J Geogr Syst (2010) 12:389–407
DOI 10.1007/s10109-010-0106-8

ORIGINAL ARTICLE

Does plot size affect the performance of GIS-based species distribution models?

Shubha N. Pandit · April Hayward · Jan de Leeuw · Jurek Kolasa

Received: 16 April 2009 / Accepted: 8 January 2010 / Published online: 29 January 2010

Abstract Species distribution models are used extensively in predicting the distribution of vegetation across a landscape. Accuracy of the species distribution maps produced by these models deserves attention, since low accuracy maps may lead to erroneous conservation decisions. While plot size is known to influence measures of species richness, its effect on our ability to predict species distribution ranges has not been tested. Our aim is to test whether the accuracy of the distribution maps produced depend on the size of the plot (quadrat) used to collect biological data in

Electronic supplementary material The online version of this article (doi:10.1007/s10109-010-0106-8) contains supplementary material, which is available to authorized users.

S. N. Pandit · J. de Leeuw
ITC, International Institute for Geo-Information Science and Earth Observation, P. O. Box 6, 7500AA Enschede, The Netherlands

S. N. Pandit · J. Kolasa
Department of Biology, McMaster University, 1280 Main St West, Hamilton, ON L8S4K1, Canada

A. Hayward
Department of Biology, Dalhousie University, 1350 Oxford St, Halifax, NS B3H 4J1, Canada

Present Address:
A. Hayward
Department of Biology, University of Florida, P. O. Box 118525, Gainesville, FL, USA

Present Address:
J. de Leeuw
International Livestock Research Institute (ILRI), P. O. Box 30709, 00100 Nairobi, Kenya

Present Address:
S. N. Pandit (✉)
Department of Biological Sciences, University of Windsor, 401 Sunset Ave, Windsor, ON N9B 3P4, Canada
e-mail: shuba.pandit@gmail.com; panditsn@mcmaster.ca

the field. In this study, the presences of four plant species were recorded in five sizes of circular plots, with radii ranging from 8 to 100 m. Logistic regression-based models were used to predict the distributions of the four plant species based on empirical evidence of their relationship with eight environmental predictors: distance to river, slope, aspect, altitude, and four principle component axes derived using reflectance values from Aster images. We found that plot size affected the probability of recording the four species, with reductions in plot size generally increasing the frequency of recorded absences. Plot size also significantly affected the likelihood of correctly predicting the distribution of species whenever plot size was below the minimum size required to consistently record species' presence. Furthermore, the optimal plot size for fitting species distribution models varied among species. Finally, plot size had little impact on overall accuracy, but a strong, positive impact on Kappa accuracy (which provides a stronger measure of model accuracy by accounting for the effects of chance agreements between predictions and observations). Our results suggest that optimal plot size must be considered explicitly in the creation of species distribution models if they are to be successfully adopted into conservation efforts.

Keywords Kappa · Map accuracy · Species distribution · Logistic regression models · Species frequency curve · Namibia

JEL Classification Q57

1 Introduction

Recent advances in remote sensing and geographic information science have aided the use of predictive vegetation models (species distribution models) in predicting the spatial distribution of plant species in un-surveyed areas at regional scales, using small amounts of observational data (e.g., Guisan and Zimmermann 2000; Miller and Franklin 2002; Zimmermann et al. 2007). Species distribution models (SDMs) are empirical models that relate species observations (presence/absence records) and environmental factors through statistical algorithms (e.g., Guisan and Zimmermann 2000; Elith et al. 2006; Hernandez et al. 2006; Miller et al. 2007). For obvious reasons, such models are becoming increasingly important in resource assessment, environmental conservation and biodiversity management (e.g., Fielding and Bell 1997; Manel et al. 1999; Austin 2002; Lütolf et al. 2009), and the degree to which predicted species distributions reflect real distributions (i.e., the accuracy of the maps) is of major concern: errors associated with predicting species distributions can lead to faulty assumptions in species occurrences, which may have profound implications for decisions regarding land acquisitions for biodiversity conservation (e.g., Hurley 1986; Schlossberg and King 2009). As such, there has been renewed focus on improving the reliability of SDM-derived species distribution maps (e.g., Segurado and Araujo 2004; Seoane et al. 2005; Allouche et al. 2006).

To this end, a number of potential sources of error in species distribution models have been identified (e.g., Stockwell and Peterson. 2002; De Leeuw et al. 2002;

Boone and Krohn 2002; Garrison et al. 2000; Segurado and Araujo 2004; McPherson et al. 2004; Fielding and Bell 1997; Hernandez et al. 2006; Guisan et al. 2007) that may diminish the accuracy of predictive maps: (1) an insufficient number of sampling points (small sample sizes); (2) misidentification of species during surveys; (3) incomplete knowledge of the range of environmental factors suitable for a given species; (4) sampling bias against rare and transient species during surveys; (5) sampling at inappropriate times or seasons; (6) failure to include other environmental variables that affect species presence in the model; and (7) mismatching the resolution of environmental and biological data. In addition to these general sources of error, predictions of species occurrences tend to be less accurate for habitat generalists (species that are found in many areas of multidimensional habitat space) than for habitat specialists (species that are restricted to a small area of the multidimensional habitat space); see Manel et al. (2001), Elith et al. (2006), Hernandez et al. (2006). Consequently, model performance varies from species to species (Venier et al. 1999), and the accuracy of maps derived from SDMs depends on number of species attributes like niche width, range size, abundance, and the degree of ecological specialization (e.g., Boone and Krohn 2002; Garrison et al. 2000; Segurado and Araujo 2004; Mcpherson et al. 2004; Elith et al. 2006).

One possible source of error that has not been addressed, however, concerns the accuracy of the data used to produce and test the maps produced by SDMs: it is generally assumed that field observations are free of errors for the purposes of both constructing models and assessing the accuracy of the predicted maps generated from SDMs using 2×2 error matrices (e.g., Fielding and Bell 1997). However, the collection of field data is a crucial step in constructing and fitting models, and the methods used to collect data in the field can be an important source of error that may diminish the level of accuracy of maps produced by SDMs. While the effect of small sample sizes (i.e., the number of samples) on the accuracy of predicted maps is well demonstrated (e.g., Stockwell and Peterson 2002), the effect of the size of the quadrat used to collect data in the field (i.e., plot size) may have an equally important effect on model accuracy. Plot size has a well-known effect on species richness estimations (e.g., Stohlgren 2007 page 52, Gray 2003; Kerr et al. 2001; Huston 1994) and, as such, it is possible that larger plot sizes may increase the probability of recording a species as present in an area while smaller plot sizes may increase the probability of recording the species as absent.

A survey of the literature suggests that the size of the plot used to collect vegetation data in the field varies among studies. For example, Leos et al. (2001) recorded distributions of vascular plants ranging from 9.8 cm^2 to 64 m^2. Leathwick and Austin (2001) used a data set collected by Leathwick (1998) using plot sizes of 0.4 and 0.04 ha to predict the spatial distribution of plant species. Miller et al. (2007) used data collected by a plot size of 4 m^2 for herbaceous species and 0.1–1 ha for tree species to predict spatial distribution of plant species. Despite the lack of consistency in the plot sizes used to predict the spatial distribution of species both within and among studies, the effect of the plot size used in the collection of field data and its effects on the accuracy and efficiency of the model and predicted maps have yet to be examined. However, there is good reason to believe that plot size may have a substantial effect in this regard: the probability of recording species presence

in an area or habitat depends strongly on the plot size used in field surveys and the reduction of plot size may increase the number of sites where species are observed as absent, even in suitable habitat areas. Such recordings of species absence may be considered false absences, since they are artifacts of sampling design, and would affect both the prediction of the range of the spatial distribution of a species in a landscape and the accuracy of the maps produced. Indeed, the effects of scale on sampling, experimental design and statistical analysis have been well documented in ecology (e.g., Dungan et al. 2002; Wu et al. 2006) and, more specifically, the effect of sampling resolution has long been recognized as one of two issues associated with the Modifiable Areal Unit Problem, wherein the spatial resolution at which some geographic phenomenon is studied affects the outcome of subsequent analyses as a consequence of the fact that an artificial spatial resolution has been imposed upon a continuous geographic phenomenon (Openshaw and Taylor 1981). Consequently, it is of paramount importance to choose a sampling resolution that minimizes the error associated with the representation of spatial phenomenon (Openshaw 1996). We therefore postulate that the quality of the biotic data used to predict the spatial extent of plant species distributions using SDMs may be affected by the size of the plot used for the collection of data in the field and that such effects may differ among species with different underlying spatial distributions. The incorporation of such data into SDMs may also affect the accuracy of the maps derived from the SDMs. Here, we test whether the spatial extent of species distributions and map accuracy depend on the size of the quadrat used to collect biological data and explore how these effects vary with the underlying spatial distribution of four Namibian plant species (*Bosnia albitrunca, Terminalia prunioides, Acacia tortilis*, and *Zygophylum simplex*) collected using five quadrat sizes.

2 Methods

2.1 Study area

The study was conducted east of Sesfontein, Namibia (Fig. 1, 13°30′–14° E; 20°–19° 30′ S). The area, which rises from 600 to 1,600 m a.s.l., has an arid climate with average annual rainfall of approximately 150 mm (Simon 1997).

2.2 Species occurrence data

Species occurrence data were collected at the end of the dry season (September–October 2001) using a clustered random sampling design that was used to select 42 cluster sites, each with eight sub-sites (with a 100 m radius) selected along the perimeter of circle with a 1,000 m radius from the center of the cluster site (see Supplementary material, Fig. A1). Among the 42 cluster sites, 6 were inaccessible due to steep slopes. At the remaining 36 cluster sites, we were unable to sample 16 sub-sites because they either occurred in a riverine area or were inaccessible to due to steep slopes. Thus, a total of 36 cluster sites including a total of 272 sub-sites were surveyed.

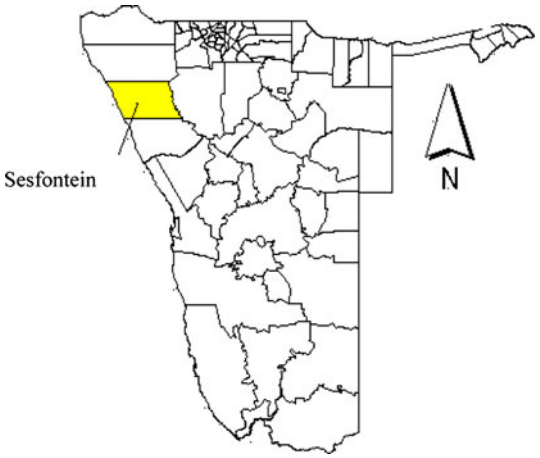

Fig. 1 Study area: Sesfontein, Namibia

We selected four plant species with differing life history strategies that are sufficiently common so as to be found in 10–50% of the study sites for further analysis: a tree, *Acacia tortilis*, a small to large tree, *Boscia albitrunca,* a tall shrub, *Terminalia prunioides*, and an herb, *Zygophylum simplex. Acacia tortilis* normally occurs in bushvelt and grassland, on clay, riverine and loamy soils, while *Boscia albitrunca* is widely distributed in the study area and is solitary in nature. *Terminalia prunioides* is found in dry, stony areas and occurs in dense clumps, while *Zygophylum simplex* is found mainly in grazed or open areas on clay and loamy soil.

At each sub-site, we recorded the distance from the center of the sub-site to the individual closest to the center of the plot for each of the four species in order to produce a new data set consisting of presence/absence data based on nested plots with radii of 8, 16, 32, 64, and 100 m, which were used in further analyses. Since the plots of different radii were nested, each plot size contained all of the smaller plot sizes and, thus, if a species was recorded at 25.5 m from the center of the 100 m plot, that species would be recorded as absent in the 8 and 16 m plots and present in the 32, 64, and 100 m plots.

The distributions of the four species were different in nature, with average densities varying across the study site in different ways. To quantify these differences and obtain a sense of how each species was distributed in the study area, we calculated the index of dispersion for each species based on the density of individuals recorded in the 100 m radius plot sizes. We obtained density estimates by counting the total number of individuals of each of *Acacia tortilis*, *Boscia albitrunca*, and *Terminalia prunioides* present in the 100 m plot. The density of the herb species, *Zygophylum simplex*, was estimated for each 100 m plot by counting all of the individuals in fifteen randomly placed 1 m^2 quadrats and extrapolating those densities to the total area of the 100 m radius plot. *Zygophylum simplex* and *Acacia tortilis* showed a clumped distribution (index of dispersion = 1.7 and 1.2, respectively) indicating that they grow densely, but only in certain locations. On the other hand, *Boscia albitrunca* and *Terminalia prunioides* exhibited low densities in most sites across the study area (index of dispersion = 0.46 and 0.51, respectively).

2.3 Environmental data

We used reflectance data from Aster satellite images (9 bands: visible and near Infra red (VNIR, bands 1–3) and short wave infrared (SWIR, bands 4–9)) of the study area taken in September 2001 to capture landscape characteristics such as general vegetation cover and soil attributes (see Fig. A2 in the Supplementary material for a general overview of how environmental data were constructed and complied). These satellite images were from the same month and year in which the biological data were collected (September 2001). The ground resolution of the Aster images differed among the bands used in the study: VNIR had 15 m ground resolution and SWIR had 30 m ground resolution. In order to the data from each band comparable, we resampled the pixel size of the images to 15 m resolution.

The various bands of satellite imagery were highly correlated, raising concerns about bias in the models due to multi-collinearity (Buckland and Elston 1993). To avoid such bias, we used principal component analysis (PCA) to reduce the dimensionality of the reflectance data from nine bands to four principle component axes. Further analysis used the first four PCA bands, which captured 99.44% of the information contained in the original nine bands. A digital elevation model (DEM) was derived from topographic maps with 15 m resolution by first digitizing the contour lines and then rasterizing the data to 15 m resolution and interpolating it using a contour interpolation operation in GIS software (ILWIS Academic 3.0 2001). The resultant DEM was used to create data layers representing altitude, slope, and aspect. The topographic map was also used to create a data layer reflecting soil moisture content based on distance from rivers at 15 m resolution. We extracted values for environmental data from the central pixel of the radial plots, assuming that the environmental data of the central pixel is representative of an average value for the environmental factors in that plot.

2.4 Species distribution model

We first partitioned the species data into a training data set and an evaluation data set using a ~51:49 ratio for training (i.e., model) and reference (i.e., testing) data, respectively. Although some studies have partitioned species data in a 75:25 ratio (e.g., Fielding and Bell 1997), others have used a ratio of 51:49 (see Dixon 2004). We chose to partition our data in a ~51:49 ratio in order optimize model performance since both processes (model fitting and accuracy) are important. The data set was partitioned by randomly assigning 0 or 1 to each of the 272 sub-sites. Using this procedure, 152 sub-sites (56% of the total data set) were assigned to model development (training data), with the remaining 120 sub-sites (~44% of the total data set) used for statistically independent cross-validation in the subsequent accuracy assessment. This procedure was reiterated independently for each species.

A variety of techniques are available for modeling the spatial correlation of environmental factors associated with species' distributions (see Austin 2002; Berg et al. 2004). Of these, logistic regression is frequently used for modeling species distributions based on species presence/absence data (see Manel et al. 1999; Pearce and Ferrier 2000; Berg et al. 2004; Johnson and Gillingham 2005; Austin 2007).

Santika and Hutchinson (2009) suggested that higher order functions are preferable when applying logistic regression models to species data. Berg et al. (2004) also found complex models to perform better. Therefore, we used second-order multiple logistic regressions to develop models of best fit describing the relationship between species presence and the environmental data for each species at each plot size (i.e., 8, 16, 32, and 100 m radii). Models took the general form:

$$Y = \frac{e^Z}{(1 + e^Z)} \tag{1}$$

where $Z = \alpha + \beta_1 \times X_1 + \beta_2 \times X_1^2 + \cdots + \beta_n \times X_n^2$, and $X_1 \cdots X_n$ represent environmental variables. α and β represent the intercept constant and the coefficient (slope parameter), respectively. Y is the probability of species present or response function always lies between 0 and 1. We used the 8 environmental variables (altitude, distance to river, slope, aspect, and the first four axes of PCA of the Aster image reflectance data) as explanatory variables to fit species presence/absence data in a backward-selected logistic regression procedure. The explanatory variables were removed from the model one at a time until all of the remaining variables contributed significantly to the explanatory power of the model ($p < 0.05$). We then used the approach of Guisan and Zimmermann (2000) to evaluate whether the model performs better using the data collected by larger or smaller quadrats. To this end, we measured the 2 log likelihood value: larger values of 2 log likelihood indicate greater predictive power (Guisan and Zimmermann 2000). 2 log likelihood is computed as:

$$2\log[(LL_N)/(LL_O)] = 2[\log(LL_N) - \log(LL_O)] = 2[(LL_N) - (LL_O)] \tag{2}$$

where LL_N is the log likelihood function for a model with N variables, and LL_O is the log likelihood function for the constants only model. The goodness of the fit of the selected model is expressed as the adjusted D^2 statistic, which is equivalent to adjusted R^2 (Guisan and Zimmermann 2000) and is computed as:

$$\text{Adjusted } D^2 = 1 - [(n-1)/(n-p)] \times \left[1 - D^2\right] \tag{3}$$

$$\text{where} \quad D^2 = \frac{LL_O - LL_N}{LL_N} \tag{4}$$

and n is the number of observations and p is the number of parameters that are significant in the model. Adjusted D^2 always lies between 0 and 1, with a perfect model (with no residual deviance) taking a value of 1.

We used the environmental variables that were retained in the logistic regression model to create GIS maps predicting the distribution of each species for each plot size (radii of 100, 64, 32, 16 and 8 m) with a GIS software (ILWIS 3.0 Academic 2001), with each pixel in the distribution map yielding the probability that a species will be present in that location (with probabilities ranging from 0 to 1). To convert the continuous probability surface to a discrete species presence/absence map, we chose a threshold of 0.5, which has been widely used in ecology (see, e.g., Manel et al. 1999, 2001; Bio et al. 2002; Stockwell and Peterson 2002; Reese et al. 2005; Mushinzimana et al. 2006; Syartinilia and Tsuyuki 2008). Although some studies have indicated that a 0.5 cut-off point may not be optimal in all cases (see Manel et al.

2001; Liu et al. 2005), we selected a threshold of 0.5 both because it is a common choice in ecology and because a cut-off value of 0.5 delivers an equal penalty when sites in which species are present are classified as absences and vice versa.

2.5 Accuracy assessment

We compared the accuracy of the predicted species distribution maps first using the conventional method to estimate overall accuracy (i.e., using an error matrix; see Fielding and Bell 1997; Hernandez et al. 2006) and second using Kappa statistics (see, e.g., Skidmore 1999; Fielding and Bell 1997; Cohen 1960). First, the overall predictive success of each model was assessed creating an error matrix by comparing model predictions against the 120 independent reference plots reserved for this purpose. Overall accuracy was calculated as the sum of the number of correct predictions of species presence and absence divided by the total number of observations. Second, we used Kappa statistics to measure the accuracy of the predicted maps. Kappa (also called 'Kappa hat') measures the agreement between the predicted maps and the reference data while accounting for the possibility of agreements occurring by chance (random agreements) and is computed as (see Skidmore 1999):

$$\text{Kappa} = \frac{N\sum_{i=1}^{r} X_{ii} - \sum_{i}^{r} X_{i+} \times X_{+i}}{N^2 - \sum_{i}^{r} X_{i+} \times X_{+i}} = \frac{\theta_1 - \theta_2}{1 - \theta_2} \tag{5}$$

$$\text{where} \quad \theta_1 = \sum_{i=1}^{r} \frac{X_{ii}}{N}, \tag{6}$$

$$\text{and} \quad \theta_2 = \sum_{i}^{r} X_{i+} \times X_{+i}. \tag{7}$$

r is the number of rows and columns in error matrix, X_{ii} is the number of observations in row i and column i, X_{i+} is the marginal total of row i, X_{+i} is the marginal total of column i. We used a z-test to determine whether Kappa deviated from zero (with the hypothesis H_o: $\kappa = 0$ and H_a: $\kappa > 0$); since κ was expected to be larger than zero, a one sided z-test was used.

3 Results

The number of sites where species were recorded as present generally increased as plot size increased (Table 1), though less-so for the herb species (*Zygophylum simplex*).

However, the cumulative frequency distributions of the four species showed differing trends (Fig. 2). *Terminalia prunioides* equilibrated at relatively high levels, at a radius of around 64 m, suggesting that plot sizes with a radius of at least 64 m are sufficient to accurately capture the presence of this species. Similarly, *Acacia tortilis* equilibrated around 32 m, at a much lower cumulative frequency; the cumulative frequency of the herb *Zygophylum simplex* was also low and reached a

Table 1 Plot size and the number of plots where species were recorded present (a total of 272 sites were included in this survey)

Plot size radius	Number of sites where species observed (*P*)			
	B. albitrunca	*T. prunioides*	*A. tortilis*	*Z. simplex*
100	181	173	61	52
64	128	166	59	52
32	68	143	51	52
16	24	96	26	49
8	14	60	12	45

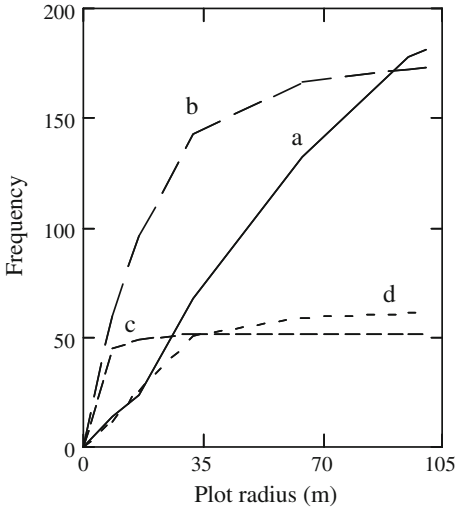

Fig. 2 Cumulative frequency of the presence of four plant species in relation to distance from the center of 272 randomly selected sites near Sesfontein, Namibia; *a Boscia albitrunca*, *b Terminalia prunioides*, *c Zygophylum simplex*, *d Acacia tortilis*

maximum at plot sizes smaller than 16 m. In contrast to the other three species, the frequency curve of *Boscia albitrunca* continued to increase up to our largest plot size (100 m radius), with no sign of impending equilibration.

The environmental variables retained in the models differed across plot size, as did the trend in the goodness of fit in the models (Table 2a–d). While the goodness of fit or adjusted D^2 for *Boscia albitrunca* increased with increasing plot size (Table 2a), the fit for *Acacia tortilis* reached an optimum at intermediate plot sizes (32 m; Table 2c). The goodness of fit for *Zygophylum simplex* reached a maximum value at a plot radius of 8 m, with adjusted D^2 decreasing as plot size increased up to 16 m, and then remaining constant as plot size was increased beyond 16 m (Table 2d).

Predictive species distribution maps (Fig. 3) were created for each plot size for each of the four species using the environmental variables that were retained in the backward logistic regression analysis (Table 2), on a case-specific basis (i.e., by species type). The likelihood of correctly predicting the distribution of species showed plot-size-related effects for three of the four plant species (Figs. 3, 4, and Supplementary material Fig. A3). While the likelihood of correctly predicting the distribution of *Boscia albitrunca* increased markedly with increasing plot size, the likelihood of correctly predicting the spatial distribution of *Terminalia prunioides* increased up to a plot size of approximately 64 m and then did not change as plot

Table 2 Coefficients and statistics describing the best fitting logistic regression model ($N = 152$) for four species at each of the five plot sizes

Variables	Plot size (radius, m)				
	8	16	32	64	100
a. *B. albitrunca*					
Constant	59.475	1501.59	37.256	2.785	76.587
Altitude				0.011	
Distance to river					0.000
PC2	−0.303	−31.587	−0.186	−0.077	−0.388
PC3		10.195			
$(PC1)^2$	0.001				−0.001
$(PC2)^2$		0.110	0.000		
$(PC3)^2$		−0.034			
2[LL(N)-LL(O)]	11.424	19.943	35.836	41.720	74.111
D^2	0.169	0.197	0.213	0.207	0.367
D^2 adjusted	0.163	0.181	0.208	0.207	0.359
b. *T. prunioides*					
Constant	−9.17	−152.4	−27.416	19.0509	18.494
Slope	0.009	0.007	0.014		
Altitude				0.000	0.000
Distance to river		1.744		0.218	−0.235
PC2		0.084		0.118	0.138
PC3			0.074		
PC4			0.447		
$(\text{Distance to river})^2$		−0.006			
$(PC1)^2$			0.000		
$(PC4)^2$			−0.003		
2[LL(N)-LL(O)]	22.499	32.845	63.846	60.805	78.325
D^2	0.157	0.172	0.314	0.306	0.405
D^2 adjusted	0.157	0.155	0.295	0.297	0.397
c. *Acacia tortilis*					
Constant	−490.6	−1177.4	−1050.4	−728.0	−799.1
Slope			0.019		0.079
Distance to river		−0.001	−0.001	−0.001	0.000
PC1	1.841	1.275	1.094	0.717	0.804
PC2	1.150	4.628	4.232	2.952	3.298
PC3		1.906	1.711	1.222	1.356
PC4		0.116			

Table 2 continued

Variables	Plot size (radius, m)				
	8	16	32	64	100
2[LL(N)-LL(O)]	19.954	54.016	80.567	75.544	82.192
D^2	0.543	0.615	0.649	0.564	0.586
D^2 adjusted	0.537	0.601	0.637	0.552	0.572
d. *Z. simplex*					
Constant	−533.0	−500.2	−500.2	−500.2	−500.2
Altitude	−0.038	−0.019	−0.019	−0.019	−0.019
Distance to river	0.000				
PC2	7.568	6.49	6.49	6.49	6.49
PC4	−0.585				
$(PC1)^2$	0.001				
$(PC2)^2$	−0.026	−0.05	−0.05	−0.05	−0.05
$(PC4)^2$	0.003				
2[LL(N)-LL(O)]	66.623	52.32	52.32	52.32	52.32
D^2	0.522	0.402	0.402	0.402	0.402
D^2 adjusted	0.503	0.393	0.393	0.393	0.393

size was further increased up to 100 m (see Figs. 3, 4). In contrast, the likelihood of correctly predicting the distribution of *Zygophylum simplex* was higher in the plot size with an 8 m radius, but declined slightly as plot size increased. The likelihood of correctly predicting the distribution of *Acacia tortilis* increased significantly when the radius of the plot reached 32 m and then increased slightly with increasing plot size.

For all species, reductions in plot size had little effect on overall accuracy as measured by the conventional method, but Kappa accuracy was negatively affected by reductions in plot size. For example, the kappa accuracy of the SDM-derived maps for *Boscia albitrunca* increased with plot size (Fig. 5). At small plot sizes (8 m radius), most sites were located in the lower left and right quadrants, representing false and true absences of *Boscia albitrunca*, respectively (Fig. 4). Kappa accuracy differed significantly from zero for 64 and 100 m radius plots, but not for the smaller plot sizes (Fig. 5).

For *Terminalia prunioides*, 84% of the sites at 100 m radius were classified as either true presence or true absence (Fig. 4). However, the number of correct predictions for the presence of *Terminalia prunioides* decreased with decreasing plot size (with a concomitant increase in the number of correct absence designations). At smaller plot sizes, the number of true presences was lower, resulting in a high number of true absences and few false absences. Type I and II error rates were 1 and 28%, respectively, at the smallest plot size (Fig. 4). Kappa, which did not significantly differ from zero at a radius of 8 m, attained higher values and significance with increasing plot size (Fig. 5).

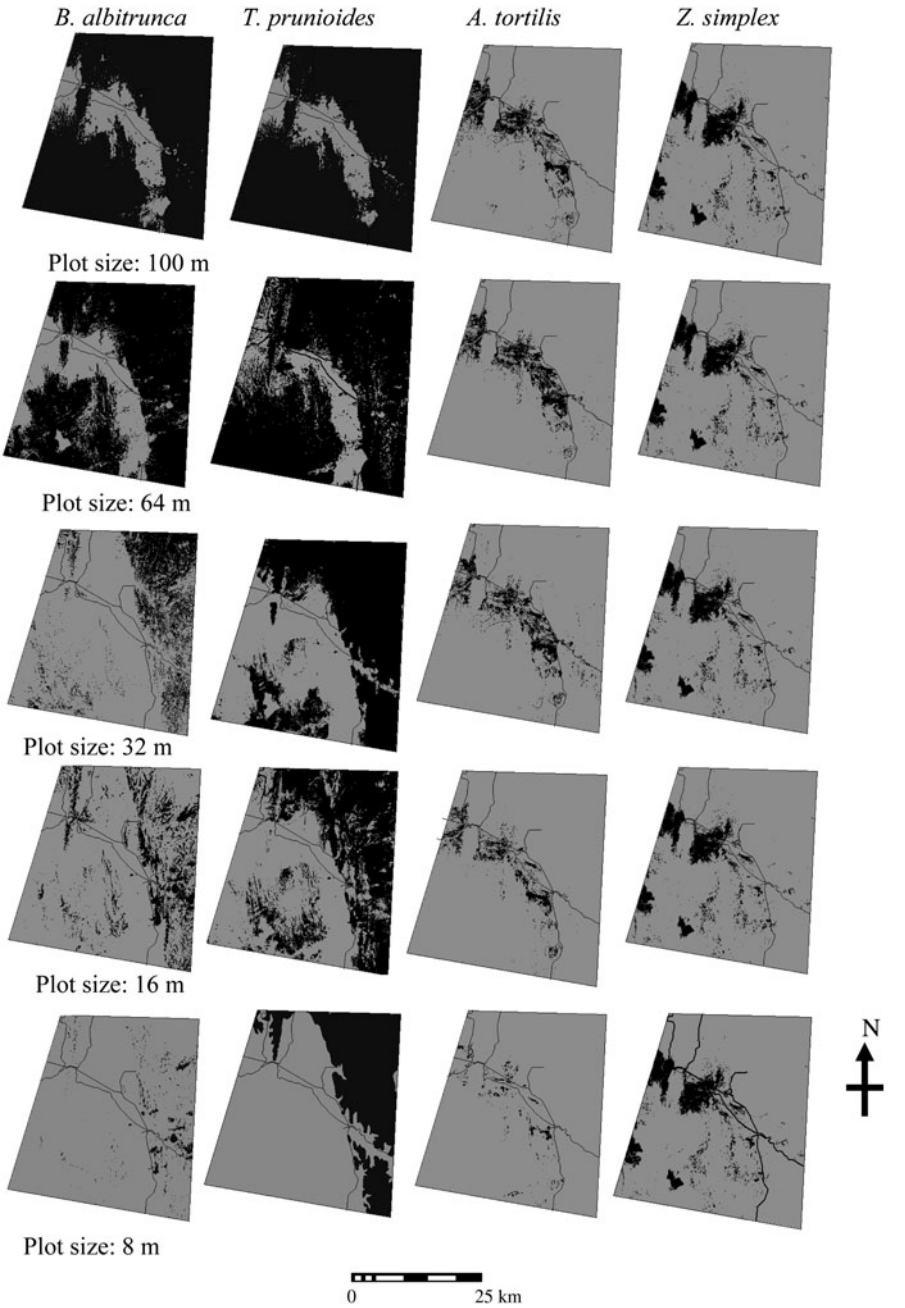

Fig. 3 Presence and absence maps for four Namibian plant species as predicted based on circular plots of variable radius (8, 16, 32, 64, and 100 m) using backward logistic regression models, implemented in a GIS. *Dark* and *light* areas denote the predicted presence and absence of the each species, respectively. The *two lines*, *dark* and *light*, in the maps represent river and road, respectively

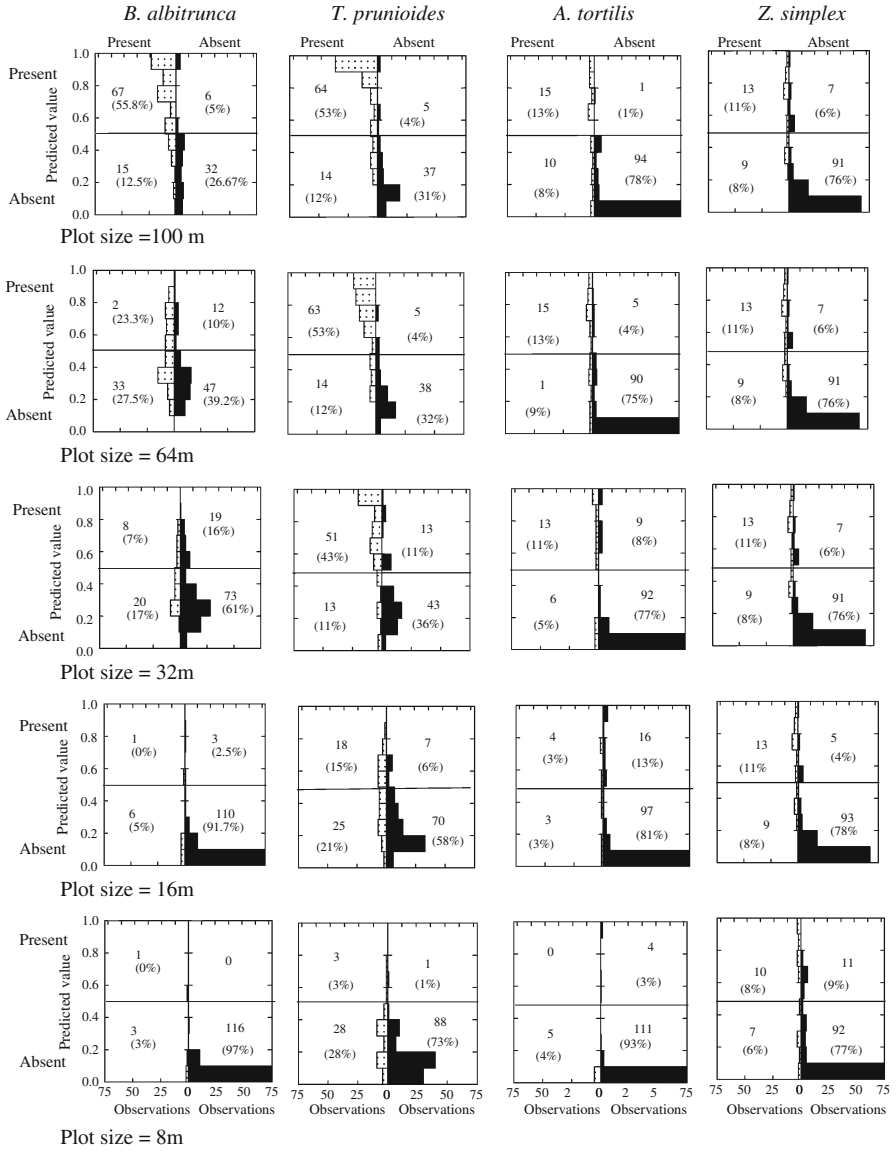

Fig. 4 Error (confusion) matrices comparing the probability of a species' presence predicted by the models and maps (*vertical axis*) against the observed species' presence (*speckled bars*) or absence (*black bars*; *horizontal axis*) for four Namibian plant species for each of *five plot* sizes. Quadrants separated by a horizontal classification threshold of 0.50 represent true presence (*upper left*), false presence (*upper right*, type I error), false absence (*lower left*, type II error), and true absence (*lower right*). The total number and percentage of correct classifications are displayed in each quadrant on each figure

A different pattern was observed in *Zygophyllum simplex:* across all plot sizes, most sites were classified as true absences, with few true presences (Fig. 4). Overall accuracy remained relatively constant across all plots sizes, but, Kappa

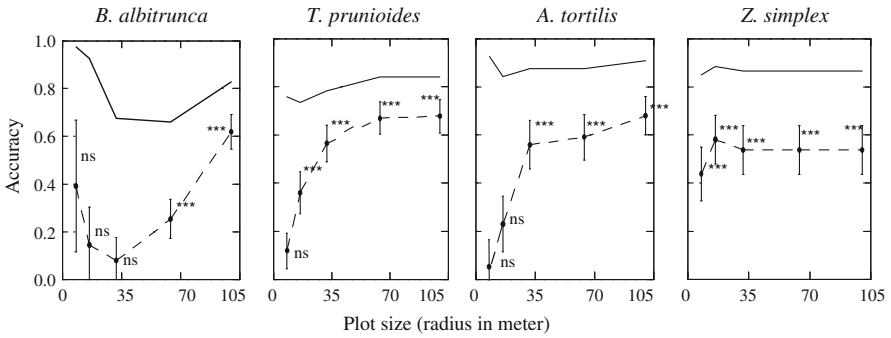

Fig. 5 The effect of the plot size used for recording species presence/absence data on map accuracy. *Vertical bars* denote standard error. *Dash line* Kappa accuracy; *Solid line* overall accuracy. *ns* Not significant, *** significant on or less than 0.001 level

accuracy decreased slightly with increasing plot size after plot size of 16 m radius (Fig. 5).

Similarly, a predominance of sites showed true absences of *Acacia tortilis,* even at the largest plot size (Fig. 4). This appears to be due to the general rarity of *A. tortilis* (Fig. 2). In fact, the number of true presences was zero at 8 m radius. Overall accuracy was more or less independent of plot size for this species (Fig. 5); Kappa accuracy increased from low values that did not differ significantly from zero at smaller plot sizes to a significant Kappa at the largest plot size (32 m radius). However, Kappa accuracy increased, though not substantially, with increasing plot size once plot size exceeded 32 m (Fig. 5).

4 Discussion

We found that the size of the quadrat used to collect biological data strongly influenced the performance of GIS-based species distribution models and that the optimum plot size for modeling the geographic distribution of plant species varied among the four plant species. More specifically, we found that the shape of the curves describing the frequency of occurrence of each of the four species with increasing plot size differed among the four species: the frequency curve equilibrated at a smaller plot size for the herb than for the shrub and tree species, and the steepness of the slope differed among growth forms (Fig. 2). However, despite differences in the shape of the curves among growth forms, curves for species with similar growth forms are likely to differentiate among locally common and rare species, with curves for common species equilibrating at a smaller plot size than those of rare species.

Similarly, while the goodness of fit (D^2 adjusted) of the logistic regression models (which are used to fit species distribution data to environmental predictor variables) generally increased with increasing plot size, the pattern differed among species. Specifically, the distribution of the species with the smallest geographic

range, *Zygophylum simplex*, was most successfully modeled using smaller plot sizes. In contrast, larger plot sizes were required to produce significant successful predictions for the species with the widest geographic ranges, *Boscia albitrunca* and *Terminalia prunioides*. In addition, the environmental variables that were found to be significant predictors of species distributions differed among plot sizes, suggesting that the parameters and processes that are important at one scale are not necessarily important predictors at other scales (Wu 2004).

Differences in the predictive capacity of the logistic regression models among species at different plot sizes may result from differences in the underlying species distribution pattern. For example, the clustered distribution pattern of *Z. simplex* (the herb species) may allow the detection of the species even in small plot sizes, since the species occurs at high densities wherever it is present. On the other hand, species with wider and more sparsely distributed populations may not be detected even in areas where the species is present and environmental conditions are suitable. In the latter case, the goodness of fit of the model may be reduced as a consequence of the species being classified as present and absent in different locations with similar environmental conditions. Importantly, when the fit of the model is low, the likelihood of correctly predicting the distribution of species can be underestimated by the model (see below).

In addition to the effects of plot size on the shape of the cumulative frequency curves and the predictive capacity of the SDMs, our results also suggest that the accuracy of species distribution maps produced using species distribution models is significantly influenced by plot size. This likely results from the fact that reducing plot size increases the frequency of absences in the observational data, which alters the regression equation, reducing the height of the species–environment response curve: if the classification threshold is held constant, a reduction in the height of the species response curve automatically reduces the area that the model predicts species presence. Consequently, reducing plot size generally reduces the probability that a species will be predicted to be present in any given area on the SDM-derived map. Notably, such an effect is not specific to logistic regression models, but will emerge any time a binary classification method is used. This problem could be circumvented by displaying the probability of occurrence rather than a binary distribution, lending support to the idea that probabilistic maps are preferable to binary ones (see Supplementary material Fig. A3) (Bonham-Carter 1994; Lenton et al. 2000).

However, the increasing number of absences with decreasing plot size had little effect on the overall accuracy of SDM-derived maps. This is likely the sole consequence of the way that overall accuracy is calculated: the increased frequency of absences in both the training and the validation data would lead to an increase in the number of true absences, which would keep overall accuracy high as plot size was reduced. Notably, however, further reductions in plot size, toward 0 m radius, would cause overall accuracy to asymptotically approach 100%, an effect that would be fully attributable to chance agreements. Kappa accuracy, or K hat (Tsoar et al. 2007; Skidmore 1999; Fielding and Bell 1997; Cohen 1960), which estimates accuracy while removing the effects of chance agreements, provides a means of assessing the accuracy of the method employed to produce species distribution

maps, rather than a simple assessment of the overall accuracy of the maps alone (which treats chance agreements as accurate predictions). Kappa accuracy was low and did not differ significantly from zero at the 8 m radius plot size for three out of the four plant species, but increased with increasing plot size, suggesting that the method used to produce the maps—the species distribution models themselves—performed significantly better with larger plot sizes. Thus, in logistic regression-based species distribution models, data collected using larger plot sizes has a greater probability of correctly predicting species presence and Kappa seems to be a more appropriate statistic for assessing the quality of species distribution maps than overall accuracy, as suggested by Manel et al. (2001).

Our results suggest that a reduction in the size of the plots used to collect biological data affects both the likelihood of correctly predicting the distribution of species and the accuracy of the maps produced. This is likely the consequence of the relatively greater number of sites where species were observed as "absent" when smaller plot sizes were employed. In fact, larger plot sizes include many more 'present' observations than 'absent' observations for three of the four species (see Table 1). The exception is the herb, *Zygophyllum simplex*. There were more "present" observations for *Z. simplex* even in the smaller plot sizes and the logistic regression model has a better fit at this scale (as demonstrated by the relatively higher adjusted r^2 in the smaller plot sizes, see Table 2d). McPherson et al. (2004) and Guisan et al. (2007) also indicated that models constructed using data with a large number of species occurrences generally perform better than models constructed from data with a large number of absences. While a large number of species occurrences are more likely to be obtained as sample size is increased, data collection efforts should be matched to the amount of data that is necessary to successfully employ the model in order to optimize the use of limited resources (time and money). In addition, the use of excessively large plot sizes carries the risk of forcing spurious correlations between species occurrences and environmental variables, since species occurrences may be inadvertently matched to unrelated environmental conditions as the environmental variables are averaged over large areas (see Guisan et al. 2007). Our results therefore suggest that it is important to determine the optimal plot size for field observations before attempting to develop SDMs. This recommendation is likely to represent a challenge for conservation biologists, however, since many SDMs are created using preexisting data sets. In fact, species distribution maps (or maps that predict the likelihood of species occurrences) are often created from SDMs using species occurrence data generated from presence only records in museum or herbarium collections, because these data are increasingly accessible electronically (see also Graham et al. 2004; Soberón and Peterson 2005; Elith et al. 2006).

4.1 Consequences for conservation

The fact that plot size influenced both the predicted species distribution range and the accuracy of these predictions suggests that plot size must be explicitly considered in the process of producing species distribution maps. In fact, underestimations of the likelihood of predicting the distribution of species will

occur whenever data is collected using a plot size that is smaller than the plot size at which the species presence-area curve equilibrates. Importantly, such a bias may lead to a failure to identify and safeguard potentially suitable and already inhabited areas. Fortunately, such errors may be avoided by selecting an appropriate sampling design, but it is important to remember that the most appropriate sampling design will differ by species: our analyses revealed considerable differences in the optimal plot size for different species (Fig. 2). This point seems particularly pertinent at present, as species distribution models are increasingly reported for larger numbers of species using a single, standard plot dimension for all species (e.g., Karl et al. 2000). Such a generalized approach neglects inter-specific variation in optimal plot size at the risk of underestimating species predictions for the all species whose optimal plot size has not yet been reached. We therefore advise establishing optimal plots sizes for every species of interest prior to the development of species distribution models and then selecting a standard plot size that minimizes bias across all the species of interest, assuming a standard plot size is necessary.

Acknowledgments We are indebted to the Polytechnic of Namibia, Windhoek, and the many people of Safoentein who assisted with field work. We are grateful to three anonymous reviewers, whose comments helped to improve this paper. We thank Dr. Iris van Duren for her insightful comments on a previous draft of this paper. This research was supported by a grant from the Netherlands Fellowship Program to SP.

References

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43(6):1223–1232

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol Model 157(2–3):101–118

Austin MP (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol Model 200(1–2):1–19

Berg Å, Gärdenfors U, von Proschwitz T (2004) Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden—importance of environmental data quality and model complexity. Ecography 27(1):83–93

Bio AMF, De Becker P, Bie ED, Huybrechts W, Wassen M (2002) Prediction of plant species distribution in low land river valleys in Belgium: modeling species response to site conditions. Biodivers Conserv 11(12):2189–2216

Bonham-Carter GF (1994) Geographic information systems for geoscientists. Pergamon Press, Oxford

Boone RB, Krohn WB (2002) Modeling tools and accuracy assessment. In: Scott JM et al (eds) Predicting species occurrences: issues of accuracy and scale. Inland Press, Washington, pp 265–270

Buckland ST, Elston DA (1993) Empirical models for the spatial distributions of wildlife. J Appl Ecol 30:478–495

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

De Leeuw J, Ottichilo WK, Toxopeus AG, Prins HHT (2002) Application of remote sensing and geographic information systems in wildlife mapping and modelling. In: Skidmore A (ed) Environmental modelling with GIS and remote sensing. Taylor and Francis, London, pp 121–145

Dixon B (2004) Prediction of ground water vulnerability using an integrated GIS-based Neuro-Fuzzy techniques. J Spatial Hydro 4(2):1–38

Dungan JL, Perry JN, Dale MRT, Legendre P, Citron-Pousty S, Fortin MJ, Jakomulska A, Miriti M, Rosenberg MS (2002) A balanced view of scale in spatial statistical analysis. Ecography 25(5):626–640

Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y,

Overton JM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberon J, Williams S, Wisz MS, Zimmermann NE (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129–151

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24(1):38–39

Garrison BA, Erickson RA, Patten MA, Timossi IC (2000) Accuracy of wildlife model predictions for bird species occurrences in California counties. Wildl Soc Bull 28(3):667–674

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol Evol 19(3):497–503

Gray A (2003) Monitoring stand structure in mature coastal Douglas-fir forests: effect of plot size. For Ecol Manage 175(1–3):1–16

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecol Model 135(2):147–186

Guisan A, Graham CH, Elith J, Huettmann F (2007) Sensitivity of predictive species distribution models to change in grain size. Divers Distrib 13(3):332–340

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29(5):773–785

Hurley JF (1986) Summary: development, testing, and application of wildlife–habitat models–the researcher's viewpoint. In: Verner J, Morrison ML, Ralph CJ (eds) Wildlife 2000: modelling habitat relationships of terrestrial vertebrates. University of Wisconsin Press, USA

Huston MA (1994) Biological diversity: the coexistence of species on changing landscapes. Cambridge University Press, Cambridge

ILWIS 3.0 Academic (2001) The integrated land and water information system (ILWIS) software. International Institute of Geo-information Science and Earth Observation, Enschede

Johnson CJ, Gillingham MP (2005) An evaluation of mapped species distribution models used for conservation planning. Environ Conserv 32(2):117–128

Karl JW, Heglund PJ, Garton EO, Scott JM, Wright NM, Hutto RL (2000) Sensitivity of species habitat-relationship model performance to factors of scale. Ecol Appl 10(6):1690–1705

Kerr JT, Southwood TRE, Cihlar J (2001) Remotely sensed habitat diversity predicts butterfly species richness and community similarity in Canada. PNAS 98(20):11365–11370

Leathwick JR (1998) Are New Zealand's Nothofagus species in equilibrium with their environment? J Veg Sci 9(5):719–732

Leathwick JR, Austin MP (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology 82(9):25–60

Lenton SM, Fa JE, Del Val JP (2000) A simple non-parametric GIS model for predicting species distribution: endemic birds in Bioko Island, West Africa. Biodivers Conserv 9(7):869–885

Leos K, Martin D, Michal H, Ivana J, Tomàš T (2001) Scale-dependent biases in species counts in a grass land. J Veg Sci 12(5):699–704

Liu C, Berry PM, Dawson TP, Pearson RG (2005) Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28(3):389–393

Lütolf M, Bolliger J, Kienast F, Guisan A (2009) Scenario-based assessment of future land use change on butterfly species distributions. Biodivers Conserv 18(5):1329–1347

Manel S, Dias JM, Buckton ST, Ormerod SJ (1999) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J Appl Ecol 36(5):734–747

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. J Appl Ecol 38(5):921–931

McPherson JM, Jetz W, Rogers DJ (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J Appl Ecol 41(5):811–823

Miller J, Franklin J (2002) Modeling the distribution of vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecol Model 157(3–4):27–47

Miller J, Franklin J, Aspinall R (2007) Incorporating spatial dependence in predictive vegetation models. Ecol Model 202(3–4):225–242

Mushinzimana E, Stephen M, Noboru M, Li L, Chen-chieh F, Ling B, Uriel K, Cindy S, Louisa B, Guofa Z, Andrew KG, Guiyun Y (2006) Landscape determinants and remote sensing of *anapheline mosquito* larval habitats in the western Kenya highlands. Malar J 5:13

Openshaw S (1996) Developing GIS-relevant zone-based spatial analysis methods. In: Longley P, Batty M (eds) Spatial analysis: modelling in a GIS environment. Wiley, New York, pp 55–74

Openshaw S, Taylor PJ (1981) The modifiable areal unit problem. In: Wrigley N, Bennett RJ (eds) Quantitative geography. Routledge and Keegan Paul Ltd, London, pp 60–69
Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecol Model 133(3):225–245
Reese GC, Wilson KR, Hoeting JA, Flather CH (2005) Factors affecting species distribution predictions: a simulation modeling experiment. Ecol Appl 15(2):554–564
Santika T, Hutchinson MF (2009) The effect of species response form on species distribution model prediction and inference. Ecol Model 220(19):2365–2379
Schlossberg S, King DI (2009) Post logging succession and habitat usage of shrubland birds. J Wildl Manage 73(2):226–231
Segurado P, Araujo MB (2004) An evaluation of methods for modelling species' distributions. J Biogeogr 3(10):1555–1568
Seoane J, Carrascal LM, Alonso CL, Palomino D (2005) Species-specific traits associated to prediction errors in bird habitat suitability modelling. Ecol Model 185(2–4):299–308
Simon E (1997) Detailed land use plan for the sesfontein constituency. Ministry of Agriculture Water and Rural Development, Windhoek
Skidmore AK (1999) Accuracy assessment of spatial information. In: Stein A, van der Meer F, Gorte BGH (eds) Spatial statistics for remote sensing. Kluwer, Dordrecht
Soberón J, Peterson AT (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers Inform 2:1–10
Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148(1):1–13
Stohlgren TJ (2007) Measuring plant diversity: lessons from the field. Oxford University Press, New York
Syartinilia, Tsuyuki S (2008) GIS-based modeling of Javan Hawk-Eagle distribution using logistic and autologistic regression models. Biol Conserv 141(3):756–769
Tsoar A, Allouche O, Steinitz O, Rotem D, Kadmon R (2007) A comparative evaluation of presence-only methods for modelling species distribution. Divers Distrib 13(4):397–405
Venier LA, McKenney DW, Wang Y, McKee J (1999) Models of large-scale breeding-bird distribution as a function of macro-climate in Ontario, Canada. J Biogeogr 26(2):315–328
Wu J (2004) Effects of changing scale on landscape pattern analysis: scaling relations. Landscape Ecol 19(2):125–138
Wu JK, Bruce J, Li H, Loucks OL (eds) (2006) Scaling and uncertainty analysis in ecology: methods and applications. Springer, New York, 351 p
Zimmermann NE, Edwards TC, Moisen G, Frescino TS, Blackard JA (2007) Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. J Appl Ecol 44(5):1057–1067