Theor Ecol (2009) 2:189–198
DOI 10.1007/s12080-009-0050-8

ORIGINAL PAPER

Profiling ecosystem vulnerability to invasion by zebra mussels with support vector machines

John M. Drake · Jonathan M. Bossenbroek

Received: 25 July 2008 / Accepted: 23 February 2009 / Published online: 11 June 2009

Abstract Decades since the initial establishment of zebra mussels (*Dreissena polymorpha*) in North America, understanding and controlling the invasion of aquatic ecosystems continues to be a problem in continent-wide conservation and landscape management. While the high economic and conservation burden of this species makes accurate predictions of future invasions a research priority, forecasting is confounded by limited data, tenuous model assumptions, and the stochasticity of the invasion process. Using a new method for niche identification, we profiled invasion vulnerability for 1,017 lakes in the Great Lakes region of the Unites States. We used a nonparametric geoadditive regression model to test for effects of two water quality variables on the present distribution of zebra mussels. We then used the support vector data description (SVDD), a support vector machine for one-class classification, to estimate the boundary of the ecological niche. By disentangling niche estimation from distributional assumptions, computational niche models could be used to test an array of fundamental concepts in ecology and evolution, while species invasions forecasting is representative of the wide range of potential applications for niche identification in conservation and management.

Keywords Invasive species · Niche · Support vector machines · Zebra mussels

J. M. Drake (✉)
Odum School of Ecology, Ecology Building,
University of Georgia,
Athens, GA 30602-2202, USA
e-mail: jdrake@uga.edu

J. M. Bossenbroek
Department of Environmental Sciences and Lake Erie Center,
University of Toledo,
Toledo, OH 43606-3390, USA

Introduction

Colonization by non-indigenous species is a leading environmental issue (Sala et al. 2000; Mooney et al. 2005; Strayer et al. 2006) and an important component of global change (Mooney and Hobbs 2000) and biotic homogenization (Rahel 2002). In North America, freshwater ecosystems support a large endemic fauna and provide valuable ecosystem services and are therefore particularly vulnerable to invasion (Vanderploeg et al. 2002). Thus, forecasting invasions in freshwater ecosystems has been identified as a leading challenge for ecological forecasting (Clark et al. 2001; Jones and Ricciardi 2005) and risk analysis (Bossenbroek et al. 2005).

Disentangling the relationships between species and their environments—species niches—is crucial for forecasting biological invasions (Peterson 2003). Though the niche is a fundamental ecological concept (Chase and Leibold 2003), niche identification has been plagued by conceptual ambiguity and technical obstacles (Pulliam 2000). Further, when identifying niches for invading species, the distribution of observations available for modeling is necessarily non-stationary, as the invading species is progressing across a landscape encountering new and different environments, violating the assumptions of most conventional statistical methods. The result is that niche models are commonly severely biased.

We used a machine-learning approach to overcome these obstacles and profile vulnerability to invasion by zebra mussels (*Dreissena polymorpha*) for 1,017 lakes and reservoirs in the Great Lakes region of North America. The zebra mussel is a nuisance in freshwater ecosystems in North America (Schloesser et al. 1996; Ricciardi et al. 1998) and Europe (Karatayev et al. 1997), dramatically altering ecosystem cycles (Strayer et al. 1999; Vanderploeg

et al. 2002), fouling underwater industrial infrastructure (O'Neill 1997), and threatening the viability of endemic clam populations (Ricciardi et al. 1998). Zebra mussels have been designated by the World Conservation Union Invasive Species Specialist Group as one of the one hundred worst invasive species worldwide (http://www.issg.org/database). Here, we set up a framework for computational niche identification and show how the support vector data description (SVDD), a machine-learning algorithm for one-class classification, can be used for niche modeling. Following this, we deploy the approach to profile ecosystem vulnerability to invasion by zebra mussels across the Great Lakes region of the United States.

Niche theory Hutchinson (1957) defined the ecological niche as the set of environments in which a species can persist. The implied subjunctive conditional—that if species *S* was introduced to environment *E*, and *E* was a member of the niche, then *S* would persist in *E*—has been a sticking point for niche identification, since data on persistence are typically available from only a fraction of possible environments. We represent the relevant features of a species' environment at a location *i* by a vector of observations or measurements corresponding to each of *n* niche axes, $x_i = [y_{i,1}, y_{i,2}, y_{i,3}, \ldots, y_{i,n}] \in \mathcal{E}$, where $\mathcal{E}$ is the set of all (possible) environments (Hutchinson 1957; Chase and Leibold 2003). Through the use of appropriately specified niche axes, any quantitative or qualitative characteristic of the environment is allowed. Thus, a niche axis may take as its domain the real numbers, non-negative integers, or binary indication {0, 1}, and possibly others. The domain of the niche axis is the set, $\mathcal{C}$, of possible niche *conditions y*. Examples are temperature (a continuous quantitative variable), the number of competitive species in the community (a discrete quantitative variable), soil type (a categorical variable), and the presence of a mutualist symbiont (a binary indicator). While not a part of our analysis below, in general other species or combinations of other species might be considered as niche axes. Historically, this has been a point of confusion, i.e., in the theory of competitive exclusion where one conception of the problem uses niche restriction to *detect* competition while another conception holds that competition is *constitutive* of the niche, being part of its definition. We remark that these differences are matters of convention and that our conception of the niche can be modified to accommodate these distinctions and others by partitioning $\mathcal{E}$ into subsets.

We now turn to the criterion of persistence. Roughly, populations persist when average individual fitness, λ, a function of the environment and population size, is greater than one, i.e., $\lambda(x, z)>1$, for some population size z, $z \ll z_{\max}$, where $z_{\max}$ is the maximum size of the population. The niche $\mathcal{N}$ may be therefore defined as $\mathcal{N} \stackrel{\text{def}}{=} \forall x(x \in \mathcal{N} \leftrightarrow \lambda(x) > 1)$. Finally, for any niche axis there will be a subset of *habitable conditions*, $\mathcal{H} \subseteq \mathcal{C}$, the set of all conditions that appear in any environment x belonging to the niche, $\mathcal{H} \stackrel{\text{def}}{=} \forall y(\exists x(y \in x, \lambda(x) > 1))$. In short, the only niche conditions which are not habitable are those that do not belong to *any* niche environment. Importantly, some regions of the niche may not occur in nature. Thus, following Jackson and Overpeck (2000) we define the *realized environmental space*, $\mathcal{R} \subseteq \mathcal{E}$, as the subset of environments realized in nature and subsequently define the *potential niche*, the subset $\mathcal{P} \stackrel{\text{def}}{=} \mathcal{N} \cap \mathcal{R}$. Niche identification, the estimation of the extension of $\mathcal{N}$, from observations drawn from $\mathcal{R}$ is generally a difficult problem.

Computational niche identification The goal of identifying $\mathcal{N}$ (or $\mathcal{P}$) is not original to this study (Grinell 1917), and is a necessary step for numerous applications in ecology, evolutionary biology, and environmental science. We submit that our approach, described below, is optimal in the sense that it eliminates unwarranted theoretical restrictions (particularly relaxing the requirement that the niche include all environments in the product space of habitable conditions; cf. Hutchinson 1957; Stockwell 2007) while retaining the intuition that the niche should be connected. Connected, in this sense, means that any niche environment x is reachable from any other niche environment by a series of operations on the elements of x in which an element is individually incremented or decremented within the local neighborhood of nearest points. While niches need not be continuous (some niche axes may be discrete), it is an empirical conjecture that they are not disconnected in the sense that two niche environments may be separated by an intermediate non-niche environment. Equivalently, niche and non-niche environments may be separated by a hyperplane.

Ideally, niche identification would proceed by estimating $\lambda(x, z)$ directly from observed data on individual growth, survival, and reproduction (cf. Pulliam 2000). Obtaining such data is costly, however, requiring intensive long-term fieldwork. Thus, one typically assumes that much less costly observations of species occurrence are indicators of local persistence and therefore a reliable surrogate for high fitness environments. In terms of our notation, numerous methods for identifying $\mathcal{P}$ (but not generally $\mathcal{N}$) based on a set of occurrence data and associated environmental measurements have been proposed in recent years (Hirzel et al. 2002; Elith et al. 2006; Pearce and Boyce 2006; Phillips et al. 2006). Some of these methods have shown remarkable accuracy when compared with validation data and reported accuracy has improved as increasingly sophisticated methods are introduced (Stockwell and Peterson 2002; Elith et al. 2006). However, all methods with which we are familiar fail to fully avoid one or both of the

following problems. First, data are typically unbalanced with observations of species occurrence vastly outnumbering confirmed locations of species absence, if any absence data are available at all. Further, misclassification errors in the data are also asymmetrical (true occurrences are more likely to be classified incorrectly as absences than vice versa), either due to sampling error, or (as in the case of a species invasion) because the current true distribution of environments inhabited by the species is transient. The severity of this problem will vary among species. Clearly, for organisms that disperse slowly, experience frequent local extinction, or are unable to thoroughly explore pockets of habitat (perhaps due to fragmentation) this problem will be severe. These and other problems pertaining to the balance of observations have come to be known as the problem of "presence-only data" since the limit case comprising only observations of species presence with no observations of species absence is the most commonly available form of data (Hirzel et al. 2002; Brotons et al. 2004; Pearce and Boyce 2006). Second, data are generally not independent. Regardless of how data are obtained (i.e., by computationally sampling from maps of species distributions with GIS, merging records from museum collections, or new field collections), they represent geographically distributed populations in which autocorrelation will be high at some spatial scales (and may not be isotropic) and for which the grain of subsampling that would ensure statistical independence is unknown. Often data exhibit other unknown investigator-induced correlations as well. In our view, these problems—the balance of observations and non-independence—conspire to seriously confound hypothesis tests and estimation. The effects of these problems are exacerbated when sampling in not even across the species distribution.

Because of the need to avoid these problems, niche identification methodology is now an active area of research (Guisan and Thuiller 2005; Moisen et al. 2006). New techniques have been introduced to overcome these problems, but to our knowledge none addresses both. Heuristic methods to circumvent the problem of presence-only data include simulating species non-occurrences (Stockwell and Peters 1999; Engler et al. 2004) or avoiding the classification formulation altogether by estimating the multivariate distribution from which observations are drawn (e.g., Hirzel et al. 2002; Phillips et al. 2006). In the former case, there are well-known theoretical problems and some practical ones (Hirzel et al. 2002; Pearce and Boyce 2006), though the procedure has sometimes proved reliable (Anderson et al. 2003). In the latter case, the estimated distribution does not in general reflect the actual distribution of eventually colonized habitats because it is affected by the distribution of habitat availability, the sampling process, the spatial trajectory of the invading species, the process of resource selection by organisms, and the spatial configuration of different environments (Keating and Cherry 2004; Manly et al. 2004; Lele and Keim 2006). Often, the autocorrelation problem is ignored (for an example from our own research, see Drake and Bossenbroek 2004; cf. Wood and Augustin 2002; He et al. 2003).

We suggest an alternative approach to forecasting species invasions that proceeds in two steps. First, to identify niche axes, use a local probabilistic regression method that controls for effects due to spatial autocorrelation to identify variables that are important predictors of the species *present transient distribution*. We use geoadditive modeling, though for a population that has achieved a stationary spatial distribution autologistic regression might be preferable. Second, use statistically significant variables to estimate *niche boundaries*, i.e., contours of extreme values of the distribution from which the observations are drawn or a decision boundary for classification of niche from non-niche environments. Importantly, however, we advocate using methods that make limited assumptions about the statistical properties of actual observations. Given the complicated conditional relationships among the processes leading to the distribution of observations, a non-probabilistic approach could be particularly useful. We applied this two-part approach to a dataset on zebra mussel presence in 1,202 lakes and reservoirs across the Great Lakes region of the United States.

Materials and methods

Data As for other mollusks, water chemistry is an important determinant of zebra mussel habitat. Specifically, dissolved calcium and pH are critical for larval survival, growth, and shell formation (Vinogradov et al. 1993). An earlier study found that these two variables predicted occurrence in Europe with 92.7% accuracy when data on both occurrence and absence were available (Ramcharan et al. 1992). Accordingly, we envision that each lake is represented by the pair $x_i=[y_1, y_2]$, corresponding to pH and dissolved calcium concentration. Water quality data were obtained from the EPA STORET data bank (http://www.epa.gov/storet/index.html) and the USGS National Water Information System (http://waterdata.usgs.gov). From each data source, we retrieved records of pH and dissolved calcium for all lakes for which they were available in each of the eight states that contain multiple inland lakes or reservoirs infested by zebra mussels (Illinois, Indiana, Michigan, New York, Ohio, Pennsylvania, Vermont, and Wisconsin). Kansas and Oklahoma together have only three reservoirs with zebra mussels and were excluded from the study. Records of pH and calcium were attributed to individual lakes using a geographic

information system (ArcGIS, ESRI, Redlands, California). To diminish the influence of spurious outliers, the top and bottom 5% of parameter values for each lake were trimmed before calculating mean dissolved calcium and pH for 185 invaded lakes and reservoirs and 1,017 non-invaded lakes and reservoirs. Dissolved calcium data were $\log_{10}$-transformed prior to analysis.

Hypothesis tests for niche axis identification We tested for an association between zebra mussel occurrence and our two putative niche axes (dissolved calcium and pH) using geoadditive models (Kammann and Wand 2003). Geo-additive models are generalized additive models (GAMs) in which the two-dimensional effect of space is included as a covariate. For the reasons discussed above, we chose to use geoadditive models to accommodate the highly probable situation in which invasion status of lakes is spatially correlated. For zebra mussels the causes of spatial autocorrelation include proximity to propagule sources, density of habitable ecosystems, stream connections, geospatial variation in the human behaviors resulting in introductions, and the effect of spatially correlated covariates, particularly large-scale features of bedrock and surface geology. Our model is fully specified by the assumption that the binary response variable (invaded or not invaded) was binomially distributed with mean

$$\eta = t(x,y) + s(z_0) + s(z_1) + s(z_2),$$

where $t(x,y)$ is a tensor-product smooth interaction (with thin-plate regression spline basis) describing the spatial effect at longitude x and latitude y, and the $s(z_i)$ are thin-plate regression splines for effects of dissolved calcium, pH, and lake area (Wood 2006). Lake area was included as a measure of lake attractiveness (Reed-Andersen et al. 2000; Bossenbroek et al. 2001), which in turn affects inbound propagule pressure and could therefore be a confounding factor (Leung et al. 2006). Models were fit in the statistical programming language R (R Development Core Team 2007) with the mgcv package (Wood 2001); tuning parameters were selected automatically minimizing the unbiased risk estimator (UBRE), which can be interpreted as an approximation to Akaike's Information Criterion for nonparametric models (Wood 2004).

Niche boundary identification and risk profiling Support vector machines (SVMs) are a class of non-probabilistic statistical pattern recognition algorithms for estimating, among other quantities, the boundary of the set from which a collection of observations is drawn. By design, SVMs are insensitive to large numbers of similar observations, thereby circumventing the autocorrelation and nonstationarity problems. Operationally, SVMs make use of a function (the kernel) to project observations with complex statistical properties in the natural input space (e.g., the niche axes) into a higher dimensional feature space in which they are more simply represented, for instance for a classification problem by a separating hyperplane (Schölkopf and Smola 2002). For niche identification, the goal is to estimate a function that distinguishes $\mathcal{N} \subseteq \mathcal{E}$ from $\neg\mathcal{N} \subseteq \mathcal{E}$ using only observations from $\mathcal{R}$—the observations of species occurrence. The function should return a set of points in the input space that represents a boundary between the two classes, while the assumptions about the data and the target class must be minimal. In statistical pattern recognition, the task of identifying counter examples to a set of training data (i.e., points belonging to $\neg\mathcal{N}$) is referred to by the nearly synonymous terms "novelty detection", "one-class classification", and "concept learning" (Manevitz and Yousef 2001; Tax 2001; Markou and Singh 2003). SVMs for novelty detection satisfy these conditions. Particularly, it is not assumed that data are independent, that data are distributed in proportion to the true distribution of the species in the environment, that the niche is a convex set, or that the observations agree in expectation with the average of the true distribution, (cf. Phillips et al. 2006). We do assume that the niche is not a disconnected subset of $\mathcal{E}$, and optimize the tuning parameter to ensure that the set of environments comprising the estimated niche is simply connected. Finally, we remark that accuracy will improve with the degree to which the distribution of observations are representative of the true potential distribution, in the sense that the range of the distribution has been sampled even if in a biased and unknown fashion. Particularly, performance will be improved when the boundary of the unknown distribution actually sampled, the boundary of $\mathcal{R}$, is coincident with the boundary of $\mathcal{N}$. We believe that this is the minimally restrictive approach to representing the ecological niche as classically defined by Hutchinson (1957). To our knowledge, the similarity between the problem of novelty detection and niche identification has not been remarked previously.

The support vector data description (SVDD) is an SVM for finding the boundary around a set of observations (Tax and Duin 2004). This boundary is the simplest boundary in the sense that it represents the smallest possible hypervolume (a hypersphere) containing a specified fraction of the observations in the projected feature space. The SVDD retains the general property of being insensitive to the distribution of the training data as long as the data are representative of the set of possible observations (Tax 2001). In particular, extreme observations provide a great deal of information about the boundary of the niche which can be exploited by the SVDD, but (if over-representative of their actual occurrence) would give a biased estimate of the distribution of the species in nature. Indeed, by avoiding

estimating the density directly, the SVDD obtains a better estimate on the boundary of the niche than would be obtained from estimating the full density and "backing out" its supporting set from contours of tail probabilities (Tax and Duin 2004). Further, consistent with Vapnik's principle to avoid solving a more general problem as an intermediate solution to a particular objective, the SVDD seeks to extract maximal information about a particular feature of a distribution (its boundary) by avoiding estimating unnecessary features (e.g., central tendency, dispersion, and skew; see discussion in Tax (2001), pp 67-ff.). For niche identification, this should result in maximally accurate estimates of niche boundaries.

Model training and performance estimation Prior to model estimation, occurrence data were randomly assigned to model training (80%) or model testing (20%) subsets. Using only the training dataset for model estimation, the SVDD was obtained by solving the quadratic programming problem

$$\min\left\{\sum_{i,j}\alpha_i\alpha_j K(x_i x_j) : \sum_i \alpha_i = 1, 0 \le \alpha_i \le C\right\} \qquad (1)$$

where $K(x_i,x_j)$ is the kernel describing the similarity between vector-valued object x_i and x_j, the α's are the Lagrange multipliers, K is the Gaussian radial basis function (RBF) $K(x_i, x_j) = \exp\left\{-\|x_i - x_j\|^2 / s^2\right\}$ with tuning parameter s, and C is a parameter that trades off the volume of the hypersphere against the errors (Tax 2001). Our rationale for choosing the Gaussian RBF kernel was twofold: (a) its theoretical underpinnings suggest this kernel can be understood as a generalization of, and compromise between, the rigid hypersphere and Parzen density estimators (Tax and Duin 2004), which alternately may be considered as the extreme cases of the boundary estimation approach adopted in this paper (the hypersphere) and the density estimation approach commonly adopted (Parzen density estimator), and (b) our experience is that it is the most numerically stable of the standard kernels.

For model estimation, we only used pH and dissolved calcium. We suspect that large lakes are invaded first, because of high rates of human visitation, and that the significant effect of lake area obtained in the geoadditive model is at least accentuated during the early phases of invasion and probably temporary altogether. The parameter C was determined implicitly by defining a tolerable error rate on the target distribution $\varepsilon \in \{0.1, 0.05, 0.025\}$. Because of the very high economic and environmental costs of zebra mussel invasion (Leung et al. 2002), we submit that ε=2.5% (representing one in forty chance of misclassifying a niche ecosystem as non-niche) is the maximum tolerable error rate in practice and focus our subsequent discussion on these analyses. The selection of s was automated to obtain the simplest description of the target set $\mathcal{N}$ subject to the specified error ε. For detailed formulas for the estimated niche boundary and the associated test procedure we refer the reader to section 2.1 of Tax and Duin (2004).

The false negative error rate was estimated by executing the trained algorithm on the 20% of occurrences in the testing dataset. Estimating the false positive rate is trickier. Since the invasion is ongoing, susceptible lakes not yet invaded and properly identified as susceptible would score as false positives, the estimated false positive rate obtained by the a posteriori classification of lakes known not to be invaded could be severely biased. Therefore, we adopted an alternative suggested by Tax and Duin (2002) and estimated the false positive rate as the fraction of the sphere with radius equal to the radius of the target data occupied by the trained classifier. We recognize that the distribution of non-niche ecosystems in nature may not reflect the uniform distribution generated using this procedure and that some generated outliers will belong to the true target distribution (so that the error rate estimated on the simulated outliers is overestimated). Tax and Duin (2002) suggest that this method will be most prone to failure when observations from the "target class" (i.e., niche environments, $\mathcal{N}$) are "scattered over the complete feature space" (i.e., the space of possible environments, $\mathcal{E}$). Equivalently, one can readily appreciate that the distribution of outliers generated this way will be most accurate when (a) the true boundary of set $\mathcal{N}$ is a highly restricted subset of $\mathcal{E}$ (so that few outliers are generated within the niche space), and (b) the set of realized environments $\mathcal{R}$ is evenly distributed over the domain $\mathcal{E}$. Recognizing these rather restrictive limitations, in this paper, we primarily use the estimated false positive rate to compute the receiver–operator curve, to facilitate evaluating our procedure and comparing with other models. Given the positive bias in the estimated false positive rate, our estimate of the summary area-under-curve statistic (AUC) should be conservative.

All data processing and estimation was performed in MATLAB version 7.2 (Mathworks, Inc., Natick, Massachusetts). Solution of the SVDD program in Eq. 1 was performed using the Data Description Toolbox (version 1.5.4, [Tax 2006]) and Pattern Recognition Toolbox (version 4.0, [Duin et al. 2004]).

Results

Invasion extent Zebra mussels were found in 185 of 1,202 lakes in our sample. Of these, five are in Illinois, two are in Indiana, 119 are in Michigan, 19 are in New York, 11 are in Ohio, five are in Vermont, and 25 are in Wisconsin. In a few instances, larger lakes are represented by more than one observation because parts of the lake belong to separate

Table 1 Output for the first geoadditive model with all effects represented as smooth terms

Effect	Term	p	est. d.f.
Spatial distribution	$t(x\,y)$	2.0e−11	9.7
Dissolved calcium	$s(z_0)$	1.9e−5	3.8
pH	$s(z_1)$	0.035	1.002
Lake area	$s(z_2)$	<2e−16	3.6

USGS polygons. Thus, as an extreme example, Lake Champlain is included in our data set seven times. Average measurements of the trimmed data series for dissolved calcium ranged from 0.83 mg L^{-1} to 586.43 mg L^{-1}, while dissolved calcium in infested lakes was between 6.2 mg L^{-1} and 436.39 mg L^{-1} (median: 31.6). Average measurements of pH ranged from 3.9 to 10.3, while pH in infested lakes was between 6.9 and 9.0 (median: 7.8).

Niche axes All effects in the geoadditive model were significant ($p<0.0001$; $R^2=0.37$; Table 1). However, the estimated degrees of freedom (e.d.f.) for the effect of pH was indistinguishable from one, so a second model was fit in which pH was included as a linear covariate. All effects retained their significance in this second model (Table 2). The effect of dissolved calcium increased noticeably from low to intermediate values, at which it levels off at a threshold around 1.5×10^1 mg L^{-1} (Fig. 1). Interestingly, the estimated coefficient for the effect of pH is negative, i.e., conditioned on dissolved calcium, pH has an inhibitory effect on mussel establishment that is constant across the observed range. Exponentiation of the fit coefficient obtains an estimate of the effect of increasing pH by one unit on the odds of having been invaded by zebra mussel, i.e., $e^{-0.66}\approx0.5$ indicates that the odds are reduced by approximately half.

Invasion risk profiles Out of 1,017 presently uninvaded lakes, 645 fall within the estimated zebra mussel niche at the most restrictive false negative error tolerance of $\varepsilon=0.025$ (Fig. 2). Relatively few of these are excluded when the error tolerance is increased to $\varepsilon=0.08$ (Fig. 3). The zebra mussel niche appears to be a relatively restricted subset of the environmental space occupied by lakes in general. Vulnerable lakes were identified in all states for which analysis was performed, but are concentrated in the Midwest, particularly in Michigan's Lower Peninsula

Table 2 Output for the final geoadditive model with linear effect of pH; all other effects represented as smooth terms

Effect	Term	p	est. d.f.
Spatial distribution	$t(x\,y)$	2.0e−11	9.7
Dissolved calcium	$s(z_0)$	1.9e−5	3.8
pH (linear fit coefficient: −0.66)	β_0	0.035	
Lake area	$s(z_2)$	<2e−16	3.6

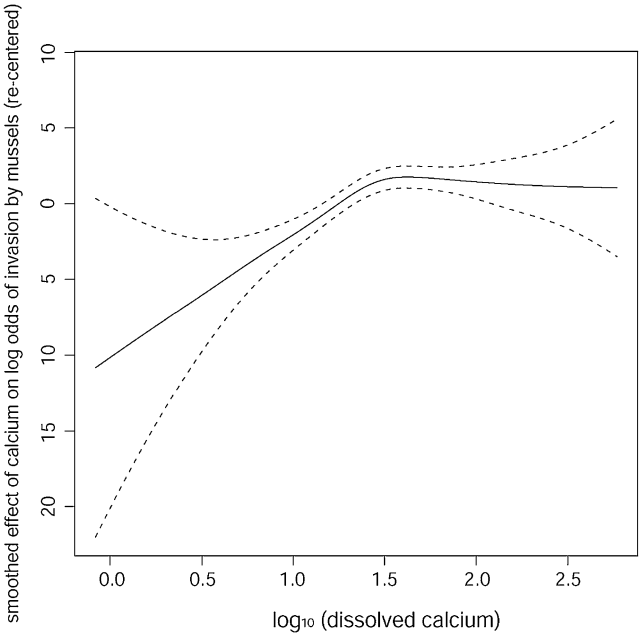

Fig. 1 The partial relationship between infestation by zebra mussels and the concentration of dissolved calcium (after $\log_{10}$ transformation) increases to a point and then levels off in a geoadditive model. Plot shows mean effect with 95% confidence intervals

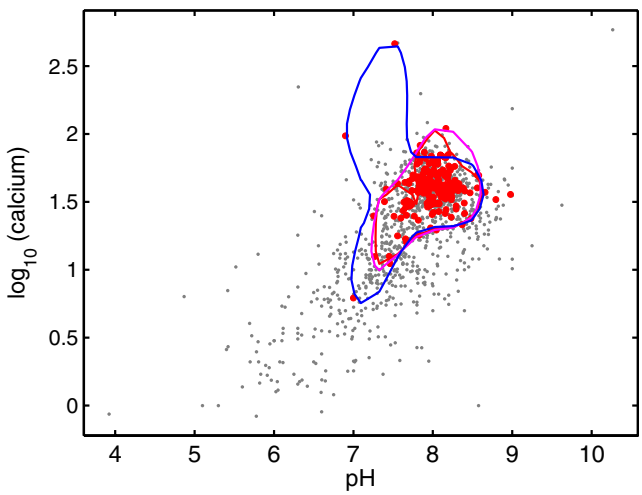

Fig. 2 The support vector data description estimated niche for zebra mussel (*Dreissena polymorpha*) in North America. Circles (*red*) represent observations of zebra mussel occurrence and dots (*grey*) represent lakes for which zebra mussel has not been reported. The model was trained only on observations of occurrence. The *blue line* is a non-probabilistic estimate of the niche boundary tuned to contain 97.5% ($\varepsilon=0.025$) of occurrence observations, the *magenta line* contains 95.0% ($\varepsilon=0.05$) of occurrence observations, and the *red line* contains 90.0% ($\varepsilon=0.08$) of occurrence observations

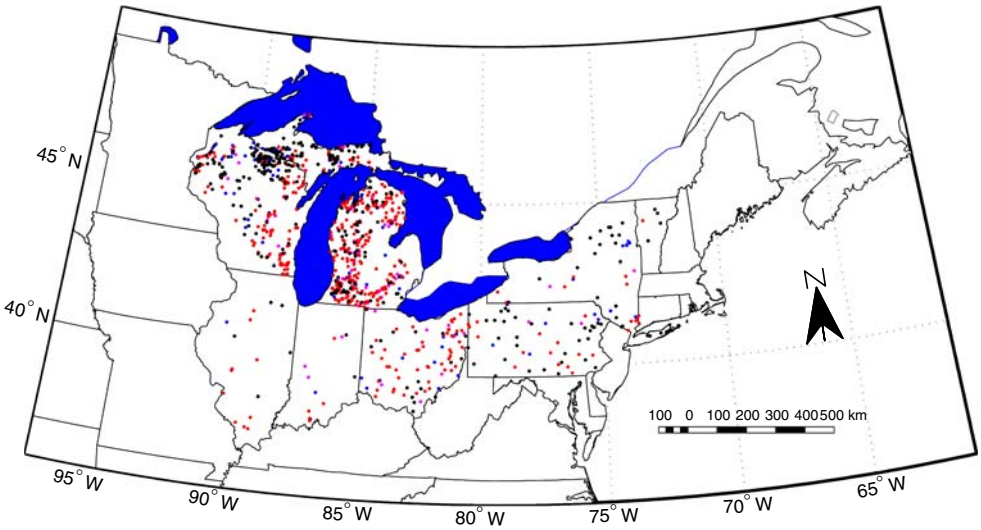

Fig. 3 Map of lakes predicted to be susceptible to zebra mussels. Susceptible (*colored*) or unsuitable (*black*) lakes are based on the ecological niche identified with a support vector machine. Lakes colored *red* occupy the most interior region of the niche (ε=0.08), followed by lakes colored *magenta* (ε=0.05) and lakes colored *blue* (ε=0.025). A table listing the individual lakes is available as Appendix S1

(Fig. 3). Individual lakes and their vulnerability status are listed in Appendix S1.

The single tuning parameter (s) was selected to obtain the simplest model consistent with the target error rate in cross-validation. It is our experience that model results are generally remarkably insensitive to the choice of this parameter. Figure 4 shows the cross-validation error rate as a function of the tuning parameter. Ideally, this figure would show a very steep shoulder with a sharp turn in the vicinity of the target error rate. In this case, a somewhat simpler model might have been chosen by tolerating an error rate twice as large, but the actual estimated niche boundary is not markedly complicated (cf. Fig. 2). That is, there are no irregular corrugations in the estimated boundary. Our confidence is further supported by inspecting the receiver–operator curve, which shows the variation in the true positive rate as ε is varied over its possible range. The estimated AUC was 0.88, while the receiver–operator curve is quite flat in the vicinity of the target error rate ε=0.025 (Fig. 5).

Discussion

Using data from publicly accessible databases, we profiled the risk of zebra mussel invasion for lakes and reservoirs of the Great Lakes region of the United States. Our method circumvents most problems associated with presence-only

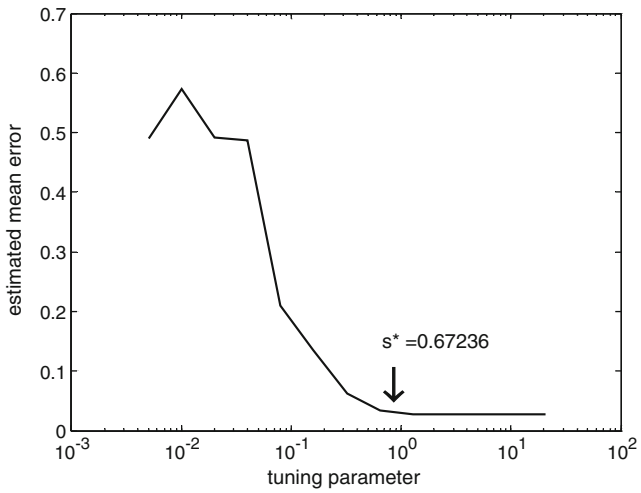

Fig. 4 Model complexity is determined by tuning the width parameter (s) to obtain the simplest model that reliably achieves an average error rate less than or equal to the target value $\varepsilon \in \{0.1, 0.05, 0.025\}$ in cross-validation. The *arrow* points to the value used for drawing the boundary in Fig. 2 and predicting lakes shown in Fig. 3 at the target error rate of ε=2.5%

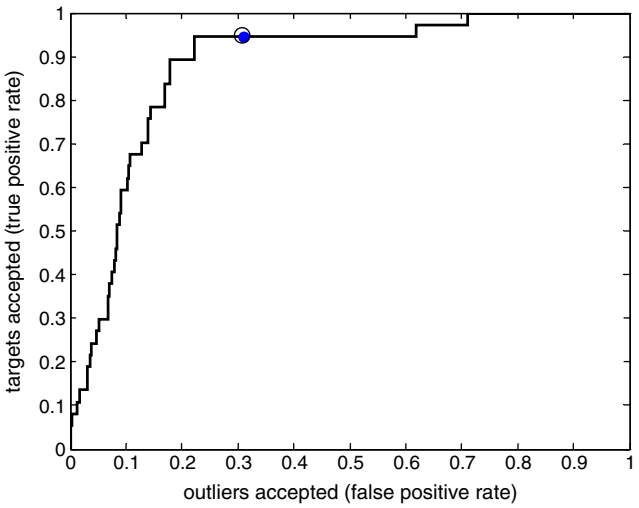

Fig. 5 The receiver–operator curve showing the rate at which truly susceptible lakes are correctly identified (true positive rate) as a function of tolerance for falsely classifying unsuitable lakes as susceptible (false positive rate). The *circle* shows the point on the curve at the target error (false negative) rate of 2.5%

data and unknown correlations, including spatial autocorrelation. Our results confirm that dissolved calcium and pH within observed ranges are determinants of zebra mussel habitat. Further, we have identified individual ecosystems at risk for zebra mussel invasion (S1). Our analysis demonstrates that the ecological niche, conceptualized as a subspace in the set of possible environments, can be estimated using support vector machines, and that computational approaches to data mining large, heterogeneous ecological datasets can be marshaled for ecological forecasting. Long-term ecological data are commonly beset by unknown dependencies, unbalanced observations, and non-random sampling. Our study shows that non-probabilistic machine-learning approaches to data analysis can avoid some of these obstacles.

We confirmed that two hypothesized niche axes (pH and dissolved calcium) significantly affected invasion status of lakes in the United States, independent of possible confounding influences of spatial correlation and lake area. Interestingly, pH was found to have an unexpected inhibitory effect, though compared with other factors (dissolved calcium, lake area, and spatial location) the evidence for an effect of pH is weak ($p=0.035$). A speculative explanation is that this effect results from reactions of the different carbon species as pH varies. As pH increases, the majority of dissolved carbon changes from free CO_2 to bicarbonate (HCO_3^-) to carbonate (CO_3^{2-}), affecting relative proportion of each (Wetzel 2001). Consequently, when pH is high, i.e., >8.5, calcium is primarily bound by carbonate and therefore unavailable for uptake by mussels. Alternatively, eutrophy is often correlated with pH (Jeppesen et al. 1990), and hypereutrophic lakes are known to be poor habitat for zebra mussels (Strayer 1991). However, given the relatively weak evidence for the association between pH and occurrence, future research should confirm that this pattern is not spurious. After identifying pH and dissolved calcium as important niche axes, we estimated niche boundaries using the SVDD to identify a large number of lakes in the Great Lakes region as vulnerable to invasion by zebra mussels. The accuracy of the estimated model was very good by disciplinary standards. Particularly, the observed area-under-curve statistic (0.88) was higher than in any comparable study of which we are aware (e.g., Elith et al. 2006); cf. discussion [Drake et al. 2006]).

Other studies have previously taken a landscape approach to quantifying zebra mussel invasion risk (Strayer 1991; Drake and Bossenbroek 2004; Whittier et al. 2008). The picture emerging from these analyses is that landscape scale management may halt or slow the westward spread of this nuisance species. Accordingly, multi-state projects like the 100th Meridian Initiative (http://100thmeridian.org/) have been developed to intercept zebra mussels and respond rapidly when infested lakes are detected. However, within the continent-wide potential distribution of zebra mussels there clearly will be unsuitable ecosystems, pockets of non-habitat, and other lakes that are highly vulnerable. Bioeconomic models have shown that prevention and control of zebra mussels at the ecosystem scale would be expedient (Leung et al. 2002; Bossenbroek et al. 2009), but until now a simple, reliable model for profiling lake vulnerability based on data from North America has not been available. Further, our approach is not limited to zebra mussels but could be used for forecasting invasion of any species in which habitat availability is at least a partial determinant of spread.

Our approach accommodates the presence-only and spatial autocorrelation problems that have often afflicted niche estimation. Our approach does not solve the problem, presented by coupled source-sink habitat patches, in which organisms are found in environments that do not belong to the niche (Gomulkiewicz et al. 1999; Pulliam 2000). In our case, the data are unlikely to be contaminated by such false positive observations because zebra mussels are rarely observed before a self-sustaining local population is established. Though this fortunate situation will not generally occur, we remark that since source habitats and sink habitats must be spatially coupled, a signature of the niche boundary will almost certainly remain in the correlations among environments in which species are observed. This will not be a simple relationship, as the source-sink dynamic further erodes the structure of an already problematic dataset. We regard this as an open problem for statistical learning.

To conclude, we remark that the SVDD approach to niche estimation could be applied to other conservation problems and in basic research. As an element of population viability analysis for threatened and endangered species, niche modeling could be used for habitat assessment (Elith and Burgman 2003). Similarly, risk assessment for release or escape of genetically modified organisms should identify the eventual range such species might occupy (Wolfenbarger and Phifer 2000). But, at best, only experimental data (not geographic distribution) will be available to such models. The non-probabilistic nature of the support vector machine admits such contrived data. Further, combined with models for regional and global climate change, our approach to niche modeling could be used to predict future species distributions, particularly since statistical downscaling of course-grained models could be embedded within the statistical learning process. Finally, combined with resampling approaches to hypothesis testing (e.g., Manly 1991), these methods could be used to shed light on such longstanding questions as the relative importance of environment versus species interactions and community drift in determining species distributions and abundance (Hubbell 2001), adaptation (Gomulkiewicz et al. 1999), evolution of species ranges (Kirkpatrick and Barton 1997), niche conservatism (Holt

1996), and habitat effects on population dynamics in heterogeneous landscapes (Engen et al. 2002). The problems of presence-only and autocorrelation data have made these questions difficult to resolve.

Acknowledgments This research was conducted while JMD was a postdoctoral fellow at the National Center for Ecological Analysis and Synthesis (Santa Barbara, California, USA). Additional support came from the University of Toledo, Department of Environmental Sciences and Lake Erie Center (to JMB) and the University of Georgia, Odum School of Ecology (to JMD). We thank D. Strayer and T. Peterson for comments on an earlier version of this paper and A. Silletti for comments and assistance preparing the manuscript.

This is publication No. 2009-07 from the University of Toledo Lake Erie Center.

Appendix S1

The file "appendix.csv" contains locations of 1,017 lakes screened for vulnerability to invasion by zebra mussels. Overall, 645 out of 1,107 (63.4%) lakes were vulnerable at the ε=0.025 level, 592 out of 1,107 (58.2%) lakes were vulnerable at the ε=0.05 level, and 514 out of 1,107 (50.5%) lakes were vulnerable at the ε=0.08 level. Columns are longitude, latitude, state name, county name, lake name, vulnerable (1/0) at the ε=0.025 level, vulnerable (1/0) at the ε=0.05 level, and vulnerable (1/0) at the ε=0.08 level.

References

Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol Model 162:211–232

Bossenbroek JM, Kraft CE, Nekola JC (2001) Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes. Ecol Appl 11:1778–1788

Bossenbroek JM, McNulty J, Keller RP (2005) Can ecologists heat up the debate on invasive species risk? Risk Anal 25:1595–1597

Bossenbroek JM, Finnoff DC, Shogren JF, Warziniack TW (2009) Advances in ecological and economical analysis of invasive species: dreissenid mussels as a case study. In: Keller RP, Lodge DM, Lewis MA, Shogren JF (eds) Bioeconomics of invasive species: integrating ecology, economics, policy, and management. Oxford University Press, pp 244–265

Brotons L, Thuiller W, Araujo MB, Hirzel AH (2004) Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27:437–448

Chase JM, Leibold MA (2003) Ecological niches: linking classical and contemporary approaches. Chicago University Press, Chicago

Clark JS, Carpenter SR, Barber M et al (2001) Ecological forecasts: an emerging imperative. Science 293:657–660

Drake JM, Bossenbroek JM (2004) The potential distribution of zebra mussels in the United States. BioScience 54:931–941

Drake JM, Guisan A, Randin C (2006) Modelling ecological niches with support vector machines. J Appl Ecol 43:424–432

Duin RPW, Juszczak P, Paclik P et al (2004) Prtools4, a Matlab toolbox for pattern recognition. Delft University of Technology, Delft

Elith J, Burgman MA (2003) Habitat models for PVA. In: Brigham CA, Schwartz MW (eds) Population viability in plants: conservation, management, and modeling of rare plants. Springer, Heidelberg, pp 203–238

Elith J, Graham CH, Anderson RP et al (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151

Engen S, Lande R, Saether BE (2002) Migration and spatiotemporal variation in population dynamics in a heterogeneous environment. Ecology 83:570–579

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

Gomulkiewicz R, Holt RD, Barfield M (1999) the effects of density dependence and immigration on local adaptation and niche evolution in a black-hole sink environment. Theor Popul Biol 55:283–296

Grinell J (1917) The niche-relationships of the California thrasher. Auk 34:427–433

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993–1009

He F, Zhou J, Zhu H (2003) Autologistic regression model for the distribution of vegetation. J Agric Biol Envir S8:205–222

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

Holt RD (1996) Adaptive evolution in source-sink environments: direct and indirect effects of density-dependence on niche evolution. Oikos 75:182–192

Hubbell SP (2001) The unified theory of biodiversity and biogeogrpahy. Princeton University Press, Princeton

Hutchinson GE (1957) Population studies—animal ecology and demography—concluding remarks. Cold Spring Harb Sym 22:415–427

Jackson ST, Overpeck JT (2000) Responses of plant populations and communities to environmental changes of the late Quaternary. Paleobiology 26:194–220

Jeppesen E, Sondergaard M, Sortkjoer O et al (1990) Interactions between phytoplankton, zooplankton, and fish in a shallow, hypertrophic lake—a study of phytoplankton collapses in Lake Sobygard, Denmark. Hydrobiologia 191:149–164

Jones LA, Ricciardi A (2005) Influence of physiochemical factors on the distribution and biomass of invasive mussels *(Dreissena polymorpha* and *Dreissena bugensis*) in the St. Lawrence River. Can J Fish Aquat Sci 62:1953–1962

Kammann EE, Wand MP (2003) Geoadditive models. J Roy Stat Soc C-App 52:1–18

Karatayev AY, Burlakova LE, Padilla DK (1997) The effects of *Dreissena polymorpha* (Pallas) invasion on aquatic communities in eastern Europe. J Shellfish Res 16:187–203

Keating KA, Cherry S (2004) Use and interpretation of logistic regression in habitat selection studies. J Wildlife Manage 68:774–789

Kirkpatrick M, Barton NH (1997) Evolution of a species' range. Am Nat 150:1–23

Lele S, Keim JL (2006) Weighted distributions and estimation of resource selection probability functions. Ecology 87:3021–3028

Leung B, Lodge DM, Finnoff D et al (2002) An ounce of prevention or a pound of cure: bioeconomic risk analysis of invasive species. Proc R Soc B 269:2407–2413

Leung B, Bossenbroek JM, Lodge DM (2006) Boats, pathways, and aquatic biological invasions: estimating dispersal potential with gravity models. Bio Invasions 8:241–254

Manevitz LM, Yousef M (2001) One-class SVM's for document classification. J Mach Learn Res 2:139–154

Manly BFJ (1991) Randomization and Monte Carlo methods in biology. Chapman and Hall, London

Manly BF, McDonald LL, Thomas DL, McDonald TL, Erikson WP (2004) Resource selection by animals: statistical design and analysis for field studies. Kluwer, Boston

Markou M, Singh S (2003) Novelty detection: a review—part 1: statistical approaches. Signal Process 83:2481–2497

Moisen GG, Edwards TC Jr, Osborne PE (2006) Further advances in predicting species distributions. Ecol Model 199:129–131

Mooney HA, Hobbs RJ (2000) Invasive species in a changing world. Island, Washington DC

Mooney HA, Mack RN, McNeely JA et al (2005) Invasive alien species: a new synthesis. Island, Washington DC

O'Neill CR (1997) Economic impact of zebra mussels—results of the 1995 National Zebra Mussel Information Clearinghouse study. Great Lakes Res Rev 3:35–42

Pearce JL, Boyce MS (2006) Modelling distribution and abundance with presence-only data. J Appl Ecol 43:405–412

Peterson AT (2003) Predicting the geography of species' invasions via ecological niche modeling. Q Rev Biol 78:419–433

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259

Pulliam HR (2000) On the relationship between niche and distribution. Ecol Lett 3:349–361

R Development Core Team (2007) R: a language and environment for sta tistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Rahel F (2002) Homogenization of freshwater faunas. Ann Rev Ecol Syst 33:291–315

Ramcharan CW, Padilla DK, Dodson SI (1992) Models to predict potential occurrence and density of the zebra mussle, *Dreissena polymorpha*. Can J Fish Aquat Sci 49:2611–2620

Reed-Andersen TE, Bennett BS, Jorgensen G et al (2000) Distribution of recreational boating across lakes: do landscapes variables affect recreational use? Freshwater Biol 43:439–448

Ricciardi A, Neves RJ, Rasmussen JB (1998) Impending extinctions of North American freshwater mussels (Unionidae) following the zebra mussel (*Dreissena polymorpha*) invasion. J Anim Ecol 67:613–619

Sala OE, Chapin FS, Armesto JJ et al (2000) Biodiversity—global biodiversity scenarios for the year 2100. Science 287:1770–1774

Schloesser D, Nalepa T, Mackie GL (1996) Zebra mussel infestation of unionid bivalves (Unionidae) in North America. Am Zool 36:300–310

Schölkopf B, Smola A (2002) Learning with kernels: support vector machines, regularization, optimization and beyond. MIT Press, Cambridge

Stockwell D (2007) Niche modeling: predictions from statistical distributions. Chapman and Hall, Boca Raton

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

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148:1–13

Strayer DL (1991) Projected distribution of the zebra mussel, *Dreissena polymorpha*, in North America. Can J Fish Aquat Sci 48:1389–1395

Strayer DL, Caraco NF, Cole JJ, Findlay S, Pace ML (1999) Transformation of freshwater ecosystems by bivalves—a case study of zebra mussels in the Hudson River. BioScience 49:19–27

Strayer DL, Eviner VT, Jeschke JM, Pace ML (2006) Understanding the long-term effects of species invasions. TREE 21:645–651

Tax DMJ (2001) One-class classification; concept-learning in the absence of counter-examples. Dissertation, Delft University of Technology. Available online: http://www.ict.et.tudelft.nl/~davidt/papers/thesis.pdf

Tax DMJ (2006) DDtools, the data description toolbox for Matlab (v 1.5.4). Delft University of Technology, Delft

Tax DMJ, Duin RPW (2002) Uniform object generation for optimizing one-class classifiers. J Mach Learn Res 2:155–173

Tax DMJ, Duin RPW (2004) Support vector data description. Mach Learn 54:45–66

Vanderploeg HA, Nalepa TF, Jude DJ et al (2002) Dispersal and merging ecological impacts of Ponto-Caspian species in the Laurentian Great Lakes. Can J Fish Aquat Sci 59:1209–1228

Vinogradov GA, Smirnova NF, Sokolov VA, Bruznitsky AA (1993) Influence of chemical composition of the water on the mollusk *Dreissena polymorpha*. In: Nalepa TF, Schloesser DW (eds) Zebra mussels: biology impacts, and control. Lewis, Boca Raton, pp 283–293

Wetzel RG (2001) Limnology: lake and river ecosystems. Academic, San Diego

Whittier TR, Ringold PL, Herlihy AT, Pierson SM (2008) A calcium-based invasion risk assessment for zebra and quagga mussels (*Dreissena* spp.). Front Ecol Environ 6:180–184

Wolfenbarger LL, Phifer PR (2000) Biotechnology and ecology—the ecological risks and benefits of genetically engineered plants. Science 290:2088–2093

Wood SN (2001) mgcv: GAMs and generalized ridge regression for R R News 1:20-25

Wood SN (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. J Am Stat Assoc 99:673–686

Wood SN (2006) Generalized additive models: an introduction with R. Chapman and Hall, Boca Raton

Wood SN, Augustin NH (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol Model 157:157–177