Effect of Expert Opinion on the Predictive Ability of Environmental Models of Bird Distribution

JAVIER SEOANE, JAVIER BUSTAMANTE,* AND RICARDO DÍAZ-DELGADO

Spatial Ecology Group, Department of Applied Biology, Estación Biológica de Doñana, Consejo Superior de Investigaciones Cientificas (CSIC), Avda María Luisa s/n, 41013, Sevilla, Spain

Abstract: *The construction of predictive models of species distribution for conservation and regional planning can be facilitated by automatic procedures employed for the selection and transformation of predictors. It has been claimed, however, that empirical predictive models benefit from the inclusion of expert opinion at different stages of the model-building process. This is a time-consuming task that is limited by the availability of experts and difficult to standardize. Automated procedures for predictor selection based on statistical criteria are faster and easier to integrate into a geographic information system and may render highly explanatory models that fit the data with which they were built. But these models do not necessarily predict independent observations well and cannot be used to extrapolate to other areas. On the contrary, supervised models may include more frequently causal relationships, and therefore may more accurately predict new observations and extrapolate better to other areas. We built predictive models for the presence/absence of 10 bird species in two areas of Andalusia (southwestern Spain) to compare three different procedures for predictor selection ranging from a completely unsupervised to a fully supervised method. We evaluated models in three ways: (1) with the same data used to build the models, (2) with a different evaluation data set, and (3) with data from a different geographic area. The increase in the degree of expert input during model construction resulted in a significant decrease of model predictive ability when evaluated with an independent data set, and did not improve the predictive ability of the model when transferred to a new area. Unsupervised models had a greater tendency to overfit the building data, but this did not negatively affect model predictive ability or transferability to a new area. Incorporating expert opinion in the model-building process neither rendered better models as measured by their predictive ability nor resulted in models that were better suited to other regions. Therefore, unsupervised fitting procedures seem to be an adequate and cost-effective way to proceed when the aim is to generate potential distribution maps of species in a regional context.*

Key Words: expert models, habitat models, model transferability, potential distribution maps

Efecto de la Opinión de Experto en la Capacidad Predictiva de Modelos de Distribución de Aves usando Predictores Ambientales

Resumen: *La construcción de modelos predictivos de la distribución de especies en los campos de la conservación y la planificación regional resulta facilitada por procedimientos automáticos de selección y transformación de variables predictoras. Se ha argumentado, sin embargo, que la construcción de modelos empíricos predictivos podría beneficiarse de la inclusión de una opinión de experto en las diferentes fases de la modelización. Esto supone una elevada inversión de tiempo, es difícil de estandarizar y está limitado por la existencia de expertos adecuados. Los procedimientos automáticos de selección de predictores son más rápidos y fáciles de integrar en un sistema de información geográfica y pueden producir modelos altamente explicativos que ajusten bien los datos usados en la construcción del modelo. Sin embargo, podrian no predecir bien sobre un conjunto independiente de observaciones y no ser útiles para extrapolar a otras áreas. Por el contrario, los modelos supervisados podrían incluir más frecuentemente relaciones causales y, de ser asi, predecirian mejor nuevas observaciones y podrian extrapolarse a otras áreas. En este trabajo generamos modelos*

*Address correspondence to J. Bustamante, email jbustamante@ebd.csic.es
Paper received August 11, 2003; revised manuscript accepted June 29, 2004.

predictivos para la presencia/ausencia de 10 especies de aves, en dos áreas de Andalucía (SO España), con el fin de comparar tres procedimientos de selección de predictores que diferian en el grado de implicación de un experto durante la construcción del modelo, y que iban desde uno automático a otro completamente supervisado. Evaluamos los modelos de tres maneras: (1) con el mismo conjunto de datos usado para construir los modelos, (2) con un nuevo conjunto de datos de la misma área y (3) con datos de un área geográfica diferente. El incremento de la implicación de un experto durante la construcción del modelo resultó en una disminución significativa de la capacidad predictiva de éste cuando se evaluó con un conjunto nuevo de datos, y no mejoró su capacidad predictiva cuando se transfirió a un área nueva. Los modelos automáticos tenían una tendencia mayor a sobreajustar los datos usados en la construcción; pero esto no afecto de manera negativa a la capacidad predictiva del modelo o su extrapolación a un área nueva. La incorporación de una opinión de experto en el proceso de modelización no genera modelos con mayor capacidad predictiva ni resulta en modelos que puedan hacer extrapolaciones más fiables. Por lo tanto, los procedimientos automáticos parecen un medio eficaz y rentable para crear mapas de distribución potencial de especies en un contexto regional.

Palabras Clave: extrapolación de modelos, mapas de distribución potencial, modelos de experto, modelos de hábitat

Introduction

There is an increasing demand for fine-grained information on the distribution of species for conservation and regional planning. To select new reserves to preserve biodiversity, perform reliable environmental impact assessments, or evaluate the effect of changes in land use, one must know how species are distributed and where suitable habitats for them are located. Species distribution atlases can provide some of this information, but they are costly to produce and update and contain relatively coarse information on distribution that may not be sufficient for many management decisions. Predictive habitat models are another alternative that is increasingly being explored to produce fine-grained distribution and habitat suitability maps (Guisan & Zimmermann 2000; Pearce et al. 2001; Scott et al. 2002). Information on species occurrences or abundance from existing wildlife surveys can be used to fit statistical models that have environmental variables as predictors (Nicholls 1989). These models would allow us to get the maximum benefit from the information available on species distribution, particularly in remote areas where data are scarce and complete wildlife surveys are expensive (Osborne & Tigar 1992; Bustamante et al. 1997; Manel et al. 1999).

Methods for modeling species occurrences have evolved in the last four decades, but there is a widening gulf between the scientists who develop these models and the managers that need them (Stauffer 2002; Wiens 2002). Scientists may be paying too much attention to the statistical assumptions of the models and trying to derive causal mechanisms from the correlations found between species distributions and environmental factors. But managers urgently need detailed and accurate distribution maps for every species. For example, in Spain wildlife habitat models for many species and areas are urgently needed to implement strategic environmental impact assessment plans and programs (Díaz et al. 2001). There are several ongoing projects in which models are being used to produce regional maps of species distribution. The Gap Analysis Program in the United States is one of the oldest and best known (Scott et al. 1993; U.S. Geological Survey GAP Analysis Program 2000), but there are different approaches. In Australia, for example, the North-East Forest Biodiversity Study and Environment Australia (Pearce & Ferrier 2000*a*) take different approaches, as does the Federal Institute for Forest, Snow and Landscape Research (Heller-Kellenberger et al. 1997) in Switzerland. Among other differences, these projects vary in the extent to which expert input is used in the development of distribution models. The Swiss project uses models based exclusively on expert knowledge, whereas the Australian projects develop statistical models from field observations and employ mostly automatic procedures for the selection and fitting of predictors.

Automatic statistical procedures for selecting and fitting predictors in ecological models have been strongly criticized (e.g., James & McCulloch 1990; MacNally 2000; Burnham & Anderson 2002). Multivariate techniques will pick up statistically significant correlations in random data sets of spurious variables if the set is big enough (Rexstad et al. 1988). Mac Nally (2000) indicates that stepwise predictor selection is highly flawed and apt to produce spurious models. Rencher and Pun (1980) show that stepwise selection inflates model-explained variance. Burnham and Anderson (2002) caution against "data dredging" methods that can lead to overfitted models with spurious parameter estimates and statistical association with ecologically unimportant variables. Also, Burnham and Anderson advocate careful thinking during the development of an a priori and well-founded set of candidate models.

Automatic predictor selection and fitting procedures are fast and easy to implement with modern computer software. When, from a management perspective, one needs to predict the distribution of hundreds of species

and has a lot of potential predictors available in a geographical information system, it is tempting to use a shotgun approach, hoping that unsupervised modeling procedures will tell which variables are important (Wiens 2002). Careful a priori selection of predictors or of potential functional relations between predictor and response variables for predicting species distributions requires the help of experts that do not necessarily exist for all species groups in any given region. Even if experts are available, the degree of their expertise is difficult to evaluate. In addition, developing expert models, or incorporating expert opinion in statistical models, is slow and tedious and must be performed on a species-by-species basis. If we have to rely on careful a priori design of species distribution models in a species-by-species and region-by-region basis to build fine-grained distribution maps, managers will not have the information they need when they need it.

When the aim of modeling species distributions is to generate distribution or habitat suitability maps for a large number of species that can be used for reserve selection or conservation planning, measures of the model predictive accuracy—like overall predictive accuracy or the accuracy of predicting species presences (sensitivity)—are the most relevant indicators of model success. The risks of automatic variable selection procedures for building models are well known (e.g., inclusion of spurious variables and overfitting), but there has been little research examining to what extent these problems influence predictive accuracy and exploring whether expert input during the variable selection and the fitting procedure improves predictions. The only work, to our knowledge, comparing statistical models fitted with automatic procedures to models fitted with expert input suggests that automatic models can be as good as those incorporating expert opinion (Pearce et al. 2001).

Predictive habitat modeling based on data from biological surveys is typically tackled with regression-like approaches. Logistic regression is often used because of its suitability for modeling a binary response variable, such as presence/absence of a species (reviewed in Guisan & Zimmermann 2000). In these cases, a response variable—the presence/absence of the species in an area—is related to predictive variables with some suspected discriminatory ability. In regression modeling, the choice of predictors may be automated by full forward or backward algorithms designed to satisfy certain statistical criteria. Automatic procedures are desirable because they are fast and easy to standardize.

On the other hand, using experts to decide a priori on potential models or using expert supervision for predictor selection and fitting by excluding or modifying relationships that do not meet some biological criteria leads to models that are more credible. Such models, however, require the existence and availability of experts and are difficult to standardize. Also, the credibility of models developed with expert input, unless tested with new independent field data, is not free, to some extent, from some circular reasoning (the model may be credible to the same experts who developed it). Another aspect of interest is that models are sometimes used to extrapolate in time or space, which is out of their universe of application. Although a violation of classical statistics, this may be the only approach available for a manager who must make a decision. If models with expert input incorporate more "causal" relationships than automatic models that may be plagued with spurious variables, the former may, on average, be more accurate when predictions are extrapolated to the future or to other areas (Lezzoni 1999).

We compared three procedures for predictor selection that differ in the degree of expert input necessary to build statistical models of bird distribution: a completely unsupervised method (automatic stepwise variable selection by statistical software) that required little expert input; a semisupervised method (in which an ornithologist selected a priori predictors for each bird species); and a completely supervised method (in which both the predictors selected and the shape of the relationships were evaluated with statistical and ecological criteria on a species-by-species basis). We tested the predictive ability of all models in three scenarios: (1) with the same data used to build the models, (2) with a different evaluation data set from the same area, and (3) with data from a different geographic area. Our aim was to explore whether including expert opinion in the building of models produced models with better predictive ability or results that could be better extrapolated to other areas.

Methods

Study Area

We examined two 70 × 70 km squares in Western Andalusia, southern Spain. We refer to them as Aracena (center: 6°21′W 37°39′N) and Grazalema (center: 5°28′W 36°44′N; Fig. 1). Both areas have roughly the same proportion of cropland (mainly wheat, sunflower, and olive groves); Mediterranean shrubland; forested areas (evergreen oak and cork oak forests, and areas of pastures with scattered oaks called *dehesas*); and numerous human settlements. The areas differ mainly in that Grazalema reaches higher elevations (0–1622 m asl) compared with Aracena (0–960 m) and that the soil type is mostly calcareous in Grazalema and mostly acidic in Aracena.

Bird Census Data

We performed 1144 unlimited distance point surveys, 15 minutes in duration and without repetition, during the springs of 1999 and 2000. Approximately half the points were surveyed each year. There were 521 point

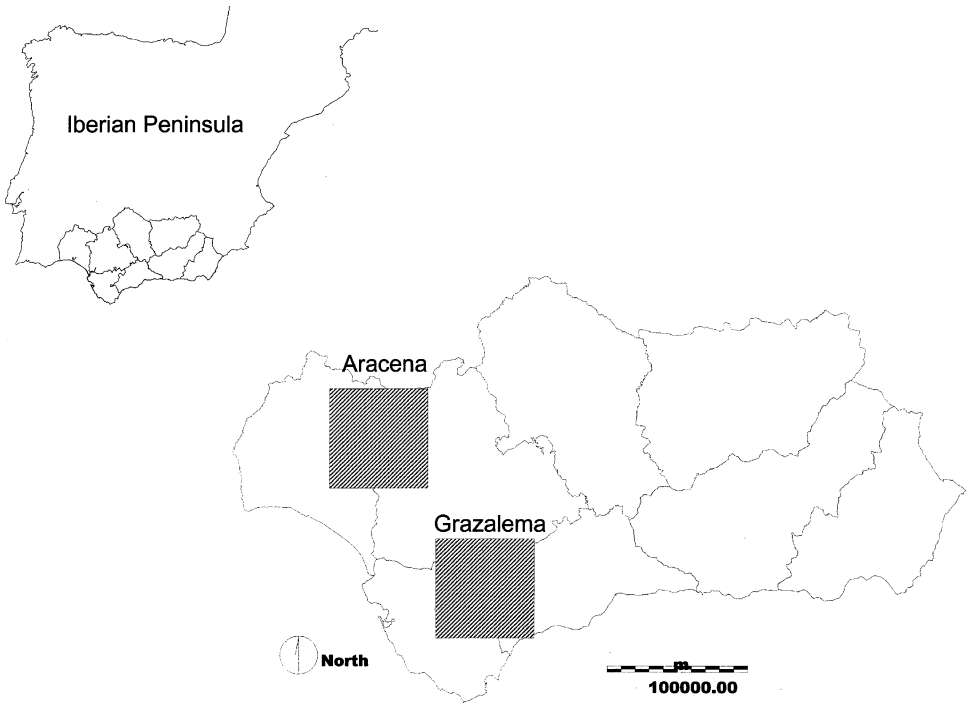

Figure 1. Location of study areas.

surveys in Aracena and 623 in Grazalema, equally distributed throughout the spring in both areas. Point surveys were on average 1 km apart. Those that were closer (100–1000 m) were always located in different habitats (e.g., a crop field vs. a forest). We selected 10 bird species (Table 1) from all those recorded in surveys ($n = 172$) according to three criteria: (1) relatively abundant in both study areas; (2) representative of a variable range of prevalences (number of recorded presences in point surveys); and (3) representative of the three main land covers (cropland, shrubland, and forest). Absences outnumbered presences for every species in both study areas (Table 1). To avoid bias resulting from this fact (Fielding & Bell 1997; Cumming 2000), we randomly selected a number of absences equal to the number of presences for each species in each study area. Our sample sizes were similar to or higher than those indicated as necessary to produce reliable estimates of accuracy in previous studies (Pearce & Ferrier 2000*a*; McKenney et al. 2002; Stockwell & Peterson 2002).

Predictive Variables

Fifty-six environmental predictors (Table 2), which summarized the most relevant environmental gradients and some landscape features, were extracted and amalgamated from a geographic information system (GIS). These predictors included variables describing vegetation, land use, landscape, topography (resolution 50 m), and climate (resolution 1 km). The variables were averaged for circles 350 m in diameter that were centered on survey points. All metrics were calculated with IDRISI 32 (Eastman 1999), IDRISI for Windows (Eastman 1997), and MIRAMON (Pons 2000).

Model Types

Generalized additive models constitute a universe of statistical models in which a response variable is related to one or several predictors by means of scatter-plot smoothing functions (Hastie & Tibshirani 1990). Because the linear regression is only one kind of scatter-plot smoothing function, generalized linear models (McCullagh & Nelder 1989) can be considered a particular subset of generalized additive models, and the logistic regression can be thought of as one particular generalized additive model.

Table 1. Bird species selected to build predictive models and number of survey points for which each species was recorded as present in each study area.

Species	*Aracena* (n = *521*)	*Grazalema* (n = *623*)
Red-legged Partridge (*Alectoris rufa*)	72	142
Linnet (*Carduelis cannabina*)	114	202
Short-toed Treecreeper (*Certhia brachydactyla*)	153	160
Robin (*Erithacus rubecula*)	37	119
Thekla Lark (*Galerida tecklae*)	85	57
Calandra Lark (*Melanocorypha calandra*)	32	44
Blue Tit (*Parus caeruleus*)	176	154
European Nuthatch (*Sitta europea*)	113	62
Sardinian Warbler (*Sylvia melanocephala*)	184	321
Wren (*Troglodytes troglodytes*)	38	136

Table 2. Variables used as candidate predictors in unsupervised distribution models for all bird species.

Variable description	*Bird species number*[a]
Mean elevation[b]	1,2,3,4,5,6,7,8,9,10
Mean slope[b]	1,2,3,4,5,6,7,8,9,10
Mean annual temperature[c]	1,2,3,4,5,6,7,8,9,10
Mean annual rainfall[c]	1,2,3,4,5,6,7,8,9,10
Mean annual potential solar radiation[b]	1,2,3,4,5,6,7,8,9,10
Percentage of crop land (crops, olive groves, vineyards)[d]	1,2,5,6,9,10
Percentage of herbaceous vegetation (including cereal crops)[d]	1,6
Percentage of olive groves[d]	
Percentage of forest (including *dehesas* and open forest)[d]	3,4,7,8,9,10
Percentage of dense forest[d]	3,4,7,8,10
Percentage of deciduous forest[d]	3,4,7,8
Percentage of coniferous forest[d]	3,4,7,8
Percentage of shrub[d]	1,2,4,5,9,10
Percentage of riparian vegetation[d]	4,7,10
Presence of sparse tree cover (e.g., included in a heterogeneous cropland area)[d]	6
Presence of dense tree cover (e.g., included in a heterogeneous cropland area)[d]	6
Presence of sparse shrub or sparse shrub-like structures (such as vineyards)[d]	5,9
Presence of dense shrub[d]	2,5,9
Length of boundaries between forested land-cover categories and rest of vegetation categories[d]	3,4,6,7,8,9
Length of boundaries between forest and shrubland[d]	2,3,4,7,8,9
Cropland heterogeneity[e]	1,6
Compactness ratio of dense forest areas (an indirect estimate of surface-perimeter ratio)[d]	3,4,7,8
Distance to nearest patch of cover type $X < 2$ ha[d]	same species as for percentage of each land use/land cover
Distance to nearest patch of cover type X between 2 and 10 ha[d]	same species as for percentage of each land use/land cover
Distance to nearest patch of cover type $X > 10$ ha[d]	same species as for percentage of each land use/land cover

[a] *Each number represents a species for which that predictor was considered in semisupervised and supervised models:1, Red-legged Partridge; 2, Linnet; 3, Short-toed Treecreeper; 4, Robin; 5, Thekla Lark; 6, Calandra Lark; 7, Blue Tit; 8, European Nuthatch; 9, Sardinian Warbler; 10, Wren.*
[b] *Digital elevation model of Andalusia at 50-m resolution.*
[c] *Raw meteorological data provided by the Instituto Nacional de Meteorología and interpolated by regression models and kriging at resolution 1 km^2 (temperature: Bustamante 2004; rainfall: J.B., unpublished data).*
[d] *The 1995 land use/land cover digital map of Andalusia from the SinambA (Consejería de Medio Ambiente, Junta de Andalucía). Three distance variables were considered for each land use/land cover type, with* X = *urban area, cropland, herbaceous vegetation, olive grove, forest, dense forest, deciduous forest, coniferous forest, shrubland, and riparian vegetation.*
[e] *Estimated with satellite images (sensor LISS-III, Indian remote sensing satellite [IRS] date: 19 July 1999 for Aracena and 16 July 1999 for Grazalema). We measured the fractal dimension of an NDVI (normalized difference vegetation index) image of areas classified as cropland in the land use/land cover map.*

Using the environmental variables as predictors, we built a generalized additive model for the presence/absence of each species in each study area with binomial errors and logit link (equivalent to a logistic regression). For each species and study area, we built three models (hereafter called "model types") that differed in the degree of expert involvement in the selection and fitting of predictors.

First, we fitted an automatic model with stepwise selection of predictors exclusively with statistical criteria (the unsupervised model). We performed a forward-backward stepwise selection from all possible predictors (with the step.gam procedure of S-PLUS 2000, MathSoft 1999). We started from a null model and tested each predictor sequentially as a smoothing spline with 3 df. The predictor that reduced the residual deviance the most was included in the model. We repeated the procedure until no more predictors improved the model. We then tried to simplify the resulting model by decreasing the complexity of each of the predictors included (by means of a smoothing spline with 2 df and a linear term; linear term is equivalent to a standard logistic regression). The criterion to enter, remove, or simplify a term was Akaike's information criterion (AIC; Sakamoto et al. 1986), which takes into account the reduction in residual deviance and degrees of freedom resulting from a certain predictor.

Second, we fitted a model in which predictors for each species were a priori selected by an expert ornithologist (J.S.), according to his field experience on habitat selection of each species in western Andalusia (Table 2). For example, he considered it appropriate to include the percentage of cropland among the predictors to be tested

with the Calandra Lark (scientific names given in Table 1), but not the percentage of forest. We called these semisupervised models. The predictors for these models were submitted to the forward-backward stepwise algorithm based on AIC, with a procedure analogous to the one described previously. To further reduce the risk of overfitting, all predictors in the final model were checked for significance by comparing the increase in residual deviance after dropping the predictor from the model with a chi-square test ($\alpha = 0.05$). This tends to be a stricter criterion than AIC.

Third, we fitted a model in which predictors were selected a priori and all significant relations were checked and modified by the expert ornithologist a posteriori. These were called supervised models. The model started from the one resulting from the semisupervised procedure. The expert performed a sensitivity analysis of each model (see Nicholls 1989) trying to detect data outliers, influential observations, and predictors with coefficients affected by a few observations. Models were refitted excluding these observations.

Because the expert cannot know the particular shape of the relationship between an environmental variable and the presence of a species, we decided to simplify the models to linear terms (which have a lower risk of overfitting than kernel smoothers such as smoothing splines). Each predictor previously selected by the automatic procedure was transformed into a linear term after visual inspection of a partial residual plot (Brown 1994; Franklin 1998). If the probability of the species presence increased or decreased with an environmental gradient, the spline was transformed into a linear regression. If the relationship seemed to have an optimum, the spline was transformed into a second-degree polynomial. And if there seemed to be no relationship with a predictor below or beyond a certain point, the relation was transformed into piecewise linear terms. The new linear terms were tested for significance with a chi-square test ($\alpha = 0.05$).

Statistically significant relationships with environmental predictors that made no sense with species habitat selection and other relationships that the expert considered spurious were excluded from the model. Other variables not selected by the automatic procedure but that the ecology of the species suggested could be good predictors were tested for inclusion as linear terms in the resulting simplified models.

Model Validation

To assess the discrimination ability of the models, we used Cohen's kappa statistic (Titus et al. 1984) and the area under the curve (AUC) of receiver operating characteristic (ROC) plots (Hanley & McNeil 1982; Murtaugh 1996; Cumming 2000). Kappa estimates the chance-corrected percentage of agreement between predictions and observations. To calculate kappa it is necessary to define a threshold of predicted probability above which to consider presence (Fielding & Bell 1997). We selected the average between the mean of probabilities for absences and the mean for presences (Fielding & Haworth 1995) for this threshold. The AUC is a threshold-independent measure of discrimination ability (Zweig & Campbell 1993) and is not affected by prevalence of presences in the sample (Manel et al. 2001); thus, it is more appropriate than kappa. Area under the curve has been used only recently in ecology, however (Fielding & Bell 1997; Cumming 2000; Pearce & Ferrier 2000*b*; Bonn & Schröder 2001; Manel et al. 2001). Thus, here we used the two indices as means of comparison. The AUC was calculated with AccuROC 2.5 (Vida 1993).

We calculated kappa and AUC for the three model types (supervised, semisupervised, and unsupervised) in three scenarios of evaluation according to an increasing degree of independence between the data used to build the model and the data used to evaluate it: (1) a building scenario, with the same data used to build the model; (2) a cross-validation scenario, a 10-fold cross-validation repeated 20 times; and (3) an extrapolation scenario, with data of the same species but from the other study area.

The first scenario is an evaluation that tends to overestimate the discrimination ability of the model (Oreskes et al. 1994). It may be used both as a maximum reference of the explanatory ability of the model and to informally assess the amount of overoptimism in the estimates by comparison with the predictive ability in the cross-validation scenario. Moreover, this scenario is the only evaluation presented in many research papers. This type of evaluation informs one of the statistical explanatory ability of the model (how well the model fits the data used in its construction).

The second scenario is an internal validation that follows one of the variety of resampling approaches used most frequently (this one according to Harrell 2001). Data were split into 10 groups of equal size; a model was built using the data of the first nine groups (90% of observations) and evaluated using the data of the tenth group (the 10% of observations not used to build the model); and the procedure was repeated 10 times with each group as an evaluation set and the remaining 9 as building set. This 10-fold cross validation was repeated 20 times, randomizing the data each time and forming new groups. The predictors included in the models were the ones resulting from the previously described fitting procedures (model types), and only new coefficients were adjusted with the building set. The result was an unbiased estimate of the predictive ability of the model within its universe of application.

The third scenario is an external validation (Harrell 2001) and was used to assess the model outside the universe in which it is statistically valid. This validation evaluated the transferability of the model (Altman & Royston 2000; Bonn & Schröder 2001). From a statistical point of

Table 3. Results of repeated-measures analysis of variance of the effect of predictive model type (unsupervised, semisupervised, and supervised); evaluation scenario (building, cross validation, and extrapolation); and study area (Aracena and Grazalema) on model discriminatory ability estimated by area under the curve (AUC).*

Variable	*df*	*SS*	*MS*	F	p
Model type	2	0.089	0.045	14.767	<0.0001
Evaluation scenario	2	1.078	0.539	178.6074	<0.0001
Study area	1	0.030	0.030	9.870	0.002
Model type ∗ evaluation scenario	4	0.112	0.028	9.259	<0.0001
Model type ∗ study area	2	0.017	0.008	2.789	0.064
Evaluation scenario ∗ area	2	0.007	0.004	1.181	0.31
Model type ∗ evaluation scenario ∗ study area	4	0.020	0.005	1.6731	0.16
Residuals	153	0.461	0.003		

**Between-species residuals: df = 9, SS = 0.667.*

view, there is no particular interest in estimating the predictive ability of a model in a universe different from that in which it was built. In conservation, however, predictive habitat models are frequently built with the expectation that they can be transferred in time (into the future) or in space (to remote areas).

Differences in predictive ability (discrimination) among models resulting from differences in model type, evaluation scenario, and study area were analyzed with a repeated-measures analysis of variance (ANOVA) of AUC and kappa values. We used a repeated-measures design to eliminate the effect of differences in predictability between species. Pairwise differences of means between factor levels were assessed estimating the least significant difference (Sokal & Rohlf 1981) with $\alpha = 0.05$.

Results

Models better than a null model were built for all species in each combination of model type, evaluation scenario, and study area. Unsupervised models included a significantly ($F_{2,57} = 48.38$, $p < 0.0001$) higher number of predictors (mean = 9, SD = 4.5) than both semisupervised (mean = 3, SD = 1.1) and supervised (mean = 4, SD = 1.3) models. The estimates of discrimination ability, both AUC (Table 3) and kappa (Table 4), differed significantly among model types and evaluation scenarios, but there was also a significant interaction between model type and evaluation scenario. This indicated that differences among model types were not constant across evaluation scenarios.

Mean estimates of predictive ability declined with higher participation of the expert in model construction (unsupervised, semisupervised, and supervised models had, respectively, 0.825, 0.785, and 0.773 mean AUC, and 0.535, 0.453, and 0.417 mean kappa). As expected, the evaluation with building data rendered higher estimates of predictive ability (AUC = 0.879, kappa = 0.631) than when the model was cross validated (AUC = 0.812, kappa = 0.491) or when the model was extrapolated to the other study area (AUC = 0.692, kappa = 0.283). Models developed in Aracena were on average superior to those from Grazalema. The significant interaction between model type and evaluation scenario (Tables 3 & 4) indicated that mean estimates of predictive ability can be misleading and that differences in predictive ability among model types can differ among evaluation scenarios.

Differences in predictive ability among model types in the evaluation with building data were not of particular interest themselves because it is well known that the evaluation with the same data used to fit the model overestimates the predictive ability, and the different model

Table 4. Results of repeated-measures analysis of variance of the effect of predictive model type (unsupervised, semisupervised, and supervised); evaluation scenario (building, cross validation, and extrapolation); and study area (Aracena and Grazalema) on model discriminatory ability estimated by kappa.*

Variable	*df*	*SS*	*MS*	F	p
Model type	2	0.441	0.221	19.626	<0.0001
Evaluation scenario	2	3.680	1.840	163.6	<0.0001
Study area	1	0.152	0.152	13.475	0.0003
Model type ∗ evaluation scenario	4	0.499	0.125	11.083	<0.0001
Model type ∗ study area	2	0.061	0.031	2.722	0.069
Evaluation scenario ∗ study area	2	0.018	0.009	0.798	0.45
Model type ∗ evaluation scenario ∗ study area	4	0.058	0.014	1.285	0.278
Residuals	153	1.720	0.011		

**Between-species residuals: df = 9, SS = 2.756.*

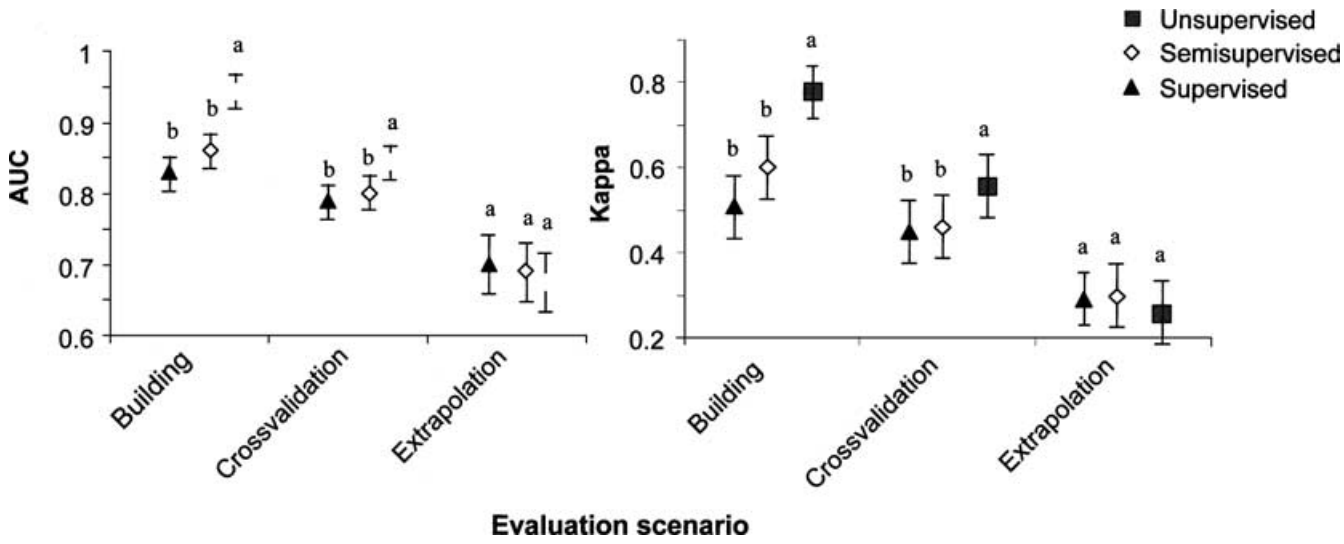

Figure 2. Mean values (and 95% confidence intervals) of estimates of discriminatory ability for each combination of evaluation scenario and model type. Letters indicate comparisons of mean model discriminatory ability within each evaluation scenario; the same letter indicates a nonsignificant difference (least significant difference, $\alpha = 0.05$); AUC is area under the curve of a receiver operating characteristic plot.

types could differ in the degree of overfitting. To test for differences in overfitting—assuming that the estimate in our cross-validation scenario was an unbiased estimate of model predictive ability—we calculated for each model type the increase in AUC (and kappa) between the cross validation and the building scenario. The predictive ability was overestimated in the building scenario for all model types, but the increase in participation of the expert in model construction reduced the degree of overestimation.

A repeated-measures ANOVA on the increase in AUC with model type and area as factors indicated very significant differences in the increase in AUC among model types ($F_{2,45} = 15.04$, $p < 0.0001$) and no significant differences between areas ($F_{1,45} = 3.31$, $p = 0.076$) or interactions between model type and area ($F_{2,45} = 0.94$, $p = 0.398$). The results were similar for the increase in kappa: a significant difference was seen among model types ($F_{2,45} = 19.11$, $p < 0.0001$) and no differences were attributed to area or its interaction with model type ($F_{1,45} = 3.85$, $p = 0.056$, and $F_{2,45} = 0.298$, $p = 0.744$, respectively). In the building scenario, AUC increased 9.4% in the unsupervised models, >6.9% in the semisupervised models, and >3.9% in the supervised models. All differences between model types were significant (least significant difference = 2.1%) and were significantly >0. The results were similar for the increase in kappa (unsupervised = 22.1% > semisupervised = 13.9% > supervised = 6.1%, least significant difference = 5.3%).

A more interesting test was whether model types differed in predictive ability according the cross-validation scenario. Unsupervised models had a mean AUC (0.852) significantly greater than that of semisupervised models (0.796) or supervised models (0.788), but the latter two did not differ significantly (least significant difference = 0.036, $\alpha = 0.05$, Fig. 2). The results were equivalent for kappa: unsupervised models were significantly superior (kappa = 0.561) to both semisupervised and supervised models (kappa = 0.460 and 0.450, respectively). Semisupervised and supervised models did not differ (least significant difference = 0.068, $\alpha = 0.05$, Fig. 2). Repeated-measures ANOVAs, considering only the cross-validation scenario, agreed with these results, indicating a significant difference in predictive ability among model types (AUC: $F_{2,45} = 12.8$, $p < 0.0001$; kappa: $F_{2,45} = 10.4$, $p = 0.0002$) and no significant differences between areas (AUC: $F_{1,45} = 0.87$, $p = 0.35$; kappa: $F_{1,45} = 1.8$, $p = 0.18$) or in the model type–area interaction (AUC: $F_{2,45} = 0.48$ $p = 0.62$; kappa: $F_{2,45} = 0.46$, $p = 0.62$).

In the extrapolation scenario, predictive ability did not differ significantly among model types according to AUC (unsupervised 0.678, semisupervised 0.694, and supervised 0.703; least significant difference = 0.036; $\alpha = 0.05$) and according to kappa (unsupervised 0.260, semisupervised 0.299, and supervised 0.289; least significant difference = 0.068; $\alpha = 0.05$; Fig. 2). Repeated-measures ANOVAs considering only the extrapolation scenarios were in agreement with these results, indicating no significant difference in predictive ability among model types (AUC: $F_{2,45} = 0.56$, $p = 0.57$; kappa: $F_{2,45} = 0.44$, $p = 0.64$).There was a slightly significant difference, however, between areas (AUC: $F_{1,45} = 4.35$, $p = 0.04$; kappa: $F_{1,45} = 4.39$, $p = 0.04$), and no significant interaction was seen (AUC: $F_{2,45} = 2.9$, $p = 0.06$; kappa: $F_{2,45} = 3.09$, $p = 0.06$). On average, models from Aracena extrapolated slightly better than those from Grazalema.

Discussion

Empirical models are typically used in three situations. First, they are used to summarize the information that was used to build them, that is, to explain the pattern in data (e.g., MacNally 2000). Second, models are used to predict a response given a new data set, with values of the predictors that must be in the range observed in the original data (see review in Guisan & Zimmermann 2000). Third, models are used to extrapolate to spatial areas (or temporal contexts) different from the ones in which they were built (Schröder & Richter 1999/2000; Bonn & Schröder 2001). In the case of predictive modeling of species distribution to construct potential distribution maps, one is usually interested in the second and third uses.

Whatever the scenario, empirical models may potentially benefit from expert opinion incorporated in some of the stages in the process of model building (Pearce et al. 2001). Our unsupervised models in fact incorporated some degree of expert opinion because the explanatory variables derived from the thematic digital cartography were selected with the prediction of bird distribution in mind. The distinction that we made between unsupervised, semisupervised, and supervised models, however, reflects a common situation very relevant to predictive habitat modeling practitioners.

Supervised models are expected to have a greater predictive ability than unsupervised models because a thoughtful selection of predictors is expected to reduce the potential inclusion of spurious relationships. They should also be able to extrapolate with less loss of accuracy because it is expected that a supervised model will more frequently include causal relationships between birds and their habitats and fewer spurious correlations.

Indeed, our results suggest that the predictive performance of unsupervised models is acceptable, being as good as or better than that of supervised models. Our unsupervised models had a greater tendency to overfit the data, as evidenced by the greater number of predictors included in the models and by the reduction in the estimate of predictive ability when comparing the building with the cross-validation scenarios. But this overfitting had no negative effect on the predictive ability of the models. In fact, unsupervised model showed a trend toward higher discrimination than the supervised models in the cross-validation scenario. This trend suggests that some of the variables disregarded by the expert in the model-fitting stage had some predictive power, at least at the resolution of our study.

Unsupervised models had a mean 0.8 cross-validation AUC, which means they classified correctly 8 in 10 pairs of presence/absence observations (Hanley & McNeil 1982). This is a fair result according to the standards posed by several authors (Monserud & Leemans 1992; Fielding & Bell 1997; Pearce & Ferrier 2000*a*), who suggest that model performance is poor when AUC is below 0.7 (or kappa is below 0.4), fair when it is between within 0.7 and 0.9, and good for values >0.9 (kappa >0.7). This discriminatory ability is similar to what is commonly found in the wildlife-modeling literature (Manel et al. 1999; Tobalske & Tobalske 1999; Cumming 2000; Bonn & Schröder 2001), and may reflect an upper limit in the accuracy of empirical models based on indirect variables (Guisan & Zimmermann 2000) to predict a response at a high spatial resolution (Fielding & Haworth 1995; Manel et al. 1999; Rico Alcázar et al. 2001). A few of the variables included in our unsupervised models were difficult to interpret in relation to the ecology of the species or suggested relationships opposite of what we expected according to our previous knowledge of habitat selection (e.g., the probability of Red-legged Partridge presence increased with distance to small patches of cropland).

The discriminatory ability of the models evaluated with external data (the extrapolation scenario) was low, although it was significantly better than that of a null model in all cases. This implies that transferring our models to neighboring areas is not advisable. Supervised models were not significantly better than unsupervised ones, suggesting that, in our situation, careful design of models to reduce the possibility of including spurious predictors did not improve their extrapolation ability. Our study areas were geographically close and similar in topography, climatology, and landscape, at least at the coarse level of variation measured in our GIS. Consequently, we expected a higher success in the extrapolation of models. This failure may be attributable to unmodeled historical factors or local processes, such as intra- and interspecific interactions, that finely adjust the habitat distribution of organisms. Currently these factors constitute an unsolved problem in wildlife-habitat relationship modeling.

Our results suggest that there is not much to gain with careful implementation of expert criteria into model building when the aim is producing maps of potential species distribution. It could be argued that perhaps our expert was not such an expert and that our supervised models were ill designed, but we were working with common bird species, and our expert was an ornithologist with 2 years of local field experience previous to building the models. This situation is much better than could be expected on average if one were building distribution models for other taxa (e.g., invertebrates) in a remote region. It is also true that distribution models can be built in many different ways (e.g., neural networks, classification and regression trees, and ordination techniques; Scott et al. 2002); that there are different ways of implementing expert criteria apart from the one we tested (e.g., habitat suitability indexes developed by experts, use of multiple experts, and literature review on the species); and that expert criteria can be incorporated at different stages of

the model building process (Pearce et al. 2001). It is important to test whether our results can be generalized to these other situations.

Incorporating expert opinion in the predictive-modeling process is time consuming and, as our results demonstrate, may not produce better predictions. Even when applied to neighboring areas, models benefiting from expert opinion did not outperform pure unsupervised models. Therefore, we conclude that unsupervised fitting procedures for building predictive habitat models seems to be an adequate, cost-effective way to proceed if the aim is generating habitat suitability maps in a regional context.

Acknowledgments

This work is a contribution to the project 1DF-97-0648 Predictive Cartography of Land Birds: A Pilot Study in Western Andalusia, funded by the Dirección General de Enseñanza Superior e Investigación Científica (Ministry of Science and Technology, Spain) and European Regional Development Funds (fondos FEDER, European Union). During the work, J.S. had a predoctoral fellowship from the Ministry of Education and Culture. The extensive fieldwork presented in this paper could not have been done without the help and sense of humor of D. López Huertas, L. M. Carrascal, and M. Díaz, who also brought intellectual insight to the project.

Literature Cited

Altman, D. G., and P. Royston. 2000. What do we mean by validating a prognostic model? Statistics in Medicine **19**:453–473.

Bonn, A., and B. Schröder. 2001. Habitat models and their transfer for single and multi species groups: a case study of carabids in an alluvial forest. Ecography **24**:483–496.

Brown, D. G. 1994. Predicting vegetation types at treeline using topography and biophysical disturbance variables. Journal of Vegetation Science **5**:641–656.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference. A practical information-theoric approach. Springer-Verlag, New York.

Bustamante, J. 2004. Cartografía predictiva de variables climáticas: comparación de distintos modelos de interpolación de la temperatura en España Peninsular. Graellsia **59**:359–376 (in Spanish).

Bustamante, J., J. A. Donázar, F. Hiraldo, O. Ceballos, and A. Travaini. 1997. Differential habitat selection by immature and adult Grey Eagle-buzzards *Geranoaetus melanoleucus*. Ibis **139**:322–330.

Cumming, G. S. 2000. Using between-model comparisons to fine-tune linear models of species ranges. Journal of Biogeography **27**:441–455.

Díaz, M., J. C. Illera, and D. Hedo. 2001. Strategic environmental assessment of plans and programs: a methodology for estimating effects on biodiversity. Environmental Management **28**:267–279.

Eastman, J. R. 1997. Idrisi for Windows. User's guide. Clark Labs, Clark University, Worcester, Massachusetts.

Eastman, J. R. 1999. Idrisi32. Reference guide. Clark Labs, Clark University, Worcester, Massachusetts.

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.

Fielding, A. H., and P. F. Haworth. 1995. Testing the generality of bird-habitat models. Conservation Biology **9**:1466–1481.

Franklin, J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. Journal of Vegetation Science **9**:733–748.

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

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**:29–36.

Harrell, F. E. 2001. Regression modeling strategies. Springer-Verlag, New York.

Hastie, T. J., and R. J. Tibshirani. 1990. Generalized additive models. Chapman & Hall, London.

Heller-Kellenberger, I., F. Kienast, M. Obrist, and T. Walter. 1997. Räumliche Modellierung der potentiellen faunistischen Biodiversität mit einem Expertensystem. Informationsblatt des Forschungsbereiches Landschaftsökologie **36**. Swiss Federal Institute for Forest, Snow and Landscape Research Birmensdorf (in German). Available from http://www.wsl.ch/land/infoblatt/Nr36/Info36.html (accessed June 2004).

James, F. C., and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annual Review of Ecology and Systematics **21**:129–166.

Lezzoni, L. I. 1999. Statistically derived predictive models: caveat emptor. Journal of General Internal Medicine **14**:388–389.

Mac Nally, R. 2000. Regression and model-building in conservation biology, biogeography and ecology: The distinction between—and reconciliation of—'predictive' and 'explanatory' models. Biodiversity and Conservation **9**:655–671.

Manel, S., J. M. Dias, S. T. Buckton, and S. J. Ormerod. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology **36**:734–747.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology **38**:921–931.

MathSoft, I. 1999. S-Plus 2000 User's guide. Data Analysis Products Division, Seattle, Washington.

McCullagh, P., and J. A. Nelder. 1989. Generalised linear modelling. 2nd edition. Chapman & Hall, London.

McKenney, D. W., L. A. Venier, A. Heerdegen, and M. A. McCarthy. 2002. A Monte Carlo experiment for species mapping problems. Pages 377–381 in J. M. Scott, P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, editors. Predicting species occurences: issues of accuracy and scale. Island Press, Washington, D.C.

Monserud, R. A., and R. Leemans. 1992. Comparing global vegetation maps with the kappa statistic. Ecological Modelling **62**:275–293.

Murtaugh, P. A. 1996. The statistical evaluation of ecological indicators. Ecological Applications **6**:132–139.

Nicholls, A. O. 1989. How to make biological surveys go further with generalised linear models. Biological Conservation **50**:51–75.

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science **263**:641–646.

Osborne, P. E., and B. J. Tigar. 1992. Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. Journal of Applied Ecology **29**:55–62.

Pearce, J., and S. Ferrier. 2000*a*. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling **133**:225–245.

Pearce, J., and S. Ferrier. 2000*b*. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecological Modelling **128**:127–147.

Pearce, J. L., K. Cherry, M. Drielsma, S. Ferrier, and G. Whish. 2001. Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. Journal of Applied Ecology **38**:412–424.

Pons, X. 2000. MiraMon: geographic information system and remote sensing software. Universidad Autónoma de Barcelona, Barcelona.

Rencher, A. C., and F. C. Pun. 1980. Inflation of R^2 in best subset regression. Technometrics **22**:49–53.

Rexstad, E. A., D. D. Miller, C. H. Flatter, E. M. Anderson, J. W. Hupp, and D. R. Anderson. 1988. Questionable multivariate statistical inference in wildlife habitat and community studies. Journal of Wildlife Management **52**:794–798.

Rico Alcázar, L., J. A. Martínez, S. Morán, J. R. Navarro, and D. Rico. 2001. Preferencias de hábitat del Águila-azor Perdicera (*Hieraaetus fasciatus*) en Alicante (E de España) a dos escalas espaciales. Ardeola **48**:55–62 (in Spanish).

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike Information Criterion statistics. KTK Scientific Publishers, Tokyo.

Schröder, B., and O. Richter. 1999/2000. Are habitat models transferable in space and time? Zeitschrift für Ökologie und Naturschutz **8**:195–205.

Scott, J. M., et al. 1993. GAP analysis: a geographic approach to protection of biological diversity. Wildlife Monographs **123**.

Scott, J. M., P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, editors. 2002. Predicting species occurrences: issues of scale and accuracy. Island Press, Washington, D.C.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry: the principles and practice of statistics in biological research. W. H. Freeman, San Francisco.

Stauffer, D. F. 2002. Linking populations and habitats: where have we been? Where are we going? Pages 53–61 in J. M. Scott, P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, editors. Predicting species occurences: issues of accuracy and scale. Island Press, Washington, D.C.

Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sampling size on accuracy of species distribution models. Ecological Modelling **148**:1–13.

Titus, K., J. A. Mosher, and B. K. Williams. 1984. Chance-corrected classification for use in discriminant analysis: ecological applications. The American Midland Naturalist **111**:1–7.

Tobalske, C., and B. W. Tobalske. 1999. Using atlas data to model the distribution of woodpecker species in the Jura, France. The Condor **101**:472–483.

U.S. Geological Survey GAP Analysis Program. 2000. A handbook for conducting gap analysis. University of Idaho, Moscow. Available from http://www.gap.uidaho.edu/handbook (accessed February 2000).

Vida, S. 1993. A computer program for non-parametric receiver operating characteristic analysis. Computer Methods and Programs in Biomedicine **40**:95–101.

Wiens, J. A. 2002. Predicting species ocurrences: progress, problems and prospects. Pages 739–749 in J. M. Scott, P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, editors. Predicting species occurences: issues of accuracy and scale. Island Press, Washington, D.C.

Zweig, M. H., and G. Campbell. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry **39**:561–577.