Journal of Applied Ecology 2004
41, 811–823

The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact?

JANA M. McPHERSON,*† WALTER JETZ†‡§ and DAVID J. ROGERS†

†Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK; ‡Ecology and Evolutionary Biology Department, Princeton University, Princeton, NJ 08544–1003, USA; and §Biology Department, University of New Mexico, Albuquerque, NM 87131, USA

Summary

1. Conservation scientists and resource managers increasingly employ empirical distribution models to aid decision-making. However, such models are not equally reliable for all species, and range size can affect their performance. We examined to what extent this effect reflects statistical artefacts arising from the influence of range size on the sample size and sampling prevalence (proportion of samples representing species presence) of data used to train and test models.

2. Our analyses used both simulated data and empirical distribution models for 32 bird species endemic to South Africa, Lesotho and Swaziland. Models were built with either logistic regression or non-linear discriminant analysis, and assessed with four measures of model accuracy: sensitivity, specificity, Cohen's kappa and the area under the curve (AUC) of receiver-operating characteristic (ROC) plots. Environmental indices derived from Fourier-processed satellite imagery served as predictors.

3. We first followed conventional modelling practice to illustrate how range size might influence model performance, when sampling prevalence reflects species' natural prevalences. We then demonstrated that this influence is primarily artefactual. Statistical artefacts can arise during model assessment, because Cohen's kappa responds systematically to changes in prevalence. AUC, in contrast, is largely unaffected, and thus a more reliable measure of model performance. Statistical artefacts also arise during model fitting. Both logistic regression and discriminant analysis are sensitive to the sample size and sampling prevalence of training data. Both perform best when sample size is large and prevalence intermediate.

4. *Synthesis and applications.* Species' ecological characteristics may influence the performance of distribution models. Statistical artefacts, however, can confound results in comparative studies seeking to identify these characteristics. To mitigate artefactual effects, we recommend careful reporting of sampling prevalence, AUC as the measure of accuracy, and fixed, intermediate levels of sampling prevalence in comparative studies.

Key-words: discriminant analysis, kappa, logistic regression, prevalence, ROC plots, sample size, satellite imagery

Journal of Applied Ecology (2004) **41**, 811–823

Introduction

Confronted by current threats to biodiversity and the difficulty of obtaining detailed, repeated species inventories for much of the world, biologists rely increasingly on distribution models to inform conservation strategies. Distribution models predict species richness (Jetz & Rahbeck 2002), centres of endemism (Johnson, Hay & Rogers 1998), the occurrence of particular species assemblages (Neave, Norton & Nix 1996) or individual species (Gibson *et al.* 2004), and the breeding habitat (Osborne, Alonso & Bryant 2001), breeding

*Present address and correspondence: Jana M. McPherson, 340 Keewatin Avenue, Toronto, Ontario, Canada M4P 2A5 (e-mail jana.mcpherson@zoology.oxford.ac.uk).

success (Paradis *et al*. 2000), abundance (Jarvis & Robertson 1999) and genetic variability (Scribner *et al*. 2001) of species.

Such models do more than fill gaps in distribution maps. By delineating favourable habitats, distribution models can help target field surveys (Engler, Guisan & Rechsteiner 2004), aid in the design of reserves (Li *et al*. 1999), inform wildlife management outside protected areas (Milsom *et al*. 2000) and guide mediatory actions in human–wildlife conflicts (Sitati *et al*. 2003). Distribution models can be used to monitor declining species (Osborne, Alonso & Bryant 2001), predict range expansions of recovering species (Corsi, Dupre & Boitani 1999), estimate the likelihood of species' long-term persistence in areas considered for protection (Cabeza *et al*. 2004) and identify locations suitable for reintroductions (Joachim *et al*. 1998). They allow biologists to identify sites vulnerable to local extinction (Gates & Donald 2000) or species invasion (Kriticos *et al*. 2003), and to explore the potential consequences of climate change (Peterson *et al*. 2002).

Distribution models will always perform better for some taxa than for others (Venier *et al*. 1999). To maximize their utility, we need to understand whether the variation in performance reveals inherent ecological differences in a species' predictability or whether it reflects statistical artefacts.

Range size is one ecological characteristic, likely to differ from species to species, that might influence the success of distribution models (Venier *et al*. 1999; Manel, Williams & Ormerod 2001; Stockwell & Peterson 2002). Such influence could have ecological roots. Species with large ranges or disjunctive distributions, for example, may exhibit subspecific variation in habitat associations because of local adaptation (Stockwell & Peterson 2002). To an automated model-fitting algorithm, such disjoint habitat preferences could appear statistically incoherent and therefore less predictable. Poor performance of models for narrow-ranging species may instead have methodological roots. Their habitat associations may be perfectly coherent at fine spatial scales, but may not manifest themselves at the spatial grain of analysis (Fielding & Haworth 1995).

Variation in model performance with species' range sizes might equally, however, reflect biases inherent in the modelling process. Range size can measure a species' extent of occurrence or its area of occupancy. Where range size measures area of occupancy, it will affect either sampling prevalence (the proportion of data points representing a species' presence) or sample size (the total number of data points, presence plus absence) in the data sets used to train (parameterize) and/or evaluate models. Both sampling prevalence (Fielding & Haworth 1995; Manel, Dias & Ormerod 1999; Cumming 2000; Olden, Jackson & Peres-Neto 2002) and sample size (Hendrickx 1999; Cumming 2000; Pearce & Ferrier 2000b; Stockwell & Peterson 2002) have been shown to influence the performance of distribution models independently of range size.

None the less, range size and prevalence are often confounded because sampling prevalence is allowed to vary with a species' 'natural' prevalence, i.e. local range size or the proportion of study sites occupied by the species (Manel, Williams & Ormerod 2001; Pearce, Ferrier & Scotts 2001; Kadmon, Farber & Danin 2003). Consequently, it is difficult to distinguish real ecological phenomenon from statistical artefact. Furthermore, it remains unclear where within the modelling procedure sample size and sampling prevalence exert their artefactual effects, and whether these effects could be avoided.

Biases could arise at two points during modelling: (i) the process of model fitting and (ii) the assessment of model performance with accuracy metrics. Among model-fitting algorithms, logistic regression, for example, is thought to bias its results towards the more prevalent category (presence or absence) (Fielding & Bell 1997). Similarly, the matching coefficient, a widely used accuracy metric, has been shown both mathematically (Henderson 1993; Fielding & Bell 1997) and empirically (Manel, Williams & Ormerod 2001; Olden, Jackson & Peres-Neto 2002) to be affected by prevalence.

We sought to address two questions. (i) To what extent does variation in model performance with species' range sizes represent statistical artefacts or ecologically meaningful patterns? (ii) Can we minimize the risk of artefacts through an informed choice of model algorithm and accuracy metric? To provide answers, we conducted three analyses using both simulated data and empirical distribution models of southern African birds based on Fourier-processed satellite data. We tested two algorithms widely used in ecological modelling: logistic regression and discriminant analysis.

Analysis 1 examined how range size will appear to influence model performance if potential artefacts are ignored.

Analysis 2 tested whether statistical artefacts relating to range size could arise at the model assessment stage. We scrutinized two increasingly popular measures of model accuracy, Cohen's kappa and the area under the curve (AUC) of receiver-operating characteristic (ROC) plots. Both have recently been advocated in the ecological literature, primarily due to their perceived independence or near-independence from prevalence (Fielding & Bell 1997; Pearce & Ferrier 2000a; Manel, Williams & Ormerod 2001).

Analysis 3 investigated whether statistical artefacts arise during the process of model fitting. Subsampling is used to decouple sample size and sampling prevalence from range size in the data sets used to train and test models, allowing us to examine the independent effect of either factor on model performance.

We discuss our findings in the context of both the ecological and epidemiological literature. Particularly concerned about the implications for comparative studies, we conclude with a number of recommendations for both the producers and users of distribution models.

Materials and methods

BIRD DISTRIBUTION DATA

We built distribution models for 32 bird species endemic or near-endemic to South Africa, Lesotho and Swaziland (for a list of species see Table 1). Distribution data for these species were taken from *The Atlas of Southern African Birds* (Harrison *et al*. 1997), with a few records added from *The Atlas of Birds of Sul do Save, Southern Mozambique* (Parker 1999). Both were provided in electronic format by the Avian Demography Unit, University of Cape Town, Cape Town, South Africa. These data have a spatial resolution of 0·25° longitude–latitude (quarter-degree squares, QDS), representing an area of approximately 24 km (east–west) by 27 km (north–south) at the latitude of South Africa. The number of QDS occupied by each species served as a measure of range size.

ENVIRONMENTAL DATA

Environmental information was derived from satellite images collected twice daily over an 18-year period (1982–99) by the National Oceanic and Atmospheric Administration's (NOAA, USA; http://www.noaa.gov) advanced high resolution radiometer satellite series. Environmental information obtained from these images included a middle infra-red signal, indices of land surface temperature, air temperature, the vapour pressure deficit, and the normalized difference vegetation index. A further index, cold cloud duration, was derived from 10 years (1989–98) of European Meteosat imagery (Hay 2000). All imagery was composited into cloud-free, monthly images and resampled from its original spatial resolution of 8 km^2 to the 0·25° resolution of bird distribution data. For each environmental index, we used temporal Fourier analysis, a data reduction technique ideal for summarizing seasonal variables (Chatfield 1996; Rogers, Hay & Packer 1996), to extract the overall mean, minimum, maximum and variance, plus the amplitude (strength) and phase (timing) of annual, biannual and triannual cycles. Furthermore, altitude, derived from a US Geological Survey's global digital elevation model, was included among the explanatory variables, yielding a total of 61 predictors.

MODEL ALGORITHMS

We tested logistic regression (LR) and non-linear discriminant analysis (DA). In LR, training data serve to

Table 1. Names, range sizes and 'natural' prevalence of 32 endemic bird species whose distributions were modelled in analyses 1, 2 and 3 as indicated. Range size measures the number of quarter-degree squares (QDS) occupied by each species. In total, the study area included 4275 QDS

Common name	Scientific name	Family	Range size	Prevalence	Analysis
Mountain pipit	*Anthus hoeschi*	Passeridae	28	0·006	1
Knysna scrub-warbler	*Bradypterus sylvaticus*	Sylviidae	36	0·008	1
Yellow-breasted pipit	*Anthus chloris*	Passeridae	44	0·010	1
Ferruginous lark	*Certhilauda burra*	Alaudidae	45	0·011	1
Drakensberg siskin	*Serinus symonsi*	Fringillidae	49	0·011	1
Rufous rock-jumper	*Chaetops frenatus*	Picathartidae	51	0·012	1
Victorin's scrub-warbler	*Bradypterus victorini*	Sylviidae	64	0·015	1
Protea seedeater	*Serinus leucopterus*	Fringillidae	74	0·017	1
Orange-breasted rock-jumper	*Chaetops aurantius*	Picathartidae	80	0·018	1
Blackcap mountain-babbler	*Lioptilus nigricapillus*	Sylviidae	84	0·020	1
Knysna woodpecker	*Campethera notata*	Picidae	108	0·025	1
Brown scrub-robin	*Cercotrichas signata*	Muscicapidae	113	0·026	1
Cape siskin	*Serinus totta*	Fringillidae	128	0·029	1
Orange-breasted sunbird	*Nectarinia violacea*	Nectariniidae	138	0·032	1
Forest buzzard	*Buteo trizonatus*	Accipitridae	149	0·034	1
Cape sugarbird	*Promerops cafer*	Nectariniidae	150	0·035	1
Melodious lark	*Mirafra cheniana*	Alaudidae	160	0·037	1
Chorister robin-chat	*Cossypha dichroa*	Muscicapidae	215	0·050	1
Forest canary	*Serinus scotops*	Fringillidae	215	0·050	1
Buff-streaked wheatear	*Oenanthe bifasciata*	Muscicapidae	227	0·053	1
Cape francolin	*Pternistis capensis*	Phasianidae	232	0·054	1
Yellow-tufted pipit	*Anthus crenatus*	Passeridae	266	0·062	1
Sentinel rock-thrush	*Monticola exploratory*	Muscicapidae	282	0·066	1, 2
Southern tchagra	*Tchagra tchagra*	Corvidae	303	0·071	1, 2
Blue bustard	*Eupodotis caerulescens*	Otididae	365	0·085	1, 2
Grey-winged francolin	*Scleroptila africanus*	Phasianidae	493	0·115	1, 2, 3
Ground woodpecker	*Geocolaptes olivaceus*	Picidae	494	0·115	1, 2, 3
Cape rock-thrush	*Monticola rupestris*	Muscicapidae	586	0·137	1, 2, 3
Southern double-collared sunbird	*Nectarinia chalybea*	Nectariniidae	680	0·159	1, 2, 3
Large-billed lark	*Galerida magnirostris*	Alaudidae	692	0·162	1, 2, 3
Cape weaver	*Ploceus capensis*	Passeridae	927	0·217	1, 2, 3
African pied starling	*Spreo bicolour*	Sturnidae	1167	0·273	1, 2, 3

establish what proportions of cases are positive (represent species' presence) at each value of the explanatory variables (Agresti 1996). A logit link transforms a linear function of predictors into response values between 0 and 1, representing the probability of occurrence of the modelled event, here species' presence (Legendre & Legendre 1998). Analyses were performed in SPlus (Insightful™ 2001); variables were selected in a forward stepwise fashion based on their ability to reduce the Akaike information criterion (AIC), a measure of model fit and parsimony (Sakamoto, Ishiguro & Kitagawa 1986). Automated stepwise variable selection, although much criticized, was applied here to reflect its wide use in distribution modelling.

In DA, training data serve to determine the multivariate mean and variance–covariance structure of predictor variables for each of the response variable's states, here presence and absence. The distribution of predictor variables is assumed to be normal, but their covariance need not be the same for all states in nonlinear DA (Rogers, Hay & Packer 1996). The posterior probability of any data point belonging to one response state or another is then calculated based on its position in *n*-dimensional space relative to each state's multivariate mean, where distance between sample point and mean is measured as Mahalanobis distance (Green 1978; Rogers, Hay & Packer 1996). For presence–absence data DA thus predicts the probability of occurrence. Nonlinear DA was implemented using custom-written programmes in QuickBasic (Microsoft®). Ten predictor variables were selected in forward stepwise fashion based on their ability to maximize training accuracy as measured by kappa (see below). Ten variables was just less than the number picked, on average, in LR models using the AIC (mean = 11; *n* = 770).

MEASURES OF MODEL ACCURACY

We focused on two measures of accuracy: Cohen's kappa and AUC of ROC plots. To facilitate comparison with other studies we also reported on sensitivity and specificity.

Sensitivity quantifies the proportion of observed presences correctly predicted as presence (the true positive fraction). Poor sensitivity therefore indicates many omission errors, i.e. erroneous predictions of absence. Conversely, specificity measures the proportion of observed absences correctly predicted as absence (the true negative fraction). Low specificity signals high commission error, i.e. erroneous predictions of presence (Fielding & Bell 1997). Both measures are mathematically independent of prevalence, because they are expressed as a proportion of all the sites with a given observed state (i.e. presence or absence; Pearce & Ferrier 2000a). None the less, these measures can be misleading. Each simply reflects how well the model predicts one category (presence or absence) without indicating how many mistakes are made in the other. Chance alone could lead to high sensitivity for particularly prevalent species or high specificity for very rare species (Olden, Jackson & Peres-Neto 2002).

		Observed		
		Present	**Absent**	**Total**
Predicted	**Present**	true positives	false positives (commission)	No. of predicted presences
	Absent	false negatives (omission)	true negatives	No. of predicted absences
	Total	No. of observed presences	No. of observed absences	*N* = total No. of observations

Fig. 1. A confusion matrix, which tabulates model results as shown.

In contrast, kappa and AUC are 'omnibus measures', designed to reflect model performance in absence and presence simultaneously (Cicchetti & Feinstein 1990). Kappa records overall agreement between predictions and observations, corrected for agreement expected to occur by chance. The statistic ranges from –1 to +1, where +1 indicates perfect agreement while values of zero or less suggest a performance no better than random (Cohen 1960). Although kappa has been reported to show some sensitivity towards prevalence, this effect has been judged negligible among ecologists (Fielding & Bell 1997; Manel, Williams & Ormerod 2001).

Kappa, sensitivity and specificity all derive from a confusion matrix (Fig. 1). Their calculation therefore requires that probabilistic predictions of occurrence be divided into concrete predictions of absence or presence, based on a single, potentially arbitrary classification threshold, here 0·5.

The area under ROC curves instead is a threshold-independent measure of model accuracy, juxtaposing correct and incorrect predictions over a range of thresholds. It ranges from 0 to 1, with values larger than 0·5 indicating a performance better than random (Fielding & Bell 1997). AUC was here calculated non-parametrically using the Wilcoxon statistic (Hanley & McNeil 1982; Pearce & Ferrier 2000a). ROC plots are thought to be independent of prevalence, because the true positive and false positive fractions determining their curve are each expressed as a proportion of all sites with a given observed state (Zweig & Campbell 1993).

ANALYSIS 1: IGNORING POTENTIAL ARTEFACTS

Distribution modelling often involves a fixed geographical study area or number of field locations from which data to train models are drawn. Consequently, total sample size is constant across species. The relative frequency of positive samples in training and test data (sampling prevalence) is determined by each species' natural prevalence, i.e. the proportion of study sites occupied by the species (Manel *et al*. 1999; Manel, Williams & Ormerod 2001; Pearce *et al*. 2001).

Our first analysis took the same approach to building distribution models for 32 bird species endemic to

South Africa, Lesotho and Swaziland. All QDS on the African mainland south of 19°S were considered study sites. For each species, data were split into test and training data sets in a geographically systematic fashion: moving west to east and north to south, every third absence and every third presence site was set aside as independent test data (1425 sites in total). All remaining sites (2850) served as model training. Both data sets covered the species' entire geographical spread, and in both the ratio of positive (presence) to negative (absence) samples reflected natural prevalence. Models built with LR and DA used training data and satellite-derived environmental indices to predict species' occurrences across the entire study region. These models were evaluated with test data.

ANALYSIS 2: INHERENT BIASES IN MEASURES OF ACCURACY

Bias in model performance with respect to prevalence could arise during model assessment if the measure of accuracy used is affected by the ratio of positive to negative cases in the sample.

Whether prevalence exerts such direct effects on kappa was assessed with simulated data, consisting of confusion matrices with three controlled characteristics.

Prevalence, here the proportion of cases simulating observed presence, was implemented at 21 levels: 0·01, 0·05–0·95 in increments of 0·05, and 0·99.

Total classification error, i.e. the percentage of cases simulating prediction errors, took one of seven values: 1%, 2%, 5%, 10%, 15%, 25% or 50%

Error allocation, i.e. the relative frequency of false positive and false negative errors, was either balanced, with misclassification of presence and absence proportional to prevalence, or biased. Bias towards error in presence (more false negatives) or absence (more false positives) was simulated at three levels: error in the chosen category exceeded the error expected in a balanced situation by 5%, 10% or 20% where this was possible without changing either total error or prevalence.

For each feasible combination of prevalence, total error and error allocation, a customized programme (in QuickBasic) randomly constructed 100 different confusion matrices. The total number of cases per confusion matrix (n) was allowed to vary, as preliminary investigations had indicated that n had no effect. To ensure that all components of the confusion matrix consisted of integers, however, n was set to be a multiple of 100, between 100 and 20 000. Kappa was calculated for all 93 100 confusion matrices created.

Given the threshold-independent nature of ROC curves, simulated confusion matrices could not be used to test the effects of prevalence on AUC. Instead, we created simulated test data sets by subsampling response surfaces produced in analysis 1 by both LR and DA. To ensure a sufficient number of presence localities, we chose predictions for the 10 most wide-ranging endemics. Each simulated data set consisted of 100 sites picked randomly (among 4275), but such that observed the species prevalence matched one of 21 levels of prevalence (as above). For each level of prevalence, 100 simulated data sets were created, yielding 42 000 in total. AUC was calculated for each.

To test the effects of sample size on both kappa and AUC, the same response surfaces were again subsampled. Simulated data sets contained 25, 50, 75 or 100 sites picked such that the observed species' prevalence was 50%. One-hundred data sets were built per sample size, yielding 8000 in total.

ANALYSIS 3: SAMPLE SIZE AND PREVALENCE EFFECTS ON MODEL FIT

To examine whether sample size and prevalence exerted influence during model fitting, we chose seven endemics (Table 1) that occurred in enough QDS to allow a sufficient range in sample sizes to be tested. Their distributions were repeatedly subsampled to yield training data sets with changing total sample size or changing sampling prevalence. In the first instance, 50, 100, 300 or 500 training locations were sampled with an invariant sampling ratio of 1 presence to 1 absence. In the second instance, total sample size remained constant at 300 but positive samples constituted 12·5%, 25%, 50% or 75% of all training locations. Each sampling regime was repeated 10 times per species.

Sampling was done via a custom-written programme (QuickBasic). Absences were selected at random. To ensure that they reflected environments that individuals of the species might encounter, however, absences were constrained to fall within 6° of the nearest presence record (an admittedly arbitrary threshold, which ideally should reflect species-specific mobility). Presence records were selected such that samples spanned the species' entire geographical range (i.e. depending on the sampling prevalence, every 2nd, 3rd, etc., presence locality was selected for training). Training data were submitted to both LR and DA.

Among the presence localities not used in training, 125 were picked at random and included in a test data set alongside three times as many absence samples (equally not used in training data). This yielded independent test data sets with a constant sampling prevalence (25%) and a constant sample size (500).

Each model was evaluated with both training data (measuring intrinsic accuracy) and test data (measuring extrinsic accuracy). Extrinsic accuracy is a stronger indicator of model performance (Fielding & Bell 1997). Intrinsic accuracy is reported here for three reasons. First, it allows us to examine the null hypothesis that sample size and sampling prevalence do not interfere with model fitting. If so, intrinsic accuracy should reflect only its effect on model assessment (i.e. mimic patterns established in analysis 2) while extrinsic accuracy should show no response as long as the size and prevalence of test data remain constant. Secondly, the divergence between intrinsic and extrinsic accuracy indicates a model's propensity to overfitting (Stockwell & Peterson 2002) and may provide insight into how potential artefacts arise. Overfitting occurs when model parameters reflect random effects in the training data as well as true patterns (Olden, Jackson & Peres-Neto 2002). Finally, the distinction between training and test data is artificial. Models should ultimately predict a species' entire distribution sufficiently well to guide scientists and managers in decision-making.

Consequently, for each species, we computed mean training and mean test accuracy per sampling regime (over the 10 replicate samples). Wilcoxon signed rank tests for matched pairs served to compare the accuracy of the two algorithms. Monotonic relationships between mean accuracy and sample size or sampling prevalence were examined with Spearman rank correlations. To test for non-linear effects, fourth-order polynomial regression models were built. Stepwise variable selection ensured that higher order terms were included only if they reduced the AIC; *t*-tests established whether coefficients of higher order terms differed significantly from zero.

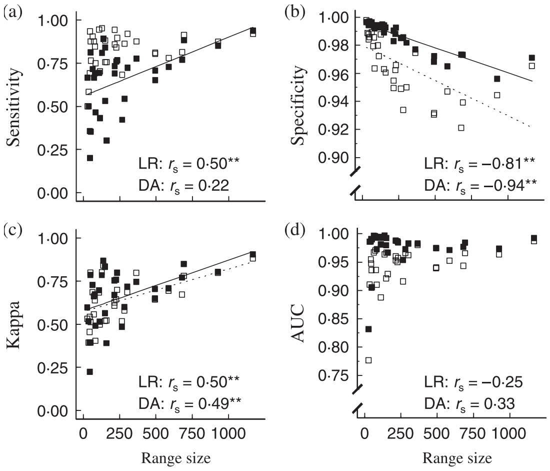

Fig. 2. The relationship between range size (number of occupied QDS) and four measures of extrinsic accuracy in analysis 1. Spearman rank correlation coefficients (r_s) and significant regression lines are displayed for both logistic regression (LR, solid symbols and solid line) and discriminant analysis models (DA, open symbols and dotted line). Significant correlations ($P < 0{\cdot}01$, $n = 32$ species) are indicated (**).

Results

ANALYSIS 1: IGNORING POTENTIAL ARTEFACTS

Natural prevalence of the 32 species in analysis 1 ranged from 0·6% to 27% (Table 1) and clearly affected the predictive power of models. As range size (and therefore sampling prevalence) increased, models tended to become better at predicting presence (i.e. sensitivity improved) but did so significantly only in LR (Fig. 2a). In contrast, their ability to predict absence correctly (specificity) deteriorated in both LR and DA (Fig. 2b). Kappa responded positively to range size in both algorithms (Fig. 2c) but no significant correlation was detected for AUC in either (Fig. 2d).

ANALYSIS 2: INHERENT BIASES IN MEASURES OF MODEL ACCURACY

In confusion matrix simulations, kappa responded to prevalence and error allocation in a systematic, curvilinear fashion (Fig. 3). Maximum kappa values occurred at 50% prevalence. Bias towards errors in presence depressed kappa values at low prevalence (< 50%) but augmented them at high prevalence (> 50%; Fig. 3b). Bias towards errors in absence had the opposite effect (Fig. 3c). This effect of bias was more pronounced when total error was large.

AUC, in contrast, remained invariable with sampling prevalence (Fig. 4a). Its value, however, tended to be unstable when sampling prevalence fell below 20% or above 75% (Fig. 4b). Larger between-species discrepancies in DA than LR corresponded to larger variation in AUC values achieved by DA models in analysis 1 (compare Figs 2d and 4a).

Sample size affected neither mean kappa nor mean AUC per species, but in both metrics standard error increased as sample size shrank (for kappa: $r_s = -0{\cdot}86$ in both LR and DA predictions; for AUC: $r_s = -0{\cdot}63$ in LR, $r_s = -0{\cdot}64$ in DA, with $n = 40$ and $P < 0{\cdot}01$ in all correlations).

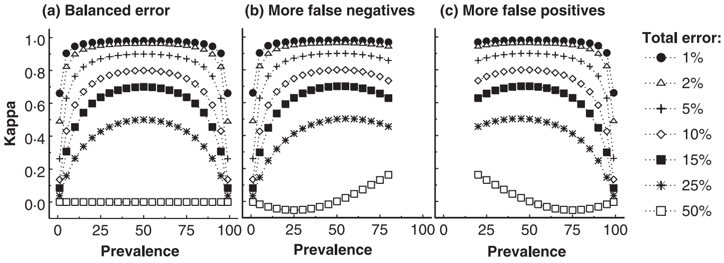

Fig. 3. Direct effects of prevalence on kappa in analysis 2, as modulated by the level of total classification error (see legend) and error allocation. Error allocation was (a) balanced (proportionate to prevalence), or biased (by 20%) towards either (b) more false negatives (more error in the prediction of presence) or (c) more false negatives (more error in the prediction of absence). Each point plotted represents the mean of 100 replicate simulations; standard error was too small for display.

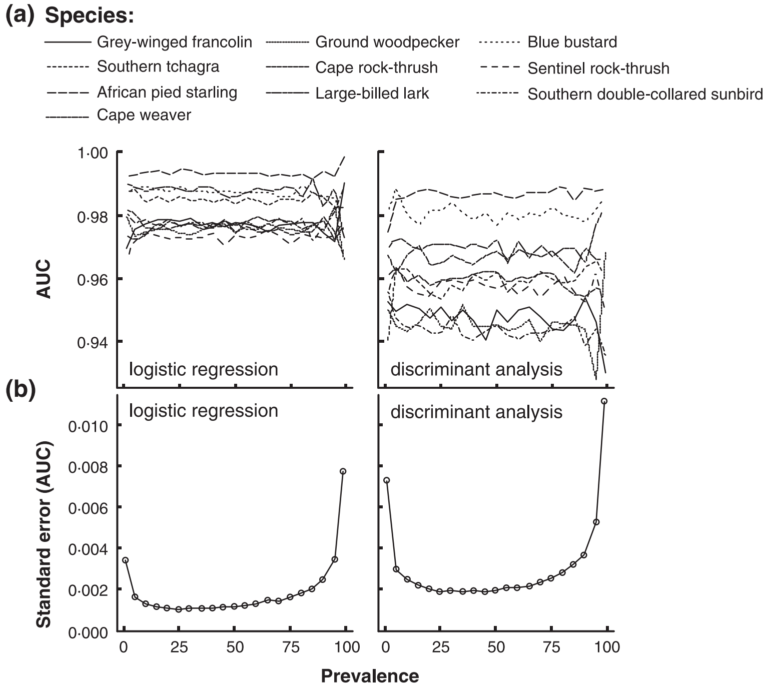

Fig. 4. Direct effects of prevalence on AUC as observed in analysis 2. The response of AUC was tested on predictive distribution models for 10 species, built with either logistic regression (left) or discriminant analysis (right). Mean AUC ($n = 100$ replicate samples) per species showed no systematic effects as sampling prevalence varied (a). Standard error, however, increased at both very low and very high sampling prevalence (b).

Table 2. Comparative performance of logistic regression and discriminant analysis in analysis 3, as indicated by the mean difference in each of four accuracy measures achieved in training data (intrinsic accuracy) and test data (extrinsic accuracy). Significant positive differences (bold, $P < 0{\cdot}05$) indicate better performance in LR, whereas significant negative differences (bold italics, $P < 0{\cdot}05$) show better performance in DA. Significance was assessed using Wilcoxon signed rank tests to compare performance in each sampling regime separately, and all sampling regimes combined. Statistical sample sizes (n) are indicated for each comparison

Sampling regime									
Sample size	50	100	300	500	Invariant sample size of 300				All
Prevalence	Invariant prevalence of 50%				12·5%	25%	50%	75%	regimes
Intrinsic accuracy	*n* = 7				*n* = 7				*n* = 49
Sensitivity	0·00	–0·01	***–0·02***	***–0·02***	0·03	0·00	***–0·02***	0·00	0·00
Specificity	0·00	**0·01**	**0·01**	**0·01**	**0·03**	**0·03**	**0·01**	**0·05**	**0·02**
Kappa	0·00	0·01	–0·01	–0·01	**0·11**	**0·05**	–0·01	**0·04**	**0·03**
AUC	0·00	**0·01**	**0·01**	**0·01**	**0·01**	**0·02**	**0·01**	**0·01**	**0·01**
Extrinsic accuracy	*n* = 7				*n* = 7				*n* = 49
Sensitivity	0·00	–0·03	***–0·02***	–0·01	***–0·05***	***–0·05***	***–0·02***	***–0·02***	***–0·03***
Specificity	***–0·03***	***–0·02***	**0·01**	**0·02**	0·00	**0·01**	**0·01**	0·00	0·00
Kappa	***–0·04***	***–0·05***	0·00	**0·03**	***–0·03***	–0·02	0·00	–0·01	***–0·02***
AUC	0·03	***–0·04***	0·00	**0·01**	–0·02	0·00	0·00	–0·01	0·00

Analysis 3: sample size and prevalence effects on model fit

Both training sample size and sampling prevalence significantly influenced model fit. Visual inspection of predictive maps suggested that increases in training sample size improved fit by reducing both false positive and false negative errors (Fig. 5a). Changes in sampling prevalence gradually shifted error from mostly omission (underprediction of the species' ranges) at low prevalence to mostly commission (overprediction) at higher prevalence (Fig. 5b). The best compromise generally occurred at 50% sampling prevalence.

The influence of training sample size and sampling prevalence was similar in both LR and DA, as indicated by the strong correlations between mean accuracy measures achieved by the two algorithms per species and sampling regime ($0{\cdot}94 \leq r_s \geq 0{\cdot}95$, $P < 0{\cdot}01$ and $n = 98$ for all four measures). A pairwise comparison revealed that LR generally performed better on training data, while DA tended to predict test data more accurately (Table 2). This suggests that LR was more prone to overfitting, potentially reflecting the different variable selection criteria used in LR and DA: the two algorithms tended to agree only on the first one or two predictor variables picked, not subsequent ones. Variable selection among

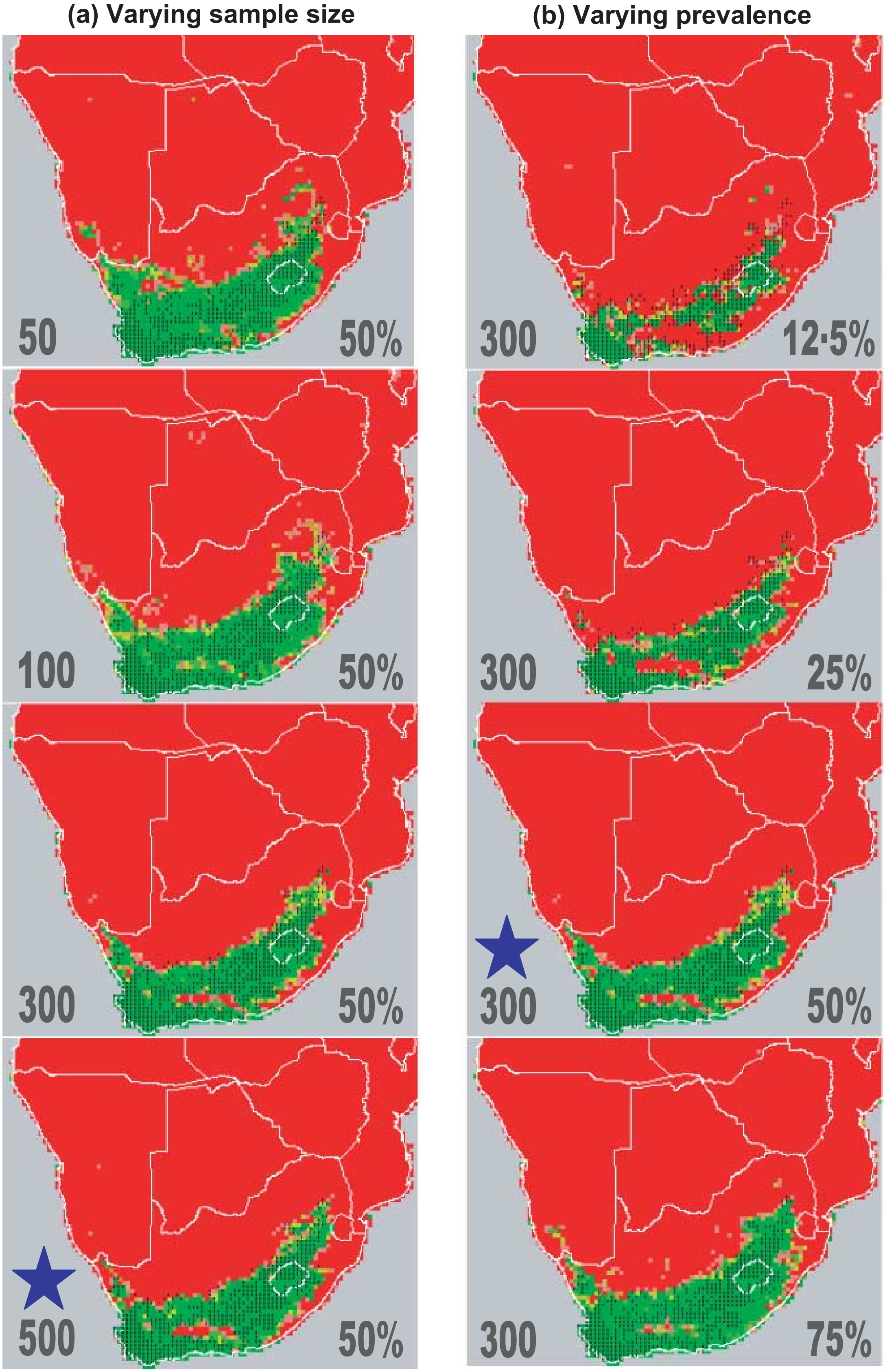

Fig. 5. An example of model predictions obtained in analysis 3. Shown are predictions of logistic regression models for the grey-winged francolin. Predicted probability of occurrence ranges from 0 (red) to 1 (green). The species' observed distribution is marked in black. In (a), training sample prevalence was constant at 50% but sample size varied as indicated in each panel. Larger sample sizes produced a better fit, with the tightest match between observed and predicted distributions at sample size 500 (blue star). In (b), training sample size was constant at 300 but sampling prevalence varied as indicated. Optimum fit (blue star) occurred at 50% prevalence.

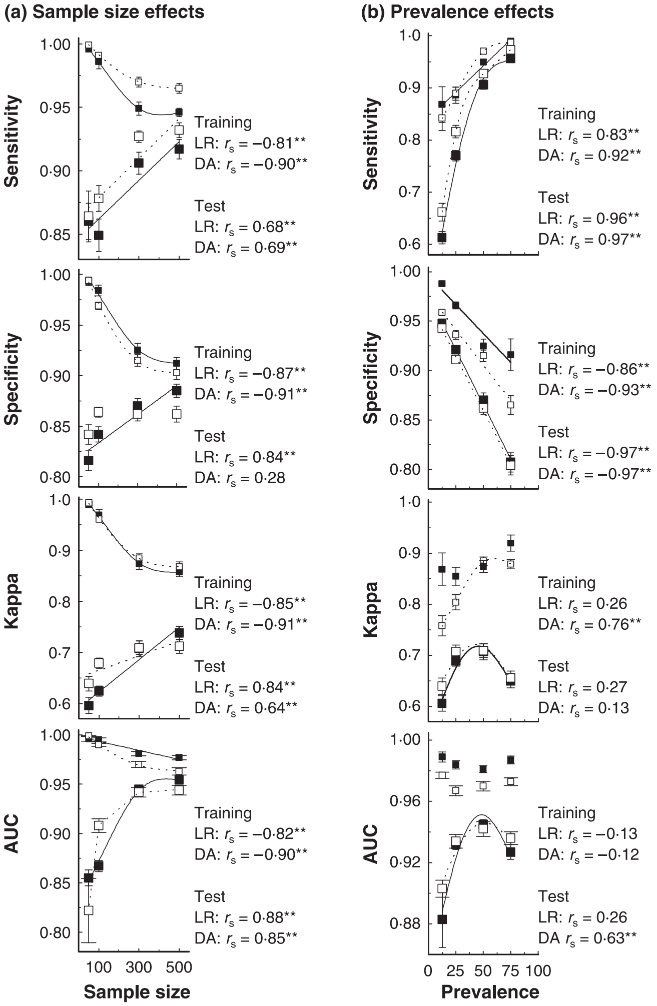

Fig. 6. Variation in model accuracy as observed in analysis 3 in response to changes in (a) training sample size and (b) training sample prevalence. Mean accuracy and standard errors (across seven species) are plotted for sensitivity, specificity, kappa and AUC measured on training data (intrinsic accuracy; small symbols) and test data (extrinsic accuracy; large symbols) in both logistic regression (LR: filled symbols and solid line) and discriminant analysis (DA: open symbols and dashed line). Regression lines illustrate significant linear or polynomial trends. Spearman rank correlations (r_s) are given to the right of each panel, with statistical significance ($P < 0{\cdot}01$) indicated (**). The larger the discrepancy between intrinsic and extrinsic accuracy, the more the model was overfit.

replicate models (per species and sampling regime) of the same algorithm, however, was equally incongruous.

Although LR appeared more susceptible, overfitting occurred in both algorithms and depended on sample size and sampling prevalence. Increases in sample size notably diminished overfitting because they reduced intrinsic accuracy while improving extrinsic accuracy (Fig. 6a). The decline in intrinsic accuracy was curvilinear for all measures but AUC in DA. Extrinsic accuracy improved linearly, with strong positive correlations evident for all measures except specificity in DA.

The effects of sampling prevalence on model performance were more complex (Fig. 6b). Higher prevalence led to better sensitivity but poorer specificity in both training and test data. Intrinsic kappa showed no significant response to prevalence in LR, but correlated positively in DA with curvilinear effects. Intrinsic AUC was not affected in either algorithm. Extrinsic kappa and extrinsic AUC both displayed a significantly convex relationship with prevalence. According to AUC, then, overfitting was minimized at intermediate prevalence.

The models with best overall predictive power for each species are listed in Table 3. Optimal models had intermediate prevalence (50%) and large sample sizes (300–500).

Discussion

In its disregard of potential statistical artefacts, conventional practice in distribution modelling can mislead: based on analysis 1 alone we might have concluded, mistakenly, that range size affected model accuracy. According to kappa, overall predictive power was greater for species with larger ranges. The lack of response in AUC might have alerted us to potential statistical artefacts. Because kappa is threshold dependent while AUC is not, we may, however, have concluded that models for species with smaller ranges should utilize a different decision threshold to separate probabilistic predictions of occurrence into predictions of presence and absence.

Instead, the response in kappa with changing range size probably reflected the direct effects sampling

Table 3. Optimal models for each species in analysis 3, built with either logistic regression (LR) or discriminant analysis (DA). A model was judged optimal if, on average (more than $n = 10$ repeat trials), it achieved the highest extrinsic AUC value for that species. Ties were solved by choosing models that also maximized extrinsic kappa. For each species, the optimal sampling regime (percentage prevalence and sample size in terms of quarter-degree squares) is indicated along with mean measures of sensitivity, specificity, kappa and AUC calculated for test data

	Sampling regime				Test accuracy							
	Prevalence		Sample size		Sensitivity		Specificity		Kappa		AUC	
Species	LR	DA	LR	DA	LR	DA	LR	DA	LR	DA	LR	DA
Grey-winged francolin	50	50	500	500	0·93	0·95	0·88	0·84	0·73	0·70	0·95	0·94
Ground woodpecker	50	50	500	500	0·94	0·93	0·86	0·84	0·72	0·69	0·95	0·94
Cape rock-thrush	50	50	500	500	0·89	0·92	0·87	0·85	0·70	0·69	0·95	0·93
Southern double-collared sunbird	50	50	300	300	0·89	0·92	0·87	0·86	0·70	0·70	0·95	0·94
Large-billed lark	50	50	500	300	0·91	0·94	0·90	0·88	0·76	0·75	0·95	0·96
Cape weaver	50	50	500	500	0·91	0·93	0·88	0·87	0·73	0·72	0·95	0·95
African pied starling	50	50	500	500	0·94	0·94	0·92	0·90	0·81	0·77	0·97	0·96

prevalence exerts on this metric. In analysis 1, the sampling prevalence of test data increased with species' range sizes. Analysis 2 showed clearly that kappa responds positively to such changes in sampling prevalence, as long as the proportion of positive cases remains below 0·5 (as was the case in analysis 1).

Kappa responded to the overall level of error, error allocation and prevalence. That kappa reflects overall error is obviously desirable. Its response to the allocation of error also seems justified. Disproportionately high error in the category (presence or absence) of lower prevalence is penalized, whereas disproportionately good performance is rewarded. Kappa's sensitivity to prevalence overall, however, renders it inappropriate for comparisons of model accuracy between species or regions unless certain precautions are taken. This has not yet been highlighted in the ecological literature.

Kappa's behaviour and implications thereof have, however, been extensively scrutinized in clinical and epidemiological contexts (Cicchetti & Feinstein 1990; Lantz & Nebenzahl 1996; Hoehler 2000). The metric suffers from two artefacts, termed bias effect (Byrt, Bishop & Carlin 1993) and prevalence effect (Thompson & Walter 1988). Kappa should therefore be used with caution in comparative studies (Thompson & Walter 1988; Byrt, Bishop & Carlin 1993) and perhaps only where experimental design can ensure 50% prevalence (Lantz & Nebenzahl 1996; Hoehler 2000).

Alternatively, analysis 2 implies that AUC permits reliable comparisons of accuracy where species' prevalence varies between models. AUC remained constant over a wide spectrum of sampling prevalence, making it a robust measure of model performance.

ROC curves first appeared in the ecological literature in the mid-1990s (Murtaugh 1996). They have, however, been used in medical analysis since the 1950s (Zweig & Campbell 1993), and AUC remains a popular measure of diagnostic accuracy (Faraggi & Reiser 2002). In a comprehensive review of ROC plots and associated statistics, Zweig & Campbell (1993) highlighted the technique's independence of prevalence. In our analysis, AUC displayed elevated standard errors at very low (< 20%) and very high (> 75%) sampling prevalence. As a safeguard, therefore, an intermediate prevalence may be advisable when measuring AUC.

Unlike AUC, model-fitting algorithms responded strongly to both sample size and sampling prevalence. The null hypothesis, that sample size and sampling prevalence exert no effect on algorithmic performance, was rejected for two reasons. First, intrinsic accuracy did not mimic patterns established in analysis 2. Secondly, extrinsic accuracy responded significantly to variations in training sample size and sampling prevalence when it was expected to remain unaffected.

The effect exerted by sample size on LR and DA in analysis 3 has been noted by other authors. Cumming (2000) reported that increasing the size of the study area, and therefore sample size, led to higher AUC in LR. Pearce & Ferrier (2000b) found that, among a number of factors tested, sample size had the largest effect on the predictive accuracy of LR. Stockwell & Peterson (2002) noted that LR performed worse at small sample sizes than two other algorithms (GARP and surrogate models) and was more prone to overfitting. Hendrickx (1999) found that, in DA, smaller sample sizes led to diminished predictive accuracy, although the relationship was not proportionate: reducing sample size by 2/3 decreased predictive power by only 10%. Williams & Titus (1988), none the less, recommended that DA models of ecological systems be trained with at least three times as many samples as the number of predictor variables to be included. Their simulation suggested that sample sizes smaller than this produced unstable canonical coefficients.

Although other authors have noted the potential influence of prevalence on model accuracy, none has tried to disentangle direct effects on measures of accuracy from sensitivities inherent in the model algorithm. Furthermore, few have separated the effects of sampling prevalence from potentially meaningful ecological effects of range size.

Among those that have, Manel, Dias & Ormerod (1999) demonstrated that sampling prevalence affected model predictions but did not quantify how performance changed. Fielding & Haworth (1995) investigated how training sample prevalence influenced sensitivity, specificity and the matching coefficient in LR and DA models, but in their study sample size changed concomitantly with prevalence. Cumming (2000), using LR models of a hypothetical species' range, showed that AUC (intrinsic) declined as sampling prevalence diminished. At very low prevalence AUC became erratic, echoing findings of analysis 2 here. Olden, Jackson & Peres-Neto (2002) randomized species distributions to demonstrate, with the help of null models, that high matching coefficients for both very rare and very common species reflect random processes rather than ecological phenomena.

Most authors studying the effects of range size on model performance have neither controlled sampling prevalence nor used null models (Manel, Dias & Ormerod 1999; Pearce & Ferrier 2000b; Manel, Williams & Ormerod 2001; Pearce, Ferrier & Scotts 2001; Kadmon, Farber & Danin 2003). The patterns they reported largely match those observed in analysis 3. Their findings therefore potentially reflect statistical artefacts rather than real range size effects.

Only one study we know of suggests that range size may have effects on model accuracy beyond those explained by statistical artefacts. Stockwell & Peterson (2002) modelled the distribution of 103 Mexican bird species with GARP, an artificial intelligence procedure. Training sample size was constant across species, and a resampling procedure internal to GARP generated an effective sampling prevalence of 50%. None the less, widespread species yielded less accurate models (lower matching coefficients). Data quality may have played a role: training data for widespread species possibly included false negatives, i.e. sites where the species occurred but had not been recorded. Yet performance for one species improved when its southern and northern populations were modelled separately, suggesting that ecologically meaningful factors, such as local variation in habitat preferences, could be responsible (Stockwell & Peterson 2002). Although a crude approach, geographical data partitioning may prove useful in exploring ecological hypotheses (Osborne & Suarez-Seoane 2002).

When attributing variation in model performance to differences in species' range sizes, consideration should be taken of (i) the measure of range size used; (ii) other ecological characteristics of the species that potentially covary with range size; and (iii) the possibility of statistical artefacts. We measured range size as area of occupancy. Extent of occurrence, an alternative measure, is potentially less entangled with sample size and sampling prevalence, but might covary with other ecological characteristics. Mobility, niche width and feeding habits may all influence how accurately models identify habitat associations (Mitchell, Lancia & Gerwin 2001; Pearce, Ferrier & Scotts 2001; Kadmon, Farber & Danin 2003).

An algorithm immune to statistical effects would be ideal. LR and DA are only two among many approaches to distribution modelling. Other algorithms may be less affected. GARP, for example, seems better able to cope with small sample sizes (Stockwell & Peterson 2002). Ironically, the algorithm might, however, suffer prevalence effects despite constituting a presence-only approach, because in addition to presence records GARP employs background samples for model training. Even pure presence-only approaches, such as BIOCLIM, may be afflicted by prevalence-related artefacts if test data involve absence records (Kadmon *et al.* 2003). The effects of sample size and sampling prevalence on models explicitly incorporating spatial autocorrelation (Augustin, Mugglestone & Buckland 1996; Hoeting, Leecaster & Bowden 2000) should also be carefully examined.

In the absence of an ideal algorithm, one option to overcome the statistical artefacts range size imposes on model accuracy is the creation of null models as suggested by Olden, Jackson & Peres-Neto (2002). Results presented here support the computationally less-demanding approach of fixing sampling prevalence across species as a viable alternative. Differences in sample size from species to species that arise in this way obviously need to be taken into account. As the effects of sample size on model performance are largely linear, however, they can be removed with relative ease through partial correlation analysis.

Pearce & Ferrier (2000a) and Vaughan & Ormerod (2003) warn that models tend to over- or underestimate a species' probability of occurrence systematically if sampling prevalence in training data is atypically high or low. No systematic bias, however, was detected in our analyses. Both natural and test sample prevalence were distinctly lower than 50%, yet training data with a sampling prevalence of 50% led to an optimal balance between false positive (commission) and false negative (omission) errors in both the full data set (Fig. 5b) and test data (Fig. 6b). A training sample prevalence of 50% appears ideal, therefore, if commission and omission entail equal ecological costs (see below).

Commission and omission errors may not always weigh equally, depending on what purpose model predictions serve (Fielding & Bell 1997). If, say, the aim of a model is to identify all remaining habitat of a critically endangered species for purposes of protection, the omission of sites where the species is present may be of more concern than the mistaken inclusion of potentially suitable but unoccupied sites. In this case, sampling prevalence might be set high to maximize sensitivity. If instead, we are using distribution models to make inferences about a species' range size and population level, excessive commission could lead to unjustified confidence in the species' conservation status. In this case, a lower sampling prevalence to maximize specificity may be more precautionary.

We need to keep in mind, however, that sensitivity and specificity can give false impressions of model performance at high and low prevalence because these measures do not correct for agreement expected to occur by chance (Fielding & Bell 1997). Brenner & Gefeller (1994) have proposed chance-corrected alternatives that should be independent of prevalence. Also of interest may be a kappa-like metric suggested by Brennan & Prediger (1981), which measures model performance over and above a best a priori strategy, such as predicting a species to be omnipresent. Like kappa and AUC, these measures were introduced in a clinical context, but may be worth exploring as tools in ecological modelling.

CONCLUSION

When comparing the performance of distribution models across species, we must distinguish ecologically meaningful patterns from statistical artefacts. Reported effects of species' rarity or range size on model accuracy appear to be largely artefactual. Both model algorithms and accuracy metrics contribute to such artefacts. The two algorithms assessed here, LR and DA, were comparable in their susceptibility to sample size and sampling prevalence. Both performed optimally at intermediate sampling prevalence. Among the accuracy metrics examined, AUC, unlike kappa, was practically immune to prevalence-related artefacts. Its standard error, however, rose towards the extremes of sampling prevalence. Consequently, we encourage researchers engaged in distribution modelling to utilize intermediate levels of sampling prevalence, obtained by subsampling where necessary. Furthermore, we recommend that authors: (i) always report sampling prevalence and distinguish it from a species' range size; (ii) use a fixed sampling prevalence for comparative studies in both training and test data; and (iii) make use of accuracy metrics such as AUC that are unaffected by prevalence and correct for agreement expected to occur by chance.

Where these recommendations are not met, measures of accuracy cannot be taken at face value. The reliability of models must then be judged with great care.

A species' ecology is likely to affect its predictability. Only once we minimize statistical artefacts, however, will we be able to detect ecologically meaningful patterns.

Acknowledgements

J. M. McPherson's research was kindly supported by a Scatcherd European scholarship from the University of Oxford. We are grateful to James H. Brown for supporting this study. Furthermore, we thank Mary Wisz for lively discussions on the topic, as well as Graeme Cumming, Duncan McPherson and two anonymous referees for helpful comments on earlier drafts of this paper.

References

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York, NY.

Augustin, N.H., Mugglestone, M.A. & Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–347.

Brennan, R.L. & Prediger, D.J. (1981) Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, **41**, 687–699.

Brenner, H. & Gefeller, O. (1994) Chance-corrected measures of the validity of a binary diagnostic-test. *Journal of Clinical Epidemiology*, **47**, 627–633.

Byrt, T., Bishop, J. & Carlin, J.B. (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, **46**, 423–429.

Cabeza, M., Araujo, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R. & Moilanen, A. (2004) Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology*, **41**, 252–262.

Chatfield, C. (1996) *The Analysis of Time Series: An Introduction*, 5th edn. Chapman & Hall, London, UK.

Cicchetti, D.V. & Feinstein, A.R. (1990) High agreement but low kappa. II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, **43**, 551–558.

Cohen, J.A. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Corsi, F., Dupre, E. & Boitani, L. (1999) A large-scale model of wolf distribution in Italy for conservation planning. *Conservation Biology*, **13**, 150–159.

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

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Faraggi, D. & Reiser, B. (2002) Estimation of the area under the ROC curve. *Statistics in Medicine*, **21**, 3093–3106.

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

Fielding, A.H. & Haworth, P.F. (1995) Testing the generality of bird–habitat models. *Conservation Biology*, **9**, 1466–1481.

Gates, S. & Donald, P.F. (2000) Local extinction of British farmland birds and the prediction of further loss. *Journal of Applied Ecology*, **37**, 806–820.

Gibson, L.A., Wilson, B.A., Cahill, D.M. & Hill, J. (2004) Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. *Journal of Applied Ecology*, **41**, 213–223.

Green, P.E. (1978) *Analyzing Multivariate Data*. Dryden Press, Hinsdale, IL.

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

Harrison, J.A., Allan, D.G., Underhill, L.G., Herremans, M., Tree, A.J., Parker, V. & Brown, C.J. (1997) *The Atlas of Southern African Birds*. Birdlife South Africa, Johannesburg, South Africa.

Hay, S.I. (2000) An overview of remote sensing and geodesy for epidemiology and public health applications. *Remote Sensing and Geographical Information Systems in Epidemiology* (eds S.I. Hay, S.E. Randolph & D.J. Rogers), Vol. 47, pp. 1–35. Academic Press, London, UK.

Henderson, A.R. (1993) Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. *Annals of Clinical Biochemistry*, **30**, 521–539.

Hendrickx, G. (1999) *Georeferenced decision support methodology towards trypanomiasis management in West Africa*. PhD Thesis. University of Gent, Gent, Belgium.

Hoehler, F.K. (2000) Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, **53**, 499–503.

Hoeting, J.A., Leecaster, M. & Bowden, D. (2000) An improved model for spatially correlated binary responses. *Journal of Agricultural Biological and Environmental Statistics*, **5**, 102–114.

Insightful™ (2001) *S-Plus 6 for Windows Guide to Statistics*, Vol. 2. Insightful Corporation, Seattle, WA.

Jarvis, A.M. & Robertson, A. (1999) Predicting population sizes and priority conservation areas for 10 endemic Namibian bird species. *Biological Conservation*, **88**, 121–131.

Jetz, W. & Rahbeck, C. (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548–1551.

Joachim, J., Cargnelutti, B., Cibien, C. & Nappée, C. (1998) Évaluation par télédétection des biotopes à gélinotte de bois (*Bonasa bonasia*) dans le Parc national des Cévennes. *Gibier Faune Sauvage*, **15**, 31–45.

Johnson, D.D.P., Hay, S.I. & Rogers, D.J. (1998) Contemporary environmental correlates of endemic bird areas derived from meteorological satellite sensors. *Proceedings of the Royal Society of London Series B, Biological Sciences*, **265**, 951–959.

Kadmon, R., Farber, O. & Danin, A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.

Kriticos, D.J., Sutherst, R.W., Brown, J.R., Adkins, S.W. & Maywald, G.F. (2003) Climate change and the potential distribution of an invasive alien plant: *Acacia nilotica* ssp. *indica*. Australia. *Journal of Applied Ecology*, **40**, 111–124.

Lantz, C.A. & Nebenzahl, E. (1996) Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology*, **49**, 431–434.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*, 2nd English edition. Elsevier, Amsterdam, the Netherlands.

Li, W.J., Wang, Z.J., Ma, Z.J. & Tang, H.X. (1999) Designing the core zone in a biosphere reserve based on suitable habitats: Yancheng Biosphere Reserve and the red crowned crane (*Grus japonensis*). *Biological Conservation*, **90**, 167–173.

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

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

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

Milsom, T.P., Langton, S.D., Parkin, W.K., Peel, S., Bishop, J.D., Hart, J.D. & Moore, N.P. (2000) Habitat models of bird species' distribution: an aid to the management of coastal grazing marshes. *Journal of Applied Ecology*, **37**, 706–727.

Mitchell, M.S., Lancia, R.A. & Gerwin, J.A. (2001) Using landscape-level data to predict the distribution of birds on a managed forest: effects of scale. *Ecological Applications*, **11**, 1692–1708.

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

Neave, H.M., Norton, T.W. & Nix, H.A. (1996) Biological inventory for conservation evaluation. II. Composition, functional relationships and spatial prediction of bird assemblages in southern Australia. *Forest Ecology and Management*, **85**, 123–148.

Olden, J.D., Jackson, D.A. & Peres-Neto, P.R. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329–336.

Osborne, P.E. & Suarez-Seoane, S. (2002) Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling*, **157**, 249–259.

Osborne, P.E., Alonso, J.C. & Bryant, R.G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38**, 458–471.

Paradis, E., Baillie, S.R., Sutherland, W.J., Dudley, C., Crick, H.Q.P. & Gregory, R.D. (2000) Large-scale spatial variation in the breeding performance of song thrushes *Turdus philomelos* and blackbirds *T. merula* in Britain. *Journal of Applied Ecology*, **37**, 73–87.

Parker, V. (1999) *The Atlas of Birds of Sul do Save, Southern Mozambique*. Avian Demography Unit and Endangered Wildlife Trust, Cape Town and Johannesburg, South Africa.

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

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

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

Pearce, J., Ferrier, S. & Scotts, D. (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. *Journal of Environmental Management*, **62**, 171–184.

Peterson, A.T., Ortega-Huerta, M.A., Bartley, J., Sanchez-Cordero, V., Soberon, J., Buddemeier, R.H. & Stockwell, D.R.B. (2002) Future projections for Mexican faunas under global climate change scenarios. *Nature*, **416**, 626–629.

Rogers, D.J., Hay, S.I. & Packer, M.J. (1996) Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology*, **90**, 225–241.

Sakamoto, Y., Ishiguro, M. & Kitagawa, G. (1986) *Akaike Information Criterion Statistics*. D. Reidel Publishing Co., Boston, MA.

Scribner, K.T., Arntzen, J.W., Cruddace, N., Oldham, R.S. & Burke, T. (2001) Environmental correlates of toad abundance and population genetic diversity. *Biological Conservation*, **98**, 201–210.

Sitati, N.W., Walpole, M.J., Smith, R.J. & Leader-Williams, N. (2003) Predicting spatial aspects of human–elephant conflict. *Journal of Applied Ecology*, **40**, 667–677.

Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Thompson, W.D. & Walter, S.D. (1988) A reappraisal of the kappa-coefficient. *Journal of Clinical Epidemiology*, **41**, 949–958.

Vaughan, I.P. & Ormerod, S.J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601–1611.

Venier, L.A., McKenney, D.W., Wang, Y. & McKee, J. (1999) Models of large-scale breeding-bird distribution as a function of macro-climate in Ontario, Canada. *Journal of Biogeography*, **26**, 315–328.

Williams, B.K. & Titus, K. (1988) Assessment of sampling stability in ecological applications of discriminant-analysis. *Ecology*, **69**, 1275–1285.

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

Received 21 February 2004; final copy received 15 June 2004