5-2008

Comparison of Machine Learning Algorithms for Modeling Species Distributions: Application to Stream Invertebrates from Western USA Reference Sites

Margi Dubal
Utah State University

Comparison of machine learning algorithms for modeling species distributions: application to stream invertebrates from western USA reference sites

by

Margi Dubal

A report submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

UTAH STATE UNVIVERSITY
Logan, Utah

2008

Comparison of machine learning algorithms for modeling species distributions: application to stream invertebrates from western USA reference sites

Abstract. Machine learning algorithms are increasingly being used by ecologists to model and predict the distributions of individual species and entire assemblages of sites. Accurate prediction of distribution of species is an important factor in any modeling. We compared prediction accuracy of four machine learning algorithms—random forests, classification trees, support vector machines, and gradient boosting machines to a traditional method, linear discriminant models (LDM), on a large set of stream invertebrate data collected at 728 reference sites in the western United States. Classifications were constructed for individual species and for assemblages of sites clustered *a priori* by similarity on biological characteristics. Predictive accuracy of the classifications was evaluated by computing the percent of sites correctly classified, sensitivity, specificity, kappa, and the area under the receiver operating characteristic curve on 10-fold crossvalidated predictions from each classification method on each individual species and assemblage of sites. The predictions from each type of classification were used to estimate the Observed over Expected (O/E) index of taxa richness. Random Forests generally produced the most accurate individual species models. However, none of the machine learning algorithms showed significant improvement over LDMs for classifications of assemblages of sites and precision of the O/E index. The performance of Support Vector Machines was particularly poor for classifying individual species and assemblages of sites, and resulted in greater bias in the O/E index. We believe that the performance of models developed for species at such large spatial scales may depend more on the predictor variables available than the classification technique.

Key words: Classification, machine learning, O/E index, species distribution modeling.

Introduction:

Predicting where individual taxa should occur under reference conditions is a critical part of biological assessments and conservation management. These predictions are typically derived from 'niche' models, which describe how taxa abundances or probabilities of presence vary with different environmental conditions. Most niche models have been based on analyses using traditional statistical methods, including logistic regression and linear discriminant models (LDMs). These traditional methods make stochastic and structural assumptions (e.g., linearity) that are not generally satisfied by ecological data and, consequently, they may not provide meaningful analyses in the situation of complex and non-linear relationships between the classes and the predictor variables. Accurate prediction is an important objective in species distribution modeling. Ecologists have recently started to evaluate models based on machine learning methods, which make almost no stochastic and structural assumptions, and they have the capacity to predict complex and highly non-linear systems (see, for example, De'ath 2007; Cutler et al. 2007). The machine learning algorithms used in this study are classification trees (hereafter CT; Breiman et al. 1984; De'ath and Fabricius 2000), random forests (hereafter RF; Breiman 2001; Cutler et al. 2007), support vector machines (hereafter SVM; Hastie et al. 2001; Drake et al. 2006), and gradient boosting machines (hereafter GBM; Hastie et al. 2001; De'ath 2007). In other applications, all these methods have been shown to have very high classification accuracy.

The Observed over Expected (O/E) index of taxa richness (see, for example, Hawkins 2001) is an important tool in assessing the biotic condition of streams. Predictions of taxa presences at selected sites are made using classification models fit to data from reference 'pristine' sites and then compared to the numbers of taxa actually observed at the site through the O/E index. The classification models that have typically been used in this kind of assessment are linear discriminant models (LDM; Hastie et al. 2001; Hawkins 2001).

The purpose of the study reported here was to evaluate the four machine learning algorithms CT, RF, SVM, and GBM, for the purposes of individual species distribution modeling, for classification of assemblages of sites, and for estimation of the O/E index. For comparison purposes, LDMs were used as a benchmark. It was anticipated that the machine learning algorithms would significantly out-perform LDM for the individual species distribution modeling and assemblage classifications, and would result in more accurate estimates of the O/E index.

Data:

the data used in the analyses was obtained from 728 reference sites in the western United States (Fig. 1). The biological data is the presence or absence of 375 stream invertebrates at each of the 728 sites. A GIS was used to generate associated environmental predictor variables for each site, including drainage area, topography, watershed geology, soils, and long-term climate variables (PRISM 2004). There were 11 continuous predictors and seven categorical variables. Variable descriptions are given in Table 1.

Table 1: Names and descriptions of predictor variables used in analyses.

Variables	Description	Type	Range
ELEV	Elevation of site	Continuous	10 - 3660
log_WSAREA	log of watershed area	Continuous	-0.33 - 4.05
GIS_LAT	Latitude	Continuous	31.63 - 48.87
GIS_LONG	Longitude	Continuous	-124.32 - 103.41
FRZ_FREE	The average annual number of days with mean air temperature above 0°C	Continuous	14 - 318
TMEAN_PT	30-year average annual air emperature at sampling site.	Continuous	-15 - 210
LOG_PPT	Log Precipitation	Continuous	2.15 - 3.57
HYDR_PT	Ratio of mean of the minimum of mean monthly flows on record (baseflow) to the mean of the maximum of mean monthly flows interpolated from USGS gauging stations: value for the sampling site.	Continuous	0 - 0.3105
GNEIS	% of gneiss geology in the watershed derived from a simplified version of Reed & Bush (2001) - Generalized Geologic Map of the Conterminous United States.	Categorical	0/1

Table 1(cont.): Names and descriptions of predictor variables used in analyses.

Variables	Description	Type	Range
GRANTIC	% of granite geology in the watershed derived from a simplified version of Reed & Bush (2001) - Generalized Geologic Map of the Conterminous United States.	Categorical	0/1
MAF_ULT	% of mafic-ultramafic geology in the watershed derived from a simplified version of Reed & Bush (2001) - Generalized Geologic Map of the Conterminous United States.	Categorical	0/1
QUART	% of quarternary geology in the watershed derived from a simplified version of Reed & Bush (2001) - Generalized Geologic Map of the Conterminous United States.	Categorical	0/1
SEDIMENT	% of sedimentary geology in the watershed derived from a simplified version of Reed & Bush (2001) - Generalized Geologic Map of the Conterminous United States.	Categorical	0/1
VOLCANIC	% of volcanic geology in the watershed derived from a simplified version of Reed & Bush (2001) - Generalized Geologic Map of the Conterminous United States.	Categorical	0/1
CARB_PT	Presence (1) / absence (0) of carbonate geology at the sampling site derived from map of merged carbonate rocks derived from state geologic maps.	Categorical	0/1
SLOPE	Slope (rise/run) of the stream channel from the National Hydrologic Dataset (NHDPlus, http://www.horizon-systems.com/nhdplus).	Continuous	0 - 0.2875
srSLOPE	Square root of SLOPE	Continuous	0 - 0.5361903
srHYDROPT	Square root of HYDR_PT	Continuous	0 - 0.5572253

Classification Methods:

There were two parts to the classification analyses. One part involved predictive classification of presences and absences of individual taxa. The other analyses involved the classification of sites that had been assembled into biologically similar groups.

Individual Species Distribution Modeling:

Presences and absences of individual species were predicted using the set of 11 continuous environmental and seven categorical variables.. Only the 111 taxa that occurred at more than 30 sites were included in these analyses. All five classification methods were used for all 111 taxa.

Assemblage of Sites Classifications:

The composite species modeling involved two distinct stages. In the first stage, the 728 reference sites were clustered into biologically similar groups based on the presence and absence of species. A similarity matrix was created using Bray-Curtis Index and then clustering algorithm was applied. The dendogram that graphically displays the degree of biotic similarity between sites and groups of sites was then used to identify similar groups of sites (Hawkins – www.cnr.usu.edu/wmc). Thus, clustering algorithm is used to group biologically similar sites into quasi-distinct classes that represent different 'types' of sites. Sites within each class are simply more similar to each other than sites from different class. In the analyses reported here two different groupings of reference sites were used. One grouping contained 28 classes (groups) of similar sites and another set contained coarser set of 11 classes (groups).

In the second stage, likelihood of group membership for each site was generated as function of environmental predictor variables using all selected classification algorithms and the set of 11 continuous and seven categorical predictor. Predicted occurrence probability of each taxon at each site was calculated by multiplying occurrence frequency of all taxa in the reference site groups and probability of group membership generated from each classification methods. Thus, a 728 x 375 matrix was obtained containing predicted occurrence probabilities of each taxon at each site.

Classification Accuracy Assessment:

The metrics used to evaluate the accuracy of the predictive classifications were the percentage of sites correctly classified (PCC), the sensitivity, the specificity, the kappa statistic, and the area under the receiver operating characteristic curve (AUC). All these metrics have their advantages and disadvantages (Fielding and Bell 1997); together they characterize overall classification accuracy. The AUC criterion is particularly widely used in ecology because it is independent of the probability threshold used to classify sites into different groups.

For both the individual species distribution modeling and the classification of assemblages of sites, 10-fold crossvalidation was employed to ensure that the models did not "overfit" the data and inflate the classification accuracies (Kohari 1995). In 10-fold crossvalidation the dataset is randomly divided into 10 equal—or nearly equal—sized pieces, which may be indexed by $i = 1, 2, 3, \ldots , 10$. The crossvalidated predictions for the i^{th} piece are obtained by fitting or "training" the classifier on the data in the remaining nine pieces and then predicting for all the sites in the i^{th} piece.

Review of Classification Methods:

Linear Discriminant Models:

Linear discriminant models are one of the oldest methods for classification. The decision boundaries between the different classes or groups of observations are linear combinations of the predictor variables. Prior probabilities of membership in the different classes may be specified. The general form of the linear discriminant function for the k^{th} class is:

$$\delta_k (x) = x^{\mathrm{T}} \sum^{-1} \mu_k - 1/2\, \mu_k^{\mathrm{T}} \sum^{-1} \mu_k + \log \pi_k \quad ,$$

where k is number of classes, π_k is prior probability of membership in the k^{th} class, $\sum$ is estimated covariance matrix for the predictors, and x is a vector of values on the predictor variables for the observation in question. The observation is classified as belonging to the class for which the value of δ_k is largest. That is, Predicted class = $\mathrm{G}(x) = \mathrm{argmax}_k\, \delta_k (x)$. A more detailed description of LDM may be found in Hastie et al. (2001).

Although they are simple and very easy to compute, LDMs have several disadvantages. The optimality of the LDM is derived assuming that the predictor variables jointly have multivariate normal distributions with a common covariance matrix for all classes. This assumption is rarely satisfied in practice. The linear form of the separators of the different classes also limits the types of problems for which LDMs are effective. Despite these shortcomings, LDMs have proved to be useful classifiers in a wide range of problems and are still widely used in ecology. The `lda` function from the MASS package in R was used for all analyses reported here.

Classification Trees:

A classification tree is built using a process of binary recursive partitioning, which splits the observations into increasingly homogeneous groups with respect to response classes. The criterion that is usually used to assess homogeneity of the subgroups of data is the Gini index (Breiman et al. 1984). At each step, an optimization with regard to the Gini index is carried out to determine the variable and cutpoint to split on. The most effective way to fit a classification tree is to fully grow the tree until no futher decrease in the Gini index is possible, and then prune the tree back by removing the lower branches to optimize crossvalidated prediction error More technical detail about classification trees may be found in Breiman et al. (1984) and Hastie et al. (2001). The classification trees of our analyses were fit using the rpart package in R. . The amount of pruning of classification trees in rpart is controlled by the complexity parameter. The value of the complexity parameter was selected by inspecting a plot of the crossvalidated error rate against value of the complexity parameter (Breiman et al. 1984).

Random Forests:

As the name suggests, RF combines the predictions from many classification trees to obtain more accurate classifications. Many (e.g., 500) samples of the same size as the original data set are draw from the dataset *with replacement.* These samples are called *bootstrap samples.* In each bootstrap sample approximately 68% of the observations in the original dataset occur one or more times. The observations in the original dataset that do not occur in the bootstrap sample are said to be *out-of-bag* for that bootstrap sample. On each bootstrap sample, a classification tree is fit. At each step in the fitting process (split) only a small number of variables (typically, the square root of the number of observations) is available to be split on. The tree is fully grown, with no pruning. The tree is then used to construct predictions for all the out-of-bag observations for that bootstrap sample. Finally, the predicted class for an observation is obtained by "voting" the predicted classes for all the trees fit on bootstrap samples for which the observation was out-of-bag. More technical detail about random forests may be found in appendix A of Cutler et al. (2007).

Random forests classifications for the analyses reported here were fit using the randomForest in R (Liaw and Wiener 2002). Although RFs are much more computationally intensive to fit than single classification trees, they may also give substantially more accurate predictions (Breiman 2001; Cutler et al. 2007).

Gradient Boosting Machine:

Gradient boosting machines is another procedure that, like RF, fits many trees to a single dataset. Gradient boosting machines differ from RF in that the trees are fit sequentially, with observation weights updated according to whether observations are correctly or incorrectly classified. The algorithm for updating and using the weights is quite complex and may be found in Friedman (2000), Hastie et al. (2001), and De'ath (2007). The last of these papers also contains some ecological examples of the use of GBMs.

The GBMs for the analysis reported here were fit using the gbm package in R.

Gradient boosting machines are very computationally intensive and require substantial tuning of parameters. However, in many applications they have proved to be the most accurate classifier that is currently available.

Support Vector Machines:

Support Vector Machine (SVM) leads to a different approach for classification other than trees. The basic idea behind support vector machine is to create non linear boundaries by generating linear boundaries on higher dimensional space. It is a computationally extensive algorithm but it works well in many situations. SVMs are stable, require less tuning and have greater prediction accuracy in ecological modeling (Drake et. al 2006). More technical details about SVM can be found in Hastie et. Al 2001. We used e1071 package in R to build SVM model.

The Observed over Expected Ratio

A second objective of this study was to see whether improved values of the observed over expected (O/E) ratio could be obtained using machine learning algorithms to obtain the expected component of the statistic instead of LDMs. We used the output (such as predicted occurrence probability of each taxon at each site) generated from each type of model to estimate the O/E value and the precision of the estimated O/E values. It is known that the O/E value can vary due to many factors. One of these factors is the threshold for the probability of presence that is used to screen taxa for inclusion in the total expected value (Yuan 2006). In the analyses reported here, two detection thresholds were used: zero and 0.5. When the threshold is set at zero, all taxa are included in the computation of the E component of the O/E statistic at a given site. With a threshold of 0.5, only those taxa that have a predicted occurrence probability greater than 0.5 are included in the calculation of E for that site. A third choice is to calculate adjusted threshold for individual taxa and use it to make decision about inclusion of that taxa in calculation of E. The optimal value was estimated by minimizing the difference between sensitivity and specificity. This assumes that predicting presences correctly is as important as predicting absences correctly. Optimization was done by checking each possible value between 0 and 1.

Different measures of model accuracy were plotted against how frequently taxa occurred among samples using SYSTAT (version 11). Taxa were only introduced in modeling, if they were observed in greater than 30 sites. Plots have been created for each metrics such as PCC, Sensitivity, Specificity, Kappa and AUC for two different thresholds (cutoff) to capture probability. Thus, only those species whose estimated capture probabilities were at least as large as the threshold were indicated as present otherwise they were counted as absent. The random selection of taxa was made such that entire range of species frequency (rare as well as more common species) was represented in any given plot. For each 0.1 interval, we randomly selected 5 taxa to plot. Only two taxa had frequencies of occurrence greater than 0.7. So the totals of 37 taxa were plotted in the graph. Trend lines were fitted by LOESS regression with tension equal to 0.5.

Results:

The first set of analyses was for the assemblages of sites. Crossvalidated and resubstitution (see Fielding and Bell 1997) estimates of classification accuracies were obtained using RF, SVM, LDM, and CT (Table 2). Gradient boosting machines were not included in these analyses because the implementations available to us only worked for binary responses. The most striking aspect of Table 2 is how poorly classification trees performed compared to all the other classifiers. The crossvalidated PCC for classification trees for the 28 class problem was about 7.1% compared to values between 30% and 38% for the other classifiers. For the 11 class analysis, the crossvalidated PCC was 17.7% for classification trees compared to values between 50% and 55% for the other classifiers. We are unable to explain why classification trees perform so poorly.

For the 28-class analyses, the PCC for RF (37.7%) was slightly higher than for SVM (31.1%) and LDM (34.3%). For the 11-class analyses, the PCCs for RF (53.5%) and SVM (53.8%) are nearly identical and slightly higher than the PCC for (LDM). Overall, there is little to choose between SVM, LDM, and RF for the classification of assemblages of sites.

Also of interest was the effects the different classifications have on the estimation of the O/E index. Because all the sites used in these analyses are reference sites the mean value of O/E should be close to 1.0 and the smaller the standard deviation of the O/E values the better. Given the poor classification performance of classification trees it is perhaps surprising that the mean O/E values using cross-validated predicted probabilities are close to 1.0 for all four classification methods (Table 3 and Figure 12(a)). The largest mean O/E value is 1.053 for RF and the smallest is 0.970 for SVM. Using pairwise *t*-tests (Tables 4 and 5) we see that the modest difference in mean O/E values for the different classification methods are all statistically significant except for the difference between CT and LDM, which have the mean crossvalidated O/E values closest to 1.0. The standard deviations of the O/E values for the different classification methods are all about 0.2 with a slightly higher value of 0.249 for SVM. The higher standard deviation and bias for SVM suggest that it is the least useful classification procedure from the perspective of estimating the O/E index.

For individual taxa modeling, most methods performed similarly with respect to the accuracy metrics, except for SVM, which performed poorly or erratically (Figures. 2, 4, 6, 8, and 10).In many cases, accuracy metrics varied substantially with frequency of occurrence. Rare species were observed to have higher accuracy measure compared to more common species for all five classification methods. There was no significant difference in classification accuracy for the entire range of species frequency in terms of PCC for all four methods excluding support vector machine (SVM) (Figure 2). Adjusting the threshold to minimize the difference between sensitivity and specificity substantially reduced the effect of commonness on perceived prediction accuracy. SVM performed poorly for the threshold value (used to classify presence or absence of the species) of 0.5 (Figure 2) and for the adjusted cutoff (Figure 3) but showed better prediction for common species than rare species.

PCC obtained by using species specific adjusted cutoff had relatively high value for random forests. There was not much difference between Classification Tree, Linear Discriminant Analysis and Gradient Boosting Machine for the value of PCC and also it had substantially low value for Support Vector Machine for the entire range of frequency. Specificity defined as percentage of correctly classified absences, was similar for all methods except SVM; however, they differed substantially for different frequency of taxa when threshold value was set to 0.5. Rare species had higher specificity prediction compared to more common species (Figure 4). SVM showed straight line across the graph showing no difference in specificity for entire range of species indicating no dependency on species range size. Again for adjusted threshold, Random Forests had relatively high value across the species distribution giving better specificity prediction for rare and common species (Figure 5). SVM again proved to be a poor method while other methods were similar in estimating specificity.

Sensitivity showed quite opposite of what we observed in specificity for rare and common species. Rare species tend to estimate lower sensitivity compare to common species. For the threshold value of 0.5, as frequency of occurrence for species increased, sensitivity also increased (Figure 6). SVM performed substantially different from other methods and resulted in moderate value for rare species. This gradually decreased for common species. Plots obtained for sensitivity and specificity looked similar for adjusted probability of detection threshold showing Random Forests as one of the best classifier among other methods (Figure 5 & 7). Ideal adjusted threshold values were calculated such that difference between sensitivity and specificity were minimized and this caused similarity between the two plots.

Kappa showed slight improvement for Random Forests compare to other methods where as SVM gave the lowest prediction accuracy (Figure 8 & 9). Kappa also gave identical plots for both threshold values which proved that we can use kappa as accuracy measure independent of the value for probability of detection threshold. Kappa varies between 0.1 and 0.5 for entire set of species distribution with few negative values which showed overall poor prediction for all methods.

Like Kappa, AUC is also independent of the choice of the threshold value. In terms of AUC, Random Forests was slightly more accurate than LDA, CT and GBM where as SVM had the lowest AUC values (Figure 10) for the entire range of distribution. Rare taxa (frequency of occurrence less than 0.2) tend to give higher classification accuracy in terms of AUC which decreased when the frequency of occurrence was approximately 0.2, increased slightly as frequency increased and stabilized for remaining frequencies (Figure 10). Thus, species that are rare were modeled with higher accuracy than the other common species.

We observed that Random Forests was slightly superior to other methods in classification accuracy. Also, SVM had the lowest performance scores across all five measures of model accuracy. However, we did not notice huge difference for Linear Discriminant Analysis, Classification Tree and Gradient Boosting Machine. Hence, it was worth comparing model precision using O/E index.

Three threshold values were used to see the performance of model precision. We found that different threshold values produced dramatically different mean and standard deviation of O/E ratio for each method. All methods showed some departure from 1 for the mean of O/E (mean (O/E)) introducing some biasness in model precision. When threshold value was chosen to be 0, values of mean (O/E) were approximately 1 producing unbiased estimate of E for all methods, except for SVM (overestimated E) (Figure 12 (d)). However, index performance was erratic at probability of detection threshold greater than 0.5. Mean (O/E) decreased immediately from 1 when threshold value was 0.5 or for any other adjusted threshold. But, we got the highest standard deviation of O/E (approximately 0.22) for all methods when threshold value is 0 (Table 7). Also, adjusted threshold had higher standard deviation of O/E for all methods (Table 8). Thus, standard deviation of O/E when threshold = 0.5 gave relatively low values compare to other threshold values. CT, SVM, and GBM generally underestimated E (Figure 11, Table 6), although these models were often equally precise as their less biased counterparts (Table 6) standard deviation Mean (O/E) for Random forest (0.96) with standard deviation (0.18) and LDA (0.94) with standard deviation (0.20) were close to a standard 1 (Table 6).

Results from pair-wise *t*-test for mean comparison of O/E indicated that all methods were statistically different from each other for adjusted cutoff value (Table 10). These results were also true when threshold value was set to 0 and 0.5 (Table 11 & 12) with the exception for the threshold value 0 where LDA and classification tree were not statistically different from each other.

Conclusion:

Overall, we found that RF was slightly superior (accuracy and precision) in predicting individual taxa as compared to that from other methods we examined. However, none of the machine learning algorithms showed significant improvements over LDMs when modeling assemblage types and estimating E from those models. We also observed no improvement in O/E index precision when RF estimates for individual taxa were aggregated to estimate O/E. SVM performed relatively poorly in both assemblage and individual taxa modeling compared with the other methods and thus resulting in higher bias in most O/E indices. Given the superior performance of RF and other machine learning algorithms for other applications, it is likely that the performance of models and associated indices developed for such large spatial scales may depend more on sample size and the availability and suitability of predictor variables than the modeling technique. Along with the predictor variables, the quality and type of data plays a large role in determining which modeling technique results accurate predictor. We also conclude that the best way to determine an ideal modeling technique is to compare the data modeling results from all the known modeling methods on various precision metrics and choose the accurate model for prediction. Modeling requirement could be different for various data sets and should not be generalized based on the outcomes and results of some other dataset. In our case even though the newer methods were more promising, upon modeling they did not provide better results than that obtained from the traditional method.

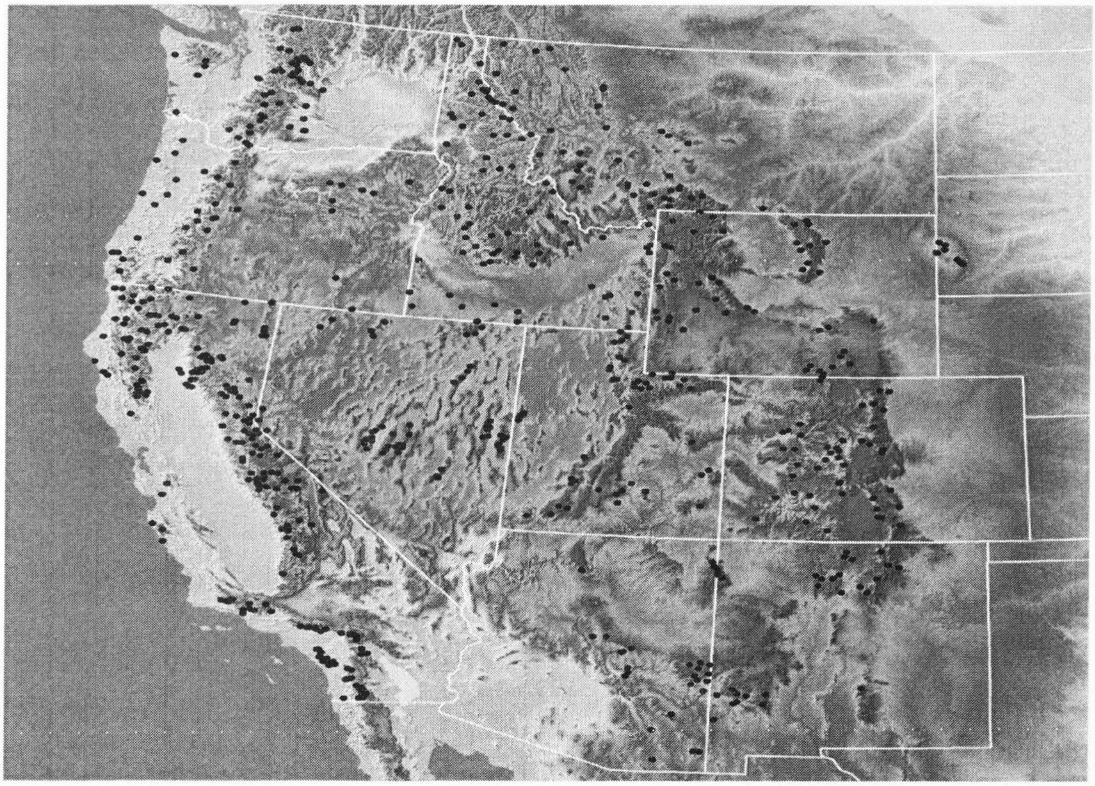

Fig. 1: Location of 728 reference sites the western United States.

Table 2: Percentage correctly classified for assemblage of sites classifications (N = 628 sites). PCC: the percentage of correctly classified presences and absences, Xval: Crossalidated estimates, Resub: Re-substituted estimates, LDA: Linear Discriminant Analysis

Classification Method	**Percentage Correctly Classified (PCC)**			
	28 Groups		11 Groups	
	Xval	*Resub*	*Xval*	*Resub*
LDA	0.343	0.437	0.507	0.573
Classification Tree	0.071	0.287	0.177	0.511
Random Forest	0.377	1	0.535	1
Support Vector Machine	0.312	0.476	0.538	0.608

Table 3: Summary for observed over expected ratio (O/E) for all four classification methods

Classification Method	**Xval**		**Resub**	
	Mean O/E	*Std O/E*	*Mean O/E*	*Std O/E*
LDA	0.993	0.193	1.009	0.191
Classification Tree	1.006	0.203	1.019	0.203
Random Forest	1.053	0.187	1.066	0.159
SVM	0.970	0.249	0.959	0.251

Table 4: Paired T-test to compare statistical significance difference in mean (O/E) for composite modeling

	Estimate	Standard Error	DF	t Value	PR > |t|
LDA * ClassificationTree	0.0093	0.15	678	1.58	0.1138
LDA * RandomForest	0.058	0.13	678	12.03	<0.0001
LDA * SVM	-0.022	0.22	678	-2.58	0.01003
ClassificationTree * RandomForest	-0.049	0.14	678	-9.24	<0.0001
ClassificationTree * SVM	0.031	0.21	678	3.83	0.00014
RandomForest * SVM	0.080	0.21	678	9.85	<0.0001

Table 5: Summary for mean difference. Methods those share same letter are not statistically different.

Methods			
LDA			A
Classification Tree			A
Random Forest		B	
SVM	C		

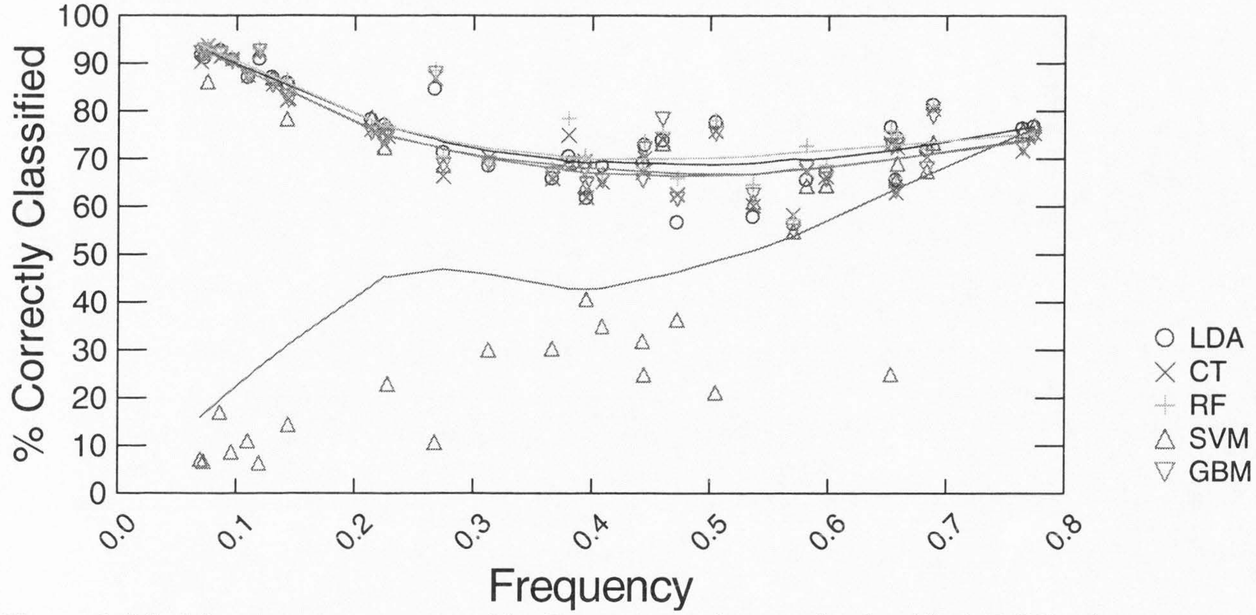

Figure 2: Model accuracy as measured by Percentage of correctly classified (PCC) when probability of detection threshold = 0.5

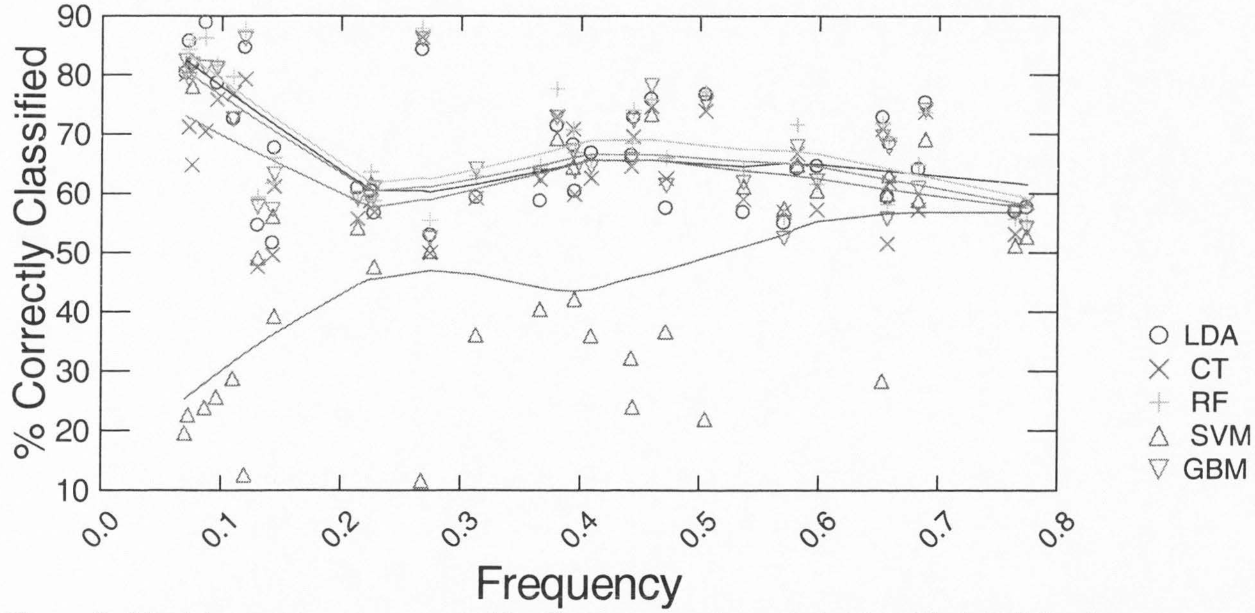

Figure 3: Model accuracy as measured by Percentage of correctly classified (PCC) when adjusted threshold is applied for probability of detection.

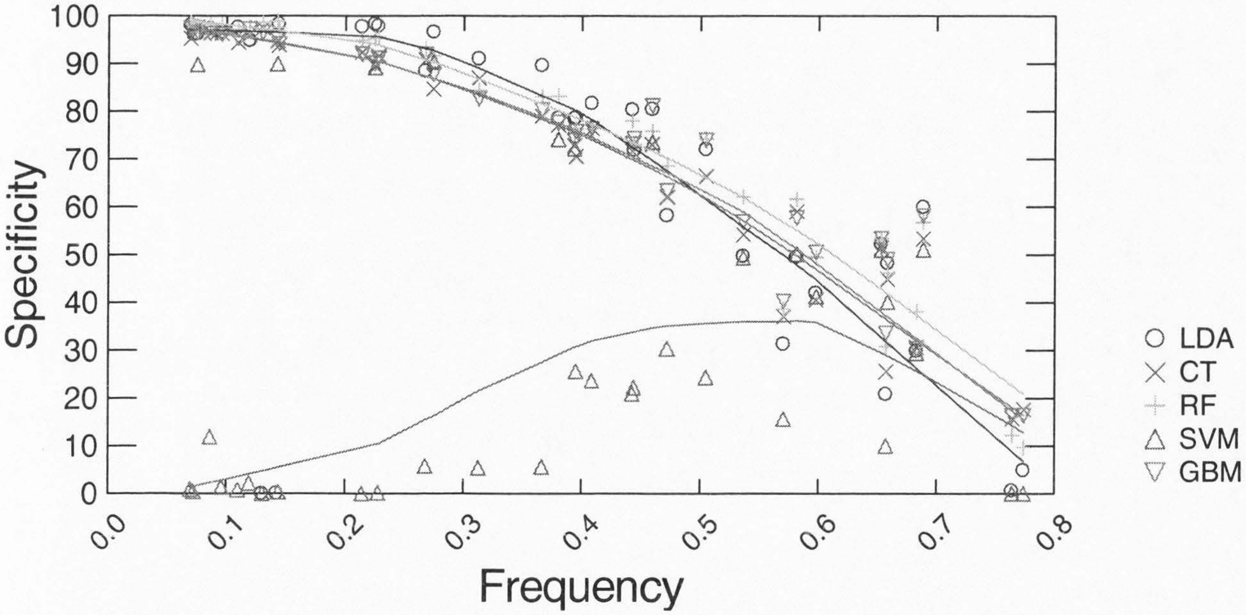

Figure 4: Model accuracy as measured by Percentage of absences correctly classified (Specificity) when probability of detection threshold = 0.5

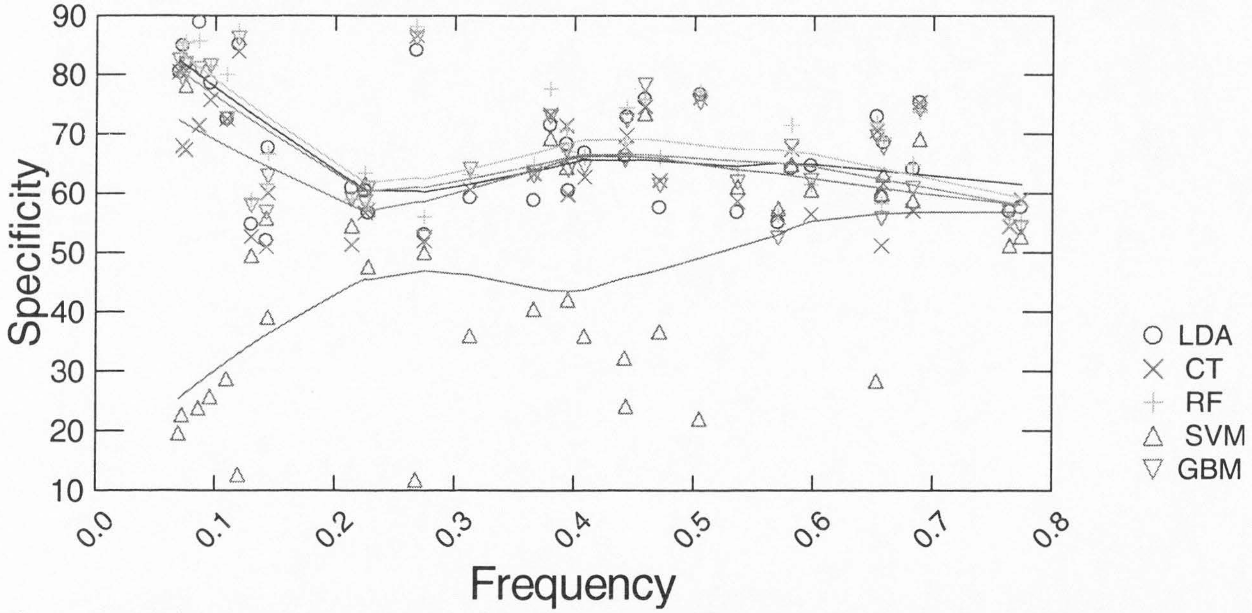

Figure 5: Model accuracy as measured by Percentage of absences correctly classified (Specificity) when adjusted threshold is applied for probability of detection.

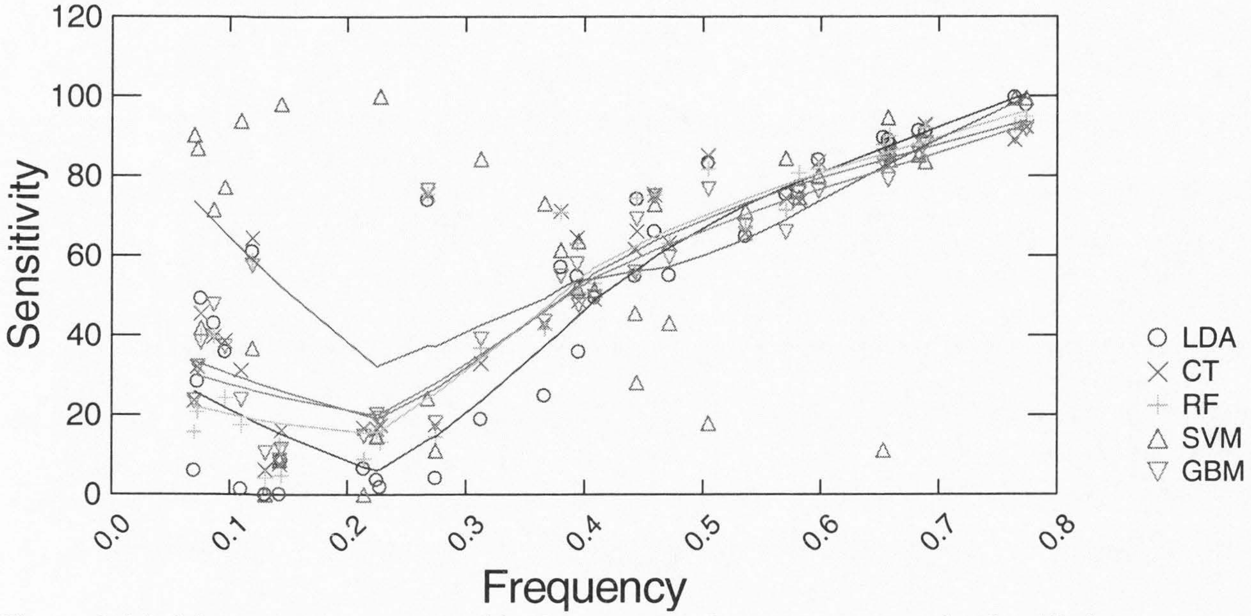

Figure 6: Model accuracy as measured by Percentage of presences correctly classified (Sensitivity) when probability of detection threshold = 0.5

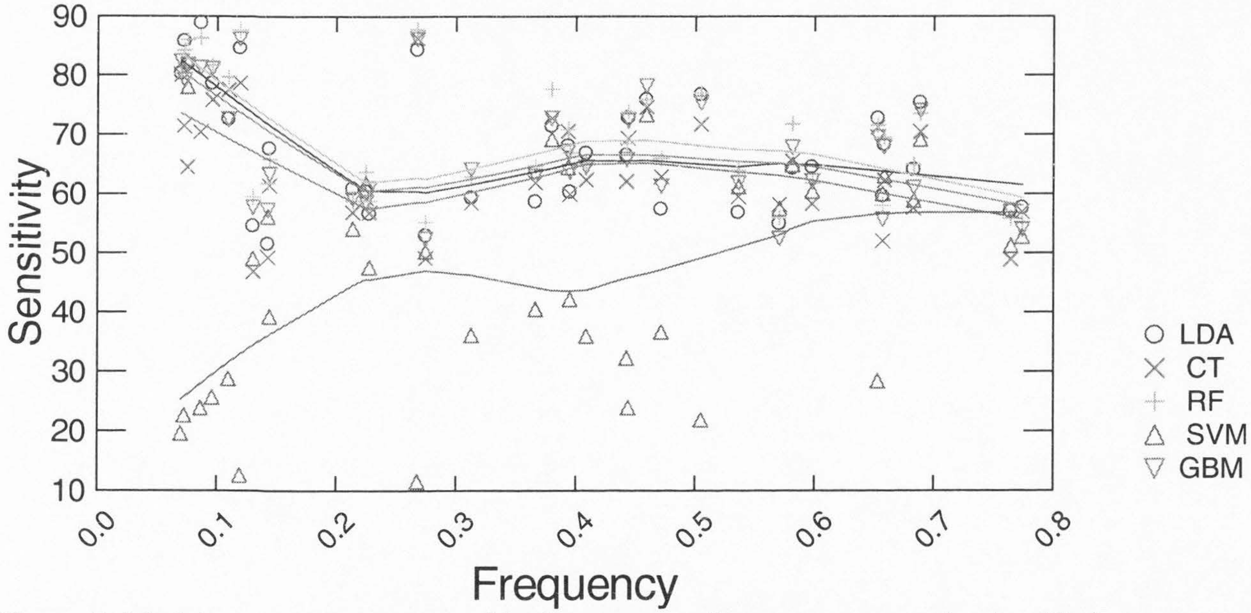

Figure 7: Model accuracy as measured by Percentage of presences correctly classified (Sensitivity) when adjusted threshold is applied for probability of detection.

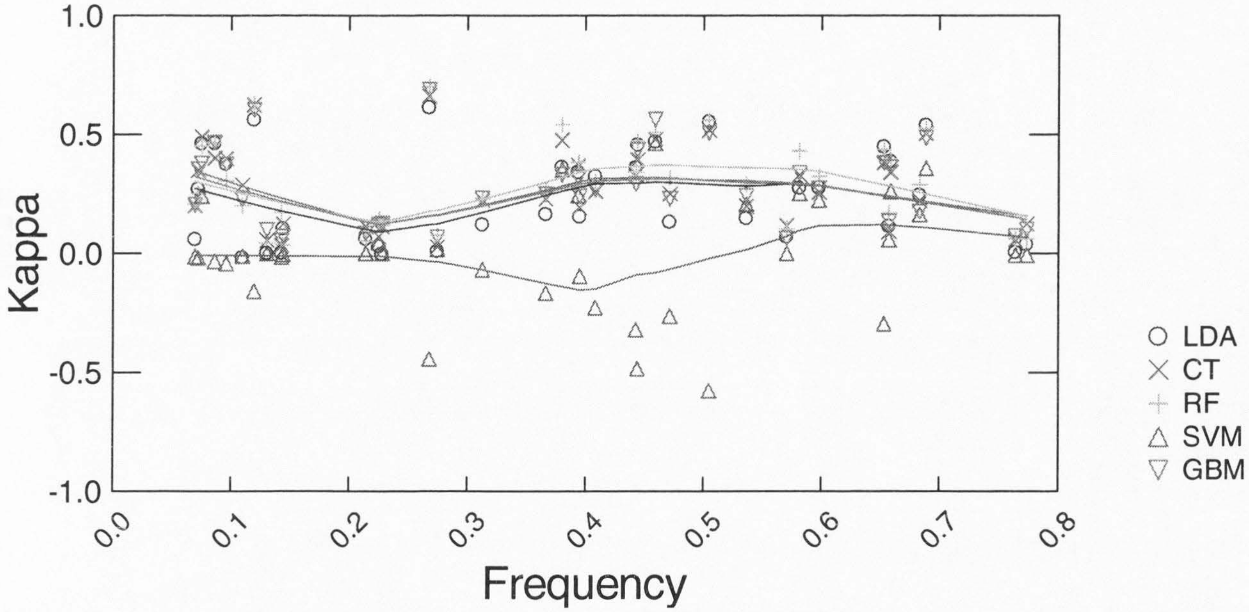

Figure 8: Model accuracy as measured by Kappa (Adjusted PCC for the agreement between presences and absences that might occur due to chance alone) when probability of detection threshold = 0.5

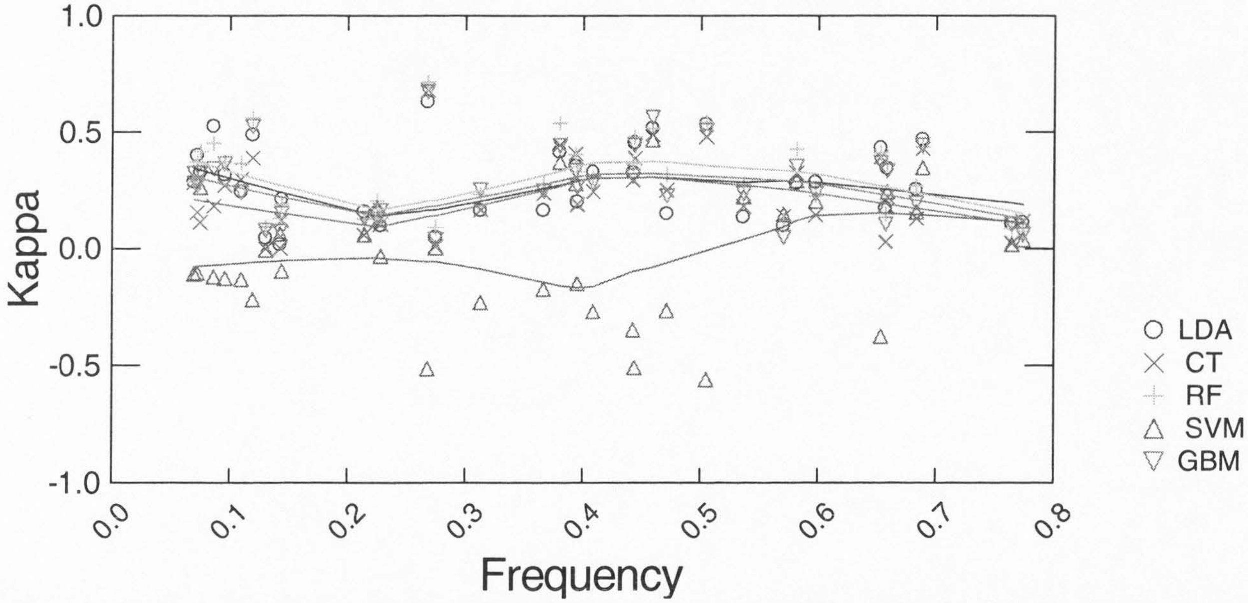

Figure 9: Model accuracy as measured by Percentage Kappa (Adjusted PCC for the agreement between presences and absences that might occur due to chance alone) when adjusted threshold is applied for probability of detection.

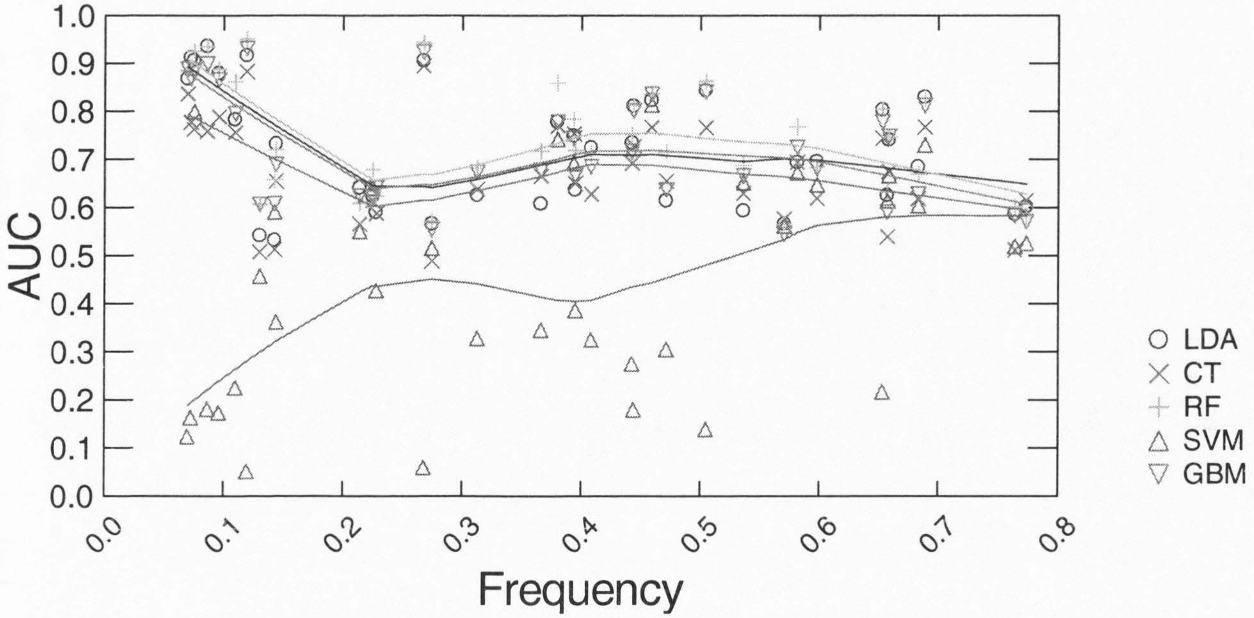

Figure 10: Model accuracy as measured by area under the curve (AUC) when probability of detection threshold = 0.5

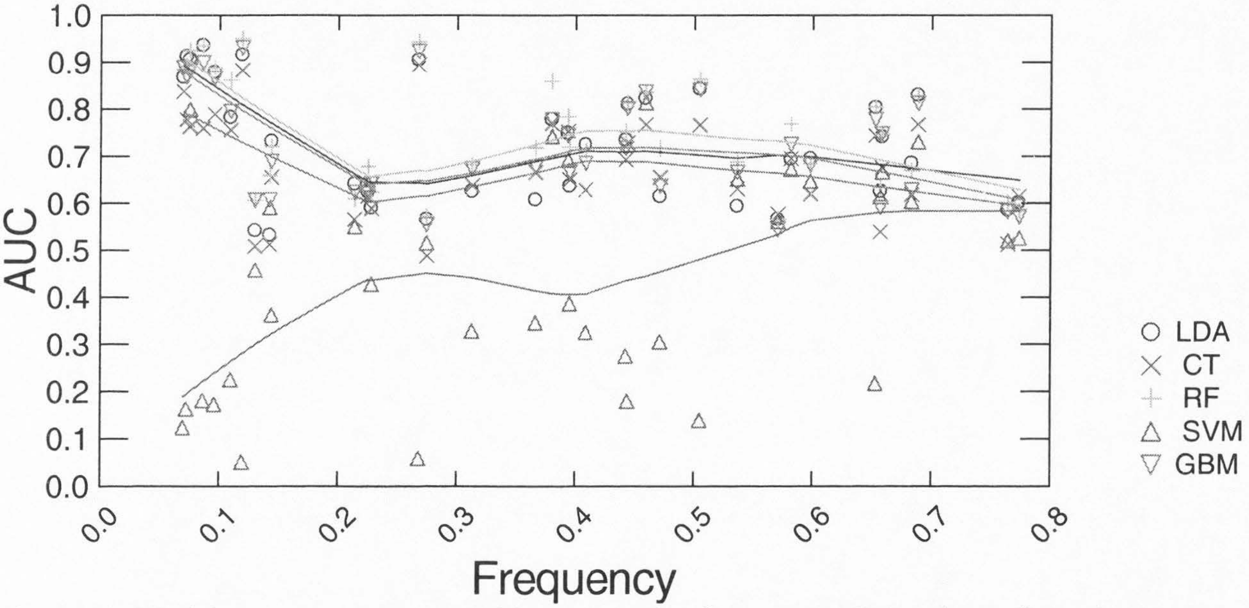

Figure 11: Model accuracy as measured by area under the curve (AUC) when adjusted threshold is applied for probability of detection.

Cutoff :0.5

	LDA	ClassificationTree	RandomForest	SVM	GBM
mean(O/E)	0.944	0.797	0.960	0.498	0.835
sd(O/E)	0.199	0.161	0.183	0.166	0.171
mean(O)	15.291	16.462	16.331	9.506	16.273
mean(E)	16.389	20.583	17.092	19.051	19.606

Table 6: Summary of O/E ratio for five classification methods (cutoff = 0.5)

Cutoff : 0

	LDA	Classification Tree	Random Forests	SVM	GBM
mean(O/E)	0.998	0.998	0.989	0.908	1.025
sd(O/E)	0.243	0.217	0.215	0.215	0.242
mean(O)	29.3923	29.3927	29.392	29.392	29.392
mean(E)	29.639	29.486	29.746	32.391	28.905

Table 7: Summary of O/E ratio for five classification methods (cutoff = 0)

Cutoff : Adjusted

	LDA	Classification Tree	Random Forest	SVM	GBM
mean(O/E)	0.974	0.849	0.932	0.742	0.856
sd(O/E)	0.2614	0.194	0.205	0.214	0.202
mean(O)	20.191	18.399	19.710	15.431	19.160
mean(E)	21.157	21.640	21.212	20.969	22.534

Table 8: Summary of O/E ratio for five classification methods (Adjusted Cutoff)

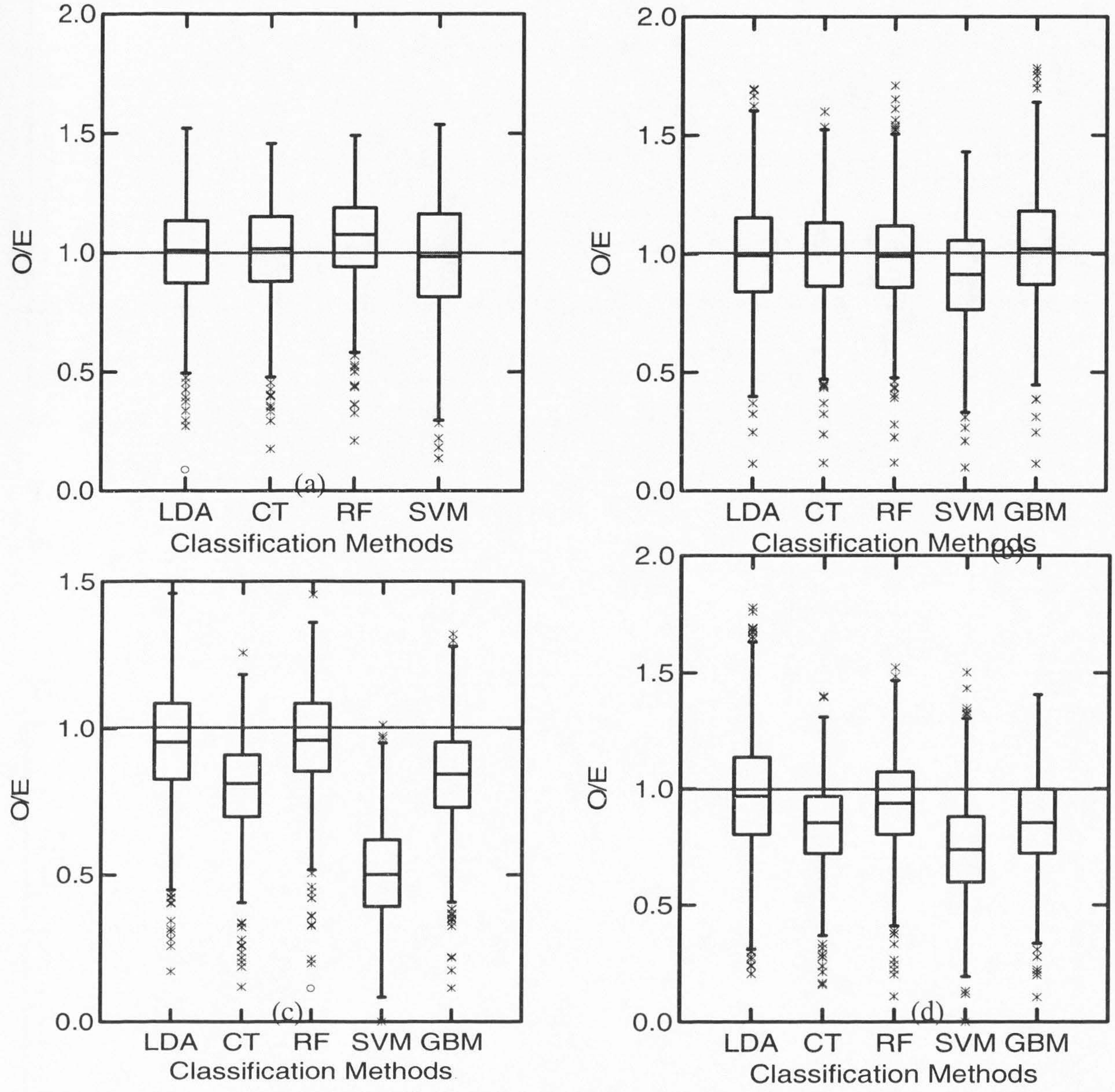

Fig. 12. Box plots of O/E index values for each prediction method. A = composite type modeling (probability of detection threshold $P_t \geq 0.5$). B = individual taxa models ($P_t \geq 0$). C = individual taxa models ($P_t \geq 0.5$). D = individual taxa models (Adjusted P_t).

Adjusted threshold

	Estimate	DF	t Value	PR > |t|
LDA * ClassificationTree	0.13	728	18.29	<.0001
LDA * RandomForest	0.04	728	5.96	<0.001
LDA * SVM	0.23	728	32.57	<0.001
LDA * GBM	0.12	728	18.34	<0.001
ClassificationTree * RandomForest	-0.09	728	-18.71	<0.001
ClassificationTree * SVM	0.11	728	15.23	<0.001
ClassificationTree * GBM	-0.01	728	-2.22	0.026
RandomForest * SVM	0.20	728	28.42	<0.001
RandomForest * GBM	0.08	728	16.63	<0.001
SVM * GBM	-0.11	728	-17.21	<0.001

Table 9: Paired T-test to compare statistical significance difference with p-value for Individual species modeling using adjusted probability of detection threshold

Threshold = 0

	Estimate	DF	t Value	PR > |t|
LDA * ClassificationTree	-0.00017	728	-0.048	0.069
LDA * RandomForest	0.0078	728	1.82	<0.001
LDA * SVM	0.09	728	22.46	<0.001
LDA * GBM	-0.029	728	-6.14	<0.001
ClassificationTree * RandomForest	0.0079	728	2.67	0.008
ClassificationTree * SVM	0.091	728	22.90	<0.001
ClassificationTree * GBM	-0.028	728	-7.25	<0.001
RandomForest * SVM	0.083	728	19.47	<0.001
RandomForest * GBM	-0.036	728	-8.93	<0.001
SVM * GBM	-0.12	728	-22.07	<0.001

Table 10: Paired T-test to compare statistical significance difference with p-value for Individual species modeling using probability of detection threshold Pt = 0

Threshold = 0.5

	Estimate	DF	t Value	PR > |t|
LDA * ClassificationTree	0.15	728	26.76	<.0001
LDA * RandomForest	-0.013	728	-2.40	0.0168
LDA * SVM	0.44	728	65.29	<0.001
LDA * GBM	0.12	728	21.25	<0.001
ClassificationTree * RandomForest	-0.16	728	-34.04	<0.001
ClassificationTree * SVM	0.29	728	47.32	<0.001
ClassificationTree * GBM	-0.039	728	-8.99	<0.001
RandomForest * SVM	0.45	728	67.43	<0.001
RandomForest * GBM	0.12	728	26.83	<0.001
SVM * GBM	-0.33	728	-52.25	<0.001

Table 11: Paired T-test to compare statistical significance difference with p-value for Individual species modeling using probability of detection threshold Pt = 0

Adjusted cutoff

Methods					
LDA					A
ClassificationTree				B	
RandomForest			C		
SVM		D			
GBM	E				

Table 12: Summary of mean difference for Individual species modeling using adjusted probability of detection threshold

Cutoff = 0

Methods				
LDA				A
ClassificationTree				A
RandomForest			B	
SVM		C		
GBM	D			

Table 13: Summary of mean difference for Individual species modeling probability of detection threshold Pt = 0

Cutoff = 0.5

Methods					
LDA					A
ClassificationTree				B	
RandomForest			C		
SVM		D			
GBM	E				

Table 14: Summary of mean difference for Individual species modeling using probability of detection threshold Pt = 0.5

References:

Breiman, L. 2001. Random Forests. Machine Learning 45:15-32.

Breiman, L. (2001) Statistical modeling: the two cultures. Statistical Science, 16, 199-231.

Breiman, L., and A. Cutler. 2005. Random Forests website: http://www.math.usu.edu/~adele/forest.

Chen, C., A. Liaw and L. Breiman. Using random forests to learn imbalanced data. Technical Report 666, Department of Statistics, University of California, Berkeley.

Clarke, R. T., M. T. Furse, J. F. Wright and D. Moss (1996) Derivation of biological quality index for river sites: comparison of the observed with the expected fauna. Journal of Applied Statistics, 23, 311-332.

Cutler, R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson and J. J. Lawler (2007). Random forest for classification in ecology. Ecology, 88(11), 2783-2792.

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. Ecology, 88(1), 243-251.

Drake, J. M., C. Randin and A. Guisan. Modelling ecological niches with support vector machines. Journal of Applied Ecology, 43, 424-432.

Ferrier, S . and A. Guisan (2006) Spatial modeling of biodiversity at the community level. Journal of Applied Ecology, 43, 393-404.

Fielding A. H. and J. F. Bell. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24, 38-49.

Graham, E. J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberson, S. Williams, M. S. Wisz and N. E. Zimmermann (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129-151.

Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J. M. Overton, R. Aspinall and T. Hastie (2006) Making better biogeographical predictions of species' distribution. Journal of Applied Ecology, 43, 386-392.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, NewYork.

Hawkins, C. P. The Western Center for Monitoring and Assessment of Freshwater Ecosystem. www.cnr.usu.edu/wmc

Hawkins, C. P. and D. M. Carlisle (2001) Use of predictive models for assessing the biological integrity of wetlands and other aquatic habitats.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (ed. C.A. San Mateo), pp. 1137–1143.

Lawler, J. J., D. White, R. P. Neilson and A. R. Blaustein (2006) Predicting climate-induced range shifts: model differences and model reliability. Global Change Biology, 12, 1568-1584.

Liaw A. and M. Wiener. (2002). Classification and regression by randomForest. R News , 2:18–22.

Morisette, J. T., C. S. Jarnevich, A. Ullah, W. Cai, J. A. Pedelty, J. E. Gentle, T. J. Stohlgren and J. L. Schnase (2006) A tamarisk habitat suitability map for the continental United States. Frontiers in Ecology and the Environment, 4(1), 11-17.

Olden, J. D., M. K. Joy, and R. G. Death (2006) Rediscovering the species in community-wide predictive modeling. Ecological Applications, 16(4), 1449-1460.

Ostermiller, J. D., C. P. Hawkins (2004) Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS- type models. Journal of the North American Benthological Society, 23(2), 363-382.

PRISM 2004. PRISM climate data, PRISM Group, Oregon State University, http://www.prismclimate.org, created 4 Feb 2004).

Rushton, S. P., S. J. Ormerod and G. Kerby (2004) New paradigms for modeling species distribution. Journal of Applied Ecology, 41, 193-200.

Yuan, L. L. (2006) Theoretical predictions observed to expected ratios in RIVPACS-type predictive model assessments of stream biological condition. The North American Benthological Society, 24(4), 841-850.

Acknowledgements:

We thank Dr. John Van Sickle for providing us a code to calculate O/E index precision (Distribution of R scripts for building and applying a RIVPACS-type predictive model). Also, we thank Ryan Hill for producing site map and providing the description of predictor variables.