Modeling Risk for SOD Nationwide: What Are the Effects of Model Choice on Risk Prediction?[1]

M. Kelly[2, 3], D. Shaari[3], Q. Guo[2], and D. Liu[2]

Abstract

Phytophthora ramorum has the potential to infect many forest types found throughout the United States. Efforts to model the potential habitat for *P. ramorum* and sudden oak death (SOD) are important for disease regulation and management. Yet, spatial models using identical data can have differing results. In this paper we examine the results from five types of models generated from common input parameters, and investigate model agreement for distribution of risk for *P. ramorum*. We examine five models: (1) Rule-based, (2) Logistic regression, (3) Classification and Regression Trees, (4) Genetic Algorithm modeling, and (5) Support Vector Machines. The models differed in terms of parametric and non-parametric requirements, necessity for presence/absence data, and whether or not the explanatory variables were determined *a priori* or revealed during the model process. Nationwide input data included vegetation/host (hardwood diversity and hardwood density), topography, and climate (e.g. precipitation, frost days, temperature, and many other layers). We developed a risk map for the conterminous United States in which probabilities for *P. ramorum* disease establishment were based not on one model, but on agreement between multiple models. The five models were consistent in their prediction of some SOD risk in coastal CA, OR and WA. All models predicted some risk in the northern foothills of the Sierra Nevada mountains in CA. Outside of the west coast, the combined models predicted highest risk for SOD in a east-west oriented band including eastern OK, central AR, TN, KY, northern MI, AL, GA and SC, parts of central NC, and eastern VA, DL and MD. The paper also discusses issues of input data accuracy, coverage, availability of nationwide host datasets, data scale, and model computational requirements. Although theoretical in nature, the results of this paper have practical and applied value for managers and regulators of this disease.

Key words: geographic information systems, spatial modeling, sudden oak death

---

[1] A version of this paper was presented at the Sudden Oak Death Second Science Symposium: The State of Our Knowledge, January 18-21, 2005, Monterey, California

[2] Department of Environmental Sciences, Policy and Management. University of California Berkeley, Berkeley. mkelly@nature.berkeley.edu

[3] Center for the Assessment and Monitoring of Forest and Environmental Resources, College of Natural Resources, University of California, Berkeley

Introduction

Phytophthora ramorum has the potential to infect many forest types found throughout the United States. Efforts to model the potential habitat for *P. ramorum* and sudden oak death (SOD) are important for disease regulation and management. For example, in California, researchers used an expert-knowledge driven Rule-based model to predict risk of SOD based on host, temperature and moisture data (Meentemeyer and others 2004) which has been used to guide sampling, aerial surveys, and other statewide monitoring efforts. Rule-based models such as this one use expert input rather than statistical inference, and thus the predictor ecological variables used are known *a priori*. Other spatially referenced ecological niche models require presence data to train the model, possibly revealing ecological niches. The variety of techniques used for ecological niche modeling is growing (Guisan and Zimmermann 2000), and there has been a corresponding increase in the spatial modeling literature in work that compares results from different models (Guisan and Zimmermann 2000; Manel and others 1999; Muñoz and Felicísimo 2004). These efforts show that spatially referenced models using identical data often have differing results, due in part to 1) the fact that models can be either parametric or non-parametric with varying reliance on variable distribution, 2) user-defined weightings placed on variables can differ by analyst, and 3) the different methods used for generating absence data for input to models can influence results, among other factors. In this research, we examined a collection of model types and produced a risk map for SOD nationwide based on agreement between these models. We examined five classes of models that differed in terms of parametric and non-parametric requirements, necessity for presence/absence data, and whether or not the explanatory variables were determined *a priori* or revealed during the model process. Our spatial models include the following: (1) Rule-based, (2) Logistic regression (LR), (3) Classification and Regression Trees (CART), (4) Genetic Algorithm (GA) modeling, and (5) Support Vector Machines (SVM).

Model Descriptions

Rule-based spatial models use research data and expert input, rather than statistical inference to determine the importance of predictor variables (Meentemeyer and others 2004). Predictor variables are given weights based on importance, and all weighted variables are summed in a geographical information system (GIS) overlay procedure to produce a mapped output. This method is straightforward and not computationally intensive. Logistic regression (LR) is a variation of ordinary regression which is used when the dependent (response) variable is a binary variable

which usually represents the occurrence or non-occurrence of some outcome event, usually coded as 0 or 1, and the independent (input) variables are continuous, categorical, or both. Resulting probabilities can be mapped over space for an easily understood cartographic representation of modeled distribution. LR is a powerful parametric yet computationally non-intensive method for ecological niche modeling (Felicisimo and others 2002; Kelly and others 2001; Mladenoff and others 1995). Classification trees are a non-parametric alternative to parametric techniques such as logistic regression (De'ath and Fabricius 2000) and Linear Discriminant Analysis (LDA) (Feldesman 2002). The tree models are developed by recursively partitioning the response variable into increasingly homogeneous binary subsets based on critical thresholds in predictor variables. The split chosen is the one that most reduces the average impurity in the resulting bins (Breiman and others 1984; De'ath and Fabricius 2000; Venables and Ripley 2002). The resulting "trees" are often displayed graphically, and are easy to understand as a series of if/then conditions, but they can be complex to render cartographically (Kelly and Meentemeyer 2002; Muñoz and Felicísimo 2004). Genetic Algorithm (GA) modeling is a an evolutionary computing system that has excellent capabilities for delineating ecological niches and geographical distributions of species (Raxworthy and others 2003; Stockwell 1999). The method uses genetic algorithms to predict the potential distribution of a species by generating a set of rules. Essentially, genetic algorithms apply the operational concepts similar to evolutionary biology such as mutation and crossover to evolve the solutions in order to find the best solutions. GA modeling is less susceptible to local maxima and able to handle various data formats (continuous and discrete). Finally, Support Vector Machines (SVM) are a new generation of learning algorithms that can perform binary classification (pattern recognition) and real valued function approximation (regression estimation) tasks. SVMs have been developed on a solid base of statistical learning theory and are designed especially to provide high flexibility for approximating class boundaries, while avoiding over-fitting phenomena. Functionally, SVMs seek to find an optimal separating hyperplane with the maximal margin between the training points for presence and absence data in multidimensional space (Cristianini and Scholkopf 2002; Huang and others 2002). SVMs are able to handle non-linear and categorical data, make no assumption on the probability density of the data, and are competitive with the best available machine learning algorithms in classifying high-dimensional data sets (Guo and others 2005). Model characteristics for these five models are summarized in *table 1*.

Table 1—*Characteristics of the five classes of models used.*

Model Name	Presence/absence data required?	Parametric/Non-parametric	Important Variable selection	Output	Computational requirements
Rule-based	No	Non-parametric	*a priori*	Ranked	Low
Logistic Regression	Yes	(semi-) Parametric	Through training	Probability	Low
CART[1]	Yes	Non-parametric	Through training	P/A based on # runs	Low
GA[2]	Yes	Both	Through training	P/A based on # runs	High
SVM[3]	No (1-class) Yes (2-class)	Non-parametric	Through training	P/A based on # runs	High

[1] Classification and Regression Tree; [2] Genetic Algorithm; [3] Support Vector Machines.

Methods

Predictor Ecological Variables

We developed five spatial models using common nationwide data. Physical variables included topography and climate; we used Digital Elevation Model (DEM) and DAYMET weather and climatologically modeled raster surfaces gridded at 1 km to summarize physical conditions for the pathogen. Daymet is an assortment of climate raster surfaces interpolated from ground-based meteorological stations on a daily basis over an 18 year period (1980 to 1997) (Thornton and others 1997). The primary climate surfaces utilized in this modeling project included total annual precipitation, total annual frost days, average minimum temperature, average maximum temperature and average maximum August temperature. Topography was derived from United States Geological Survey (USGS) National Elevation Dataset (USGS. 1999b).

Host/Vegetation Data

We had a considerable challenge finding a detailed vegetation map for the conterminous United States with sufficient floristic and spatial detail to allow modeling. We explored the utility of several datasets. The first among these was the National Land Cover Data (NLCD) 1992 dataset, a 21-category classification derived primarily from Landsat Thematic Mapper (TM) imagery from 1992, containing categories for several different forest types including deciduous, conifer and mixed (Vogelmann and others 1998; Vogelmann and others 2001). The NLCD classification supplies high spatial resolution (30-m) but poor floristic detail, providing only three general vegetation categories relevant to our project: deciduous forest, evergreen forest, and mixed forest. Using this information, we created a 1-km gridded vegetation dataset depicting "hardwood density" based on the percent of deciduous and mixed forest found within each 1-km cell. The second dataset we investigated was the digital tree range maps for North America created by the USGS for a vegetation climate modeling study (USGS. 1999a), which provided more floristic detail, but was coarse in spatial detail. This product was based upon a series of tree range maps assembled by Elbert L. Little, Jr. in the 1970s as the "Atlas of the United States Trees" (Little 1971; Little 1976; Little 1977; USGS. 1999a). Of the 58 digital oak species maps created by USGS, we determined that 34 were potentially susceptible to *P. ramorum*. These 34 oak species maps were then combined with digital maps of 12 other tree range maps of species found to either be directly susceptible to the pathogen or to be related (*i.e.*, in the same taxonomic genus as a susceptible species). The 46 tree range maps were then combined to form a "hardwood diversity index" map. It should be emphasized that the hardwood diversity index as calculated for this study was limited to a portion of the tree range maps made available by the USGS and contains only a minimal number of shrub or understory species. It is only intended to represent areas within the United States that potentially contain high numbers of susceptible *P. ramorum* host species (both foliar and terminal hosts) with the recognition that there are many more species not included. The final vegetation layer examined was provided by the USDA-Forest Service Northeastern Research Station (Gottschalk and others 2002). In this product, Forest Inventory and Analysis (FIA) plot data were used to calculate the percentage of forest basal area composed of the red and live oak groups and these points were kriged to create a continuous raster surface for the eastern United States. Percent basal area estimates were adjusted for forest density using the NLCD dataset.

Model Creation

We first developed our nationwide rule-based model using similar input data to those used in the California model. Climate variables for six winter months were parameterized and placed into weighted classes in accordance with the methods of Meentemeyer and others (2002). Two different coarse-resolution vegetation maps (hardwood diversity and hardwood density) were used as a surrogate for the detailed vegetation map used in the California modeling case. The remaining models (with the exception of GA) required that presence and absence data be generated for model training. For presence data we used the database of locations (n = 169) of confirmed *P. ramorum* existing in 12 counties in California and one in Oregon provided through the University of California, Berkeley (Kelly and others 2004). Absence data (or in this case 'pseudo-absence' data, as we do not have samples of *P. ramorum* absence) were generated in the following manner. We created a zone of infestation within California consisting of the 12 infested counties and six border counties. We then constrained the algorithms to search for and generate pseudo-absence data from locations in California outside this zone.

We began with CART modeling. Using Splus v. 6.2 for Windows, we generated 100 classification trees using 100 unique pseudo-absence point distributions and the 169 *P. ramorum* presence points. Each "tree" was pruned by examining a plot of deviance and tree complexity (Feldesman 2002), and the resulting tree models were implemented in ArcInfo using Arc Macro Language (AMLs). We then developed LR equations using Splus v. 6.2 for Windows and implemented the results in ArcInfo. Desktop GARP (Genetic Algorithm for Rule-set Production) (Stockwell and Peters 1999) software was used for the application of the GA model. One hundred model runs were performed using presence data only. Finally, we developed SVM models using Matlab and LIBSVM software (Chuang and Lin 2001). Cross-validation was used for each of the 100 runs to optimize parameter selection. It became clear early on that the vegetation information would not be useful in the initial modeling exercises, as detailed vegetation maps are not available for every state. The initial model inputs (with the exception of the Rule-based model) were limited to climate variables.

The different predicted *P. ramorum* risk map results of the models examined in this project displayed significant levels of variation depending on the model in question and the input climate variables used. Therefore, a combination of several different model input formulas was used to create a cumulative and weighted map result. The four predictive models (excluding Rule-based) were run 100 times each with the

following three collections of input predictor variables: 1) precipitation total and frost days, 2) precipitation total, frost days and average maximum temperature, and 3) precipitation total, average minimum temperature and average maximum temperature. Model results were added together along with proportional values from the Rule-based models to create a finalized *P. ramorum* risk grid. Finally, in an effort to eliminate areas of non-hardwood forest in the final risk map, we filtered the combined model results through the NLCD and red oak basal area vegetation maps by multiplying the map by each vegetation map rescaled from 0 to 1. We did not include the USGS vegetation map in this exercise as the spatial fidelity of the product seemed problematic. The representation of hardwood diversity in the southeast United States may not be accurate due to the relatively small number of tree ranges used.

All model results were normalized for visual display purposes using the following technique: the mean value was calculated and then arbitrary boundaries were set for plus or minus two standard deviations. Any values above or below were reclassed to the minimum or maximum of the 95th percentile. This grid was then rescaled from 0 to 100, and classed into five classes of risk: 0 to 20 percent low, 20 to 40 percent, 40 to 70 percent medium, 70 to 90 percent, and 90 to 100 percent. The exception to this approach was the Rule-based model, which was scaled and displayed according to Meentemeyer and others (2004).

Results

The results from each of five models are shown in *Figure 1*. The rule-based model shows a high risk for *P. ramorum* spread across the southeast from eastern TX to VA, with declining risk to the north. The model also shows high risk in the northwest and risk in the northern foothills of the Sierra Nevada mountains in CA. While this model is a replica of that provided by the Meentemeyer and others (2004) model, this is a different result and likely due to qualitative differences in input data and spatial resolution. The LR results show a broadly similar pattern to the Rule-based model, with less risk on the west coast and less overall risk in the southeast: the model constrains the highest risk to the deep southern states of LA, AL and MS. The LR formula determined that precipitation total, average minimum temperature and average maximum temperature were important in risk prediction. CART and GA results are similar, with a band of highest risk occurring throughout the middle southeast from OK in the west to VA and NC in the east. This is likely due to the importance both models placed on latitudinally controlled temperature variables. Both models found precipitation total, frost days and average maximum temperature to be the most important predictors. The SVM result is coarser than the others; the

algorithm required resampling of the input data to a coarser spatial resolution (~12 km) due to computational limitations. The SVM algorithm predicted risk as a result of precipitation total, frost days and average maximum temperature.

The final risk map – a combination of the results from five models – is shown in *Figure 2*. The five models were consistent in their prediction of some *P. ramorum* risk in coastal CA, OR and WA, although LR and GA show less risk than do the other models. All models predicted some risk in the northern foothills of the Sierra Nevada mountains in CA. Outside of the west coast, the combined models predicted highest risk for *P. ramorum* in an east-west oriented band including hardwood forested areas of OK, AR, TN, KY, northern portions of MI, AL, GA and SC, parts of central NC, eastern VA, DL and MD. This area includes portions of several ecoregions, including the Piedmont ecoregion of NC, SC, GA, AL and MI, a transitional area between the mostly mountainous ecoregions of the Appalachians to the northwest and the relatively flat coastal plain to the southeast. Much of this region has reverted to successional pine and hardwood woodlands, with an increasing conversion to an urban and suburban land cover (ECOMAP 1993; Omernik 1987; Omernik 1995). There is also predicted risk for *P. ramorum* in hardwood forests of the Southwestern Appalachains in TN and in the primarily oak-hickory forests of the Interior Plateau in KY and TN. Eastern OK and central AR shows high risk in the oak-hickory-pine forests of the Ouachita Mountains, and in the red oak, white oak, and hickory dominated forests of the Boston Mountains, and in the predominantly oak forests of the Southern Ozarks (ECOMAP 1993; Omernik 1987; Omernik 1995). Coastal MD and DL, part of the Middle Atlantic Coastal Plain, are climactically susceptible and have forests at risk in riparian areas.

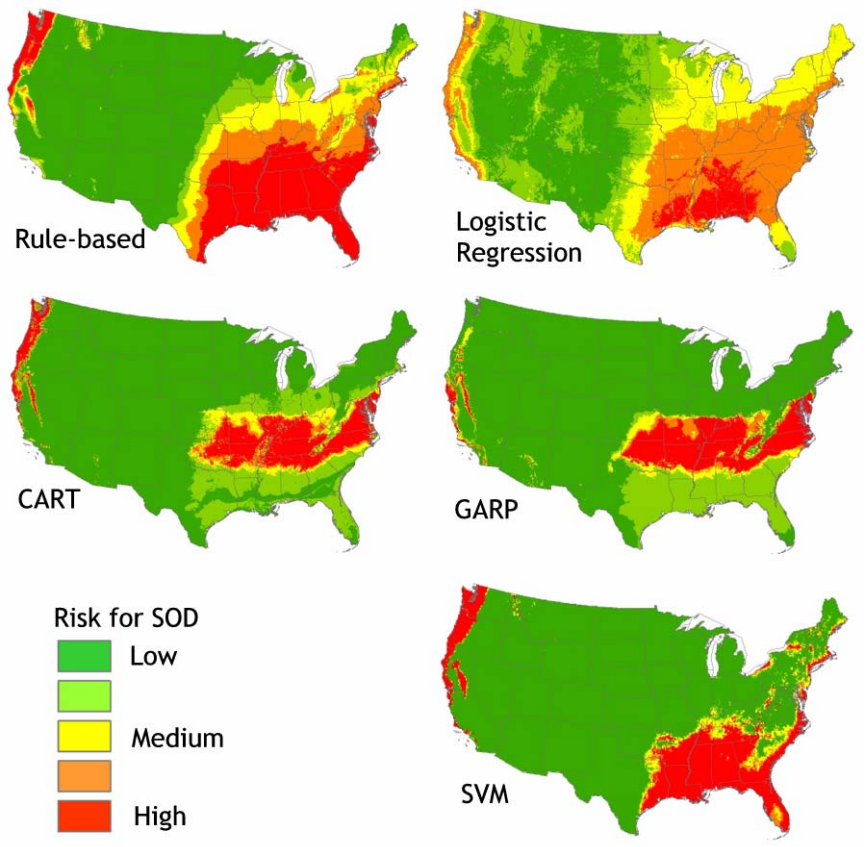

Figure 1—Risk for Sudden Oak Death in the conterminous United States: results from five spatially referenced models**.**

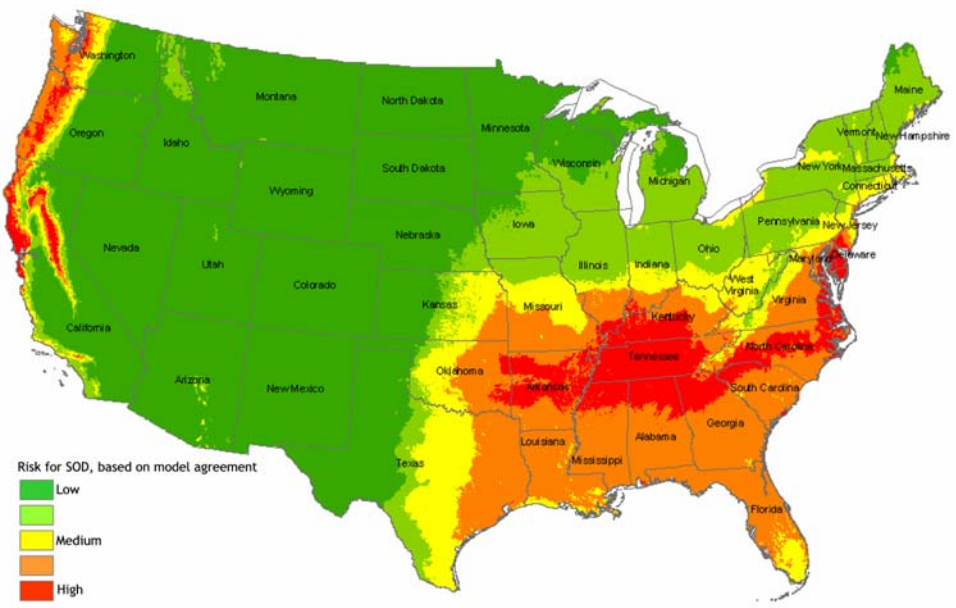

Figure 2—Risk for Sudden Oak Death in the conterminous United States based on agreement between five spatially referenced models.

Discussion and Conclusions

Because there are no wildland cases of SOD outside of California and Oregon, none of these models can be assessed for accuracy. This is an unfortunate but not uncommon situation when modeling invasive species (Muñoz and Felicísimo 2004). Several of the models allow for some form of cross-fold validation tools for assessing accuracy, but these tools can be problematic due to the small number of training samples and their concentrated distribution (Graham and others 2004). The generation of pseudo-absence data must also be examined. We do not have reliable negatives for *P. ramorum*, so we used a common method for generation of pseudo-absence data, constraining the pool of possible absence points to be taken from outside the zone of infestation in California. Experiments with pseudo-absence data generated within the zone of infestation resulted in models that over-predicted the risk of the disease. One-class SVMs are able to model species distributions without absence data (Guo and others 2005), and this method might be of use in this case. We only examined two-class SVMs in this paper, for consistency. Finally, the host data available for the entire United States was the largest limiting factor in our modeling exercise; all nationwide vegetation layers we used had significant drawbacks. For example, the NLCD data was the most spatially comprehensive layer, with complete coverage at a high spatial resolution; however, specific floristic detail was absent, and the vegetation classes were much too broad to be of use in the modeling exercise. The USGS layer had sufficient floristic detail, but was tremendously course in resolution. Finally, the FIA product only covered the east coast area, and thus could not be used in the models that required training. A similar west coast product is not available. Computationally, the Rule-based, Logistic regression and CART were the least computationally demanding of the algorithms. GARP and SVM require more CPU time. This work examined common ecological niches for *P. ramorum*, but an investigation of the human component to disease establishment and spread should also be considered. The locations of nurseries might be an additional important component to this research. Attempts at finding similar climate datasets for European locations of *P. ramorum* were not successful prior to completion of this paper but should be considered. Although theoretical in nature, the results of this paper have practical, applied value for managers and regulators of this pathogen.

References

Breiman, L.; Friedman, J.H.; Olshen, R.A.; and Stone, C.J., 1984. **Classification and Regression Trees**. Chapman and Hall, New York.

Chuang, C.C. and Lin, C.J., 2001. **LIBSVM - A library for Support Vector Machines**.

Cristianini, N. and Scholkopf, B. 2002. **Support vector machines and kernel methods - The new generation of learning machines**. Ai Magazine, 23(3): 31-41.

De'ath, G. and Fabricius, K. 2000. **Classification and regression trees: a powerful yet simple technique for ecological data analysis**. Ecology, 81(11): 3178-3192.

ECOMAP, 1993. **National hierarchical framework of ecological units**, USDA Forest Service, Washington, DC.

Feldesman, M.R. 2002. **Classification trees as an alternative to Linear Discriminant Analysis**. American Journal of Physical Anthropology, 119: 257–275.

Felicisimo, A.M.; Frances, E.; Fernandez, J.M.; Gondalez-Diez, A.; and Varas, J. 2002. **Modeling the potential distribution of forests with a GIS**. Photogrammetric Engineering & Remote Sensing, 68(5): 455-462.

Gottschalk, K.W.; Morin, R.S.; and Liebhold, A.M., 2002. **Potential susceptibility of eastern forests to Sudden Oak Death, *Phytophthora ramorum***, USDA Forest Service Forest Health Monitoring (FHM) Conference.

Graham, C.H.; Ferrier, S.; Huettman, F.; Moritz, C.; and Peterson, A.T. 2004. **New developments in museum-based informatics and applications in biodiversity analysis**. TRENDS in Ecology and Evolution, 19(9): 497-503.

Guisan, A. and Zimmermann, N.E. 2000. **Predictive habitat distribution models in ecology**. Ecological Modelling, 135: 147-186.

Guo, Q.; Kelly, M.; and Graham, C. 2005. **Support vector machines for predicting distribution of Sudden Oak Death in California**. Ecological Modelling, 128(1): 75-90.

Huang, C.; Davis, L.S.; and Townshend, J.R.G. 2002. **An assessment of support vector machines for land cover classification**. International Journal of Remote Sensing, 23(4): 725-749.

Kelly, M. and Meentemeyer, R.K. 2002. **Landscape dynamics of the spread of Sudden Oak Death**. Photogrammetric Engineering & Remote Sensing, 68(10): 1001-1009.

Kelly, M.; Tuxen, K.; and Kearns, F. 2004. **Geospatial informatics for management of a new forest disease: sudden oak death**. Photogrammetric Engineering and Remote Sensing, 70(1): 1001-1004.

Kelly, N.M.; Fonseca, M.; and Whitfield, P. 2001. **Predictive mapping for management and conservation of seagrass beds in North Carolina**. Aquatic Conservation: Marine and Freshwater Ecosystems, 11(6): 437-451.

Little, E.L., 1971. **Atlas of United States trees, volume 1, conifers and important hardwoods**. Miscellaneous Publication 1146, U.S. Department of Agriculture.

Little, E.L., 1976. **Atlas of United States trees, volume 3, minor Western hardwoods**. Miscellaneous Publication 1314, U.S. Department of Agriculture.

Little, E.L., 1977. **Atlas of United States trees, volume 4, minor Eastern hardwoods**. Miscellaneous Publication 1342, U.S. Department of Agriculture.

Manel, S.; Dias, J.M.; and Ormerod, S.J. 1999. **Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird**. Ecological Modelling, 120: 337-347.

Meentemeyer, R.; Rizzo, D.; Mark, W.; and Lotz, E. 2004. **Mapping the risk of establishment and spread of sudden oak death in California**. Forest Ecology & Management, 200(1-3): 195-214.

Mladenoff, D.J.; Sickley, T.A.; Haight, R.G.; and Wydeven, A.P. 1995. **A regional landscape analysis and prediction of favorable grey wolf habitat in the northern great lakes region**. Conservation Biology, 9: 279-294.

Muñoz, J. and Felicísimo, Á.M. 2004. **Comparison of statistical methods commonly used in predictive modelling**. Journal of Vegetation Science, 15: 285-292.

Omernik, J.M. 1987. **Ecoregions of the conterminous United States. Map (scale 1:7,500,000)**. Annals of the Association of American Geographers, 77(1): 118-125.

Omernik, J.M., 1995. **Ecoregions: A spatial framework for environmental management**. In: W.S. Davis and T.P. Simon (Editors), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, FL, pp. 49-62.

Raxworthy, C.J.; Martinez-Meyer, E.; Horning, N.; Nussbaum, R.A.; Schneider, G.E.; Ortega-Huerta, M.A.; and Peterson, A.T. 2003. **Predicting distributions of known and unknown reptile species in Madagascar**. Nature, 426(18/25 December): 837-841.

Stockwell, D.R., 1999. **Genetic Algorithms II**. In: F. A.H. (Editor), Machine learning methods for ecological applications. Kluwer Academic Publishers, Boston, pp. 123-144.

Stockwell, D.R.B. and Peters, D.P. 1999. **The GARP modelling system: Problems and solutions to automated spatial prediction**. International Journal of Geographic Information Systems, 13: 143-158.

Thornton, P.E.; Running, S.W.; and White, M.A. 1997. **Generating surfaces of daily meteorological variables over large regions of complex terrain**. Journal of Hydrology, 190: 214-251.

U.S.G.S., 1999a. **Digital representation of "Atlas of United States Trees" by E.L. Little Jr.**, U.S. Geological Survey.

U.S.G.S., 1999b. **National Elevation Database (USGS)**, U.S. Geological Survey, Sioux Falls, SD.

Venables, W.N. and Ripley, B.D., 2002. **Modern Applied Statistics with S**. Springer, New York, 495 pp.

Vogelmann, J.; Sohl, T.; and Howard, S. 1998. **Regional characterization of land cover using multiple sources of data**. Photogrammetric Engineering and Remote Sensing, 64(1): 45-57.

Vogelmann, J.E.; Howard, S.M.; Yang, Y.; Larson, C.R.; Wylie, B.K.; and Van Driel, N. 2001. **Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper Data and ancillary datasources**. Photogrammetric Engineering and Remote Sensing, 67: 650-652.