Diversity and Distributions (2001) **7**, 15–27

BIODIVERSITY RESEARCH

A PCA-based modelling technique for predicting environmental suitability for organisms from presence records

M. P. ROBERTSON*[1], N. CAITHNESS[2] and M. H. VILLET[1] [1]*Department of Zoology and Entomology, Rhodes University, Grahamstown, 6140, South Africa, *Corresponding author, E-mail: m.robertson@ru.ac.za,* [2]*Natural History Museum, London SBD SW7, U.K.*

Abstract. We present a correlative modelling technique that uses locality records (associated with species presence) and a set of predictor variables to produce a statistically justifiable probability response surface for a target species. The probability response surface indicates the suitability of each grid cell in a map for the target species in terms of the suite of predictor variables. The technique constructs a hyperspace for the target species using principal component axes derived from a principal components analysis performed on a training dataset. The training dataset comprises the values of the predictor variables associated with the localities where the species has been recorded as present. The origin of this hyperspace is taken to characterize the centre of the niche of the organism. All the localities (grid-cells) in the map region are then fitted into this hyperspace using the values of the predictor variables at these localities (the prediction dataset). The Euclidean distance from any locality to the origin of the hyperspace gives a measure of the 'centrality' of that locality in the hyperspace. These distances are used to derive probability values for each grid cell in the map region. The modelling technique was applied to bioclimatic data to predict bioclimatic suitability for three alien invasive plant species (*Lantana camara* L., *Ricinus communis* L. and *Solanum mauritianum* Scop.) in South Africa, Lesotho and Swaziland. The models were tested against independent test records by calculating area under the curve (AUC) values of receiver operator characteristic (ROC) curves and kappa statistics. There was good agreement between the models and the independent test records. The pre-processing of climatic variable data to reduce the deleterious effects of multicollinearity, and the use of stopping rules to prevent overfitting of the models are important aspects of the modelling process.

Key words. Bioclimatic modelling, biological invasions, *Lantana camara*, potential distribution, principal components analysis, *Ricinus communis*, *Solanum mauritianum*.

INTRODUCTION

In response to the needs of environmental managers, a wide variety of recent biogeographical distribution models have been applied to a selection of biological problems. They have been used to predict the potential distribution of problem organisms such as weeds (Panetta & Dodd, 1987; Panetta & Mitchell, 1991; Sindel & Michael, 1992; Beerling *et al.*, 1995), and disease vectors, including Tsetse flies (Rogers & Williams, 1993; Rogers *et al.*, 1996; Robinson *et al.*, 1997) and ticks (Rogers & Randolph, 1993; Cumming, 2000). They have been used to assess the potential impacts of climate change on species distributions (Lindenmayer *et al.*, 1991; Rogers & Randolph, 1993; Beerling *et al.*, 1995; Schulze & Kunz, 1995), and to determine where new populations of a

threatened species could be established (Pfab & Witkowski, 1997) or where extinct populations may have occurred (Bauer *et al.*, 1994). These models have found applications in conservation (Lindenmayer *et al.*, 1991; Osborne & Tigar, 1992; Austin *et al.*, 1996; Lloyd & Palmer, 1998) and ecoclimatic site matching of forestry species (Richardson & McMahon, 1992).

The above models rely on strong, often indirect, links between species' locality records and predictor variables and can thus be termed *correlative models* (Beerling *et al.*, 1995). Correlative models use locality (distribution) records as surrogates for explicit performance parameters. They can be classified as either *group discrimination techniques*, which use both presence and absence locality records, or *profile techniques*, which use only presence locality records (Caithness, 1995).

Examples of group-discrimination techniques include those models based on discriminant analysis (Rogers & Randolph, 1993; Rogers & Williams, 1993; Rogers *et al.*, 1996), logistic regression (Osborne & Tigar, 1992; Cumming, 1999; Higgins *et al.*, 1999) and decision-tree-based methods (Walker, 1990; Lees, 1994; Michaelsen *et al.*, 1994; Williams *et al.*, 1994). Examples of profile techniques include the approaches used in the modelling packages known as BIOCLIM (Nix, 1986; Busby, 1991) and DOMAIN (Carpenter *et al.*, 1993).

We describe a profile technique for predicting suitability based on principal components analysis. We then use the technique in combination with climatic predictor variables to illustrate the prediction of bioclimatic suitability for three alien plant species in South Africa, Lesotho and Swaziland.

THE PRINCIPLE COMPONENTS ANALYSIS (PCA) TECHNIQUE

Principal components analysis (PCA) is a multivariate technique that produces a set of abstract variables (called principal components) which are weighted linear combinations of the original variables (James & McCulloch, 1990). The components are constructed so as to maximize the variance explained by each component and in such a manner that they are uncorrelated (orthogonal).

A map of the area for which one wants to predict an organism's distribution is subdivided into regular grid cells. This allows the map to be represented as a matrix of values. The values of the predictor variables associated with grid-cells in which the target organism has been recorded as present are referred to as the 'training dataset'. The values of the predictor variables associated with all the grid-cells within the map region comprise the 'prediction dataset' (i.e. the training dataset plus the values associated with the remaining unsampled grid-cells).

The essence of the method is as follows. A PCA is performed on the training dataset to construct a mathematical hyperspace in which each orthogonal axis is defined by an orthogonal principal component axis. The value of the component score on a principal component axis associated with a particular observation defines the position of that observation on that axis. The n component scores of an observation thus define the position of that observation as a point in the n-dimensional hyperspace. The origin of this hyperspace is taken to characterize the centre of the niche of the organism in terms of the predictor variables. The Euclidean distance from any point to the origin gives a measure of the 'centrality' of the point in the hyperspace defined by the values of the observations in the training set.

If all the values of the predictor variables associated with the prediction set are mapped into the hyperspace defined by the training set, then one can calculate the distance from each unsampled site to the multivariate origin of the hyperspace. The squared Euclidean distance between any two points in a n-dimensional space can be calculated by taking the sum of squares of the Manhattan distances (using Pythagoras' theorem), where the number of terms in the equation is equal to the number of dimensions defining the space. The squared distance between a point and the origin of the n-dimensional hyperspace is thus calculated by taking the sum of squares of the component scores.

This distance can be used to calculate a probability of bioclimatic suitability for each locality (grid-cell) as follows. Based on the assumption that the fundamental niche of an organism is generally considered to follow a broad Gaussian curve (Austin & Meyers, 1996), a normal distribution would be most appropriate for this purpose. As the distance of a point from the origin of the hyperspace is calculated from the sum of its squared component scores, and as the sum of squares of n standard normal random variates is distributed as chi-square with n degrees of freedom

(Sokal & Rohlf, 1987), a chi-square distribution can be used instead of a normal distribution. This assumes that the further a point is from the origin of the hyperspace, the less suitable it is for the target species. The probability associated with each chi-square value can thus be determined by referring to a chi-square distribution (a chi-square distribution is equivalent to a squared normal distribution). These values can be mapped back to the cells of the original real-world map. An output of the model is therefore a map of grid-cells, with each grid-cell containing a probability value, and these probability values can be interpreted as an indication of the suitability of that grid-cell for the target organism.

METHODS

The target species

The target species were selected due to a combination of: their weed status; their priority ranking using a prioritization system (Robertson *et al*. in prep.) and data availability in existing databases. In addition, these species were selected because they could be identified easily and were unlikely to be confused with other species of similar appearance. This is likely to have resulted in fewer false positive and false negative errors as a result of misidentification. Data obtained from existing databases would be particularly prone to misidentification errors because the data housed in these databases are supplied by large numbers of volunteers.

The data

Data sources

Digital predictor variable maps (climatic variables and altitude) of South Africa, Lesotho and Swaziland developed by Schulze *et al*. (1997) were selected for the purpose of illustrating this method of predictive modelling. Each of the climatic predictor variables was interpolated from point data obtained from a network of weather recording stations distributed throughout South Africa, to produce continuous digital maps at a resolution of 60 pixels per degree (Schulze *et al.*, 1997). Localities representing species presence were obtained from the Southern African Plant Invader's Atlas (Henderson, 1998) and the National Herbarium's Computerized Information System (PRECIS). Additional records of presence or absence were collected, using a GPS, on road transects selected to sample major climatic gradients represented in the map region. If a target species occurred continuously along any part of a transect then its position was recorded approximately every 2–4 km to represent that species' presence. Absence records were only recorded if they were at least more than 10 km from any presence localities. The absence records were used only for model assessment and not for model building. The presence data used to predict the distribution of an organism obviously represent records collected from that organism's realized niche.

Climatic variable preprocessing

To reduce the dimensionality of available climatic variable data, principal components analyses (PCAs) were performed on each of 12 mean monthly rainfall maps, 12 monthly potential evaporation maps, 12 mean daily maximum temperature and 12 mean daily minimum temperature maps. PCA has previously been employed as a pre-analytical data reduction technique used in distribution modelling (Osborne & Tigar, 1992; Buckland & Elston, 1993; Robinson *et al.*, 1997).

Those principal component axes whose eigenvalues were greater in magnitude than eigenvalues obtained from datasets of random numbers of the same sample size were retained as predictor variables. This follows the 'broken stick' stopping rule for PCA (Jackson, 1993). Ten predictor variables were selected (Table 1).

Locality data

Localities where *Lantana camara* L., *Ricinus communis* L. and *Solanum mauritianum* Scop. were present were partitioned randomly into a set of training localities and a set of testing localities in a ratio of 3 : 1, based on Huberty's (1994) recommendations. For each species, the values of the predictor variables (Table 1) corresponding with the training localities comprised the training dataset for the model.

Implementation

In the first step, the values of the training set were standardized by subtracting the mean and dividing by the standard deviation for each variable.

Table 1 Predictor variables selected for building the distribution models

No.	Predictor variable
1	Digital elevation model
2	Number of days with frost
3	Component axis 1 of a PCA on 12-monthly potential evaporation surfaces
4	Component axis 2 of a PCA on 12-monthly potential evaporation surfaces
5	Component axis 1 of a PCA on 12-monthly maximum temperature surfaces
6	Component axis 2 of a PCA on 12-monthly maximum temperature surfaces
7	Component axis 1 of a PCA on 12-monthly minimum temperature surfaces
8	Component axis 2 of a PCA on 12-monthly minimum temperature surfaces
9	Component axis 1 of a PCA on 12-monthly rainfall surfaces
10	Component axis 2 of a PCA on 12-monthly rainfall surfaces

This is equivalent to performing the eigenanalysis on the correlation matrix instead of the covariance matrix (Fig. 1), and removes the effects of differing measuring units. The matrix of standardized values (U) is arranged so that the n variables are in columns and the x observations are in rows. The means and standard deviations are kept for the third step of the analysis.

Next, one performs a PCA on the matrix U, which gives a matrix (V), in which the n columns (eigenvectors) are the component loadings for each axis of the model. Each eigenvector has a corresponding eigenvalue (denoted by λ) describing its variance (Fig. 1).

In the third step, the observations of the prediction dataset were standardized by the means and standard deviations calculated from the training dataset in the first step of this analysis to produce matrix W (i.e. the mean and standard deviation calculated for each variable from the training set were used to standardize the corresponding variables from the prediction set). The effect of standardizing the prediction set (using means and standard deviations of the training set) is to centre it on the origin of the hyperspace, which allows the origin to be viewed as the niche optimum for the target organism.

This matrix was then multiplied by the matrix V (containing the n columns of component loadings) to produce a matrix (Z) of component scores for all map localities in the model (Fig. 1). Conceptually this step projects the prediction set into the hyperspace defined by the training set.

The principle components of a PCA are constructed so that most of the variance in the original variables is accounted for in the first few components. Using too many components results in overfitting of the model which usually results in loss of generality. In the fourth step of the modelling process, a stopping rule was used to determine the optimum number of principal components that should be included in the model so that overfitting is avoided. In a review of stopping rules, Jackson (1993) found that the 'broken stick' method was the most reliable of a range of methods for deciding how many principal components to include. This method estimates the distribution of eigenvalues obtained from random data and admits only components with eigenvalues that exceed these estimates. To make our model more conservative, only those components whose eigenvalues exceeded the mean plus two standard deviations of these estimates were used in our models (Fig. 1), following Caithness (1995).

Because the variance on each PCA axis is different, spherical probability contours, concentric about the origin of the hyperspace, can only be assumed if the variance on each component axis is first standardized. In the fifth step (Fig. 1), the variances of each component axis were therefore standardized by dividing the component scores of each component (in Z) by their respective eigenvalues (λ) to produce a matrix of standardized component scores (Z). In step six, the probability associated with each observation was obtained by summing the squares of the standardized component scores and substituting this value into the chi-square probability distribution function (Fig. 1). In the final step, the probability values for each grid cell were mapped back to their associated original geographical coordinates of each observation (Fig. 1). The calculations were performed using MATLAB (a numerical computation and visualization software package) and the maps were produced using IDRISI32 (a raster-based GIS software package).

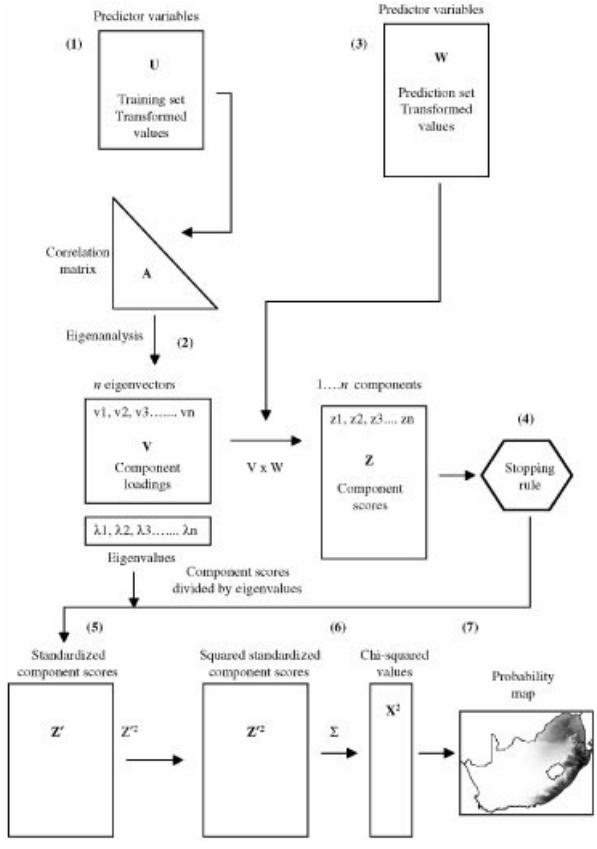

Fig. 1 Implementation of the PCA modelling technique. The numbers in round brackets correspond with the steps described in the methods section.

Model assessment

In order to have confidence in a predictive model or in the approach used to build it, the model's predictions should be assessed by some objective means. This is usually performed qualitatively by an expert who is familiar with the species or quantitatively using a set of independent testing locality records and an accuracy assessment measure. Fielding & Bell (1997) reviewed a number of model assessment measures for quantitatively assessing a model's prediction success. One of the most robust measures described by them is derived from a receiver operator characteristic (ROC) plot. For a recent application of ROC see Cumming (2000).

ROC plots

If those testing localities where a target species has been recorded as present are termed 'positives' and those localities where it has been recorded as absent are termed 'negatives', then sensitivity is defined as the probability that the model produces a positive result in a positive locality and specificity is the probability that the model produces a negative result in a negative locality (Table 2). A ROC plot is obtained by plotting all sensitivity values on the *y*-axis against their equivalent (1-specificity) values for all available decision thresholds on the *x*-axis (Fielding & Bell, 1997). The area under the ROC function (AUC) provides a single measure of overall accuracy that is not dependent on a particular decision threshold (Fielding & Bell, 1997). The value of the AUC ranges between 0.5 and 1, where 0.5 indicates randomness and 1 indicates a perfect fit.

Table 2 A confusion matrix used to define sensitivity and specificity (Fielding & Bell, 1997) where + indicates presence and – indicates absence, sensitivity = a/(a + c) and specificity = d/(b + d)

		Observed +	–
Predicted	+	a	b
	–	c	d

Area under the curve (AUC) values of ROC curves were calculated for each species using a set of testing localities. These calculations were performed using Analyse-It Clinical Laboratory software. The set of testing localities used to calculate AUC values comprised a set of localities representing species presence (obtained from the partition described above) as well as a set of localities where the species was recorded as absent (Fig. 2). Although absence data are not used to build the model they are required by the ROC accuracy assessment measure for model testing.

As the ROC accuracy measure is considered to be relatively new to ecology (Packer *et al.*, 1999) and may not be well known, we also provide kappa statistics for each of the species (Fielding & Bell, 1997). To calculate kappa values, we used a probability threshold of 0.3 for assigning probabilities to presence or absence categories (i.e. probabilities greater than 0.3 were assign presence and values less than or equal to 0.3 were assigned absence) for calculating the parameters in the confusion matrix (Table 2). Monserud & Leemans (1992) suggested the following ranges of agreement for the kappa statistic (K): no agreement < 0.05; very poor 0.05–0.20; poor 0.20–0.40; fair 0.40–0.55; good 0.55–0.70; very good 0.70–0.85; excellent 0.85–0.99 and perfect 0.99–1.00.

RESULTS

Although most of the standardized component scores calculated from the training sets for the three species differ significantly from a normal distribution (Table 3) they do not appear to deviate radically from normality (Fig. 3).

Regions of high bioclimatic suitability for *L. camara* include the coastal regions of the Eastern Cape, parts of KwaZulu-Natal, Mpumalanga, Gauteng, Northern Province and Swaziland (Figs 2 and 4). The Free State Province, Lesotho, North-West, Northern Cape and Western Cape provinces demonstrate low bioclimatic suitability. The regions of high suitability correspond approximately with the Savanna (excluding the Kalahari Thornveld) and Forest biomes (Low & Rebelo, 1996). Those areas of lower suitability appear to

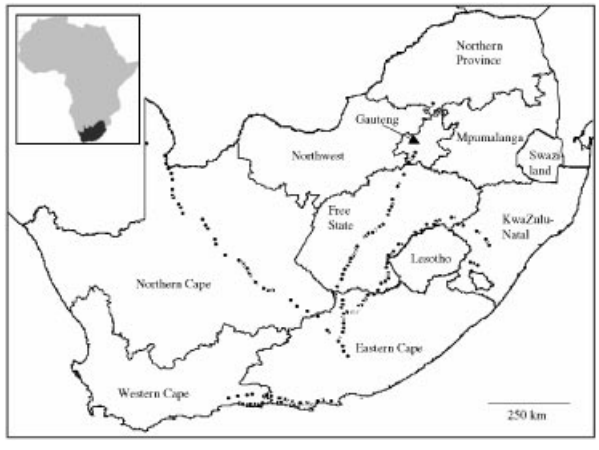

Fig. 2 Map of South Africa (indicating the provinces), Lesotho and Swaziland. Black symbols indicate localities from which the absence test data were drawn for model testing. In the inset, black indicates southern Africa relative to Africa.

Table 3 Shapiro–Wilks' W statistics and Kolmogorov–Smirnov one-sample D statistics with Lilliefors probabilities calculated from compontent scores (for components 1–3) of the training sets of each species. If the W statistic or D statistics are significant (indicated by *), then the hypothesis that the respective distribution is normal should be rejected

	Comp.	W statistic	*P*	D statistic	*P*
Lantana camara	1	0.851	0.000*	0.186	$P < 0.01$*
	2	0.963	0.000*	0.089	$P < 0.01$*
	3	0.740	0.000*	0.158	$P < 0.01$*
Ricinus communis	1	0.857	0.000*	0.112	$P < 0.01$*
	2	0.854	0.000*	0.202	$P < 0.01$*
	3	0.973	0.000*	0.056	$P < 0.10$
Solanum mauritianum	1	0.988	0.015	0.050	$P < 0.10$
	2	0.689	0.000*	0.206	$P < 0.01$*
	3	0.916	0.000*	0.107	$P < 0.01$*

be associated with the Grassland biome (Low & Rebelo, 1996). *L. camara* is reported to invade forests, plantation margins, savanna and watercourses (Henderson, 1995), which would explain the correspondence between the model's predictions and the Savanna and Forest biomes. An AUC value of 0.991 was calculated for this species (using 78 presence and 172 absence localities) which indicates a good fit between the distribution predicted by the model and the independent test localities. An AUC value of 0.991 indicates that in 991 of 1000 cases, random selection of a point from the group of known occurrences will be associated with a probability that is greater than that of a random selection from the negative group (Fielding & Bell, 1997). A kappa value of 0.909 was calculated, which can be considered to indicate 'excellent' agreement between the model and the test data (Monserud & Leemans, 1992).

Regions of high bioclimatic suitability for *R. communis* include the coastal regions of the Eastern Cape, parts of KwaZulu-Natal, Mpumalanga, Northern Province and Swaziland (Fig. 5). The river valleys, particularly in the Eastern Cape and KwaZulu-Natal, appear to be particularly suitable for this species, and the high-altitude central plateau appears to be less suitable. The regions of high suitability appear to correspond approximately with the Savanna and Forest biomes and those of lower suitability with the Grassland biome (Low & Rebelo, 1996). *R. communis* is reported to invade riverbanks, riverbeds, roadsides and wasteland (Henderson, 1995). This would largely explain the high suitability predicted for the river valleys in the Eastern Cape and KwaZulu-Natal (Fig. 5). An AUC value of 0.948 was calculated for this species (using 68 presence and 134 absence localities). This AUC value (0.948) also indicates a good fit between the model and the independent test localities, although the *L. camara* model (AUC 0.991) performed slightly better. A kappa value of 0.799 was calculated, which can be considered to indicate 'very good' agreement between the model and the test data (Monserud & Leemans, 1992).

Regions of high bioclimatic suitability for *S. mauritianum* include the higher altitude regions of Eastern Cape, KwaZulu-Natal, Mpumalanga and Swaziland (Fig. 6). The coastal regions appear to be less suitable for this species than the higher altitude regions although the high-altitude regions of Lesotho and the Free State are unsuitable. The highest suitability areas appear to be associated with the Forest biome (Low & Rebelo, 1996). *S. mauritianum* is reported to be associated with forest margins, plantations and wooded valleys (Henderson, 1995) which may explain the correspondence between areas predicted as high suitability for this species and the Forest biome. In addition, this species is considered to be the principal weed of South Africa's timber plantations (Bromilow, 1995), which are situated within the areas of high predicted suitability. An AUC value of 0.950 was calculated for this species (using 97 presence and 149 absence localities)

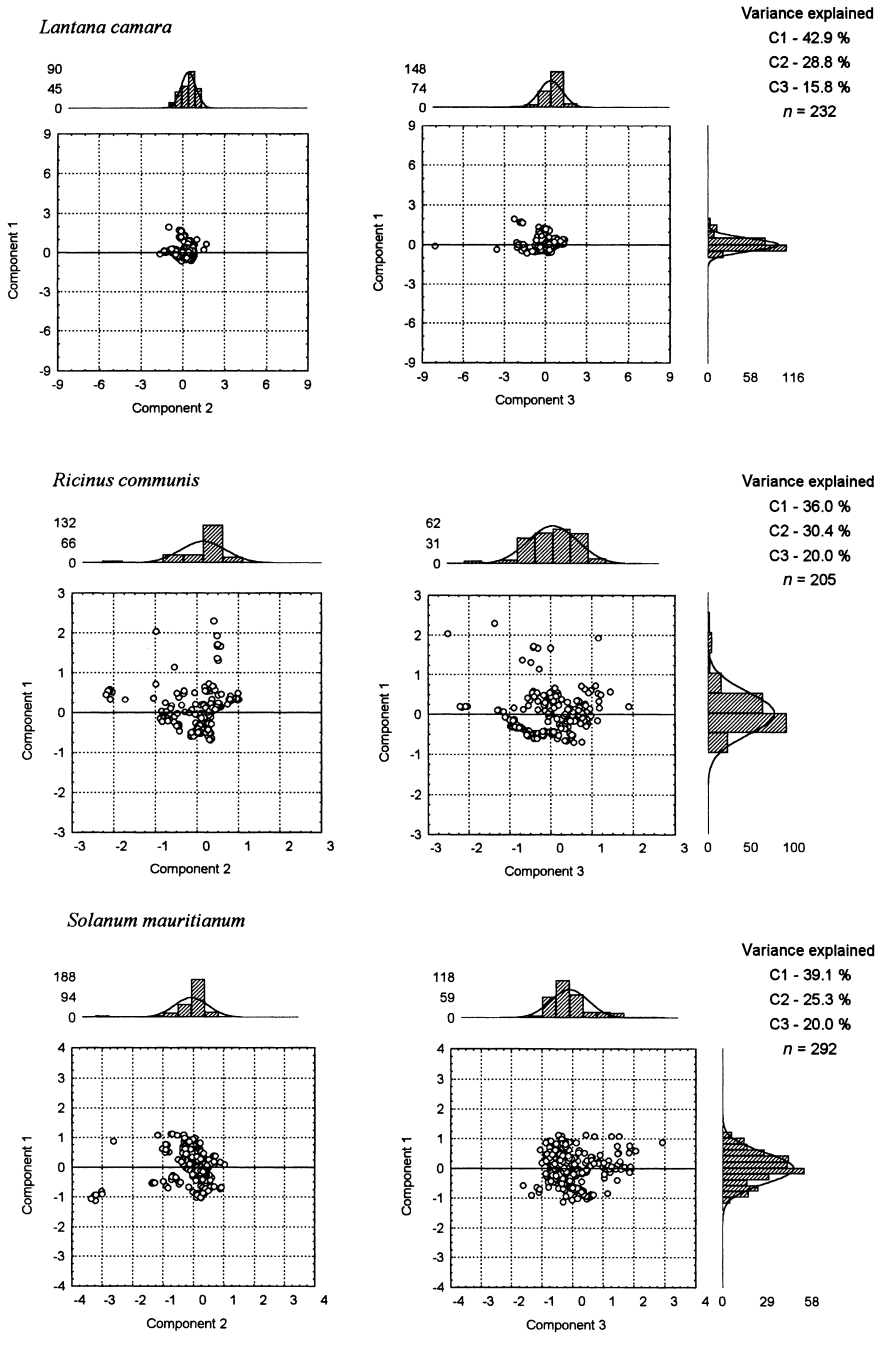

Fig. 3 Plots of component scores and histograms (with normal distribution curves), calculated from the training sets for each species. Plots are given for components 1 vs. 2 and components 1 vs. 3 for *L. camara*; *R. communis* and *S. mauritianum*. Only the first three components were included as the remaining components were excluded by the stopping rule. The percentage variance explained by each component is given for each species.

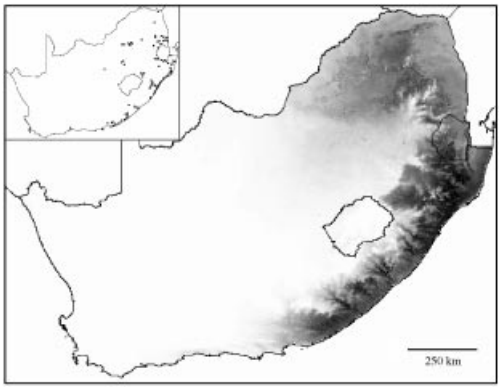

Fig. 4 Bioclimatic suitability map for *L. camara* in South Africa, Lesotho and Swaziland produced from 232 localities (see inset) where the species was recorded present (condition number = 10). Darker shades indicate higher probabilities.

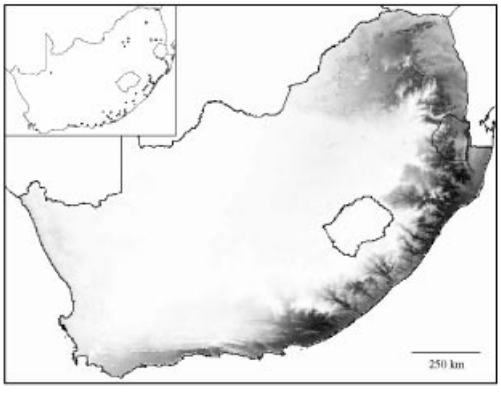

Fig. 5 Bioclimatic suitability map for *R. communis* in South Africa, Lesotho and Swaziland produced from 205 localities (see inset) where the species was recorded present (condition number = 5). Darker shades indicate higher probabilities.

which indicates a good fit between the model and the independent test localities. This model (AUC = 0.950) performed slightly better than the *R. communis* model (AUC = 0.948) but not as well as the *L. camara* model (AUC = 0.991). A kappa value of 0.726 was calculated, which can be considered to indicate 'very good' agreement between the model and the test data (Monserud & Leemans, 1992).

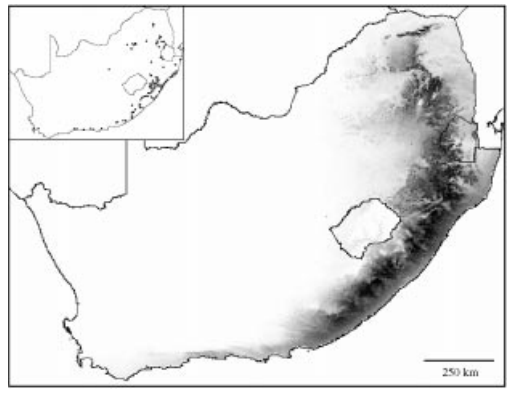

Fig. 6 Bioclimatic suitability map for *S. mauritianum* in South Africa, Lesotho and Swaziland produced from 292 localities (see inset) where the species was recorded present (condition number = 11). Darker shades indicate higher probabilities.

DISCUSSION

The modelling process described here can be summarized in a set of steps: climatic variable pre-processing; partitioning of locality records into training and testing sets; building the PCA model using the training set; and model assessment using independent testing locality records.

Climatic variable pre-processing

In addition to data reduction, pre-processing of the original variables is intended to remove or considerably reduces multi-collinearity in the predictor variables eventually used to build the models. When one or more linear relationships exist among the original variables they are said to be linearly dependent or multicollinear (Bernstein *et al.*, 1988). Multi-collinearity produces highly unstable results, especially in factor analysis and multiple regression, with the result that slight differences in sampling error or rounding may lead to substantially different results (Bernstein *et al.*, 1988). While this may not be considered to be a serious problem when PCA is used for data reduction, it becomes particularly important when it is used as a predictive tool and when one intends to analyse the resulting principal components further, as we have done.

Data that are multi-collinear have ill-conditioned covariance or correlation matrices (matrices that

are singular or nearly singular; Bernstein *et al.*, 1988). Multi-collinearity can be detected by means of the condition number (cn) that is calculated by dividing the square root of the largest eigenvalue by the square root of the smallest eigenvalue (Johnston, 1984). Condition numbers in the range of 20–30 indicate serious multi-collinearity (Johnston, 1984). The condition numbers calculated for the models produced for each species were below this (*L. camara* cn = 10; *R. communis* cn = 5; *S. mauritianum* cn = 11).

The PCA model

The predictive technique presented here has the advantage that it does not require absence locality data for the purposes of prediction, in contrast to group discrimination techniques. While group discrimination techniques should not be dismissed, there are a number of data quality issues associated with absence data that make it less desirable than presence data for the purposes of model training. Absence data are often not available (Margules & Austin, 1994) and may be considered to be less reliable than presence data (Fielding & Bell, 1997). Absence records are likely to be unreliable due to survey errors (particularly false absence errors) arising from local extinction, seasonal migration, hibernation, taxonomic errors or because insufficient time has elapsed for the species to colonize the area, e.g. alien invasive organisms. In the case of alien plants the chance of recording false absence records is high in cases where the plant is recorded absent at a site because insufficient time has elapsed for the plant to invade that area rather than because the area is climatically unsuitable. The technique described here is suited to cases where absence data are not available, are of low quality, or are difficult to acquire (for example, alien organisms).

The fundamental niche of an organism was defined by Hutchinson (cited in Schoener, 1990) as a *n-dimensional hypervolume* defined by *n* environmental dimensions within which the organism can survive and reproduce. The organism may be excluded from parts of its fundamental niche due to competition or other biotic interactions. The reduced hypervolume in which the organism can survive is its realized niche. The organism's occurrence along each axis of the fundamental niche is generally considered to follow a broad Gaussian curve (Austin & Meyers, 1996). In contrast, occurrence in the realized niche has been shown to exhibit various skewed shapes (Austin *et al.*, 1990), which is often attributed to competition. In a correlative model such as the one presented here, the locality records used to build the model are drawn from the realized niche of the organism and as a result are likely to demonstrate skewed responses which will differ among predictor variables as well as among species (Austin *et al.*, 1990). In the modelling technique described here, we use a normal distribution to describe the shape of the response on each component axis as a compromise among several possible responses. [The chi-squared distribution (which is equivalent to a squared normal distribution) can be used instead of the normal distribution because the component scores have to be squared in order to calculate the distance of a point from the origin of the hyperspace.] This technique is based on a fundamental niche concept because it uses a normal distribution to describe environmental responses and because it does not explicitly take biotic factors such as competition into account. As the technique is based conceptually on the fundamental niche, the predictions produced using this approach could be said to describe the fundamental niche of the target organism. However, as the prediction is based on locality records that are drawn from the realized niche, the resulting prediction would not entirely describe the fundamental niche of the target organism. Although the responses of the species modelled here are not normal, they do not appear to deviate radically from normality (Fig. 3). For this reason and the one outlined above, the use of a normal distribution is justifiable. In addition, the models have performed well against independent test records and also correspond with known habitat associations, indicating that the departures of the data from the modelling assumptions may not be serious. The data certainly seem to occupy a central cluster in the hyperspace, and do thin out away from the origin (Fig. 3).

Model assessment

Model assessment is an important component of the modelling process as it allows the user to objectively assess the quality of the model's predictions. The best means of objectively assessing model performance is to use an independent set

of locality records and a quantitative accuracy measure (Fielding & Bell, 1997). While model assessment using only presence data would be preferable, as the model is built using only presence data, accuracy assessment measures that use only presence data tend to be less rigorous and less objective than accuracy measures that rely on both presence and absence locality data (Fielding & Bell, 1997). As we are evaluating a new modelling technique we have used rigorous accuracy measures that use both presence and absence data (Fielding & Bell, 1997). The presence-only measures described in the literature are threshold measures based on a confusion matrix (Table 2). When absence data are not available, then parameters *b* and *d* in the confusion matrix cannot be calculated, thus limiting the measures to those containing parameters *a* and *c* only. These measures include Sensitivity [equation: a/(a + c)] and False Negative Rate [equation: c/(a + c)] (Fielding & Bell, 1997). These measures can only test for false negative errors but not for false positive errors, and for this reason are less rigorous.

Quantitatively, the high AUC values indicate a good fit between the models and the independent test localities, which in turn suggests that the modelling technique performs well. Kappa values indicate that the model performance could be classified as 'very good' (*R. communis* and *S. mauritianum*) to 'excellent' (*L. camara*) according to ranges defined by Monserud & Leemans (1992). In addition, the models have successfully identified areas corresponding to known habitat preferences, e.g. the correspondence between areas predicted as highly suitable for *S. mauritianum* and the Forest biome.

The major advantage of this technique is that it produces a statistically justifiable probability response surface using presence data instead of presence and absence data, as required by most other multivariate techniques. The technique is, however, unlikely to perform well when small samples (< 40) of locality records are used. Future research should compare the performance of profile and group discrimination models to investigate problems associated with the use of absence data for predictive modelling.

ACKNOWLEDGMENTS

We thank the Department of Agricultural Engineering at the University of Natal for making the climatic predictor variable data available to us in digital form; Lesley Henderson at the Southern African Plant Invaders Atlas for locality data; Craig Peter (Rhodes University) for collecting locality data; the National Botanical Institute for the use of data from the National Herbarium's (Pretoria) Computerized Information System (PRECIS); Adrian Craig and two anonymous referees for commenting on previous drafts of this manuscript; and Sarah Radloff for statistical advice. This work was funded by the National Research Foundation and Rhodes University.

REFERENCES

Austin, M.P. & Meyers, J.A. (1996) Current approaches to modelling the environmental niche of eucalypts: implications for management of forest biodiversity. *Forest Ecology and Management* **85**, 95–106.

Austin, M.P., Nicholls, A.O. & Margules, C.R. (1990) Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. *Ecological Monographs* **60**, 161–177.

Austin, G.E., Thomas, C.J., Houston, D.C. & Thompson, D.B.A. (1996) Predicting the spatial distribution of buzzard *Buteo buteo* nesting areas using a Geographical Information System and remote sensing. *Journal of Applied Ecology* **33**, 1541–1550.

Bauer, I.E., McMorrow, J. & Yalden, D.W. (1994) The historic ranges of three equid species of north-east Africa: a quantitative comparison of environmental tolerances. *Journal of Biogeography* **21**, 169–182.

Beerling, D.J., Huntley, B. & Bailey, J.P. (1995) Climate and the distribution of *Fallopia japonica*: use of an introduced species to test the predictive capacity of response surfaces. *Journal of Vegetation Science* **6**, 269–282.

Bernstein, I.H., Garbin, C.P. & Teng, G.K. (1988) *Applied multivariate analysis*, 1st edn. Springer-Verlag, New York.

Bromilow, C. (1995) *Problem plants of South Africa*, 1st edn. Briza Publications, Arcadia.

Buckland, S.T. & Elston, D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology* **30**, 478–495.

Busby, J.R. (1991) BIOCLIM — a bioclimatic analysis and prediction system. *Nature conservation: cost effective biological surveys and data analysis* (ed. by C.R. Margules and M.P. Austin), pp. 64–68. CSIRO, Melbourne.

Caithness, N. (1995) Pattern, process and the evolution of the African antelope (Mammalia: Bovidae). PhD Thesis, University of the Witwatersrand, Johannesburg.

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* **2**, 667–680.

Cumming, G.S. (2000) Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography* **27**, 441–445.

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.

Henderson, L. (1995) *Plant Invaders of Southern Africa*, 1st edn. Agricultural Research Council, Pretoria.

Henderson, L. (1998) Southern African plant invaders atlas (SAPIA). *Applied Plant Sciences* **12**, 31–32.

Higgins, S.I., Richardson, D.M. & Cowling, R.M. (1999) Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. *Conservation Biology* **13**, 303–313.

Huberty, C.J. (1994) *Applied discriminant analysis*, 1st edn. Wiley Interscience, New York.

Jackson, D.A. (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**, 2204–2214.

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

Johnston, J. (1984) *Econometric methods*, 3rd edn. McGraw-Hill, Auckland.

Lees, B.G. (1994) Decision trees, artificial neural networks and genetic algorithms for classification of remotely sensed and ancillary data. *7th Australian Remote Sensing Conference Proceedings* **1**, 51–59.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F. & Tanton, M.T. (1991) The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography* **8**, 371–383.

Lloyd, P. & Palmer, A.R. (1998) Abiotic factors as predictors of distribution in southern African bulbuls. *Auk* **115**, 404–411.

Low, A.B. & Rebelo, A.G. (1996) *Vegetation of South Africa, Lesotho and Swaziland*, 1st edn. Department of Environmental Affairs and Tourism, Pretoria.

Margules, C.R. & Austin, M.P. (1994) Biological models for monitoring species decline: the construction and use of databases. *Philosophical Transactions of the Royal Society, London, Series B* **344**, 69–75.

Michaelsen, J., Schimel, D.S., Friedl, M.A., Davis, F.W. & Dubayah, R.C. (1994) Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetation Science* **5**, 673–686.

Monserud, R.A. & Leemans, R. (1992) Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*. **62**, 275–293.

Nix, H.A. (1986) A biogeographical analysis of Australian elapid snakes. *Atlas of Elapid snakes of Australia* (ed. by R. Longmore), pp. 4–15. Australian Government Publishing Service, Canberra.

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

Packer, M.J., Canney, S.M., McWilliam, N.C. & Abdallah, R. (1999) Ecological mapping of a semiarid savanna. *Mkomazi: the ecology, biodiversity and conservation of a Tanzanian savanna* (ed. by M.J. Coe, N.C. McWilliam, G.N. Stone and M.J. Packer), pp. 43–68. Royal Geographical Society (with The Institute of British Geographers), London.

Panetta, F.D. & Dodd, J. (1987) Bioclimatic prediction of the potential distribution of skeleton weed *Chondrilla juncea* L. in Western Australia. *Journal of the Australian Institute of Agricultural Science* **53**, 11–16.

Panetta, F.D. & Mitchell, N.D. (1991) Bioclimatic prediction of the potential distributions of some weed species prohibited entry to New Zealand. *New Zealand Journal of Agricultural Research* **34**, 341–350.

Pfab, M.F. & Witkowski, E.T.F. (1997) Use of Geographical Information Systems in the search for additional populations, or sites suitable for re-establishment, of the endangered Northern Province endemic *Euphorbia clivicola*. *South African Journal of Botany* **63**, 351–355.

Richardson, D.M. & McMahon, J.P. (1992) A bioclimatic analysis of *Eucalyptus nitens* to identify potential planting regions in southern Africa. *South African Journal of Science* **88**, 380–387.

Robertson, M.P., Villet, M.H., Palmer, A.R., Fairbanks, D.H.K., Henderson, L., Higgins, S., Hoffmann, J.H., Le Maitre, D.M., Riggs, I., Shackleton, C.M. & Zimmermann, H.G. A proposed prioritization system for the management of weeds in South Africa, in preparation.

Robinson, T.P., Rogers, D.J. & Williams, B.G. (1997) Mapping tsetse habitat suitability in the common fly belt of southern Africa using multivariate analysis of climate and remotely sensed vegetation data. *Medical and Veterinary Entomology* **11**, 235–245.

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.

Rogers, D.J. & Randolph, S.E. (1993) Distribution of tsetse and ticks in Africa: past, present and future. *Parasitology Today* **9**, 266–271.

Rogers, D.J. & Williams, B.G. (1993) Tsetse distribution in Africa: seeing the wood and the trees.

Large-Scale Ecology and Conservation Biology (ed. by P.J. Edwards and R. May), pp. 247–271. Blackwell Scientific Publications, Oxford.

Schoener, T.W. (1990) The ecological niche. *Ecological concepts: the contribution of ecology to an understanding of the natural world* (ed. by J.M. Cherrett), pp. 79–113. Blackwell Scientific Publications, Oxford.

Schulze, R.E. & Kunz, R.P. (1995) Potential shifts in optimum growth areas of selected commercial tree species and subtropical crops in southern Africa due to global warming. *Journal of Biogeography* **22**, 679–688.

Schulze, R.E., Maharaj, M., Lynch, S.D., Howe, B.J. & Melvil-Thomson, B. (1997) *South African atlas of agrohydrology and climatology*, 1st edn. Water Research Commission, Pretoria.

Sindel, B.M. & Michael, P.W. (1992) Spread and potential distribution of *Senecio madagascariensis* Poir. (fireweed) in Australia. *Australian Journal of Ecology* **17**, 21–26.

Sokal, R.R. & Rohlf, F.J. (1987) *Introduction to biostatistics*, 2nd edn. Freeman, New York.

Walker, P.A. (1990) Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. *Journal of Biogeography* **17**, 279–289.

Williams, B.G., Rogers, D.J., Staton, G., Ripley, B. & Booth, T. (1994) Statistical modelling of georeferenced data: mapping tsetse distributions in Zimbabwe using climate and vegetation data. *Modelling vector-borne and other parasitic diseases* (ed. by B.D. Perry and J.W. Hansen), pp. 267–280. ILRAD, Nairobi.