Available online at www.sciencedirect.com
SCIENCE DIRECT®
Ecological Modelling 182 (2005) 75–90
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Support vector machines for predicting distribution of Sudden Oak Death in California

Qinghua Guo[a,*], Maggi Kelly[a,1], Catherine H. Graham[b,c]

[a] *Department of Environmental Science, Policy and Management, 151 Hilgard Hall #3110, University of California, Berkeley, CA 94720-3110, USA*
[b] *3101 Valley Life Science Building, Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720-3160, USA*
[c] *Department of Ecology and Evolution, 650 Life Sciences Building, Stony Brook University, NY 11794-5245, USA*

Received 1 May 2003; received in revised form 22 June 2004; accepted 12 July 2004

Abstract

In the central California coastal forests, a newly discovered virulent pathogen (*Phytophthora ramorum*) has killed hundreds of thousands of native oak trees. Predicting the potential distribution of the disease in California remains an urgent demand of regulators and scientists. Most methods used to map potential ranges of species (e.g. multivariate or logistic regression) require both presence and absence data, the latter of which are not always feasibly collected, and thus the methods often require the generation of 'pseudo' absence data. Other methods (e.g. BIOCLIM and DOMAIN) seek to model the presence-only data directly. In this study, we present alternative methods to conventional approaches to modeling by developing support vector machines (SVMs), which are the new generation of machine learning algorithms used to find optimal separability between classes within datasets, to predict the potential distribution of Sudden Oak Death in California. We compared the performances of two types of SVMs models: two-class SVMs with 'pseudo' absence data and one-class SVMs. Both models performed well. The one-class SVMs have a slightly better true-positive rate (0.9272 ± 0.0460 S.D.) than the two-class SVMs (0.9105 ± 0.0712 S.D.). However, the area predicted to be at risk for the disease using the one-class SVMs (18,441 km^2) is much larger than that of the two-class SVMs (13,828 km^2). Both models show that the majority of disease risk will occur in coastal areas. Compared with the results of two-class SVMs, the one-class SVMs predict a potential risk in the foothills of the Sierra Nevada mountain ranges; much greater risks are also found in Los Angles and Humboldt Counties. We believe the support vector machines when coupled with geographic information system (GIS) will be a useful method to deal with presence-only data in ecological analysis over a range of scales.

Keywords: Geographic information systems; Support vector machines; Potential disease spread; Sudden Oak Death

* Corresponding author. Tel.: +1 510 642 7898; fax: +1 510 643 5098.
E-mail addresses: gqh@nature.berkeley.edu (Q. Guo), mkelly@nature.berkeley.edu (M. Kelly), cgraham@life.bio.sunysb.edu (C.H. Graham).
[1] Tel.: +1 510 642 7272; fax: +1 510 642 1477.

0304-3800/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2004.07.012

1. Introduction

In the central California coastal forests, a newly discovered virulent pathogen (*Phytophthora ramorum*) has killed hundreds of thousands of native trees including tanoak (*Lithocarpus densiflorus*), coast live oak (*Quercus agrifolia*), and black oak (*Quercus kelloggii*) (Rizzo et al., 2002; Rizzo and Garbelotto, 2003). The disease was quickly and convincingly dubbed "Sudden Oak Death" (SOD) by both the popular press and the research community (Garbelotto et al., 2001; Rizzo et al., 2002). The state of California has dedicated millions of dollars for management of the disease, and a monitoring system has been implemented for the state's forests. As of June 2004, the disease existed in 13 coastal counties in the state, and hosts for the disease exist throughout the Californian coastal and foothill forests (Fig. 1). Predicting the potential

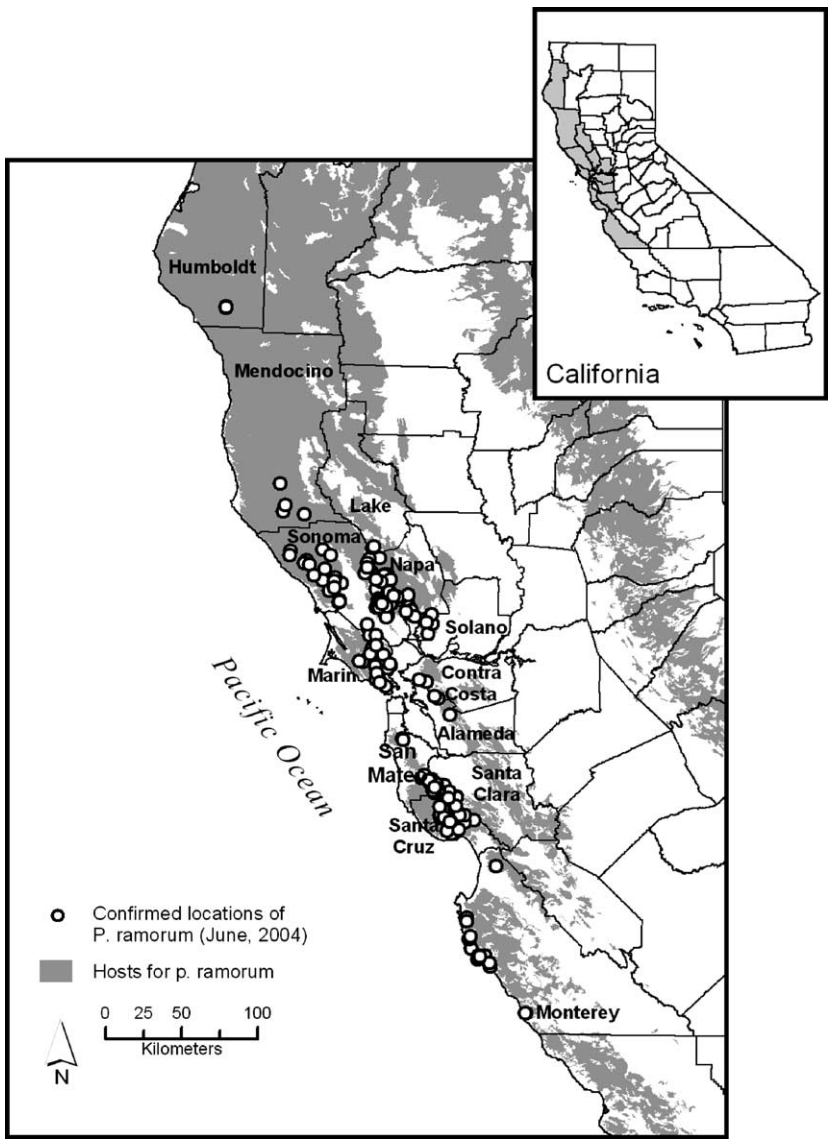

Fig. 1. Sudden Oak Death in California as of June 2004.

distribution of the disease in California remains an urgent demand of regulators and scientists alike, and requires innovative approaches to modeling its potential spread.

Since the existence and dispersal of any pathogen is likely influenced by environmental conditions such as humidity, temperature, and elevation, niche models that combine known localities of a given species with layers of meaningful ecological data can be used to extrapolate suitable environmental parameters for a given species (Franklin, 1995). Researchers have demonstrated that environmental niche models are powerful tools for predicting the potential distribution and spread of a disease or invasive species (Peterson and Vieglais, 2001; Peterson et al., 2002; Welk et al., 2002; Beard et al., 2003). Methods used to predict species distributions in niche models consist of various statistical approaches such as linear, multivariate, and logistic regression (Mladenoff et al., 1995; Bian and West, 1997; Kelly et al., 2001; Felicisimo et al., 2002; Fonseca et al., 2002), generalized linear modeling and generalized additive modeling (Frescino et al., 2001; Guisan et al., 2002), discriminant analysis (Livingston et al., 1990; Fielding and Haworth, 1995; Manel et al., 1999), classification and regression tree analyses (De'ath and Fabricius, 2000; Fabricius and De'ath, 2001; Kelly, 2002), genetic algorithms (Stockwell and Peters, 1999), and artificial neural networks (Manel et al., 1999; Spitz and Lek, 1999; Moisen and Frescino, 2002).

These methods require data on species presence and absence to establish a statistical relationship. However, in reality, many types of ecological datasets (e.g. those collected by museums or on wildlife surveys) lack reliable absence data. Such is the case with Sudden Oak Death. Confirmation of *P. ramorum* is a time-consuming process that involves culturing the pathogen from material removed from the border of an infected canker. After about a week, the pathogen can be identified based on morphological traits (Garbelotto et al., 2001). However, while negative samples are reported, these are not used for disease management, nor are they used to determine regulation such as quarantine boundaries, due to the large potential for false negatives caused by seasonality of sample, species of host, and time to lab. Because of the false-negative rate, negative samples are not used as proxies for absence data in this case.

The traditional approach to statistical modeling when only presence data is collected is to generate 'pseudo' absence data (Zaniewski et al., 2002). There are problems with this, however, among them the fact that in generating 'pseudo' absence data, one will likely affect the prediction accuracy by sampling the potential distribution area. As an alternative approach, some researchers have proposed the direct modeling of presence-only data, avoiding the generation of 'pseudo' absence data. Examples of this approach include BIOCLIM (Busby, 1986), DOMAIN (Carpenter et al., 1993) and ENFA models (Hirzel et al., 2002), which have been successfully applied in various ecological studies. There are other newer modeling methods that show promise, among these are support vector machines (SVMs).

1.1. Support Vector Machines

In recent years, with the advance of computational efficiency combined with sophisticated statistical methods, machine learning methods have been increasingly used and shown as powerful tools in a wide variety of science disciplines including planetary science, computer science, bioinformatics, and environmental science (Mjolsness and DeCoste, 2001). Among many machine-learning methods, SVMs, originally developed by Vapnik (1995), are considered to be a new generation of learning algorithms. SVMs have several appealing characteristics for modelers, including: they are statistically based models rather than loose analogies with natural learning systems, and they theoretically guarantee performance (Cristianini and Scholkopf, 2002). SVMs have been applied successfully to text categorization, handwriting recognition, gene-function prediction, and remote sensing classification, demonstrating the utility of the method across disciplines, and proving that SVMs produce very competitive results with the best available classification methods, and require just a minimum amount of model tuning (Joachims, 1998; Brown et al., 2000; Cristianini and Scholkopf, 2002; Decoste and Scholkopf, 2002; Huang et al., 2002).

Typically, SVMs are designed for two-class problems where both positive and negative objects exist. For these classification problems, two-class SVMs seek to find a hyperplane in the feature space that maximally separates the two target classes. In reality, as

mentioned above, we often do not have negative data, and thus commonly have only a one-class dataset. For example, in a handwritten number recognition problem, suppose that we are trying to classify the handwriting number "3" when we only have as a sample a set of handwritten "3"s. Here we need to develop a classifier to identify whether the target-handwritten number is a "3," without examples of what is not a "3." A parallel exists in ecology, where many museum-collected records exist in presence-only format; these data are often used to predict the potential distribution of a species. These are common one-class problems, which require the separation of a target class from the rest of the feature space. Computationally, the one-class problem is more difficult to solve than the traditional two-class problem, because the latter has positive and negative data to train and constrain the statistical learning models, while the former only has positive data to constraint the model (Tax and Duin, 1999b).

Recently, Scholkopf et al. (1999) developed one-class SVMs to deal with the one-class problem. In Scholkopf's experiment, he used one-class SVMs to classify the handwriting numbers "0." He achieved 91% accuracy in recognition of handwritten "0s," with a low false positive rate of 7%. The method has proved useful in other venues, and other applications of one-class SVMs include document classification (distinguishing one specific category from other categories) (Manevitz and Yousef, 2002), texture segmentation (distinguishing one specific texture from other textures) (Tax and Duin, 2002), and image retrieval (retrieving a subset of images based on the similarity between given query images) (Lai et al., 2002).

We wanted to use these machine learning techniques to approach the Sudden Oak Death problem, and evaluate whether these new methods might be useful in modeling potential risk for the disease, as well as helping solve the one-class problem we have as a result of only having reliable presence data. Thus, the objective of this study is to evaluate the use of SVMs in mapping the potential distribution of Sudden Oak Death in California. Specifically, we will compare the results provided by one-class SVMs with that provided by two-class SVMs used with 'pseudo' absence data, and provide a discussion of the relative merits of the SVM models in comparison to other more traditional statistical models.

2. Method and materials

2.1. Data

The training data used in this study were locations of confirmed *P. ramorum* in California. These data are routinely provided to the monitoring community as part of the management and regulatory function of the California Oak Mortality Task Force (COMTF). We used the distribution of *P. ramorum* samples in 13 California counties as of June 2004 (Fig. 1). A hand-held GPS was used in the field with each sample to collect location information. Accuracy was reported with the sample, and when an offset greater than 100 m was reported, a second visit with a GPS was completed. All spatial location data are stored in a common projection system (Kelly and Tuxen, 2003). Spatial distribution of host species was provided by the California GAP dataset. The California GAP Analysis Project produces maps at relatively low spatial detail (e.g., 1:100,000 map scale, 100 ha MMU) to provide a broad overview of the distribution of biota and their management status, and to identify landscapes that contain large numbers of potentially unprotected vegetation types and vertebrate species (Davis et al., 1998). Vegetation types are identified by one to three overstory species, each contributing greater than 20% of relative canopy cover. These are labeled primary (most abundant), secondary (second-most abundant), and tertiary (third-most abundant) levels. The vegetation map was produced for the state using summer 1990 Landsat Thematic Mapper (TM) satellite imagery, 1990 high altitude color infrared photography (1:58,000 scale), VTM maps based on field surveys conducted between 1928 and 1940, and miscellaneous recent vegetation maps and ground surveys (Davis et al., 1998).

Selection criteria for the host layer were all polygons containing any of the host species listed at http://www.suddenoakdeath.org/ in any of the three levels (primary, secondary, and tertiary). These include Big leaf maple (*Acer macrophyllum*), California bay laurel (*Umbellularia californica*), California black oak (*Quercus kelloggii*), California buckeye (*Aesculus californica*), California coffeeberry (*Rhamnus californica*), California hazelnut (*Corylus cornuta*), Canyon live oak (*Quercus chrysolepis*), Cascara (*Rhamnus purshiana*), Coast live oak (*Q. agrifolia*), Huckleberry (*Vaccinium ovatum*), Madrone (*Arbutus menziesii*),

Rhododendron (*Rhododendron* spp.), Tanoak (*L. densiflorus*) and Toyon (*Heteromeles arbutifolia*). This list does not contain all hosts, as some (e.g. *Viburnum*, California honeysuckle) are not listed as a primary, secondary, or tertiary species in any of the top three co-dominant species lists in the Cal GAP dataset.

Our goal was to predict if trees in presently uninfested areas are potentially likely to be infected by the pathogen. We thus made the assumption that if the host plants share similar conditions (both environmental and anthropogenic conditions) with those in areas with confirmed Sudden Oak Death, they are more likely to be potential targets for *P. ramorum*. We used 14 environmental variables to train the model and predict the potential distribution for the pathogen. The variables included: annual mean temperature, annual mean precipitation, mean temperature in January, April, July, and October, mean precipitation in January, April, July, and October, annual mean solar radiation, distance to main roads, distance to the edge of patches of hosts, and elevation. California climate data were extracted from the DAYMET conterminous United States database (http://www.daymet.org/). DAYMET database was developed by Numerical Terradynamic Simulation Group at the School of Forestry, University of Montana. DAYMET is a model that generates temperature, precipitation, and solar radiation by using digital elevation models and ground-based meteorological data. The spatial resolution of DAYMET database is 1 km by 1 km. All the following analyses were based on 1 km^2 resolution. Multiple confirmed Sudden Oak Death samples that are distributed within a same grid were removed and only one sample retained, yielding 166 sample points used in this study.

Our rationale for the choice of the environmental variables is described as follows. We used the temperature and precipitation variables because we know that *P. ramorum* has temperature and moisture requirements (Rizzo et al., 2002). Because the seasonality of climate affects species distributions (Paruelo and Lauenroth, 1996; Weltzin and McPherson, 2000), we chose mean temperature and precipitation in January, April, July, and October to capture such seasonality. In addition, we know that seasonality of climate may play a significant role in the distribution of this pathogen. For example, *P. ramorum* does not favor dry and hot summer conditions found in inland and southern California (Rizzo et al., 2002). We also know that solar radiation and elevation play a role in disease occurrence (Kelly, 2002; Rizzo et al., 2002). We chose to use distance to roads as a proxy for the anthropogenic dispersal variables. Distance to edges of forest patches has been shown to strongly correspond to the presence of the disease (Kelly, 2002). Because the unit dimensions among those variables vary dramatically, each layer was rescaled to values from −1 to 1 by using minimum and maximum values (Chang and Lin, 2001). All 14 layers were used in the calculation of the SVMs as described below.

2.2. One-class SVMs

Assuming we have l training points x_i ($i = 1, 2, \ldots, l$), we want to find a hypersphere as small as possible to contain the training points in multidimensional space. Meanwhile, we also allow a small portion of outliers to exist using a slack variable (ξ_i):

$$\text{Min } R^2 + \frac{1}{vl}\sum_i \xi_i \tag{1}$$

Subject to:

$$(x_i - c)^{\mathrm{T}}(x_i - c) \le R^2 + \xi_i, \quad \xi_i \ge 0 \ \text{ for all i} \in [l] \tag{2}$$

where c and R are the center and radius of the sphere, T is the transpose of a matrix, and $v \in (0, 1]$ is the trade-off between volume of the sphere and the number of training points rejected. When v is large, the volume of the sphere is small so more training points will be rejected than when v is small, where more training points will be contained within the sphere. v can be roughly explained as the percentage of outliers in the training dataset (Scholkopf et al., 2001). We will describe how to choose the v value in the later section.

This optimization problem can be solved by the Lagrangian:

$$\begin{aligned} L(R, \xi, c, a_i, \beta_i) &= R^2 + \frac{1}{vl}\sum_{i=1}^{l}\zeta_i \\ &\quad - \sum_{i=1}^{l} a_i\{R^2 + \zeta_i - ({x_i}^2 - 2cx_i + c^2)\} - \sum_{i=1}^{l}\beta_i\xi_i \end{aligned} \tag{3}$$

where $a_i \geq 0$ and $\beta_i \geq 0$. Setting the partial derivative of L with respect to R, a_i, and c equal to 0, we get:

$$\sum_{i=1}^{l} a_i = 1 \tag{4}$$

$$0 \leq a_i \leq \frac{1}{vl} \tag{5}$$

$$c = \sum_{i=1}^{l} a_i x_i \tag{6}$$

Substituting Eqs. (4)–(6) to Eq. (3), we have the dual problem:

$$\min_a \sum\nolimits_{i,j} a_i a_j (x_i \cdot x_j) - \sum\nolimits_i a_i (x_i \cdot x_i) \tag{7}$$

Subject to:

$$0 \leq a_i \leq \frac{1}{vl}, \quad \sum_{i=1}^{l} a_i = 1$$

To determine whether or not a test point (x) is within the sphere, we can calculate the distance between the test point and the center C. It can be expressed as:

$$(x \cdot x) - 2\sum_i a_i (x \cdot x_i) + \sum a_i a_j (x_i \cdot x_j) \leq R^2 \tag{8}$$

So far, we have assumed that the data are spherically distributed. In reality, the data are often not spherically distributed. To make the method more flexible to account for this issue and capture the non-linearity such as multi-mode distribution, the kernel function $K(x_i, x_j)$ can be introduced. Basically, we express the inner product in Eq. (8) as the kernel function:

$$K(x, x) - 2\sum_i a_i k(x, x_i) + \sum_{i,j} a_i a_j K(x_i, x_j) \leq R^2 \tag{9}$$

Two types of kernels are often used: polynomial and Gaussian kernels, however, the former usually does not produce a tight description of the data and is sensitive to outliers when the polynomial degree is high (Tax and Duin, 1999a). A more robust way is to construct the Gaussian kernel, which has been commonly used for one-class SVMs (Tax and Duin, 1999a; Scholkopf et al., 2001):

$$K(x_i, x_j) = \mathrm{e}^{-(x_i - x_j)^2 / S^2} \tag{10}$$

where S is the kernel width. The Gaussian kernel was applied in this study. It should be noted that the one-class SVMs method discussed above was proposed by Tax and Duin (1999); Scholkopf et al. (1999) proposed another version to find a hyperplane to separate the training data from the origin with maximum margin. For the Gaussian kernel, these two methods are equivalent (Scholkopf et al., 2001). We implemented the one-class SVMs by the modified version of LIBSVM—a library for support vector machines developed by Chang and Lin (2001). A more detailed mathematical derivation of one-class SVMs can be found in Scholkopf et al. (1999), and Tax and Duin (1999a).

2.3. Two-class SVMs

Consider a set of training points x_i (i = 1, 2, …, l) which are assigned to one of two classes with corresponding labels $y_i = \pm 1$. The goal of the two-class SVMs is to find an optimal separating hyperplane with the maximal margin between the training points for class −1 and class +1. Define a discriminant function:

$$g(x) = (w \cdot x) + b \tag{11}$$

where $w = (w_1, \ldots, w_n)$ is a vector of n elements, n is the dimension of the feature space; b is a scalar. $(w \cdot x)$ represents the inner product between w and x.

The classification rule is:

$$f(x) = \mathrm{Sign}((w \cdot x) + b) \tag{12}$$

$$f(x) > 0 \Rightarrow x \in \text{class } y_i = +1$$

$$f(x) < 0 \Rightarrow x \in \text{class } y_i = -1$$

Hence, the optimization problem can be formulated as:

$$\text{Minimize } \frac{1}{2}||w||^2 \tag{13}$$

Subject to:

$$y_i((w \cdot x_i) + b) \geq 1 \tag{14}$$

The problem can be solved by the Lagrangian:

$$L = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} a_i(y_i((w \cdot x_i) + b) - 1) \quad (15)$$

where a_i: $i = 1, \ldots, l$; $a_i \geq 0$ are the Lagrange multipliers. Taken the derivative with respect to w and b, and set to zero, we get:

$$w = \sum_{i=1}^{l} a_i y_i x_i \quad (16)$$

$$\sum_{i=1}^{l} a_i y_i = 0 \quad (17)$$

Substituting (16) and (17) into (15) gives the dual form of the Lagrangian:

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j (x_i \cdot x_j) \quad (18)$$

Subject to:

$$a_i \geq 0, \quad \sum_{i=1}^{l} a_i y_i = 0 \quad (19)$$

So far, what we have discussed above is suitable for linearly separable cases. We now turn to more general cases:

(1) For linearly non-separable cases (e.g. some classes overlap in the feature space), a slack variable ξ_i ($i = 1,\ldots, l$) is introduced into the constraints to give:

$$y_i((w \cdot x) + b) \geq 1 - \xi_i \quad (20)$$

$$\xi_i \geq 0$$

An additional term is introduced to the cost function by replacing (13) by:

$$\text{Minimize } \frac{1}{2}||w||^2 + C \sum_{i=1}^{l} \xi_i \quad (21)$$

where C is a parameter to measure the amount of penalty for misclassification.

(2) For the non-linear decision boundary, similar to one-class SVMs, kernel functions are introduced.

The optimization problem of Eq. (18) becomes:

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \quad (22)$$

And the decision rule can be expressed as:

$$f(x) = \text{Sign}\left(\sum_{i=1}^{Ns} a_i y_i K(x_i, x) + b \right) \quad (23)$$

where Ns is the number of support vectors.

More detailed mathematical description about two-class SVMs can be found in Hastie et al. (2001) and Webb (2002).

3. Model implementations and evaluation

3.1. Cross-validation method

A five-fold cross-validation method was used to estimate the accuracy of the predicted model. The cross validation was implemented as follows (Hastie et al., 2001). First, the training data were randomly split into five subsets of equal size. Second, each subset was in turn used for accuracy testing and the remaining four subsets for training. Finally, the total accuracy was estimated by averaging the accuracy of each test. It should be noted that the accuracy reported in this study by the cross-validation method represents true-positive rate (= 1 − false-negative rate). Ideally, a good model should produce the results with high accuracy of both true-positive and true-negative rate (=1 − false-positive rate). Due to the lack of true absence data when dealing with presence-only data, we were unable to estimate the true-negative rate. However, we used the predicted area to aid the evaluation of the model performances. The idea is that it is easy to overfit the model, resulting in 100% true-positive rate, and consequently overpredicting the potential distribution. For example, a predicted area covering the whole study area will have 100% true-positive rate, but this result is most likely incorrect. Engler et al. (2004) propose that a good model prediction with presence-only data should predict a potential area as small as possible while still covering a maximum number of the species occurrences.

3.2. Implementing one-class SVMs

For the one-class SVMs, it is important to choose $\upsilon \in (0, 1]$ in Eq. (1). υ is the trade-off between volume of the sphere and the number of training points rejected. υ can be roughly explained as the percentage of outliers in the training dataset (Scholkopf et al., 2001). To decide the suitable υ value for our one-class SVMs model, we plotted a range of υ values against true-positive rate. In addition, the relationship between the υ value and predicted area was also examined to guide the selection of the parameter. As shown in Fig. 2, the true-positive rate increase linearly when υ decreases from 1.0 to 0.1. The true-positive rate levels off at around 92% when υ decreases to 0.04. Similarly, the predicted area increases with the decrease of υ values. Because the risk of Sudden Oak Death is our most important concern, we seek to minimize the false-negative rate (potential Sudden Oak Death that is rejected). We are interested in upper left corner of Fig. 2 where the true-positive rates (=1 − false-negative rate) are high (>90%). In this study, we chose υ to be 0.04 because the true-positive rate does not increase when υ further decreases, while the predicted area will continue to grow with a smaller υ.

3.3. Implementing two-class SVMs

The two-class SVMs require both absence and presence data to train the model. Since there is no absence data for presence-only data, one solution is to generate 'pseudo' absence data. The 'pseudo' absence data were generated by randomly sampling 166 uninfested grid points (equal number of presence data) out of the total host data (n = 101045, 1 km × 1 km grid points). Because the final prediction results will vary with each generation of 'pseudo' absence data, we repeated the sampling procedure 1000 times. The common accuracy measures between the one-class and two-class SVMs are the true-positive rate and prediction areas. We reported the mean and standard deviation of these two measures for direct comparisons between the SVMs models.

We also compared the final prediction results between the one-class SVMs and two-class SVMs. However, unlike the one-class SVMs that produce a unique potential map, the two-class SVMs will result in different potential distribution maps in each generation of 'pseudo' absence data. In order to obtain a single map from the results of two-class SVMs, we applied the majority rule to decide whether an area will be poten-

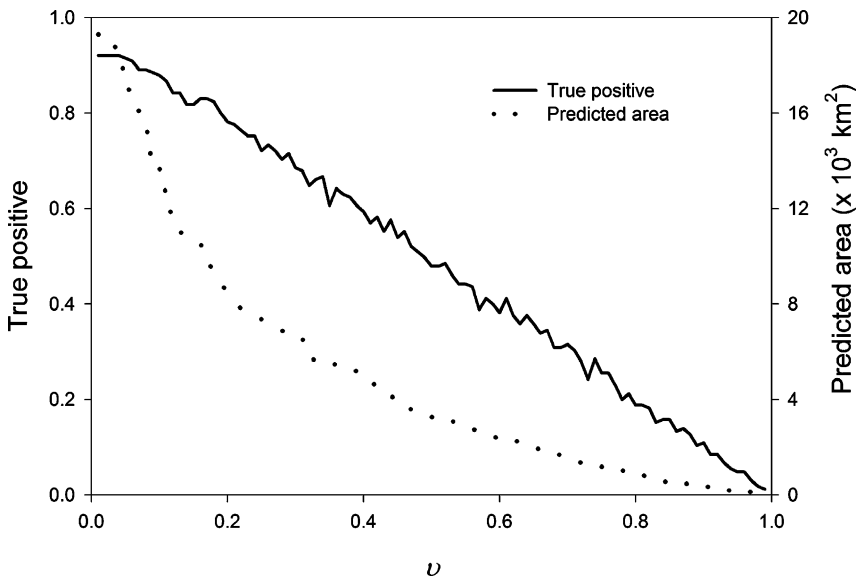

Fig. 2. Relationship between υ and true-positive rate/predicted area.

tially at risk for Sudden Oak Death, and assigned a grid pixel to a class (either presence or absence) reflecting the majority class type in 1000 prediction maps.

4. Results

The average (±standard deviation) true-positive rate of the five-fold cross-validation method for the one-class SVMs method was 0.9272 ± 0.0460; the model predicted the potential Sudden Oak Death over 18,441 km^2. For the two-class SVMs method, the average (±standard deviation) true-positive rate in 1000 simulations is 0.9105 ± 0.0712, and the predicted area at risk for the disease was 13,828 ± 1316 km^2. The average true-positive rate of the one-class SVMs was greater than that of the two-class SVMs. However, the area predicted to be at risk for the disease using the one-class model was also greater than that of the two-class method. The standard deviation reported for the one-class method is derived from the five-fold cross-validation results only, while the standard deviation reported for the two-class method includes the variance from both the five-fold cross validation, and the 1000 simulations of 'pseudo' absence data. After completing the five-fold cross validation, we used all the training data to produce the final risk maps, consequently, for the one-class method, we only have one mapped result predicting risk, but for the two-class method, we are able to report the mean value and standard deviation of the predicted area due to the multiple simulations required.

The results from each method are mapped for the entire state in Fig. 3. The one-class model predicts continued risk for the disease in the coastal areas, as far north as Humboldt County and as far south as Los Angeles County. The Sierra Nevada foothills also show potential risk for the disease (Fig. 3A). The two-class method shows less risk in these areas, and displays a concentration of risk in the central coastal area of the state. Both models agree in the central coastal area of the state, suggesting a suitable environmental niche for *P. ramorum* through Monterey County's redwood/tanoak forests into San Luis Obispo, Ventura and Santa Barbara Counties. These currently uninfested counties share very similar environmental conditions with the northern California counties that are already infested.

There are important regional differences between the models. In the north part of the state (Fig. 4) the one-class method resulted in several distinct and significant areas of risk in Humboldt, Trinity, and Mendocino Counties (Fig. 4A), while the two-class method mirrored that of the one-class method in Mendocino County, with the exception of an increase toward the north of the County, and a decrease in risk prediction toward the coast. In the south portion of the state, the one-class method predicted significantly more risk in San Luis Obispo woodlands and predicted a new pocket of risk in Los Angeles County, but predicted less risk in Santa Barbara County than did the two-class method (Fig. 5). The Sierra Nevada foothills provide the greatest difference between the model results (Fig. 6). The Counties of Nevada, Placer, El Dorado, Amador, Calaveras, Tuolumne, Mariposa, and Madera Counties, all hosting oak woodlands flanking the Sierra Nevada mountain range, show predicted risk for SOD according to the one-class model (Fig. 6A). None of these areas are predicted to be risky by the two-class method (Fig. 6B). Interestingly, the foothills contain oak woodlands with a large component of black oaks—one of the hosts for *P. ramorum*. In 2001, *P. ramorum* was cultured from a sample taken from the Sierra Nevada foothills, but the results have not been repeated to date.

5. Discussion

Sudden Oak Death has the potential to affect biodiversity, fire risk, soil erosion and aesthetic value of oak landscapes in areas similar to the California forests where it has reached epidemic proportions (Kelly, 2002; Rizzo and Garbelotto, 2003). As with other diseases, the mechanisms underlying the dispersal SOD can be pursued from a variety of approaches and spatial scales, and spatial modeling of the dispersal pathways of the causal pathogens remains an important pursuit (Thrall and Burdon, 1999). Indeed, developing a spatially-explicit model of potential SOD spread is an important step in unraveling the nature of the epidemic. A map of pathogen risk is an important backdrop for other research and monitoring efforts. Thus, the maps produced here can be used to target SOD monitoring in high risk, but currently uninfested areas, alert citizen groups in areas likely to host *P. ramorum* to watch

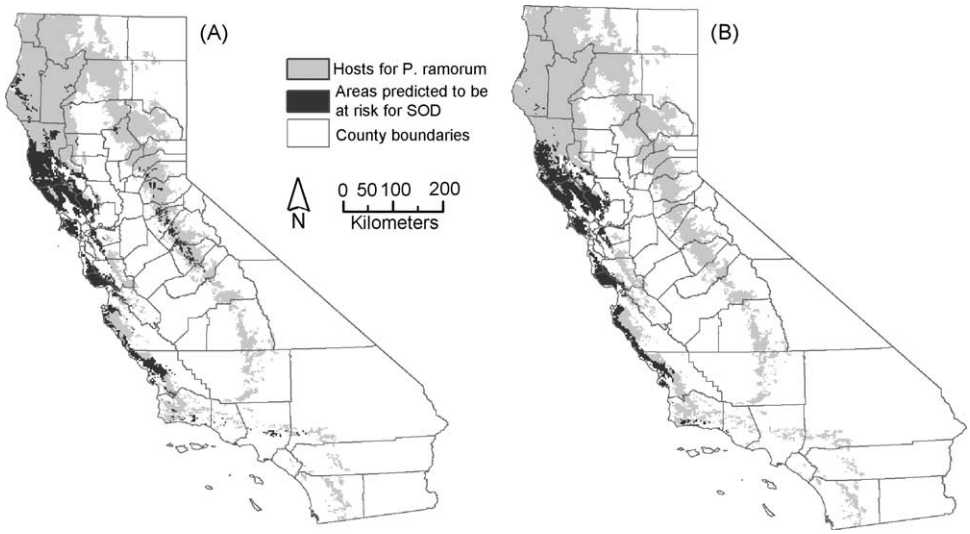

Fig. 3. Predicted area of SOD risk in California. Mapped results from: (A) one-class SVMs; and (B) two-class SVMs.

for symptoms of the disease, and strengthen quarantine regulations in areas of greater risk while lessening the regulation in areas without the necessary environmental factors for harboring the disease.

Many ecological datasets (e.g. those collected by museums or on wildlife surveys) lack reliable absence data. Or if available, the absence data are often unreliable (e.g. from a mobile species) or meaningless (e.g. invasive species) (Hirzel et al., 2002). A variety of methods have been proposed to deal with these presence-only data. The majority of these methods use traditional statistical approaches (e.g. multivariate or logistic regression) by generating 'pseudo' absence data. Others seek to model the presence-only data directly (e.g. BIOCLIM, ENFA, or DOMAIN). In this study, we presented alternative methods to those approaches by using SVMs, and a discussion of their relative merits is useful here.

When compared with traditional statistical or learning models which are based on generations of 'pseudo' absence data in predicting species distributions, two-class SVMs have two main advantages. First, the methods are easy to use. Unlike many other machine learning algorithms, which rely on creativity and extensive tuning of parameters by users, SVMs require a minimum of tuning (Cristianini and Scholkopf, 2002). Second, because SVMs are theoretically-based models, combining optimization, statistics and functional analysis to achieve maximum separation, they have many appealing characteristics: SVMs are free from local mimima, they are computationally efficient, and they provide outstanding performance (Cristianini and Scholkopf, 2002). When compared with methods which model the presence-only data directly, one-class SVMs have two advantages. First, one-class SVMs seek to find an optimal hypersphere which contains all or most of the training points, at the same time tightly constraining the presence data in feature space. By using kernels, one-class SVMs are able to represent various data distribution shapes in feature space (e.g. banana shapes, sphere shapes, or even very irregular shapes; Tax and Duin, 2002). Methods such as BIOCLIM only use hyperboxes to contain the presence data, and are thus often unsuitable for other forms of data that have irregular distributions in feature space. Second, because one-class SVMs seek to find the boundaries of the hypersphere to contain presence data, they make no assumption on the probability density of the data (Tax and Duin, 2002). This characteristic is useful when the data do not follow an underlying probability distribution (such as a normal distribution), or insufficient data are available to test the distribution. For example, ENFA models require normality in the input variables, and violation of this could potentially decrease accuracy in the resulting prediction of species distributions (Engler et al., 2004). We

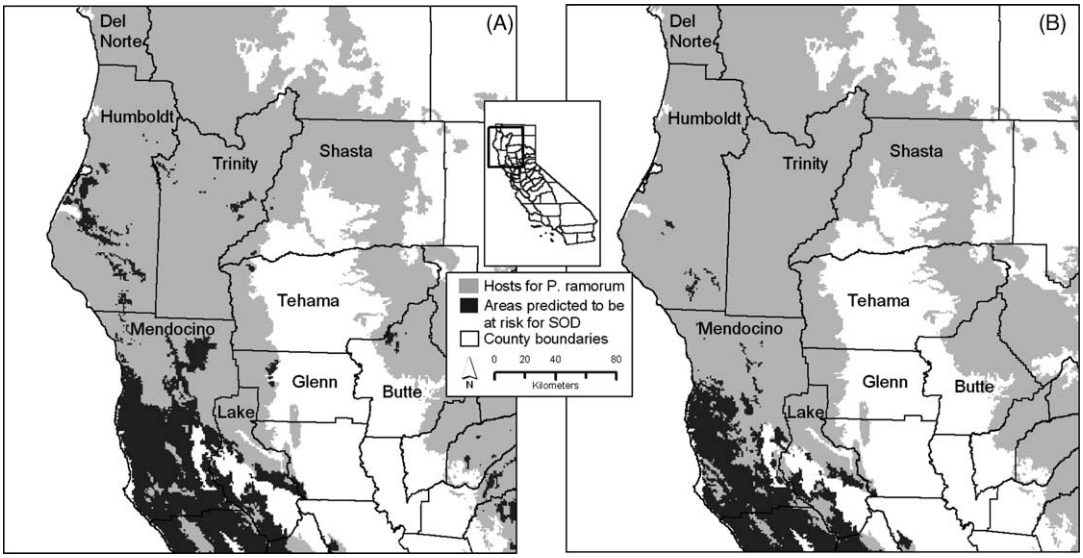

Fig. 4. Predicted area of SOD risk in northern California. Mapped results from: (A) one-class SVMs; and (B) two-class SVMs.

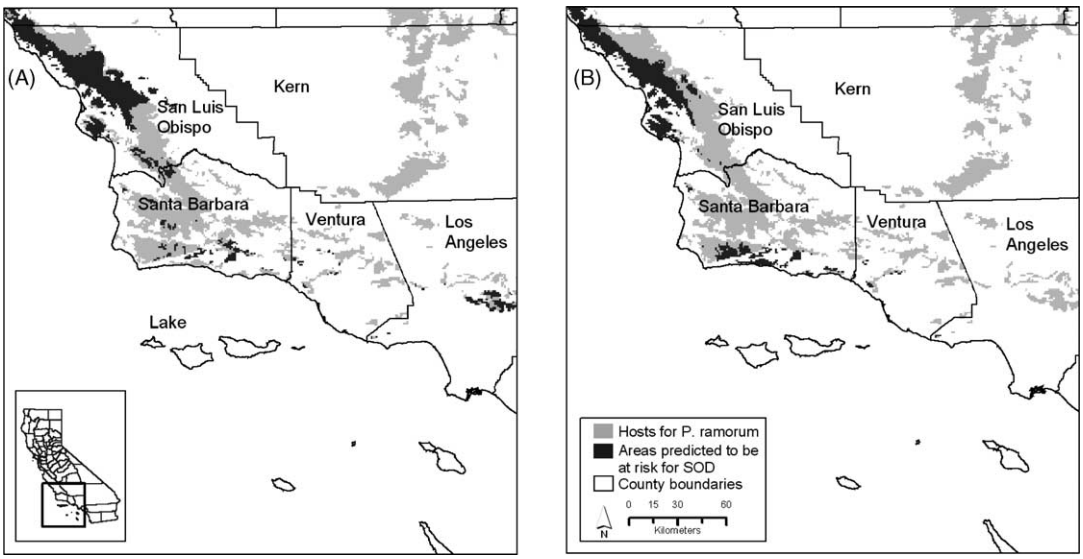

Fig. 5. Predicted area of SOD risk in southern California. Mapped results from: (A) one-class SVMs; and (B) two-class SVMs.

have discussed the relative merits of SVMs versus other model methods, but we would also like to compare the relative strengths of the one- and two-class SVMs themselves. When compared with two-class SVMs, one-class SVMs have several advantages: first, one-class SVMs do not need to generate 'pseudo' absence data. Because each generation of random 'pseudo' absence data will produce different results, a significant number of simulations are needed to give a robust estimation of the potential distribution of negative data. Hence, presence-only models are computationally more efficient. Second, one-class SVMs are particularly suitable for cases where absence data are unreliable (e.g. mobile animals) or meaningless (e.g.

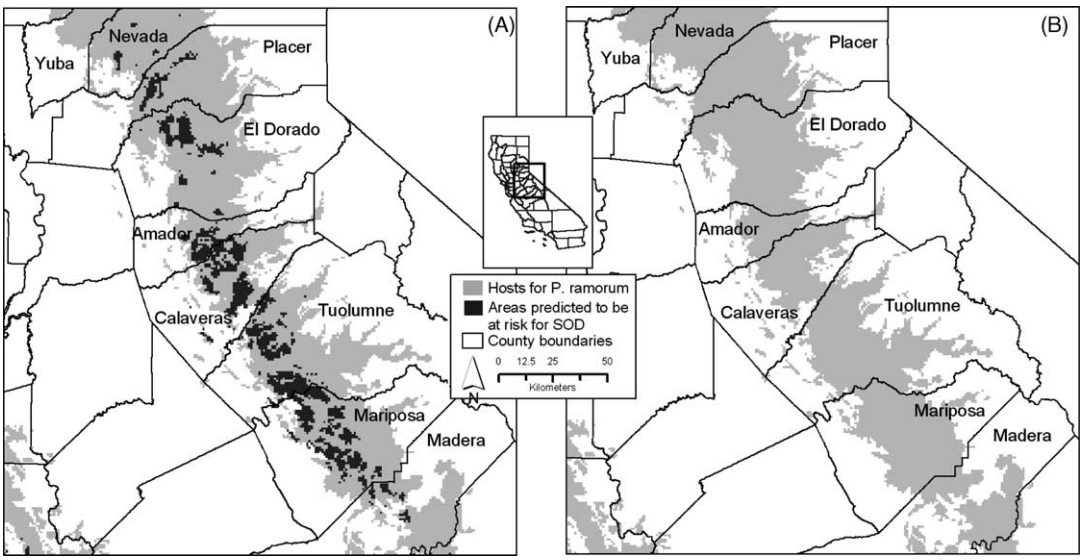

Fig. 6. Predicted area of SOD risk in the Sierra Nevada area. Mapped results from: (A) one-class SVMs; and (B) two-class SVMs.

invasive species), therefore, the generation of pseudo-absence data will inevitably sample areas suitable for the species and could result in a prediction of species distributions that is too restrictive.

The average true-positive rate of the one-class SVMs (0.9272) is better than that of the two-class SVMs used with 'pseudo' absence data (0.9105). However, the predicted area of disease risk provided by the one-class SVMs (18,441 km^2) also is much greater than that by the two-class SVMs (13,828 km^2), which may suggest over-prediction. Indeed, there is a trade-off between the over-prediction provided by an expansive one-class model, and under prediction produced by the necessary selection of 'pseudo' absence data in the two-class model. Engler et al. (2004) and Zaniewski et al. (2002) have found that the lack of absence data in one-class models can result in an over-prediction of species distributions. While it is often true that a higher true-positive rate is associated with higher false-positive rate and result in larger prediction area, in our study, one-class SVMs had a lower false-negative rate (potential SOD that is falsely rejected by the model) than did the two-class method.

Another explanation of the difference between the models is that methods based on generating 'pseudo' absence data will likely sample both 'true absence' and 'potential presence' habitats. The mis-sampling the 'potential presence' data as 'pseudo' absence data in models will necessarily push the decision boundary in the SVM model toward the potential presence habitats and tend to produce a more restrictive prediction. For invasive species that have not yet reached all their potential habitats, 'pseudo' absence models are likely to erroneously sample areas with potential habitats, and result in a more constrained mapped prediction. Hirzel et al. (2001) showed that models using presence-only data have proven to be superior to ones using both presence and absence data in the case of invasive species modeling. Conversely, without absence data, models using presence-only data could include areas where the presence habitats overlap with the true absence habitats as expressed by the environmental features. One way to reduce the over-prediction in one-class SVMs is to control the parameter $v \in (0, 1]$ in equation 1, which can be roughly explained as the percentage of outliers or non-representative samples (e.g. share the similar feature spaces with true absence habitats) in the training dataset. By tuning the v value, we are able to control decision boundary between the presence habitats and true absence habitats. The greater the v value, the more training points will be rejected, and consequently the area of potential distribution contracts.

It should be noted that the models described so far are mainly statistical models, which are used to predict

species distributions based on similarity between test points and confirmed samples in the environmental space. Alternatively, process-based models, which use detailed ecological parameters to simulate the interaction process between species and the environment, have also been successfully applied in predicting potential species distributions (Sutherst and Maywald, 1985; Yonow et al., 2004). Process-based models can be used to test hypothesis, estimate important ecological parameters, and provide insights on ecological process, however, process-based models often do not satisfy the immediate needs of conservationists faced with insufficient information on ecological mechanisms or a lack of detailed parameters needed to run the models (Carpenter et al., 1993), in which case, researchers often rely on statistical models.

It is important to note that mapped potential distribution of the pathogen does not mean that an area will necessarily become infected. Areas of predicted high risk indicate that those areas share similar environmental niches with areas of confirmed Sudden Oak Death, and therefore are possible locales for *P. ramorum* to survive. But management and regulation can impact the spread of the disease. The SOD pathogen can spread through aerial dispersal (Davidson et al., 2001; Davidson et al., 2002), in combination with high levels of inoculum production, and high levels of virulence (Swiecki, 2001). The presence of certain host species such as California Bay has been shown to assist in pathogen dispersal (Kelly and Meentemeyer, 2002; Rizzo and Garbelotto, 2003). Other long and short range anthropogenic vectors might include activities associated with the rhododendron and camellia trade (both ornamental species are hosts for the pathogen), movement of infected plant material, and movement of soil. Worth noting is the fact that of the currently known movement pathways for the spread of the disease, regulatory efforts have a greater impact on the anthropogenic spread mechanisms (e.g. activities associated with trade in ornamental host plants such as rhododendron and other ornamentals that are hosts for *P. ramorum*; Rizzo and Garbelotto, 2003). We mention this for two reasons. First, there might be additional inputs to the models not used here, and second, that regulation can limit spread of the disease into the areas predicted by the model.

The regional differences expressed by the models are interesting. The one-class method produced several entirely new areas of risk (e.g. the scattered patches of risk in Trinity and Mendocino Counties, and extensive risk in Humboldt County, shown in Fig. 4A, the sparse woodland in Los Angeles County shown in Fig. 5A, the Sierra Nevada foothills region shown in Fig. 6A, and small patches of risk in Contra Costa, Alameda, and

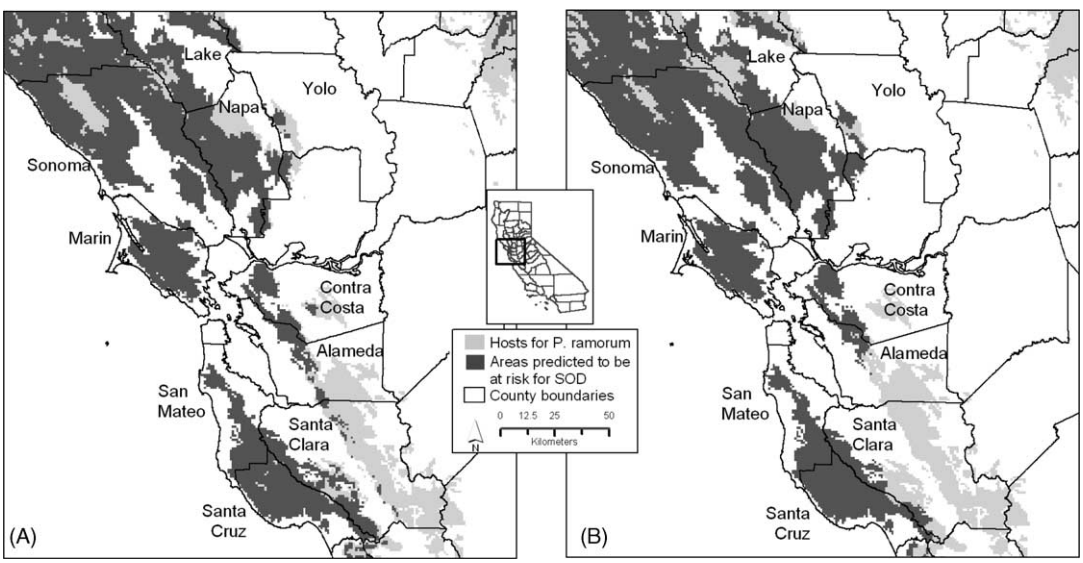

Fig. 7. Predicted area of SOD risk in the San Francisco Bay area. Mapped results from: (A) one-class SVMs; and (B) two-class SVMs.

Santa Clara woodlands shown in Fig. 7A). The two-class method tended to expand areas of core risk (those that were predicted to be at risk by both models). See, for example, areas in Mendocino County (Fig. 4B), Santa Barbara County (Fig. 5B), and small fringe areas in Sonoma, Napa, and Yolo Counties (Fig. 7B). These patterns of difference need further discussion. Controls on this disease spread are complex, and involve interactions between environmental factors and host distribution, which are modeled here, and other factors not modeled here, such as genetic variability and resistance, and human assisted spread of the disease. In this paper we are not attempting to explain all the regional variations in the predicted risk maps; we plan on investigating that later. We can surmise here that the model prediction differences have largely to do with the use of 'pseudo' absence data. Our 'pseudo' absence data was constrained to the area with hosts for the disease, and so 'pseudo' absence data were located throughout the host range. This perhaps caused the two-class model to predict less overall risk in the study area.

6. Conclusions

In this study, we used one-class and two-class SVMs to predict the potential distribution of a new forest disease called Sudden Oak Death in California. Two-class SVMs with 'pseudo' absence data and one-class SVMs with presence data alone were compared. One-class SVMs have a slight better true-positive rate (0.9272) than two-class SVMs (0.9105) when evaluated using a five-fold cross validation. However, the area predicted to be at risk for the disease by the one-class SVMs method is much larger than that by the two-class SVMs method (18,441 and 13,828 km^2, respectively). Both models show that the majority of potential SOD will occur in coastal areas. The two methods agreed in SOD prediction in the San Francisco Bay area with some minor differences; however, the one-class SVMs predict a greater potential risk for the disease in the foothills of the Sierra Nevada mountain ranges and in Los Angeles and Humboldt Counties. Since the risk of Sudden Oak Death is our most important concern, we feel that the one-class method, with its higher true-positive rate and lower false-negative rate, is an appropriate method to use.

We believe that support vector machines, while not used commonly in ecology, are a useful addition to ecological niche modeling. When coupled with geographic information systems, SVMs will be a useful method to deal with presence-only data in ecological analysis over a range of scales. We plan to further investigate the differences between the models in regions, to refine our understanding of the complex interaction between the environmental variables in areas such as the Sierra Nevada foothills, where the two models predicted different results. We also plan to expand this modeling approach to cover the conterminous US, and compare the results of this work with other SOD modeling approaches in the near future.

Acknowledgement

This work was partially supported by grants from USDA-FS, and a NASA New Investigator Program Award to Kelly. The authors would like to thank Professor Chih-Jen Lin for his help on modifying the source codes of LIBSVM and comments on the manuscript.

References

Beard, C.B., Pye, G., Steurer, F.J., Rodriguez, R., Campman, R., Peterson, A.T., Ramsey, J., Wirtz, R.A., Robinson, L.E., 2003. Chagas disease in a domestic transmission cycle in southern Texas, USA. Emerg. Infect. Dis. 9, 103–105.

Bian, L., West, E., 1997. GIS modeling of Elk calving habitat in a prairie environment with statistics. Photogrammetric Eng. Remote Sens. 63, 161–167.

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America 97, 262–267.

Busby, J.R., 1986. A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. Aust. J. Ecol. 11, 1–7.

Carpenter, G., Gillison, A.N., Winter, J., 1993. Domain—a flexible modeling procedure for mapping potential distributions of plants and animals. Biodiversity Conservation 2, 667–680.

Chang, C., Lin, C., 2001. LIBSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Cristianini, N., Scholkopf, B., 2002. Support vector machines and kernel methods—the new generation of learning machines. Ai Mag. 23, 31–41.

Davidson, J., Rizzo, D., Garbelotto, M., 2001. Transmission of *Phytophthora* associated with Sudden Oak Death in California. Phytopathology, S108.

Davidson, J.M., Rizzo, D.M., Garbelotto, M., Tjosvold, S., Slaughter, G.W., 2002. Phytophthora ramorum and sudden oak death in California: II. Transmission and survival. In: Fifth Symposium on Oak Woodlands, USDA-Forest Service, San Diego, CA.

Davis, F.W., Stoms, D.M., Hollander, A.D., Thomas, K.A., Stine, P.A., Odion, D., Borchert, M.I., Thorne, J.H., Gray, M.V., Walker, R.E., Warner, K.J.G., 1998. The California GAP Analysis Project—Final Report. University of California, Santa Barbara, California.

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

Decoste, D., Scholkopf, B., 2002. Training invariant support vector machines. Machine Learn. 46, 161–190.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274.

Fabricius, K., De'ath, G., 2001. Environmental factors associated with the spatial distribution of crustose coralline algae on the Great Barrier Reef. Coral Reefs, 303–309.

Felicisimo, A.M., Frances, E., Fernandez, J.M., Gondalez-Diez, A., Varas, J., 2002. Modeling the potential distribution of forests with a GIS. Photogrammetric Eng. Remote Sens. 68, 455–462.

Fielding, A.H., Haworth, P.F., 1995. Testing the generality of bird-habitat models. Conserv. Biol. 9, 1466–1481.

Fonseca, M.S., Whitfield, P.E., Kelly, N.M., Bell, S.S., 2002. Statistical modeling of seagrass landscape pattern and associated ecological attributes in relation to hydrodynamic gradients. Ecol. Appl. 12, 218–237.

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. Prog. Phys. Geogr. 19, 474–499.

Frescino, T.S., Edwards, T.C., Moisen, G.G., 2001. Modeling spatially explicit forest structural attributes using Generalized Additive Models. J. Veg. Sci. 12, 15–26.

Garbelotto, M., Svihra, P., Rizzo, D., 2001. Sudden oak death syndrome fells three oak species. California Agric., 9–19.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining Inference and Prediction. Springer, New York.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data. Ecology 83, 2027–2036.

Huang, C., Davis, L.S., Townshend, J.R.G., 2002. An assessment of support vector machines for land cover classification. Int. J. Remote Sens. 23, 725–749.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of European Conference on Machine Learning. Springer-Verlag, Berlin, pp. 137–142.

Kelly, M., Tuxen, K., 2003. WebGIS for monitoring "sudden oak death" in coastal California. Comput. Environ. Urban Sys. 27, 527–547.

Kelly, N.M., 2002. Monitoring sudden oak death in California using high-resolution imagery. USDA-Forest Serv., 799–810.

Kelly, N.M., Fonseca, M., Whitfield, P., 2001. Predictive mapping for management and conservation of seagrass beds in North Carolina. Aquat. Conserv. Marine Freshwater Ecosystems 11, 437–451.

Lai, C., Tax, D., Duin, R., Pekalska, E., Paclik, P. (Eds.), 2002. On combining one-class classifiers for image database retrieval. Multiple Classifier Systems. Springer-Verlag, Berlin, pp. 212–221.

Livingston, S.A., Todd, C.S., Krohn, W.B., Owen, R.B., 1990. Habitat models for nesting bald eagles in Maine. J. Widlife Manag. 54, 644–665.

Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J. Appl. Ecol. 36, 734–747.

Manevitz, L.M., Yousef, M., 2002. One-class SVMs for document classification. J. Machine Learn. Res. 2, 139–154.

Mjolsness, E., DeCoste, D., 2001. Machine learning for science: state of the art and future prospects. Science 293, 2051–2055.

Mladenoff, D.J., Sickley, T.A., Haight, R.G., Wydeven, A.P., 1995. A regional landscape analysis and prediction of favorable grey wolf habitat in the northern great lakes region. Conserv. Biol. 9, 279–294.

Moisen, G.G., Frescino, T.S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157, 209–225.

Paruelo, J.M., Lauenroth, W.K., 1996. Relative abundance of plant functional types in grasslands and shrublands of North America. Ecol. Appl. 6, 1212–1224.

Peterson, A., Sanchez-Cordero, V., Ben Beard, C., Ramsey, J.M., 2002. Ecologic niche modeling and potential reservoirs for Chagas disease. Mexico Emerg. Infect. Dis. 8, 662–667.

Peterson, A.T., Vieglais, D.A., 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. Bioscience 51, 363–371.

Rizzo, D., Garbelotto, M., Davidson, J.M., Slaughter, G.W., Koike, S.T., 2002. *Phytophthora ramorum* as the cause of extensive mortality of *Quercus* spp. and *Lithocarpus densiflorus* in California. Plant Dis., 205–213.

Rizzo, D.M., Garbelotto, M., 2003. Sudden oak death: endangering California and Oregon forest ecosystems. Front. Ecol. Environ. 1, 197–204.

Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 1999. Estimation the Support of a High-dimensional Distribution. Technical Report MSR-TR-99-87, Microsoft Research.

Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471.

Spitz, F., Lek, S., 1999. Environmental impact prediction using neural network modelling. An example in wildlife damage. J. Appl. Ecol. 36, 317–326.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inform. Sci. 13, 143–158.

Sutherst, R.W., Maywald, G.F., 1985. A computerised system for matching climates in ecology. Agric. Ecosys. Environ. 13, 281–299.

Swiecki, T.J., 2001. Observations and Comments on Oak and Tanoak Dieback and Mortality in California, Phytosphere, Inc.

Tax, D., Duin, E., 1999a. Support vector domain description. Pattern Recognit. Lett. 20, 1191–1199.

Tax, D., Duin, E., 1999b. Support vector domain description. Pattern Recognit. Lett., 1191–1199.

Tax, D.M.J., Duin, R.P.W., 2002. Uniform object generation for optimizing one-class classifiers. J. Machine Learn. Res. 2, 155–173.

Thrall, P.H., Burdon, J.J., 1999. The spatial scale of pathogen dispersal: consequences for disease dynamics and persistence. Evol. Ecol. Res., 681–701.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Webb, A., 2002. Statistical Pattern Recognition. John Wiley & Sons, New York.

Welk, E., Schubert, K., Hoffmann, M.H., 2002. Present and potential distribution of invasive garlic mustard (*Alliaria petiolata*) in North America. Divers. Distributions 8, 219–233.

Weltzin, J.F., McPherson, G.R., 2000. Implications of precipitation redistribution for shifts in temperate savanna ecotones. Ecology 81, 1902–1913.

Yonow, T., Zalucki, M.P., Sutherst, R.W., Dominiak, B.C., Maywald, G.F., Maelzer, D.A., Kriticos, D.J., 2004. Modelling the population dynamics of the Queensland fruit fly Bactrocera (*Dacus*) tryoni: a cohort-based approach incorporating the effects of weather. Ecol. Model. 173, 9–30.

Zaniewski, A.E., Lehmann, A., Overton, J.M.C., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol. Model. 157, 261–280.