The use of logistic regression in modelling the distributions of bird species in Swaziland

V. Parker
Avian Demography Unit, Department of Statistical Sciences, University of Cape Town, Rondebosch, 7701 South Africa

Received 21 December 1995; accepted after revision 19 February 1999

The method of logistic regression was used to model the observed geographical distribution patterns of bird species in Swaziland in relation to a set of environmental variables. Reporting rates derived from bird atlas data are used as an index of population densities. This is justified in part by the success of the modelling process. For each species the variables which were significantly related to its distribution were identified. Various methods for coding environmental variables from maps were investigated. A jack-knifing technique was used to demonstrate the predictive power of the logistic models. A criterion for assessing the goodness of fit of the logistic models was derived.

Introduction

Logistic regression (McCullagh & Nelder 1989, Collett 1991) was previously used to predict bird distributions in Lesotho based on a binary response variable representing the presence or absence of a species in a geographical grid unit by Osborne & Tigar (1992). They reduced their set of explanatory variables describing habitat and land use to four principal components, so that it was not possible to relate bird distributions to the explanatory variables that were observed. This analysis extends their work in three ways: firstly, it uses full binomial (rather than binary presence/absence) modelling, thus taking into account the relative abundance of species. Secondly, it relates the distributions of the bird species to the individual environmental explanatory variables. Thirdly, the availability of comprehensive data on the distribution of birds in the study area (Parker 1994) is made use of to assess the fit of the models critically and to derive a criterion for measuring goodness of fit.

Logistic regression was also used to model the distribution of three kangaroo species in Australia in relation to a set of climatic variables (Walker 1990). Here again, binary (presence/absence) modelling was used rather than binomial modelling. Cluster analysis was used to relate the distribution of Elapid snakes in Australia to climatic regions (Nix 1986) and visual inspection of distribution maps to relate the distributional boundaries of wintering birds in North America to environmental variables (Root 1988). In the latter study, estimates of relative densities were available, but were converted to presence/absence data for the analysis.

Study area and methods

Data collection – bird distributions

The Kingdom of Swaziland covers an area of 17 364 km^2 and has a diverse natural environment. The altitude ranges from 200 to 1800 m, the average rainfall varies from 500 to 1300 mm per annum and eleven distinct vegetation zones are recognized (Goudie & Price-Williams 1979).

Data on the distribution of bird species in Swaziland were accumulated for the Swaziland Bird Atlas (Parker 1994) in the form of more than 2600 checklists listing the species observed within a 1/8 degree grid cell (1/8 degree latitude by 1/8 degree longitude) within a calendar month. Most grid cells falling only partly within Swaziland were omitted from the analysis. The data were summarised in the form of reporting rates for each species for each grid cell. The reporting rate is the proportion of field cards for a grid cell on which the species was recorded and is regarded as an estimate of the relative abundance of the species between grid cells (a species is believed to be most numerous where it was recorded most often) (Underhill *et al.* 1992). The fact that coverage of Swaziland was both comprehensive and far more even than that for other atlas schemes in the region removes some of the possible problems related to considering the reporting rates as an index of relative densities (Underhill *et al.* 1992). A remaining problem was that of observer bias. Inconspicuous and more difficult to identify species are recorded less often by inexperienced observers (Underhill *et al.* 1992). This problem was eliminated by using a subset of the checklists consisting of approximately 1700 checklists compiled by the author. This subset consisted of at least 12 checklists per grid square (except for one grid square with six checklists) with at least 35 species recorded per checklist.

Environmental variables

Data on the environmental variables were obtained from a series of 1 in 250 000 maps (Government of Swaziland 1980) and from the Atlas of Swaziland (Goudie & Price-Williams 1979) (Table 1). Rainfall data were obtained in the form of the estimated mean annual rainfall for each one minute of latitude by one minute of longitude from the Computing Centre for Water Research, University of Natal, Pietermaritzburg.

Three alternative ways for coding altitude were used. It was coded as a continuous variable, as a factor with eight levels, or a set of eight separate binary variables corresponding to the levels of the factor. The latter method was introduced because it allowed some of the variables to be omitted from the model when their coefficients were found to be not significant, thus yielding a more parsimonious model.

The number of checklists was included as an explanatory variable because for the few grid cells where the number of

Table 1 Coding of the environmental variables

Variable	Explanation
ALTITUDE	The median of the altitude read at the north eastern corner of each of 40 random 1x1 km quadrats
ALTITUDE RANGE (ABS)	The range of the altitudes read at the north eastern corner of each of the 40 random quadrats
ALTITUDE RANGE (IQ)	The interquartile range of the altitudes read at the north eastern corner of each quadrat
RAIN	The mean of the estimated mean annual rainfall values for each minute of latitude by longitude
STREAMS	The number of random quadrats which contain at least 0.5 km of stream
CARDS	The number of field cards accumulated for the grid cell
GEOLOGY	The value for each of the six variables is the proportion of the grid cell which is assigned to the corresponding geological type in the map by Goudie & Price-Williams (1979)
RIVERS	The number of random quadrats which intersect a river at least 5m in width
PLANTATIONS	The proportion of the grid cell which is covered by exotic timber plantations
AGRICULTURE	The proportion of the grid cell which is utilized for intensive cultivation of sugar, cotton or citrus
LATITUDE	The latitude in minutes of the southern boundary of the grid cell
LONGITUDE	The longitude in minutes of the western boundary of the grid cell
VEGETATION TYPE	The value for each of the 11 variables is the proportion of the grid cell which is assigned to the corresponding vegetation type in the map by Goudie & Price-Williams (1979)
DAMS	A binary variable representing the presence or absence of artificial impoundments
FORESTS	A binary variable representing the presence or absence of natural forests
NATURE RESERVES	The proportion of the grid cell which falls within a nature reserve

checklists was considerably greater than the minimum, the additional checklists related to specific localities within those grid cells, so that reporting rates were biased in favour of species occurring at those localities.

Three alternative methods of representing the vegetation types as explanatory variables were assessed. The vegetation types occurring within each grid cell were represented by a set of 11 continuous variables corresponding to the 11 'veld types' of the natural vegetation map used (Goudie & Price-Williams 1979). For each grid cell, the proportion of its area falling within each veld type was recorded. The second method was to code the vegetation types as 11 levels of a single factor. To achieve this, each grid cell was assigned to the single vegetation type category which covered the largest area within the square. The third method was to represent each vegetation type as a binary variable reflecting either presence or absence in each grid cell.

Trials were made to compare Acocks (1975) veld-type classification with that of Goudie & Price-Williams (1979). The latter classification is less widely known than the former, but is based on more extensive fieldwork in the region (l'ons 1967) and was considered to be possibly a more accurate representation of the vegetation of the country. Logistic regression on vegetation types for nine species of the Cisticola family using each veld-type classification was carried out and the results compared. The Cisticola family was chosen for this comparison because it has a wide variety of distribution patterns, including species with widespread ranges and those restricted to single topographic regions.

The geological data were coded in a manner similar to that adopted for vegetation types. A set of six variables with values representing the proportion of the grid cell assigned to the respective geological class in the map by Goudie & Price-Williams (1979) was used. An alternative set of 18 variables representing the geological classes in the more detailed Government of Swaziland (1982) map was also used and results of using the two different classifications were compared for nine species of the Cisticola family.

Statistical methods

The Genstat statistical package was used to carry out the logistic regression analysis (Payne *et al.* 1987). A set of environmental variables (Table 1) were entered as possible explanatory variables in the logistic model, with reporting rates of the bird species, expressed as a binomial random variable as the response variable. This use of fully binomial logistic regression was used by Underhill *et al.* (1992) to describe seasonality; it is here used to model distribution. The same caveats as described by Underhill *et al.* (1992) are relevant to this application. Although Osborne & Tigar (1992) used arcsine and square root transformations to improve the normality of some of the explanatory variables, no transformations were used in this study because the method of logistic regression does not require that the explanatory variables be normally distributed (McCullagh & Nelder 1989). For each species, the significant explanatory variables were identified by first running the regression program with each variable alone. In the light of experience with fitting and cross validating the models, criteria were established whereby variables were classified as significantly or not significantly associated with the response variable at the univariate stage.

When using altitude to model the distribution of a species, a decision was made as to whether to use altitude as a continuous variable, which involves one explanatory variable, or whether to use the factored variable or the separate variables, which both involve up to seven explanatory variables. It was felt that this decision should not be based solely on the change of deviance associated with each option because the latter two options involved models with a greater number of variables and should not necessarily be regarded as fitting better when they were associated with a greater change in deviance. The following procedure was therefore adopted. Models were fitted using the first two options, and the resulting changes in deviance compared. This comparison was made between models including all the significant explanatory variables. (Comparisons made between the univariate models yielded inconsistent results because sometimes the difference in deviance between the continuous model and the factor model was accounted for by other variables in the full

model.) The continuous variable was selected whenever it was associated with a larger change in deviance. However, when the factored variable was associated with a greater change in deviance than the continuous variable, the cross validation step was used to find the model with the smallest sum of prediction residuals using each of the three coding methods; this enabled the model with the best fit to be selected. In these cases, the results were tabulated against the values for the difference in change in deviance in order to establish how great the difference should be to offset the disadvantage of the greater number of variables in the models.

In cases where the response variable had zero values for more than one level of the factor (zero cells), the fitting process was unstable as indicated by large standard errors associated with the coefficients. Attempts were then made to adjust the limits of the levels of the factor to amalgamate the zero cells into one level. This model was then compared with the model using the continuous variables described above.

A systematic forward selection procedure was used to fit a 'combined' model, including a subset of all the available variables. Vegetation-type models were also fitted, which included only those variables representing vegetation types in which the species was known to occur. When including vegetation-type variables in the 'combined' model, variables with similar coefficients in the vegetation-type model were combined as a single variable after checking that the vegetation-type model with the composite variables did not have a significantly smaller change in deviance. In addition, a model containing only abiotic variables (that is excluding vegetation types) was fitted and compared to the vegetation types only model.

The standardised residual of a grid square in the 'combined' model was considered high if it exceeded 2.5 in absolute value, which identifies approximately 1% of cases as outliers. The number of bird species for which each grid square had a high residual was counted.

In the initial model-fitting process, a dispersion parameter of 1 was assumed in all cases, as for the binomial distribution (Collett 1991). In order to check whether overdispersion (variability greater than that anticipated) could affect the models, for the two species with the highest mean deviance of the residual, the actual dispersion parameter was estimated (Pearson's chi-squared/degrees of freedom) and the models refitted using the estimated dispersion parameter.

Cross validation

A jack-knifing technique was applied to test the predictive power of the models (Quenouille 1949, Miller 1974). For each species, the reporting rate data for each of the grid cells in turn were omitted and the regression coefficients calculated for the restricted model. The new coefficients were then used to calculate a predicted value for the reporting rate for the omitted grid cell and this could then be compared to the observed value. The deviance residuals between the observed and predicted reporting rates for each grid cell were calculated (Hosmer & Lemeshow 1989) and used to identify possible outliers. The sum over the 98 squares of deviance residuals (prediction residuals) was used to assess the goodness of fit of the predicted distributions.

The predicted values of the response variable were represented on a map using the display method of the Swaziland Bird Atlas (Parker 1994) and compared to the corresponding representation of the observed values. On the maps, a circle appears in each grid cell in which the species occurs (or is predicted to occur) with a radius proportional to the reporting rate. This display method was also used by Hockey *et al.* (1989). No circle appears when the predicted number of records is less than one, although the corresponding predicted reporting rate is not zero.

For each bird species, the distribution maps representing the predicted and observed distributions were compared in relation to the sum of prediction residuals, to establish a criterion for assessing the goodness of fit of the models in relation to the prediction residuals.

Results

Comparison of coding methods for explanatory variables

In selecting a coding method for altitude it was found that the model using the continuous variable always produced a better fit than that using the factored variable in terms of the sum of prediction residuals in the cross validation whenever the difference in change in deviance was less than 27 and sometimes produced a better fit when the difference was less than 55. Only when the difference in change of deviance was greater than 55 in favour of the factored variable did the latter invariably produce a better fit (Table 2).

Table 2 Comparison of goodness of fit of the models using continuous and factored variables for altitude. The entries in the table represent the number of times each coding method resulted in a better fit (as determined by sum of prediction residuals) for each range of values for the difference in deviance

Coding method	Difference in deviance (Factored – Continuous)		
	<27	27–55	>55
Continuous	257	8	0
Factored	0	15	32

In all cases where the factored variable was preferred to the continuous, the method of using separate variables yielded a slightly better fit, but the improvement was not significant (less than 1%) and the factored variable was used for convenience. However, in a total of three cases, both the continuous and factored variables were found to be not significantly associated with the response, but a subset of the separate variables was significant and its inclusion improved the fit of the model. In all cases where the factored variable was unsuitable because it contained zero cells, the variable obtained by readjusting the levels was not preferable to the continuous variable.

For vegetation-type coding methods, the continuous method performed better than the binary method in all cases and better than the factor method in all but two cases (Table 3) and in these cases the differences were insignificantly small. When using the factor method, the factor representing the vegetation types often contained several zero cells, with the result that

Table 3 Comparison of coding methods of vegetation types with respect to change in deviance of the logistic model. The table gives values of the change in devianTce associated with the vegetation types coded as: levels of a single factor (FACTOR), as 11 binary variables (BINARY) and as 11 variables with values in the range (0-10) (CONTINUOUS). The full data set was used in these comparisons, which accounts for discrepancies in the values of the change in deviance between this table and table 4, where a restricted data set consisting only of field cards compiled by the author was used

Species	Factor	Binary	Continuous
Wailing Cisticola	420	399	415
Rattling Cisticola	1158	1198	1226
Redfaced Cisticola	327	291	374
Levaillant's Cisticola	924	1015	1035
Croaking Cisticola	240	243	266
Lazy Cisticola	760	680	747
Neddicky	600	537	676
Water Dikkop	100	48	118
Purplecrested Lourie	571	556	689

the fitting process was unstable and the associated standard errors were large. The continuous method was adopted as the most suitable way of coding the vegetation-type data and was used exclusively in the subsequent model-fitting processes.

In the comparison of veld-type classifications, the regression analysis invariably produced greater changes in deviance using the Goudie & Price-Williams (1979) classification compared to Acocks (1975) classification and the differences were significant (with one exception) after taking into account the greater number of categories (11 *versus* 8) (Table 4). It is therefore likely that the former classification describes the vegetation of Swaziland more accurately.

In the comparison of geological classifications, using the classification based on the map by Goudie & Price-Williams (1979), the variables representing geological classes in which the species predominantly occurs were found to be significant for all but one of the species. By contrast, when using the alternative more detailed classification (Government of Swaziland 1982), none of the individual variables were found to be significant in the logistic models. The simpler classification was therefore adopted for use in the modelling process.

Table 4 Comparison of veld-type classifications (Acocks *vs* Goudie & Price-Williams) with respect to the change in deviance of the logistic model

Species	Acocks	G&P-W	Difference
Desert Cisticola	40	66	26
Ayre's Cisticola	318	360	42
Wailing Cisticola	456	463	7
Rattling Cisticola	1068	1166	102
Redfaced Cisticola	363	378	15
Levaillant's Cisticola	688	945	257
Croaking Cisticola	189	219	30
Lazy Cisticola	687	723	36
Neddicky	490	580	90

Criteria for significance of variables

Variables whose inclusion in the combined model was found to improve the fit of the model as measured by the sum of prediction residuals in the cross validation step, were found to be almost invariably among those which had Wald statistic (t) values in excess of two and were associated with changes in deviances of at least five in the univariate models (McCullagh & Nelder 1989). These were then adopted as criteria for identifying which variables were significantly associated with the response variable.

Over-dispersion

The over- or under-dispersion of the models, as measured by the mean deviance of the residual, was found to be closely related to the number of observations of the species concerned. The mean deviance was large for species recorded most often and was considerably less than one for the least frequently observed species (Table 5).

For the two species for which the model was refitted using the estimated dispersion parameter, namely (nomenclature follows Clancey 1980) Redfaced Cisticola *Cisticola erythrops* and Croaking Cisticola *Cisticola natalensis*, it was found that although the values of the Wald statistic for each variable were smaller, nevertheless all of the variables selected in the initial model fitting process remained significant. The over-dispersion, therefore, did not appear to make any real difference to the model fitting process.

Each of the environmental variables entered were significant for at least some of the species and vegetation type was a significant variable for all but four of the 335 species (Table 6). Variables were denoted as highly significantly associated when the variable was associated with a change in deviance which was more than half the change for the combined model. The combined models were found to account for an average of 62.6% of the total deviance for passerine and near-passerines (*sensu* Maclean 1985: xxiv) and 58.7% for non-passerines (Table 7).

Models consisting of abiotic (climatic, topographic and geologic) variables only were better (in terms of change in deviance) than the vegetation type models in 321 cases out of 335. This indicates that the relative densities of the bird species vary within vegetation types and that these differences are at least partially accounted for by the abiotic variables.

Each grid cell had a large residual for a minimum of four and a maximum of 39 out of 335 of the models (Figure 1). It is apparent that high residuals occurred least often in the lowveld, which is the most homogenous of the topographic regions (Goudie & Price-Williams 1979). The modelling process assumes that the explanatory variables are constant within a grid cell and therefore the models are expected to fit less well where these variables vary most rapidly.

Cross validation

For the Lazy Cisticola *Cisticola aberrans*, one grid cell was found to contribute 84 to the total deviance of 449, with no

Table 5 Goodness of fit statistics (Cisticola family & selected species)

	OBS	TOT DEV	COMB %	ABIOTIC %	VEG %	PRED RESID	MDR	GF1	GF2
Fantailed CisticolaT	452	237	48	48*	35	168	1.3	1.4	1.4
Desert Cisticola	20	120	63*	56	55	64	0.5	1.9	1.5
Ayre's Cisticola	240	571	78	75*	71	294	1.4	2.3	1.8
Wailing Cisticola	209	663	85	83*	70	152	1.1	4.8	1.4
Rattling Cisticola	954	1458	87	87*	80	238	2.0	6.1	1.3
Redfaced Cisticola	571	953	55	48*	39	639	5.5	1.5	1.3
Levaillant's Cisticola	408	1219	86	83*	77	302	1.9	4.0	1.5
Croaking Cisticola	532	625	45	41*	35	516	3.9	1.4	1.3
Lazy Cisticola	679	1121	74*	72	64	494	3.6	2.5	1.6
Neddicky	674	980	72*	69	59	331	3.1	3.0	1.2
Cape Turtle Dove	1327	546	78	78*	47	190	1.4	2.9	1.6
Yellowfronted Tinkerbarbet	74	378	90	90*	47	197	0.4	1.9	5.6
Forest Weaver	66	404	98*	98	79	20	0.1	20.2	2.5

Key:

OBS – Number of observations of the species

TOTAL DEV – Total deviance

COMB(%) – Percentage of total deviance accounted for by the combined model

ABIOTIC(%) – Percentage of total deviance accounted for by the abiotic model

VEG (%) – Percentage of total deviance accounted for by the vegetation-types model

PRED RESID = Sum of prediction residuals

MDR – Mean deviance of the residual (best model)

GF1 = (Total deviance)/(Sum of prediction residuals)

GF2 – (Sum of prediction residuals)/(Sum of residuals of the full model)

* – Indicates which model yielded the best fit for each species

other grid cell contributing more than 30. This point was then deleted and the analysis repeated. The exclusion of the grid cell was found to make a negligible difference to the fit of the remaining grid cells. For the other species, no single grid cell was found to contribute much more than the others to the total deviance.

The prediction residuals calculated by jack-knifing for some species are included in Table 5. Comparison of the prediction residuals to percentage points of the Chi-square distribution (Hosmer & Lemeshow 1989) was found to be inappropriate as a criterion for goodness of fit because the prediction residuals cannot be less than the residual deviance of the full model. When the total deviance of the model was large, these values frequently exceeded the relevant Chi-squared value even though the models accounted for high proportions of the total deviance. In many cases, visual examination of the

Table 6 Numbers of species distributions with which each variable is significantly associated.

Variable	T	p	n	P	N
Vegetation type	331	331	0	272	0
Geology	324	324	0	96	0
Altitude	315	108	207	58	97
Rainfall	295	94	201	17	48
Dams	225	147	78	14	0
Rivers	164	108	56	1	0
Cards	214	76	138	1	7
Plantations	250	84	166	1	19
Streams	281	102	179	5	11
Latitude	194	109	85	4	4
Longitude	292	190	102	52	28
Agriculture	228	168	60	5	0
Forests	250	86	166	7	2
Nature reserves	249	83	66	2	0
Altitude range	206	118	88	4	0

Key

T: No. of species significantly associated

p: No. of species significantly positively associated

n: No. of species significantly negatively associated

P: No. of species highly significantly positively associated

N: No. of species highly significantly negatively associated

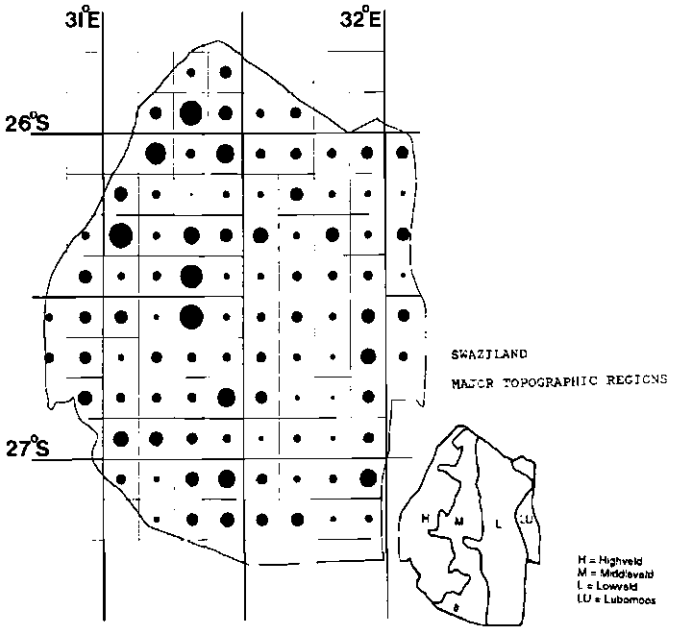
Table 7 Goodness of fit data

Summary statistics:	Non Passerines			Passerines and near passerines		
	MEAN	SD[1]	RANGE	MEAN	SD	RANGE
Percentage of variation accounted for by:						
Combined model	58.7	19.6	5–92	62.6	17.5	12–98
Abiotic model	55.6	19.2	5–91	60.0	18.1	12–98
Vegetation	33.5	14.7	0–66	47.0	17.2	0–85
Goodness of fit statistics						
GF1[2]	1.85	0.86	0.42–6.26	2.23	1.57	0.82–20.20
GF2[3]	1.73	1.31	1.03–12.46	1.68	1.82	1.02–26.14

1 SD = Standard deviation

2 GF1 = (Total deviance/sum of prediction residuals)

3 GF2 = (Sum of prediction residuals/sum of residuals for the full model)



The size of the circles is proportional to the number of species models for which the grid square had a high residual. Minimum= 4/335; Maximum=39/335.

Figure 1 Distribution of high residuals within Swaziland. Inset shows major topographic regions of Swaziland

predicted and observed distributions suggested that the fit of the models was in fact excellent. On the other hand, when the total deviance was very small, the prediction residual was found to be considerably less than the Chi-square value even when the fit of the predictions did not look particularly good.

Comparison of the ratio of the total deviance of the full model to the prediction residuals with the observed similarity of the predicted and observed distribution maps yielded a more appropriate measure of goodness of fit. It was observed that the fit always appeared to be good whenever this ratio exceeded one.

A value of one for this ratio implies that the model fits no better than a model representing a constant reporting rate throughout. However, when the predicted distribution maps for species where the ratio was close to one were examined, it was observed that where there was a large difference between the predicted and observed reporting rates, the grid cell was often contiguous with cells whose observed values matched the predicted value (Figure 2d). The actual fit of the model in these cases was therefore generally better than the ratio would suggest.

This ratio had a value greater than one for 324 out of 335 species and a mean value of 2.2 for passerines and near passerines (n=242) and 1.9 for non passerines (n=93). Eight of the 11 species for which the ratio was less than one were water birds.

An additional measure of the goodness of fit of the models is the ratio of the sum of prediction deviance residuals to the residual deviance of the full model. This ratio indicates how much the predicted values differ from the fitted values of the full model without reference to the observed values (Table 5).

The maps representing predicted and observed distributions for four species (Figure 2) illustrate the fact that the models have predicted both the limits of the distributions of the species as well as their reporting rates within these limits with reasonable accuracy.

Discussion

This investigation has shown that logistic regression can be used to identify the environmental variables which are significantly associated with the geographical distributions of bird species. This is a stronger result than that of Osborne & Tigar (1992), who showed that species distributions could be related to latent variables derived from the environmental variables using principal components.

Comparison of the predicted distribution maps to the observed distributions has established that if bird distribution information was available for some centres and lacking in the intervening areas for some region, then the distributions in the intervening areas could be reasonably accurately predicted.

The success of the modelling process is a justification of the use of reporting rates as an index of population density because it is difficult to conceive of an alternative explanation for the association between reporting rates and the environmental variables.

The model selection process used is in contrast to the

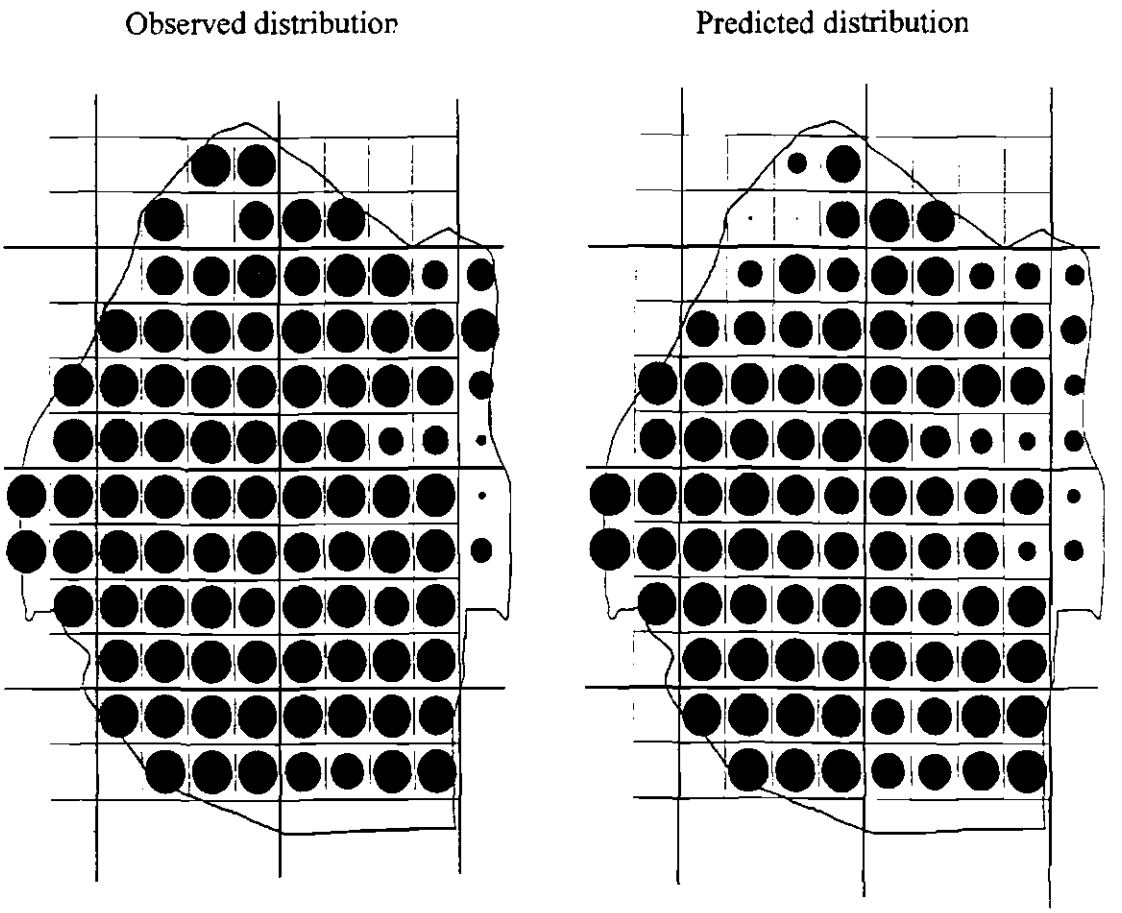


Figure 2a Observed and predicted distributions: Cape Turtle Dove (Goodness of fit stat.: gf1=2.9)

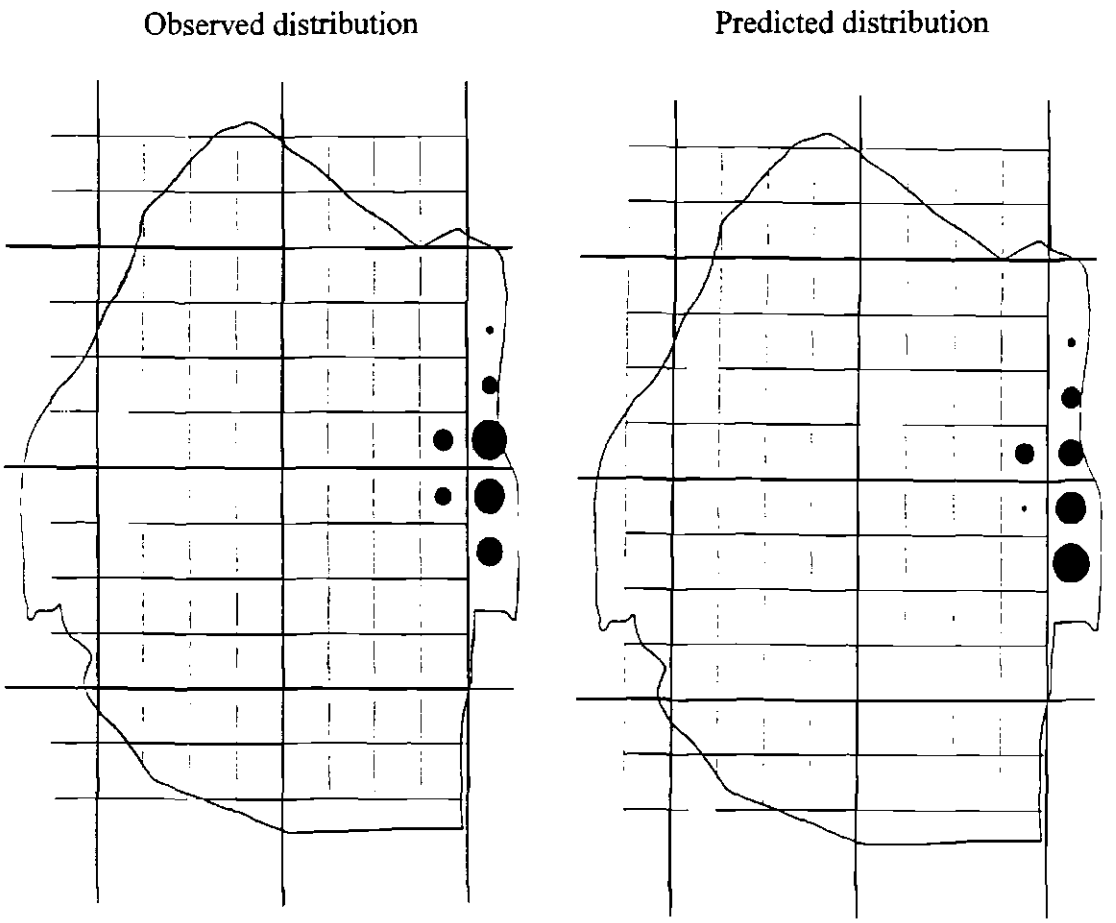


Figure 2b Observed and predicted distributions: Forest Weaver (Goodness of fit stat.: gf1=20.2)

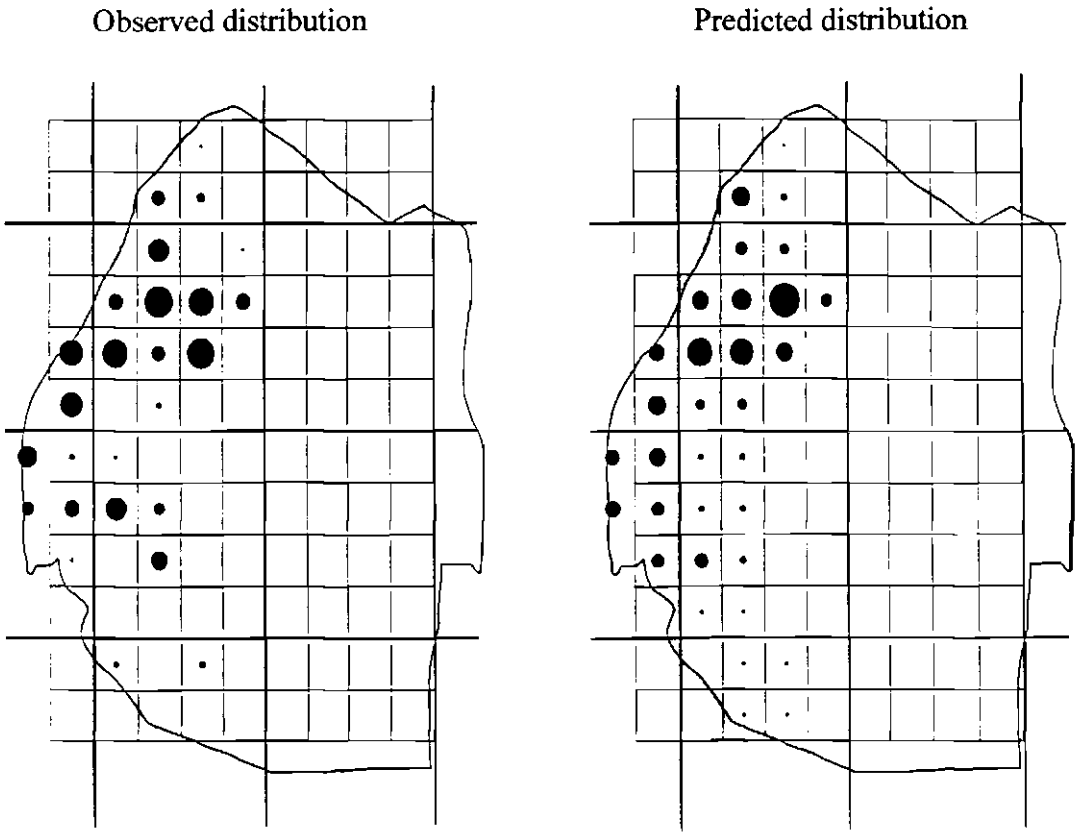


Figure 2c Observed and predicted distributions: Wailing cisticola (Goodness of fit stat.: gf1=4.8)

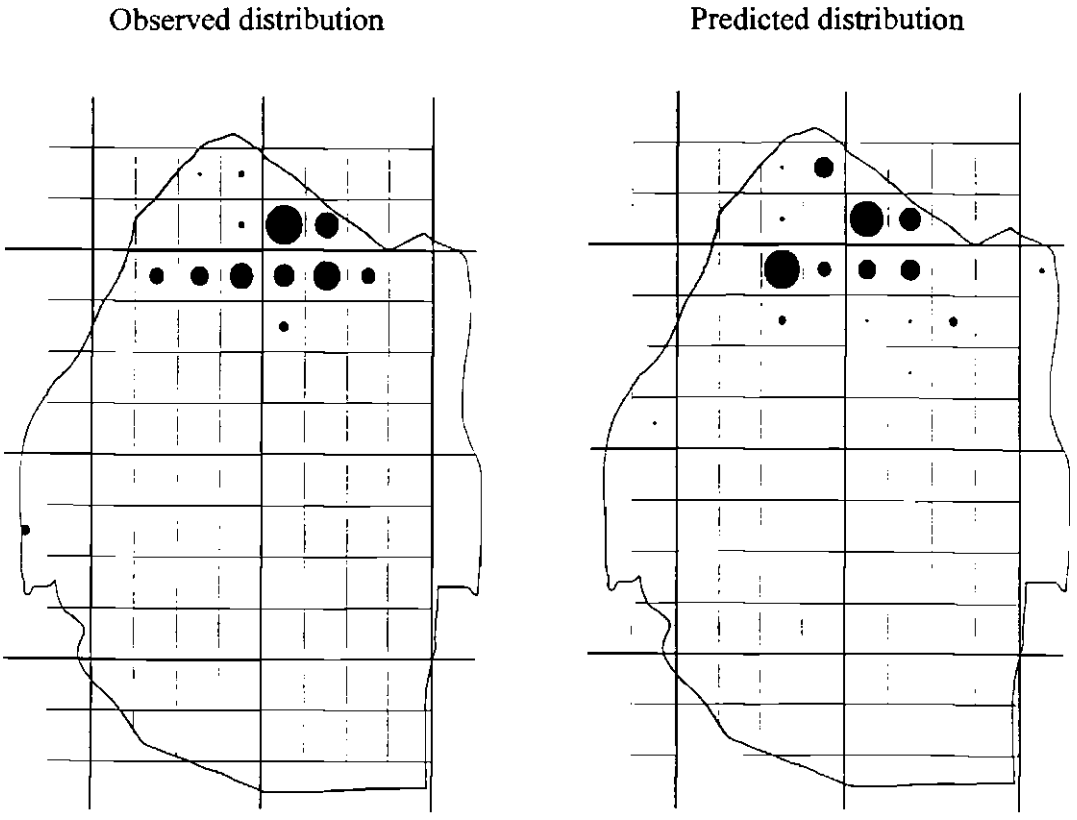


Figure 2d Observed and predicted distributions: Yellowfronted Tinker Barbet (Goodness of fit stat.: gf1=1.02)
Note: This predicted distribution does not represent the best fitting model derived for this species (Table 5) but was selected to illustrate the fit of a model for which gf1=1

method of Osborne & Tigar (1992) who reduced all the available variables to the first four principal components. The first four principal components of the variables do not necessarily coincide with the components which are most significantly related to the response variable, as demonstrated for example by Cuadras (1993). Moreover, during the model selection process, the inclusion or exclusion of a single variable was often observed to make a large difference to the fit of the model, as reflected in the value of the prediction residual. Noting also that the models for very few of the species were identical in respect of the variables included, the principal components method was considered to be inadequate.

The method used by Osborne & Tigar (1992) of identifying as outliers individual points with high residuals was found to be potentially flawed. In this study, it was often found that when the fit of a model was poor, there were one or two points which appeared to be outliers on the basis of their excessive residuals. However, in most of these cases it was possible to obtain a better fitting model by using a different combination of variables, and in the new model the points in question no longer had excessive residuals. (When the model did not fit, the model was wrong, not the data.)

More work needs to be done on the distribution of the sum of prediction residuals before a criterion for goodness of fit can be proposed for general use in logistic regression. However, for the purposes of this study, the criterion derived here was found to be consistent with the observed goodness of fit and was useful in ranking the models and identifying those for which the fit was relatively poor.

Underdispersion of the models for which the number of obsevations is small was expected because in these cases most of the values for reporting rates for the grid cells are zero and consequently the variability is less than that expected for a binomially distributed variable. On the other hand, overdispersion for models where the number of observations is large is probably attributable at least in part to the fact that the distributions are partly determined by variables other than those which were available for inclusion in the models.

The fact that most of the birds for which the fit was relatively poor were water birds indicates that the models for these species would probably be improved by the inclusion of further variables representing the occurrence, nature and extent of wetlands. The variables used also probably do not account adequately for the effects of human activities on the environment. It is possible that the species most affected by human activities will be among those for which the fit is relatively poor.

One of the aims of this study was to explore methods for coding environmental variables from maps, and to determine which method is the most suitable at least for Swaziland. Further studies in other regions and at different scales may help to establish whether these methods are generally applicable.

Temperature could not be included as an explanatory variable because no data were available at the appropriate scale. As temperature is highly correlated with altitude (Goudie & Price-Williams 1979), the inclusion of temperature might not make a significant difference to the models. Moreover, the variables latitude and longitude are expected to act as surrogates (together with altitude) for temperature. Temperature decreases with latitude (though this effect is small over the 1.5 degrees of latitude) and increases with longitude due to the influence of the Indian Ocean which lies less than 50 km away to the east.

The appropriateness of logistic regression in this context is dependent on the assumption of independence of observations of birds in different grid cells. This assumption would be violated if the same individual were observed in different grid cells. It is believed that the assumption is reasonable for most small passerines. The assumption might not be valid for some larger, more mobile species, especially those of the family Accipitridae and water birds. Use of the regression models also assumes that a species occurs wherever environmental conditions are favourable irrespective of its occurrence or non occurrence in neighbouring squares. Further research into the impact of these assumptions is needed.

Acknowledgements

Fieldwork in Swaziland was supported by the Conservation Trust of Swaziland, the Southern African Ornithological Society and the Natural History Society of Swaziland. I acknowledge the assistance of Prof L.G. Underhill. Advice was received from Prof J.M. Juritz. J. Harrison and R. Navarro assisted with transforming the bird atlas data into the required format. Mrs F. Train assisted with word processing. Rainfall data were obtained from the Computing Centre for Water Research, University of Natal, Pietermaritzburg.

References

ACOCKS, J.P.H. 1975. Veld types of South Africa. 2nd ed. Government Printer, Pretoria.

CLANCEY, P.A. 1980. S.A.O.S. Checklist of Southern African birds. Southern African Ornithological Society, Pretoria.

COLLETT, D. 1991. Modelling binary data. Chapman and Hall, London.

CUADRAS, C.M. 1993. Interpreting an inequality in multiple regression. *Amer. Stat.* 47: 256–258.

GOUDIE, A. & PRICE-WILLIAMS, D. 1979. The atlas of Swaziland. Swaziland National Trust Commission, Mbabane.

GOVERNMENT OF SWAZILAND. 1980. Swaziland 1:50 000 map series. Government of Swaziland, Mbabane.

GOVERNMENT OF SWAZILAND. 1982. 1:250 000 Geological map of Swaziland. Government of Swaziland, Mbabane.

HOCKEY, P.A.R., UNDERHILL, L.G., Neatherway, M. & Ryan, P.G. 1989. Atlas of the birds of the Southwestern Cape. Cape Bird Club, Cape Town.

HOSMER, D. & LEMESHOW, S. 1989. Applied logistic regression. Wiley, New York.

I'ONS, J.H. 1967. Veld types of Swaziland. Ministry of Agriculture, Mbabane.

MACLEAN, G.L. 1985. Roberts' birds of Southern Africa. John Voelcker Bird Book Fund, Cape Town.

MILLER, R.G. 1974. The jackknife – a review. *Biometrika.* 61. 1–17.

MCCULLAGH, P. & NELDER, J.A. 1989. Generalized linear models. Chapman and Hall, London.

NIX, H. 1986. A biogeographic analysis of Australian elapid snakes. In: R. Longmore. Atlas of elapid snakes of Australia. Australian Flora and Fauna Series Number 7. 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. *J. Appl. Ecol.* 29: 55–62.

PARKER, V. 1994. Swaziland bird atlas 1985–91. Webster's, Mbabane.

PAYNE, R.W. *et al.* 1987. Genstat 5 reference manual. Clarenden Press, Oxford.

QUENOUILLE, M.H. 1949. Approximate tests of correlation in time series. *J. Royal Stat. Soc. B.* 11: 68–84.

ROOT, T. 1988. Environmental factors associated with avian distributional boundaries. *J. Biogeog.* 15: 489–505.

UNDERHILL, L.G., PRYS-JONES, R.P., HARRISON, J.A., MARTINEZ, P. 1992. Seasonal patterns of occurrence of Palearctic migrants in southern Africa using atlas data. *Ibis.* 134 suppl.1: 99–108.

WALKER, P.A. 1990. Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. *J. Biogeog.* 17: 279–289.