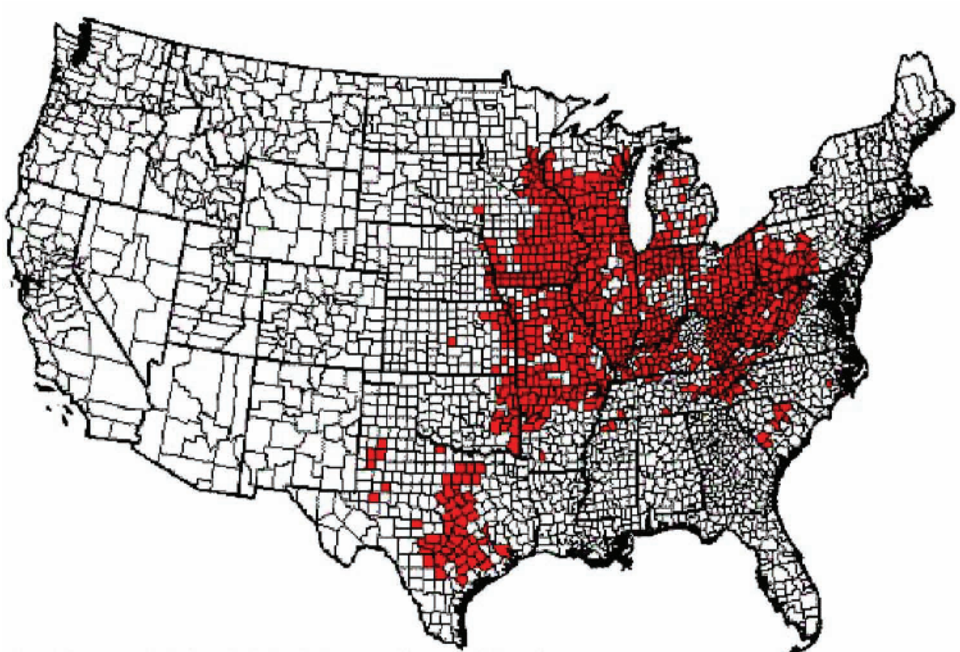

Figure 1. U.S. distribution of oak wilt by County—1999 USDA Forest Service.

Materials and Methods

Study Area

Anoka County, just north of the Minneapolis/St. Paul urban area, is comprised of 110,000 hectares, and has a population of nearly 300,000 (U.S. Census Bureau, 2003). The northwest corner of the county lies at the latitude of 45.41 degrees north and longitude of 93.51 degrees west.

The county is considered to be relatively homogeneous with primarily sandy soils and oak forest landcover. It falls within the Anoka Sand Plains (MNDNR, 1999), a subsection of the Minnesota and northeast Iowa Morainal Oak Savanna Section (Bailey, 1994), a part of the Eastern Broadleaf Forest Continental Province (Bailey, 1995), in the Hot Continental Division (Bailey, Ecoregion 220). The climate consists of cold winters and hot summers. The natural vegetation is characterized by winter deciduous forest dominated by tall broadleaf trees (Bailey, 1996). The landform is a broad sandy lake plain containing small dunes, kettle lakes, and level to gently rolling topography (Wright, 1972). Elevation ranges from 243 to 342 meters above sea level (USGS 30-meter resolution, 1:24000 scale DEM).

Anoka County's Land Type Associations (LTAs) are: Anoka Lake Plain (72 percent of the land mass); Mississippi Sand Plain (12 percent); Burns Till Plain (9 percent); Forest Lake Moraine (4 percent); Maplewood Moraine (2 percent); Elk River Moraine (1 percent) (MNDNR, 2000). Anoka County is made up of oak forest (5 percent), other hardwood forest (9 percent), coniferous forest (1 percent), shrubs and grasslands (17 percent), agriculture (22 percent), water bodies (5 percent), and urban areas (41 percent) (U of MN, 2002; Ward, 2002).

Spatial Information Database

Our database consisted of two components: 1) the dependent variable, and 2) the independent variables. The sample point field location data served as the dependent variable. There were twenty-two independent variables used in the analysis, which were collected, aggregated, or re-sampled to a 30 x 30 meter spatial resolution.

Dependent Variable datasets were merged to become the Sample Point Theme against which the independent variables were tested for correlation:

- Oak wilt presence and absence field sample data from the Land Management Information Center (LMIC), Forest Health, Oak Wilt, treated site polygon data, 1998.
- Field visits by the USDA FS NCRS Forest Disease Unit and the Forest Health Technology Enterprise Team (FHTET) to the LMIC, Forest Health, and "active" oak wilt sites.
- Field visits by the USDA FS NCRS Forest Disease Unit and FHTET to randomly selected healthy oak forest sites.

Independent variables used in the binary Classification Tree to determine the level of correlation with the dependent variable:

- Landsat TM Path 27 Row 29 bands 1-7, acquired May 1998.

- Landsat TM Path 27 Row 29 bands 1-7, acquired September 1998.
- Elevation, derived from the USGS 30 meter resolution DEM (1:24000 scale).
- Slope degrees, derived from the USGS DEM using ArcView Spatial Analyst (ESRI) slope function.
- Aspect (compass direction), derived from the USGS DEM using ArcView Spatial Analyst (ESRI) aspect function.
- Landform (independent of slope), created from a custom ArcView Avenue application, which uses an irregular 3 x 3 kernel, where positive values indicate concavity and negative values indicate convexity, to calculate landform from a USGS DEM. A zero value indicates flat terrain (McNab, 1989).
- Distance to Streams USGS 1:100,000 DLG data, measured using ArcView Spatial Analyst, distance in meters from line feature function.
- Distance to Lakes USGS 1:100,000 DLG data, measured using ArcView Spatial Analyst, distance in meters from feature function.
- Road Density was measured using ArcView Spatial Analyst, distance in meters from line feature function. It was calculated as the sum of roads within 400 x 400 meter grid surfaces. Roads include City Streets, County Roads, and TWP Roads from USGS 1:24,000 data and Major and Ramp roads from MN Department of Transportation data.
- Stream Density, calculated as the sum of all stream surface area within 400 x 400 meter surface grids.

A dependent variable GIS Sample Point Theme was created for this study using the LMIC oak wilt database as our primary data source. Many sample locations were acquired from the 1998 LMIC oak wilt "treated" polygon data. Additional sample locations, coded as "possible active" oak wilt sites during the 1998 growing season, from the same database, were also randomly selected and visited in July and August, 2002, by USDA FS NCRS Forest Disease and FHTET personnel. GPS location points were collected if evidence suggested these "possible active" sites had active oak wilt infection centers in 1998. Healthy oak site locations also were acquired during this time period. All field visited sample point locations were collected using a GPS (Garmin E-Trex Legend) at tree locations. Of the 422 sample points collected, 121 were identified as being healthy oak sample points, and 301 were identified as having been active oak wilt sites in 1998.

All polygon centroid locations and sample point locations (from LMIC, NCRS and FHTET) were merged to create the final dependent variable GIS Sample Point Theme in ArcView 3.2a. Location points were acquired from polygon centroids using the ArcView command "ReturnCenter" (ESRI, 2000) with customized ArcView Avenue code to automatically create points from the polygon centroid. Healthy oak wilt sample point locations were assigned a value equal to zero, and oak wilt sample point locations were assigned a value equal to 1.

Twenty-two auxiliary, or independent variable, GRID themes were also constructed for this study: Fourteen were created from two, multi-temporal Landsat TM images using the ERDAS Imagine software Grid Export function, which un-stacks each band to an individual grid. (ERDAS 2001); eight were constructed using the Spatial Analyst extension for ArcView. The DEMs were used to create the slope and aspect GRID themes. Anoka county lakes and streams data were used to construct the distance-to-streams, and distance-to-lakes GRID themes. Customized ArcView functionality, created with ArcView Avenue, was incorporated to create the GRID themes of landform, road density, and stream density.

The twenty-two independent variable attributes were added to the Sample Point Theme to produce a Spatial Information Database. This was accomplished using an automated ArcView function (written with ArcView Avenue script language), to extract the grid cell values from each of the twenty-two independent data themes at each of the 422 oak wilt presence or absence sample point locations. These grid cell values, taken at the sample point locations, were then used to populate the Spatial Information Database (USDA Forest Service, FHTET, 2003).

Stratification

The southern section of Anoka County has a higher degree of urban coverage than the northern section of the county. More oak wilt sample point locations were found in the more urbanized southern half of Anoka County. To determine whether spatial correlation exists between oak wilt and urban or natural landscape features, and to ensure that the urban condition in the south was not affecting the results of the model for the non-urban area to the north, the county was stratified into urban and non-urban datasets and two models were created (fig. 2).

Classification Tree Analysis and Oak Wilt Presence/Absence Surface Map

The Spatial Information Database for Anoka County, comprised of 30 x 30 meter biological, geological, hydrological, landscape, and physical information, was the basis for our classification tree analyses to predict the distribution of oak wilt for the urban and non-urban models.

S-PLUS© statistical software was used to fit the classification tree to the Spatial Information Database

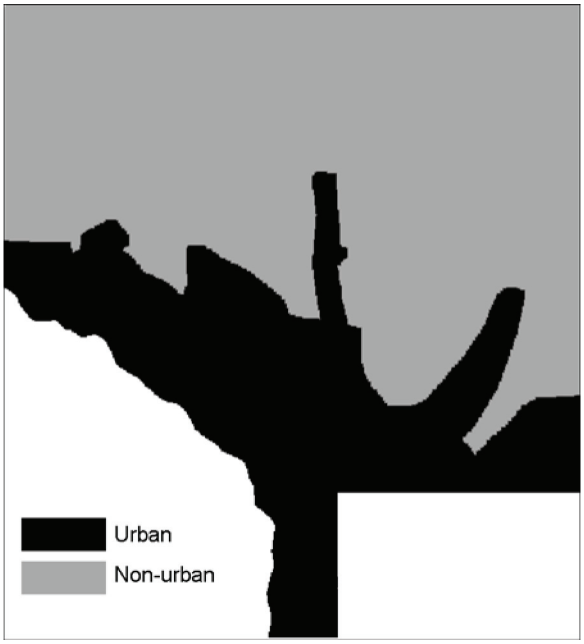

Figure 2. Urban (black) and non-urban (gray) spatial stratifications for Anoka County.

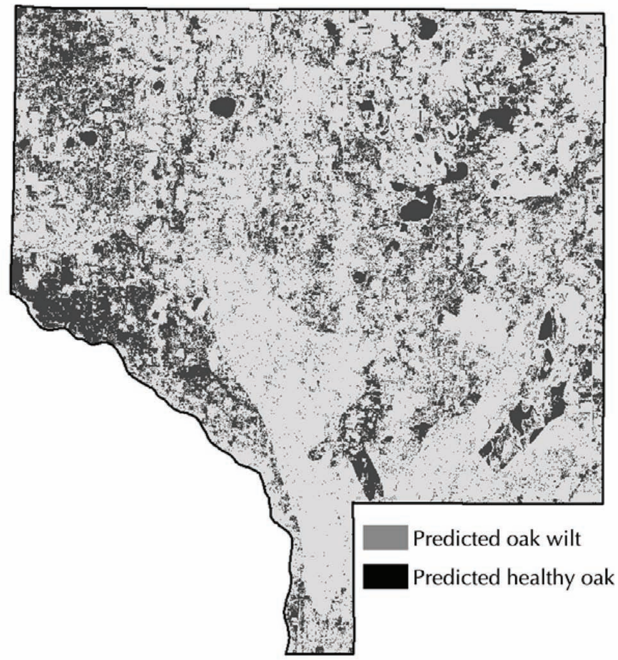

Figure 3. The predictive oak wilt surface for all of Anoka County.

for both the urban and non-urban models (TREE; S-PLUS©, Statistical Sciences, 2000). The output from the classification tree was the input for conditional statements (CON statements), (ESRI ArcView, 2000). CON statements were used to create an oak wilt presence or absence raster grid surface. The CON request performs a conditional if/else evaluation on a cell-by-cell basis. Grid theme cells with values of 0 indicated lower probabilities of oak wilt presence (defined as absence). Grid theme cells with values of 1 indicated higher probabilities of oak wilt presence.

After modeling the urban and non-urban sections of Anoka County, the two models were merged according to the predicted binary output for the entire study area into a single Potential Distribution of Oak Wilt grid surface (fig. 3).

Accuracy Assessment

There were two accuracy assessments performed. The initial accuracy was estimated as a sample-based misclassification error rate, the tenfold cross-validation, (Efron and Tibshirani 1993), calculated in S-PLUS© as part of the classification tree procedure. The cross-validation procedure validates the tree sequence by shrinking and/or pruning the tree by portioning the data into a number of subsets, fitting subtree sequences to these, and using a subset previously held out to evaluate the sequence. This procedure was used to identify the tree size that minimized the prediction error.

The second assessment estimated the accuracy of the urban and non-urban models, using actual treated oak wilt polygon data from a later date (LMIC data, 1999-2000). The assessment data contained polygons of known and verified oak wilt locations. To assess the accuracy of the urban model, a total of 164 known oak wilt polygons, with a mean size of 0.76, minimum size of .07, and a maximum size of 10.01 acres, were used. To assess the non-urban model, a total of 65 known oak wilt polygons, with a mean size of 1.94, minimum size of 0.14, and a maximum size of 13.16 acres, were used. These known oak wilt polygons were intersected with the predicted urban and non-urban models to determine the rate at which we accurately predicted the presence of oak.

ArcView Avenue code was used to measure the percentage of predicted oak wilt area within each assessment polygon (fig. 4). The data was assessed in three ways:

- presence of oak wilt in 50 percent or more of the assessment polygon.
- presence of oak wilt in 67 percent or more of the assessment polygon.
- presence of oak wilt in 100 percent of the assessment polygon.

To determine the presence of oak wilt in 50, 67, and 100 percent or greater of the assessment polygon, we intersected the assessment polygon with the results from the PDOW, then divided the area of predicted oak wilt within the assessment polygon by the total area of

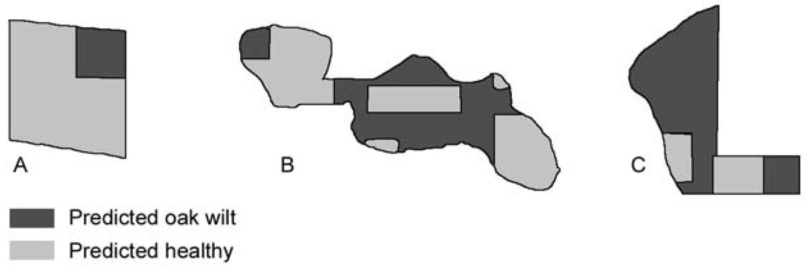

Figure 4. Examples of accuracy assessment polygons and the potential amounts of predicted oak wilt within any individual polygon. The dark gray denotes areas of predicted oak wilt and the light gray denotes areas predicted to be healthy oak. Figure A is an assessment polygon that has less than 50 percent predicted oak wilt by area; figure B has at least 50 percent predicted oak wilt; figure C has at least 75 percent predicted oak wilt.

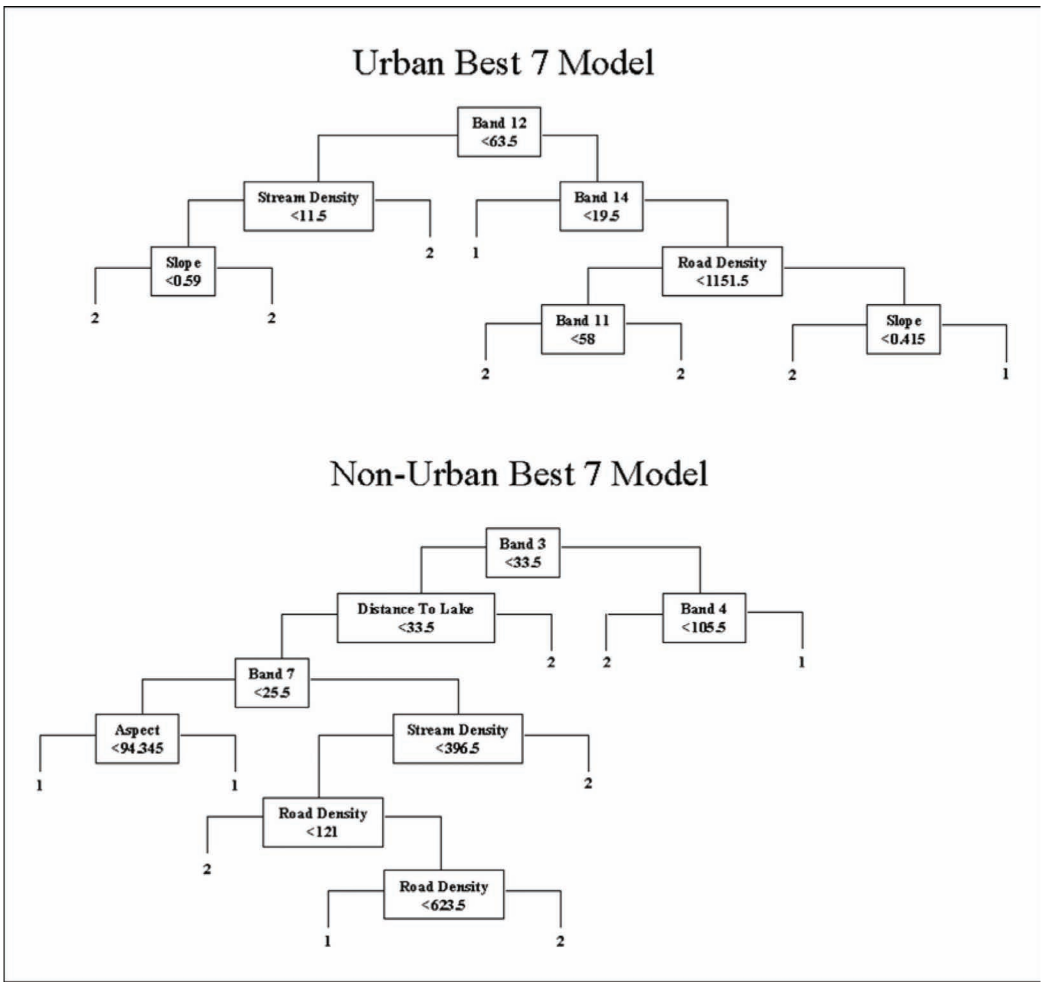

Figure 5. Urban and Non-urban decision tree models.

the assessment polygon. For an overall percentage of accuracy in predicting the presence of oak wilt, the number of assessment polygons that had 50 and 67 percent or more, and 100 percent predicted oak wilt by area, were totaled and divided by the total number of assessment polygons.

Results

Oak Wilt Potential Distribution Surface

The combined PDOW model predicted a potential for oak wilt presence in 79 percent of the County. The remaining 21 percent of the county had less potential for oak wilt presence and was identified as healthy. There were eight terminal end nodes, which accounted for 87.5 percent of the variability in the urban model. The combination of independent variables important in predicting the presence or absence of oak wilt were: Band 5 (September 1998), Stream Density, Slope, Band 7 (September 1998), Road Density, and Band 4 (September 1998).

For the non-urban model, there were nine terminal end nodes, which accounted for 82.4 percent of the variability in the non-urban model. The combination of independent variables important in predicting the presence or absence of oak wilt were: Band 3 (May 1998), Distance to Lakes, Band 7 (May 1998), Aspect, Stream Density, Road Density and Band 4 (May 1998) (fig. 5).

Proportion of Predicted Oak Wilt and Healthy Oak in Urban/Non-Urban Models

The proportions of predicted oak wilt and non-oak wilt of both models were calculated and compared for the entire study area (table 1). The comparison was made between the urban and non-urban models in relation to the proportion of the area predicted for oak wilt and healthy forested area. Results show there was 17 percent more predicted oak wilt in the urban model than in the non-urban model (table 1).

Accuracy Assessment

The classification tree selected through cross validation for each of the urban and non-urban models had

Table 1. Proportion of predicted oak wilt to healthy oak in the urban and non-urban models for the entire study area.

	Urban Model	Non-Urban Model	County Average
Predicted Oak Wilt	87.8 percent	70.7 percent	79.25 percent
Predicted Healthy Oak	12.2 percent	29.3 percent	20.75 percent

Table 2. A) Non-Urban Error matrix for the tenfold cross validation accuracy assessment. Non-urban errors of omission: Absent 24/62+24 = 27.9 percent; Present 15/15+121 = 11.0 percent. Non-urban errors of commission: Absent 15/62+24 = 17.4 percent; Present 24/15+121 = 17.6 percent. **B)** Urban errors of omission: Absent 21/14+21 = 60.0 percent; Present 4/4+161 = 2.4 percent. Urban errors of commission: Absent 4/14+21 = 11.4 percent; Present 21/4+161 = 12.7 percent.

A. Non-Urban	Classified Absent	Classified Present	B. Classified Urban	Classified Absent	Classified Present
Actual Absent	62	24	Actual Absent	14	21
Actual Present	15	121	Actual Present	4	161

Table 3. Accuracy assessment results using verified oak wilt assessment polygons from the 1999 and 2000 LMIC database. These assessment polygons were not used in the modeling process. **Column A:** Portion of assessment polygon predicted as having the potential for Oak Wilt; **Column B:** Proportion of assessment polygons that were accurately predicted with having oak wilt by the Urban Model; **Column C:** Proportion of assessment polygons that were accurately predicted with having oak wilt by the Non-Urban Model.

A	B	C
Half of Assessment Polygon	93.3 percent	87.7 percent
Three quarters of Assessment Polygon	90.9 percent	64.6 percent
Entire Polygon	70.1 percent	24.6 percent

misclassification errors of 0.125 and 0.1757, respectively. The error matrices are shown in table 2.

A second accuracy assessment was conducted on the oak wilt predictions for both the urban and non-urban sections of the county. Accuracy assessment results of the urban and non-urban models, using the LMIC actual oak wilt locations from 1999 and 2000 are found in table 3. The oak wilt locations from 1999 and 2000 were not part of the dataset used to develop the oak wilt models.

Discussion and Conclusions

We showed that using a classification tree on commonly acquired datasets could reliably predict the distribution of oak wilt across Anoka County, MN. Although our potential distribution of oak wilt might be considered a theoretical construct, (Felicisimo and others 2002) our accuracy assessment using additional oak wilt locations establishes that the classification tree analysis of large-scale, commonly acquired data can be successfully used to construct a model for the spatial distribution of oak wilt in both the urban and non-urban landscapes.

Land use, specifically the urban and non-urban categories, played an important role in modeling oak wilt. The total area of predicted oak wilt in the urban model differed from that in the non-urban model by 17 percent with more oak wilt predicted in the urban model than in the non-urban model. These results seem to suggest that stratifying the landscape according to urbanized and non-urbanized land cover produces more accurate predictive oak wilt models.

We showed that a classification tree analysis of large-scale data, combined with location data collected in the field, is useful for identifying indicator variables that are correlated with the presence and/or absence of oak wilt in Anoka County. Our model shows that Landsat TM Bands 3, 4, 5, and 7 are significant for predicting the presence or absence of oak wilt in Anoka County. Bands 3 and 4 are commonly used for vegetative analyses and are logical variables for predicting healthy or stressed vegetation. Bands 5 and 7 have also been related to vegetative condition but also to soil moisture retention. To substantiate the importance of these variables, further application of these methods in another study area dissimilar in characteristics to Anoka County is suggested.

Time spent in the field was limited to the collection of only one healthy sample site for every 2.5 oak wilt sites. We believe this biased our results: The greater percentage of oak wilt sites likely skewed the model, causing it to over predict the potential for oak wilt. An appropriate sample design is critical for modeling success. Acquiring more of the healthy sample sites, selected with a systematic sampling routine, will result in a model that is equally robust at predicting the location of oak wilt and healthy locations.

We were mindful of project costs as we constructed a methodology to be used by forest managers. We restricted our analysis to variables that were easily obtained and with minimum costs. In so doing, we reduced our ability to illuminate factors affecting oak wilt on a smaller spatial scale. Given additional resources, additional variables could be collected in the field at the sampling locations, such as: type of oak species, other tree species present, percent oak species, dominant over and under story, tree diameters, tree height, soil samples. These variables tend to change over short distances and may explain some residual conditions essential to the presence or absence of oak wilt. Future studies should also include soil information collected at the sample locations.

Acknowledgments

The authors would like to thank the following individuals from the USDA Forest Service, North Central Research Station, Forest Disease Unit, for their participation in this project: Jennifer Juzwik for Kathy Kromroy for their intellectual contributions during the conceptualization and execution of the project; Kathy Ward and Paul Castillo for their technical assistance in collecting oak wilt field data and the acquisition of appropriate data layers for use in the analysis. We would also like to acknowledge: Joseph O'Brien, from the USDA Forest Service, Forest Health and Protection, Northeastern area, for his general support of this work; and Susan Burks, Minnesota Department of Natural Resources, Division of Forestry, for her participation as a stakeholder.

References

Appel, D.N. and R. C. Maggio. 1984. Aerial survey for oak wilt incidence at three locations in central Texas. Plant Dis. 68:661-664.

Bailey, Robert G. 1996. Ecosystem Geography. New York: Springer. 90 p.

Bailey, Robert G. 1995. Description of the ecoregions of the United States (2nd ed.). Misc. Pub. No. 1391, Map scale 1:7,500,000. USDA Forest Service. 108pp.

Bailey, Robert G., Peter E. Avers, Thomas King and W. Henry McNab. 1994. Ecoregions and Subregions of the United States. USDA Forest Service. Map.

De'ath, G., and K.E. Frabricus. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81, 3178-3192.

Efron, B., R.J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall: New York.

Environmental Services Resources Inc. (ESRI). 1992-2000. ArcView GIS 3.2a. Redlands, CA, USA.

ERDAS, Inc. ERDAS Imagine V8.5. 2001. Atlanta Georgia

Felicisimo, A.M., E. Frances, J.M. Fernandez, A. Gonzalez-Diez, and J. Varas. 2002. Modeling the Potential Distribution of Forests with a GIS. Photogrammetric Engineering & Remote Sensing 68(5) 455-461.

Friedl, MA and CE Brodley 1997. Decision tree classification of land cover from remotely sensed data, Remote Sensing of Environment, 61:399-409.

Isaaks, E.H., and R.M. Srivastava. 1989. An Introduction to Applied Geostatistics. Oxford: Oxford University Press.

Joy, S.M. 2002. Modeling forest structure on the Kaibab National Forest in Arizona. Chapter 2. In: Northern goshawk habitat on the Kaibab National Forest in Arizona: factors affecting nest locations and territory quality. Ph.D. Dissertation, Colorado State University, Fort Collins.

Joy, S.M., R.M. Reich, and R.T. Reynolds. 2003. A non-parametric, supervised classification of vegetation types on the Kaibab National Forest using decision trees. International Journal of Remote Sensing 24(9):1835-1852.

Juzwik, J. 2000. An Oak Wilt Primer. The Journal of the International Oak Society 11:14-20.

Kelly, M. and R.K. Meentemeyer. 2002. Landscape Dynamics of the Spread of Sudden Oak Death, Photogrammetric Engineering and Remote Sensing. 68(10)1001-1011.

MathSoft, 1999. S-PLUS 2000: Modern Statistics and Advanced Graphics; Guide to Statistics, volume 1, 369-401.

McNab H.W. 1989. Terrain shape index: Quantifying Effect of Minor Landforms on Tree Height. Forest Science. 35(1):91-104.

Michaelson, J., Schmiel, D.S., Friedal, M.A., Davis, F.W., and Dubayah, R.O., 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. Journal of Vegetation Sciences, 5, 673-686

MNDNR, Minnesota Department of Natural Resources Division of Forestry. 1999. Ecological Subsections of Minnesota. [System ecssbne2]. http://deli.dnr.state.mn.us/metadata/full/ecssbne2.html. (August 27, 2003).

MNDNR, Minnesota Department of Natural Resources Division of Forestry - MIS Bureau. 2000. Land Type Associations of Minnesota. [System ecsltpy2]. http://deli.dnr.state.mn.us/metadata/full/ecsltpy2.html. (August 27, 2003).

O'Brien, J.G., M.E. Mielke, D. Starky, J Juzwik. 2003. How to Identify, Prevent, and Control Oak Wilt. USDA Forest Service, Northeastern Area State and Private Forestry. NA-PR-03-00. http://www.na.fs.fed.us/spfo/pubs/howtos/ht_oakwilt/toc.htm. (August 28, 2003).

Reich, R. M., J. E. Lundquist, V.A. Bravo. 2004. Spatial models for estimating fuel loads in the Black Hills, South Dakota, USA. International Journal of Wildland Fire, 13, 119-129.

U.S. Census Bureau. 2003. State and County QuickFacts: Anoka County, Minnesota. http://quickfacts.census.gov/qfd/states/27/27003.html. (November 2003).

USDA Forest Service Diseases of Trees in the Great Plains. 1986. USDA Forest Service Pub RM-129.

USDA Forest Service, Forest Health Protection, Forest Health Technology Enterprise Team (FHTET). Spatial Statistics: Spatial data creation and collection. http://www.fs.fed.us/foresthealth/technology/spatialstatistics/. (November, 2003)

Wright, H.E. Jr. 1972. Physiography of Minnesota in Geology of Minnesota. Eds P.K. Sims and G.B> Morey. St. Paul, Minnesota Geological Survey. Pp 561-78.