4

ANALYSIS OF DISTRIBUTION MAPS FROM BIRD ATLAS DATA:

Dissimilarities between Species, Continuity within Ranges and Smoothing of Distribution Maps

Birgit Erni

Dissertation for the Degree of Master of Science in Mathematical Statistics

Supervisor: Professor Les Underhill

Department of Statistical Sciences
University of Cape Town

December 1998

ACKNOWLEDGEMENTS

I would especially like to thank Les Underhill for his supervision and encouragement throughout the two years. Many thanks also to the staff of the Avian Demography Unit, UCT, for their always willingly offered help and assistance, especially Felicia Stoch and Rene Navarro. Professor June Juritz gave useful support with the theoretical side.

The Foundation for Research and Development (FRD) provided financial support. The Percy Fitzpatrick Institute, UCT, made the Gordon Sprigg scholarship available to me for two years.

The ADU provided me with a subset of their excellent data.

ABSTRACTS

Chapter 1

The Dissimilarity between Avian Distributions

A dissimilarity coefficient for estimating the dissimilarity between two bird atlas distributions is developed. This coefficient is based on the Euclidean distance concept. The atlas distributions are compared over all quarter degree grid cells. Existing coefficients are not suitable for the comparison of distributions with different total areas and species with different mean reporting rates. In each grid cell the reliability of the reporting rates depends on the number of checklists collected for the grid cell. Weights are used to solve this problem. To solve the problem of different levels of abundance and conspicuousness of species, the reporting rates are sorted into percentiles, using five or 10 categories for the strictly positive reporting rates. Each grid cell is weighted by a function of the number of checklists collected for the grid cell. The coefficient is scaled by the maximum possible sum of the differences which would occur if there is no overlap between the two distributions, so that the dissimilarity coefficient lies between zero (a perfect match) and one (no overlap). A variety of these coefficients are investigated and compared.

Chapter 2

The Continuity of Bird Distributions

The continuity of observed reporting rates in a spatial cellular map is an indication of spatial autocorrelation present, especially between observations which are in close vicinity. We are particularly interested in measuring and comparing the continuity of the reporting rates in the bird distributions from *The Atlas of Southern African Birds*. The variogram, developed in geostatistics, estimates this spatial autocorrelation. The classical variogram estimator, however, is dependent on the scale of measurement and assumes that the data are intrinsically stationary. The bird atlas distribution maps contain trend and the variance of each observation (reporting rate) is a function of the number of checklists collected for the grid cell and the underlying probability of encountering the species in the grid cell. The approach of removing this binomial measurement error from the variogram developed by McNeill (1991) is investigated but not found satisfactory. A weighted variogram, where each squared difference is weighted by a function of the smaller number of checklists, is developed. To make

the variogram values comparable between species a function of the mean reporting rates is used to scale the variogram. We were particularly interested in the first variogram value of each species distribution, $2\gamma(1)$.

Chapter 3

Smoothing of Bird Atlas Distribution Maps, Based on Reporting Rates

The bird distribution maps in *The Atlas of Southern African Birds* show the raw observed reporting rates. Each of these reporting rates is a random variable dependent on sampling error due to binomial variation based on the number of checklists collected for the grid cell and on the underlying probability of encountering the species. The distribution maps show this measurement error. It is believed that a smoothed version of the bird distribution maps will to some extent improve the statement these observed distributions are aiming to make. Single-step regression methods are investigated for a fast approach to this problem. These cause problems because of frequent 'zero' observed reporting rates and because they smooth the maps too heavily. Generalized Linear Models are investigated and this iterative procedure is applied to model the reporting rates with a binomial distribution on square blocks of nine grid cells where a value for the central cell is 'predicted' in each regression. This approach is especially suited to accommodate the binomial distribution characteristics and is found to smooth the bird atlas distributions well. Because only a local window is taken for each regression, the spatial autocorrelation is adequately included in the spatial explanatory variables.

TABLE OF CONTENTS

CHAPTER 1

The Dissimilarity between Avian Distributions

INTRODUCTION

How can the similarity between two bird distribution maps be measured? How similar are the distributions of two species? Which species have the most similar distributions?

These are the questions that will be discussed in this chapter. Species of birds that have a similar distribution should match with respect to the area in which they occur and the peak places of occurrence should correspond.

For example the distributions of the Cape Weaver and the Cape Canary (Figs. A1 and A2, Appendix) are almost equal, similar, while the distribution of the Masked Weaver (Fig. A3, Appendix) is different to those of the above two species in overall shape and in that the cores of the distributions do not coincide. With more distribution maps it becomes more difficult to rank them according to similarity. The aim here is to find a mathematical measure to represent the similarity between two distributions.

A similarity measure may be of interest when

- comparing distributions of species within a region,
- investigating habitat requirements
- investigating dependencies between species, such as host-parasite interactions (e.g. between cuckoos and the species in whose nests they lay their eggs).

There are a number of available similarity and dissimilarity coefficients, none of which however appears to be suitable for the comparison of bird distribution maps. The particular problems introduced by the atlas maps are that the different degrees of commonness of species distort any absolute difference between observed reporting rates. Secondly, the different total areas covered by bird distributions introduce scaling problems, so that the commonly used coefficients do not lie in the desired interval [0, 1].

METHODS

SIMILARITY OF SPATIAL MAPS

What we are planning to do is to compare the shapes of distributions, particularly bird distributions taken from *The Atlas of Southern African Birds* (Harrison *et al.* 1997a, b). This comparison should not only take into account the subjective visual perception one has of maps but the method should be able to objectively compare patterns occurring within the maps. One possible application for such a comparison is model checking in maps, residuals could be compared to geographic maps to explain any residual patterns (Carstensen 1987).

A feature that is of interest to ornithologists is which species have the most similar distributions. This might provide an indication of which species have similar ecological requirements. For example the distribution map for the European Bee-eater (Fig. A4, Appendix) showed a strikingly unusual range through the central area of southern Africa. This was thought to be unique, until a long search showed that the Wattled Starling (Fig. A5, Appendix) had a rather similar-shaped distribution (L.G. Underhill pers. comm.). A measure of similarity is also useful for studying the relationships between parasitic species and their hosts (eg. cuckoos and weavers, robins, cisticolas, etc., honeyguides and barbets, etc., and whydahs and waxbills).

In general terms, the way to measure similarities between two units is to compare their different characters. The more these characters match, the more similar the units are. This matching of characters can be done to different degrees of accuracy, depending on the coding or form of the information and on the accuracy of the data. For nominal data the difference can either be a match or a non-match. Therefore one needs only to count the number of matches and relate this to the total number of characters that have been compared. But for quantitative data it is common practice to rather take the direct difference between the character values as an idea of the degree to which the units match in this character. These differences need to be combined in some way to form an overall measure of similarity. Continuous data can be reduced to nominal data when the observed values are placed into categories.

For cellular maps the distributions need to be compared over each of the cells. In the bird atlas the cells would be the quarter degree grid cells. These would be the 'characters' of the distribution. The more grid cells two distributions have in common, the more similar they are. But also the more the actual observed reporting rates for the grid cells match, the more do the internal patterns of the distributions coincide.

In the bird distribution maps in Harrison *et al.* (1997a, b) the relative frequency or probability of occurrence in a grid cell is of interest, relative to the rest of the distribution and relative to other species in this area. Maps like this can only be developed from data in the form of proportions, the number of successes out of a total number of trials. When referring to the bird atlas data, these proportions are called reporting rates. Epidemiological studies may also use this form of mapping, e.g.

Cressie & Read (1985), but mostly maps are produced by simple measurements on a zero to, theoretically, infinity scale.

SIMILARITY AND DISSIMILARITY COEFFICIENTS

To estimate the extent to which two units resemble each other, one can either find the similarity or the dissimilarity. Similarity coefficients are commonly used for qualitative or nominal data, for example when the response falls into either of the categories male or female. More relevant to the bird atlas data, the reporting rates can be expressed as a qualitative variable: the species is either present or absent in a given grid cell, coded as zero (absent) and one (present). In this form the variable is said to be binary.

If the characters are measured quantitatively, either continuous or discrete, it makes more sense to use a dissimilarity measure. This is because the difference between values will give a direct indication as to the magnitude of dissimilarity (Gower 1985). Dissimilarity coefficients normally take on values between zero (identical) and one (completely different, dissimilar).

To find the dissimilarity between two units, each of their characters are compared. A general structure of the dissimilarity coefficient (between unit *j* and unit *k*) could be the sum of the absolute differences over all n characters

$$D_{jk} = \sum_{i=1}^{n} | x_{ij} - x_{ik} |$$

This form however underestimates the real distance (Sneath & Sokal 1973) and it is therefore more common to sum the squared differences and then only to take the square root of the sum.

$$D_{jk} = \sum_{i=1}^{n} (x_{ij} - x_{ik})^2$$

This is how distances in euclidean space are found.

Here follows a discussion of a selection of existing similarity and dissimilarity coefficients.

THE KHAT COEFFICIENT OF AGREEMENT

Carstensen (1987) shows how the KHAT coefficient, usually used in psychology and remote sensing, can be used to compare map patterns. His particular interest was to find a measure that corresponds to how people perceive similarity of maps. He argued that people are unlikely to estimate similarity in the same way as for example the correlation coefficient does. Correlation coefficients are what has mostly been used to compare maps (Carstensen 1987).

The KHAT coefficient should be used on cellular maps with nominal data, but it is often not difficult to convert other maps to such. The categories used have to be the same on both maps, or rather on the group of maps.

Any disagreement in cell values is seen as a difference, no distinctions are made between different magnitudes of disagreement. For example when working with shaded maps, a difference between black and white is the same as the difference between white and grey. Cells either are the same or they are not. This means that meaningful values and ranks in the data would be ignored.

An 'error matrix' is calculated in the following way. The rows and columns represent the categories of map B and map A respectively. The entries n_{jk} will be the number of cases where the cell in map A is of category 'j' and of category 'k' in map B.

The KHAT coefficient compares the observed agreement P_o to the expected agreement P_e under the hypothesis that the two maps have uncorrelated values. P_o and P_e only consider exact agreement. Po is the proportion of units that agree out of the total number of units, the proportion of values on the diagonal of the error matrix. P_e is the proportion of all cases that would be expected to agree if the values in two maps are assumed to be uncorrelated. P_e is calculated from row and column totals as is done to calculate expected probabilities in contingency tables. Then $(1 - P_e)$ is the proportion of cells that is expected to have different values.

$$\text{KHAT} = \frac{P_o - P_e}{1 - P_e}$$

The KHAT coefficient measures how much more agreement is present than would be expected if the maps are assumed independent. Its value falls into the interval [-1, 1]. The KHAT coefficient requires that all categories are mutually exclusive and exhaustive (Carstensen 1987). That means that this coefficient could not be used for the atlas bird distribution maps where the distributions have different sizes. Two distributions can cover completely different areas. If a zero reporting rate would be taken as a possible category there would be too much agreement in cases where two small distributions are compared and the rest of the area is blank. If zeroes are not included the error matrix would not be square and the categories would not be exhaustive.

The 'error matrix' is a useful summary of the components of the difference. We develop the concept of decomposing the dissimilarity coefficient we devise, later on.

SIMILARITY COEFFICIENTS

There are coefficients for binary data in which only presence or absence of a character are recorded (as one or zero). Most of these are based on a table which contains the number of compared characters of the two units which fall into one of the following four categories:

- **a** is the number of characters that are present in both units *j* and *k*
- **b** is the number of characters that are present in unit *j* but absent in unit *k*
- **c** is the number of characters that are absent in unit *j* but present in unit *k*
- **d** is the number of characters that are absent in both units *j* and *k*

It depends on the type of data whether 'd' (character absent in both units) should be seen as a match of characters or not.

JACCARD

The Jaccard similarity coefficient (Sokal & Sneath 1973) is given by

$$S_J = \frac{a}{(a + b + c)} \qquad (1)$$

This coefficient can be used for qualitative data and for presence/absence data. When the Jaccard coefficient is used for the bird atlas data, much of the available information is lost because the observed reporting rates would have to be converted to presence/absence data. In the case of bird distributions the Jaccard coefficient would calculate the extent of overlap.

SIMPLE MATCHING

The simple matching coefficient is given by Sokal & Sneath (1973) to be

$$S_M = \frac{a + d}{(a + b + c + d)}$$

This coefficient is similar to the Jaccard coefficient because it is also used with qualitative or presence/absence data. In addition it regards a common absence of a character (both species have a zero reporting rate in some grid cell) as a similarity. In the context of our application, this coefficient depends too much on the overall sizes of the distributions. For an example, if the Forest Canary were compared with the Protea Canary the large number of grid cells where neither species occurs, d, would dominate the coefficient. Here d would be much larger than a, the area where the distributions really overlap. These two species would appear to be more similar than they are.

For the presence/absence data the Jaccard coefficient would certainly be better than the Simple Matching Coefficient when comparing bird distribution maps.

CZEKANOWSKI COEFFICIENT

This coefficient is also referred to as the Sørensen or Dice coefficient (Cox & Cox, 1994) and is given by

$$S_C = \frac{2a}{2a + b + c} \qquad (2)$$

Clifford (1975) remarked that if a (then number of characters present in both units) is large, the Jaccard coefficient is more attractive because the calculated values will have a wider spread. If a is relatively small the Czekanowski coefficient is preferred because the calculated similarities will be larger. This however is merely a matter of preference. The Czekanowski coefficient will turn out to be a special case of the dissimilarity coefficient we will devise.

CORRELATION COEFFICIENT

The Pearson product-moment correlation coefficient is defined as

$$r_{jk} = \frac{\sum_{i=1}^{n} (X_{ij} - X_{.j})(X_{ik} - X_{.k})}{\sqrt{\sum_{i=1}^{n} (X_{ij} - X_{.j})^2 \sum_{i=1}^{n} (X_{ik} - X_{.k})^2}} \qquad (3)$$

It is not easy to interpret this coefficient because the value obtained does not directly give a distance but will lie in the interval [-1, 1]. It is not always clear however whether -1 (perfect negative correlation) or a correlation of zero should equal zero dissimilarity. The correlation coefficient is not metric (Sneath & Sokal 1973).

An advantage, elsewhere considered a disadvantage, is that perfect correlation would also occur between units that are not identical but where one column is a scalar multiple of the other (Sokal & Sneath 1973). This would solve the problem of different magnitudes of average reporting rates discussed later.

It is not clear what the mean, required to calculate the correlation coefficient, should be in the case of the atlas distributions. The average reporting rate is not meaningful, a weighted average would have to be used instead. It is also not clear whether such an overall mean has any justification.

However for the comparison of maps this coefficient has been most widely used (Carstensen 1987).

COEFFICIENT OF DIVERGENCE

The coefficient of divergence is given by

$$CD_{jk} = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_{ij} - X_{ik}}{X_{ij} + X_{ik}} \right)^2 \right]^{1/2} \qquad (4)$$

An advantage of the coefficient of divergence is that the difference between each pair of characters is scaled immediately, which ensures firstly that the coefficient will not exceed the maximum value of one and secondly prevents outliers from making too much contribution to the sum. Each of the summands will be a value between zero and one. The coefficient of divergence is a metric when all the data are positive (Gower 1985, p. 401).

Let j and k be the two species to be compared. If species j occurs in some area where k does not occur, then $X_{ij} \neq 0$ but $X_{ik} = 0$ (the reporting rate of species k in cell i equals zero). But then the value added to the sum will equal one, no matter what the observed value for species j was. This may give too much weight to the areas where only one of the two species occurs relative to the areas where both occur.

An advantage of this coefficient is that it not only considers absolute differences but also relates the difference to the original values (Sokal & Sneath 1973). A difference of 0.2 from (0.3 - 0.1) is relatively larger than a difference of 0.2 observed from (0.8 - 0.6).

VARIATIONS ON THE EUCLIDEAN DISTANCE

The Euclidean distance between two units with n characters is calculated as the distance in p-dimensional space as follows

$$D_{ij}^2 = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \qquad (5)$$

where D_{ij}^2 is the squared distance. Because this distance increases with an increasing number of characters being compared, when calculating dissimilarities, an average difference is taken

$$D_{ij}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \qquad (6)$$

where p is the number of characters over which the two units were compared. If the characters are of different importance for establishing the dissimilarity, then each character can be weighted according to the relative contribution it should make. The Weighted Euclidean distance, given by Cox & Cox (1994) is

$$D_{ij}^2 = \sum_{k=1}^{n} w_k (x_{ik} - x_{jk})^2 \qquad (7)$$

If characters are measured at different scales, each character should be scaled so that the contribution does not depend on the scale at which the character has been measured. This variation of the Euclidean distance is called Taxonomic Distance and is given by

$$D_{ij}^2 = \frac{1}{n} \sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})^2}{r_k^2} \qquad (8)$$

The usual choice of the scaling factor r_k in the taxonomic distance is the standard deviation of the variable or alternatively the range of the variable in either the sample or the population (Gower 1985).

The Euclidean distance has valuable properties: it is metric and euclidean, it is easy to compute and has a simple geometric interpretation. These properties are useful, for example, in multidimensional scaling and other ordination techniques. Metric and and euclidean properties are important conditions for the use of some ordination methods (Digby & Kempton 1987). A coefficient is said to be a metric if its calculated dissimilarities obey the triangular inequality.

See Cox & Cox (1994), Sokal & Sneath (1963, 1973), (Gower 1985) and Clifford (1975) for more detailed summaries and discussions of existing dissimilarity and similarity coefficients and their applications.

THE DISSIMILARITY OF BIRD ATLAS DISTRIBUTION MAPS

We will discuss some aspects that make existing coefficients unsuitable for the particular case of the bird atlas distribution maps. This has to do with the properties of reporting rates. We then begin to develop a new coefficient based on the Euclidean distance, adjust it to achieve desirable properties and also investigate whether the data should not be transformed to a more stable form.

The data available for the project consisted of the reporting rates in 1540 grid cells, the region of southern Africa south of 27°S. A reporting rate is made up of the number of checklists collected for a grid cell and the number of successes out of these, i.e. the number of times that the species has been observed. The reporting rate is the ratio of the number of successes and the number of trials. Shown on the bird atlas maps are only the reporting rates in four categories, represented as different shades. It is hoped that reporting rates estimate the relative frequency of occurrence of a species, relative to other grid cells and then also in relation to other species. This has been shown by all studies that have tested this concept; these are listed by Harrison & Underhill (1997).

Each grid cell represents one character on which the two species can be compared. In the case of the bird atlas maps the units are the two bird distributions to be compared and their 'characters' are the reporting rates in individual grid cells.

In the previous section some of the existing coefficients were discussed. The most attractive of these is the Euclidean distance coefficient. There are however some problems that are caused by the particular form of the bird atlas data.

MAGNITUDE OF AVERAGE REPORTING RATES

The aim is to estimate the extent to which any two species occupy the same area and, furthermore, whether the cores of their ranges coincide. Which of the two species is more common, more dominant or easier to observe are different issues, not related to the shape of the distribution.

A reporting rate of *r*, say, is an estimate of the probability of sighting a species in a given grid cell. But seen relative to the remaining distribution of the species, this value may be the maximum observed reporting rate for the species or it may be below the average reporting rate. Therefore a reporting rate of *r* has different meanings for different species and needs to be seen as a value relative to the rest of the distribution for that species and not relative to the reporting rates of other species. By this is meant that the observed reporting rates differ largely between species. Rare species have smaller reporting rates than common species over most of their distributions. For one species the average reporting rate may be 10 %, for another species the average reporting rate may be 40 %.

The problem does not lie with one distribution but is introduced when comparing two distributions, of two different species. For an example consider the Cape Sugarbird and the Cape Siskin (Figs. A29 and A6, Appendix A). Their distributions appear to be almost equal with the exception of a few grid cells. Even the shades in the grid cells seem to match. But the numerical reporting rates on which these shades were based differ. The average reporting rates recorded in the atlas are 25.9 % and 9.5 % for the Cape Sugarbird and the Cape Siskin respectively (Harrison *et al.* 1997b, pp. 485 and 659). This means that even when the shades in a cell are equal, the difference between reporting rates (when using the Euclidean distance for example) will be larger than they should be, the distance or dissimilarity should be close to zero. The darkest shades for the two above species occur when the observed reporting rates were larger than 32 % and larger than 17.5 % respectively. These two distributions when looked at are what we consider to be very similar. But these differences in the magnitudes of the reporting rates obscure the real differences that should be measured. There would be a constant large difference between the species not caused by a difference in shape of distributions but by a difference in their average observed frequency.

This suggests that the reporting rates should not directly be used but rather should be converted back to the percentile categories used in the atlas maps. Comparing such categories would also allow us to establish whether the internal patterns of the distributions match. In other words, the coefficient must be able to take into account that some species are generally rare, even in the areas where they are at their most abundant, while other common species have relatively high reporting rates throughout their distributions. When comparing distributions between species, one can thus not merely take the differences between raw reporting rates because this takes no account of the possible differences between species with small and generally large reporting rates.

Some existing coefficients make use of presence/absence data. This would eliminate the above problem of the reporting rates. It would then be a measure of the overlap of the distributions. But much of the valuable reporting rate information would be lost.

For any species the reporting rates establish where it is most common and where it is rare. In other words reporting rates provide information on where the core and the edges of a distribution are and this needs to be incorporated into the similarity or dissimilarity coefficient.

CONVERSION INTO PERCENTILES

If the reporting rates are left untransformed, then even if two species, the one rare or inconspicuous and the other common, occupy the same area, there will throughout the distribution be a considerable difference in reporting rates. Even if the difference is one of 0.3 only (0.6 - 0.3 or 0.4 - 0.1) a continuous difference between the reporting rates of the two species will add up to a large sum. This would result in a larger difference than we would want.

How can the reporting rates be converted so that common and rare species are comparable? They can be transformed so that the grid cells where a species is most frequent are assigned a value of 10 and the cells where the species was least frequently observed is assigned a value of one.

We ranked the reporting rates from smallest to largest. The 10% of grid cells with the highest reporting rates were assigned a value of 10, the 10% with the lowest reporting rates were assigned a value of one and intermediate reporting rates were assigned the appropriate value from two to nine. Grid cells with reporting rates of zero were left unchanged. The use of 10 intervals is an example; more or fewer intervals can be used. If only one category is used this would result in presence/absence data. After the reporting rates have been converted to categories based on percentiles, each category will have an equal number of cells, apart from minor discrepancies resulting from tied reporting rates and integer arithmetic.

The fewer categories are used, the less will the internal structures of the distributions be preserved. If only a few intervals are used, the coefficient will tend to become a measure of the degree of overlap. A compromise between too many categories, which may not be justified for small distributions, and too few categories, where most of the information is lost, has to be found. The notation for the transformed reporting rates used here is R_{ij}, the transformed reporting rate for species j in grid cell i. We investigated transforming the positive reporting rates into one, five and 10 categories.

DIFFERENT RELIABILITIES OF REPORTING RATES

The only possible reporting rates when one checklist was collected are zero and one, (0% or 100%). When two checklists are available, the possible reporting rates are 0%, 50% and 100%, probably none of which is close to the true probability of observing the species in the grid cell, which is to be estimated. Values obtained from grid cells with very few checklists (less than five, but especially one or two checklists) are unreliable and should therefore not be allowed to have the same influence on the estimation of the dissimilarity between distributions as more reliable reporting rates, which stem from grid cells for which many checklists were obtained.

This suggests that in the calculation of dissimilarities between distributions, the reporting rate for grid cell *i* should be weighted by the number of checklists n_i from which it was calculated or some function $f(n_i)$ of the number of checklists, so that the distance coefficient contains the expression

$$\sum_i f(n_i)\,(R_{ij} - R_{ik})^2$$

where $f(n_i) = n_i$ was initially chosen. At a later stage in the project, $f(n_i) = \sqrt{n_i}$ was considered as an alternative weight function.

SPATIAL AUTOCORRELATION

The location of a grid cell in two-dimensional space induces spatial autocorrelation. If a species has a large reporting rate in one grid cell, it is more likely to have a large reporting rate in the neighbouring grid cells and vice versa. We ignored correlation between the grid cells and simply compared the bird distributions grid cell by grid cell. This is an area where further work might be needed.

DENOMINATOR, SCALING FACTOR

The issue of differing total areas of distributions also needs consideration. The size of the sum of differences does not only depend on the differences between the reporting rates but also on the total areas covered by the two distributions which are compared. The sum is larger when comparing larger distributions than it would be when comparing two smaller distributions, because the sum is taken over more grid cells.

The sum of the squared differences between species *j* and species *k* is

$$\sum_i n_i\,(R_{ij} - R_{ik})^2 \qquad (9)$$

in the case where each squared difference is weighted by the number of checklists n_i for grid cell *i*. The sum is taken over all grid cells where at least one of the two species has a strictly positive reporting rate. R_{ij} is the transformed reporting rate, as described in the previous section.

The distance coefficient should only take into account those areas where either or both of the two species occur. The total areas of the bird distributions in the atlas differ largely. If cells, where both species do not occur, are regarded as similar, then for small distributions this blank area would have a larger influence on the coefficient than the areas in which the species really occur. This would also have the effect that distributions are compared to the entire area under consideration instead of just to the one other distribution. It is also desirable in this context to scale the dissimilarity by some factor so that the final dissimilarity coefficient has a value between zero and

one. These two problems can be solved by taking the average, i.e. dividing by the number of characters over which the sum was taken, or in the weighted case the denominator should be the sum of weights. If the compared values (R_{ij} and R_{ik}) do not all equal zero and one (as in the case of presence/absence data), this scaling method does not ensure that the coefficient equals one in the case of no overlap. A second option is to divide the total sum by a scaling factor which ensures directly that the coefficient will equal one if the two distributions do not overlap.

The maximum dissimilarity between any two distributions must be obtained if there is no overlap at all between the two distributions. Then the above sum (eq. 9) becomes

$$\sum_i n_i R_{ij}^2 + \sum_i n_i R_{ik}^2 = \sum_i n_i (R_{ij}^2 + R_{ik}^2) \qquad (10)$$

because all terms $R_{ij}R_{ik}$ equal zero.

Two distributions with no overlap should have the maximum possible dissimilarity of one. This is ensured if the sum of differences is scaled by the maximum possible value that can be obtained by comparing those specific two distributions. If all transformed reporting rates of the two distributions are equal, then the sum in equation 9 (the numerator) equals zero and the dissimilarity between the two distributions will be zero as it ought to be.

With this denominator the coefficient becomes

$$D_{jk} = \frac{\sum_i n_i (R_{ij} - R_{ik})^2}{\sum_i n_i (R_{ij}^2 + R_{ik}^2)} \qquad (11)$$

where the R_{ij}'s are the transformed reporting rates. If the cells are not weighted by the number of checklists, and only two categories are used, present ($R_{ij} = 1$) or absent ($R_{ij} = 0$), the above coefficient reduces to the Czekanowski similarity coefficient when subtracted from one. With weights all equal to one, we have

$$D_{jk} = \frac{\sum_i (R_{ij} - R_{ik})^2}{\sum_i (R_{ij}^2 + R_{ik}^2)} \qquad (12)$$

Consider first the numerator $\sum_i (R_{ij} - R_{ik})^2$. Let $\delta = (R_{ij} - R_{ik})^2$. Then the following four categories arise:

Case:		Squared Difference:	Frequency:
1.	$R_{ij} = R_{ik} = 0$ (ignored)	$\delta = 0$	d
2.	$R_{ij} = R_{ik} = 1$	$\delta = 0$	a
3.	$R_{ij} = 0, R_{ik} = 1$	$\delta = 1$	b
4.	$R_{ij} = 1, R_{ik} = 0$	$\delta = 1$	c

The numerator is the sum of all those cases that fall into the categories 3 and 4 (the sum of differences) and therefore the numerator sum is (b + c).

For the denominator, $\sum_i (R_{ij}^2 + R_{ik}^2)$, let $\lambda = (R_{ij}^2 + R_{ik}^2)$. The following cases arise:

Case:		Sum of Squares:	Frequency:
1.	$R_{ij} = R_{ik} = 0$ (ignored)	$\lambda = 0$	d
2.	$R_{ik} = R_{ij} = 1$	$\lambda = 2$	a
3.	$R_{ij} = 0, R_{ik} = 1$	$\lambda = 1$	b
4.	$R_{ij} = 1, R_{ik} = 0$	$\lambda = 1$	c

It follows that the denominator sums to (2a + b + c).

Therefore, for binary data the above coefficient (eq. 11), provided that no weighting is used, reduces to

$$D_{jk} = \frac{\sum_i (R_{ij} - R_{ik})^2}{\sum_i (R_{ij}^2 + R_{ik}^2)} = \frac{b + c}{2a + b + c}$$

If this dissimilarity is converted to a similarity by $S_{jk} = 1 - D_{jk}$, the above becomes

$$S_{jk} = 1 - D_{jk} = 1 - \frac{b + c}{2a + b + c} = \frac{2a}{2a + b + c}$$

This similarity coefficient is known in the literature as the Czekanowski, Sørensen or Dice coefficient (Cox & Cox 1994).

SUM COMPONENTS

From the overall calculated dissimilarity it is unclear whether most of the difference was caused by different patterns inside the distribution, while the overall shape was similar, or whether the difference was caused due to the fact that the one distribution covers a larger area than the other. There is relevant information for examining this in each value of the sum forming the calculated distance between two distributions.

It is therefore of interest to decompose the overall distance to find out how large a contribution is made by each component to the distance coefficient. For example, when comparing an extensive distribution with a small one, much of the difference between the two distributions is due to reporting rates of the larger distribution. How much of the difference is caused merely by this difference in size? How much of the difference comes from areas where the two species co-occur but their reporting rates do not match?

In order to investigate these contributions to the dissimilarity, the distance was split up into four components, depending on the conditions in each grid cell. Suppose that the distribution of species *j* is compared to that of species *k*. Then the observed difference falls into one of the following four categories:

1.) species *j* occurs but species *k* does not
2.) species *k* occurs but species *j* does not
3.) species *j* occurs at a higher transformed reporting rate than species *k*
4.) species *k* occurs at a higher transformed reporting rate than species *j*

Mathematically, D_{jk} is decomposed as follows:

$$D_{jk} = E + F + G + H \tag{13}$$

where

$$E = \frac{1}{M} \sum_{j>0,\,k=0} n_i R_{ij}^2$$

$$F = \frac{1}{M} \sum_{j=0,\,k>0} n_i R_{ik}^2 \tag{14}$$

$$G = \frac{1}{M} \sum_{j>k} n_i (R_{ij} - R_{ik})^2$$

$$H = \frac{1}{M} \sum_{k>j} n_i (R_{ij} - R_{ik})^2$$

where

$$M = \sum_i n_i (R_{ij}^2 + R_{ik}^2)$$

and where

$\sum_{j=0, k>0}$ is taken to mean the sum over those grid cells where $R_{ij} > 0$ but $R_{ik} = 0$,

$\sum_{j>0, k=0}$ means the sum over the grid cells where $R_{ij} = 0$ and $R_{ik} > 0$,

$\sum_{j>k}$ means the sum over those grid cells where $R_{ij} > 0$, $R_{ik} > 0$ and $R_{ij} > R_{ik}$ and

$\sum_{k>j}$ means the sum over those grid cells where $R_{ij} > 0$, $R_{ik} > 0$ and $R_{ij} < R_{ik}$

Note that these differences are not caused by species *j* being more common than species *k*; this factor was removed when transforming the reporting rates to percentile categories.

The last two components are a measure of how much the cores of the two distributions correspond. They should be small if the two species have the cores of their distributions in the same area. The first two components measure how large the area is where the one species occurs but the other one does not.

PRESENTATION OF RESULTS

Different coefficients were compared, in particular the Jaccard coefficient and variations of the modified Euclidean distance (different numbers of categories for transforming the reporting rates and different weights):

a.) The Jaccard similarity coefficient. The similarities were transformed to dissimilarities by $D_{jk} = 1 - S_{jk}$.

b.) The unweighted (weights equal one) form of the Euclidean distance (eq. 12) with one, five and 10 categories for the strictly positive reporting rates. These coefficients are denoted as UW1, UW5 and UW10.

c.) The weighted form of the Euclidean distance, where the weights equal the number of checklists n_i for the respective grid cell, was investigated with one, five and 10 categories for the positive reporting rates. These coefficients are denoted as W1, W5, W10.

d.) A weighted form of the Euclidean distance with weights the square root of the number of checklists n_i for the respective grid cell. This coefficient, denoted as

SQRTW10, was only investigated with 10 categories for the strictly positive reporting rates.

The coefficients were calculated by writing programs in C++. For computational reasons similarities and dissimilarities are expressed as values in the range of zero and 10000 instead of values between zero (for perfect equality) and one (completely dissimilar). Values between zero and one will only be used occasionally to simplify explanations. A calculated dissimilarity of 5670 therefore means the same as 0.5670 or a 56.7 % dissimilarity. All similarities were converted to dissimilarities, especially the similarities calculated by the Jaccard similarity coefficient. If S_{jk} represents the calculated similarity then the dissimilarity was calculated as $D_{jk} = 1 - S_{jk}$.

Dissimilarities are compared through the aid of 'distance diagrams'. For each species a bar graph shows the distances between this species and the other species for which the distances have been calculated. This is done in the form of a scaled rectangle where the calculated dissimilarities are shown as vertical lines representing distances from the left hand side of the bar.

RESULTS

For this project data were only available from the area south of 27° S, therefore the comparisons are restricted to this area.

THE MODIFIED EUCLIDEAN COEFFICIENT W1

The Coefficient W1 (weighted and scaled Euclidean, all strictly positive reporting rates transformed to one) produces some abnormal results. For example, in Fig. 1 (figures and tables can be found at the end of this chapter), the closest species to the Cape Siskin is the Protea Canary for all seven coefficients. The second most similar is generally the Bully Canary. This is the case for all coefficients except for W1. This coefficient claims that the distributions of the Whitethroated Canary and Cape Siskin are more similar than are the distributions of the Cape Siskin and the Bully Canary. Although both of these distributions differ markedly from that of the Cape Siskin, it is clear when visually inspecting the atlas maps (Figs A6-A9, Appendix A), that the distribution of the Bully Canary is the one that is more similar.

Fig. 2 shows dissimilarities of distributions of some canaries to that of the Yellow Canary. For all the coefficients either Blackheaded or Blackthroated Canary are calculated to have the second most similar distribution to that of the Yellow Canary, again excepting the W1 coefficient, where the Cape Canary is second. It also claims that the distributions of the Streakyheaded and Yellow Canaries are more similar than the distributions of Yellow and Blackheaded Canary. Again it is apparent from the atlas maps that the Blackheaded (Fig. A11, App. A) and Blackthroated Canaries (Fig. A12, App. A) should be ranked more similar to the distribution of the Yellow Canary (Fig. A10, App. A) than the Cape Canary (Fig. A2, App. A). Except for the Western Cape, the distributions of Cape and Yellow Canary are almost complementary, and so are the distributions of the Yellow and the Streakyheaded Canaries (Fig A10 & A13, App. A).

The Coefficient W1 calculates in general smaller dissimilarities between distributions than the other coefficients. But because of the many obscure results it produces, this coefficient is not considered worthy of further investigation.

YELLOWEYED CANARY

All coefficients presented in Fig. 3 find that the Streakyheaded Canary has the closest distribution to that of the Yelloweyed Canary (see also Table 3). The second most similar distribution is in all cases the Bully Canary distribution. After that the Jaccard and UW1 coefficients rank the Cape Canary closer, the weighted methods and UW5 and UW10 rank the Forest Canary closer.

The Jaccard coefficient for the Cape Canary is smaller than that for the Forest Canary, 5957 and 6278 respectively (Table 3), when comparing these distributions with the distribution of the Yelloweyed Canary. That means that there is more overlap between the distributions of the Cape and the Yelloweyed Canaries, because the Jaccard coefficient is a measure of overlap in the case of bird distributions. However

from the maps it appears that the core parts of the distributions correspond to a larger extent in the case of Forest and Yelloweyed Canary (Figs A14 & A15, App.). The Cape Canary has a central part of its distribution in the south-western Cape (Fig. A2, App.), where the Yelloweyed Canary does not occur and the Forest Canary occurs only in a few grid cells.

This may be the reason why coefficients, that sort the strictly positive reporting rates into a larger number of intervals, estimate the Forest Canary to be more similar to the Yelloweyed Canary than the Cape Canary, because these coefficients take into account the internal structures of the distributions. This shows the advantages and also the purpose of using a larger number of categories. The magnitudes of the dissimilarities of the Forest and Cape Canary distances from the Yelloweyed Canary do not differ by much, the differences in calculated dissimilarities roughly equal 300 for all coefficients (Table 3).

It is difficult, by visual inspection of the atlas maps, to conclude which of the two species, Forest or Cape Canary, should be the closer to the Yelloweyed Canary (Figures A2, A14, A15, App.). What can however be suggested is that the dissimilarity between Cape and Forest Canaries should be smaller than the dissimilarity between Yelloweyed Canary and either of Cape and Forest Canaries. A reason for this is that these two species occur at least roughly in the same area even though the distribution of the Forest Canary is more restricted to the coastal areas.

Excepting the outcomes of the W1 coefficient and the exchange between Cape and Forest Canary mentioned above, the ordering of species in the distance diagrams is constant up to a dissimilarity of about 8000 (Table 3 and Fig. 3). The methods produce slightly different magnitudes in the dissimilarities but the distances between the ranked values are approximately proportionally preserved (Fig. 3).

YELLOW CANARY

The coefficient that produces the minimum dissimilarity when the distribution of the Yellow Canary is compared to distributions of other canaries (Fig. 2 and Table 2), is UW1 (excepting W1, which is ignored here).

For the three unweighted Euclidean coefficients (UW) and the Jaccard coefficient the seven species with the most similar distributions to that of the Yellow Canary are ranked in the same order. For the weighted methods this changes in that the distribution of the Blackthroated instead of the Blackheaded Canary is ranked second most similar to the distribution of the Yellow Canary (Fig. 2).

One would expect that the smaller the difference between two dissimilarities calculated with one coefficient, the more likely it will be that the ranks will be exchanged when other coefficients are used. For the Jaccard and the unweighted coefficients there is not a large difference between the calculated dissimilarities for the Blackheaded and Blackthroated Canaries (5449 and 5687). For W5 and W10 this difference is slightly larger (Fig. 2 and Table 2), (5524 and 4081 in the case of W10).

East of 25° E and south of 26° S there is a sudden increase in the number of checklists that were collected per grid cell (see fig. 5 of Harrison & Underhill 1997). This is the area where the Blackthroated Canary has the core of its distribution (Fig. A12, App.).

The Jaccard coefficient is useful as a reference because it measures the proportion of overlap of the two distributions. It is not influenced by the actual reporting rates or by the accuracy of the measurements. From the Jaccard coefficient, converted to a dissimilarity, it can be observed that there is more overlap between the Blackheaded and Yellow Canary distributions (Jaccard = 5449) than between those of the Blackthroated and the Yellow Canaries (Jaccard = 5687) (Table 2). The Blackheaded Canary (Fig. A11, App.) occurs mostly in the area where ten or less checklists were recorded except in the eastern parts of its distribution, where the number of checklists is mostly less than 20. The area where the Blackheaded Canary does not occur, especially east of 25° E, is where the Blackthroated Canary (Fig. A12, App.) has the core of its distribution. At the same time this area has, in general, more checklists. This means that the area of the Yellow Canary where the Blackthroated Canary also occurs is weighted more heavily than the area where the Blackthroated Canary does not occur. The weighting effect is opposite in the case of the Blackheaded Canary. The effect of the weighting therefore is that the distribution of the Blackthroated Canary is estimated to be more similar to that of the Yellow Canary than is the distribution of the Blackheaded Canary. This is confirmed by Coefficients W5 and W10 (Fig. 2). For the unweighted coefficients the Blackheaded Canary is closer to the Yellow Canary.

It is however difficult to judge visually which of the two, Blackheaded or Blackthroated, should in distribution be closer to the Yellow Canary if only the maps are considered and the weights are ignored (Figs A10 – A12, App.).

SPOTTEDBACKED WEAVER

The four distributions ranked most similar to the distribution of the Spottedbacked Weaver are in the same order for all six coefficients (Fig 4 and Table 4). Then UW1 and the Jaccard coefficient rank the Golden Weaver fifth while in the other coefficients the Cape Weaver is ranked fifth. By inspection of the maps, however, (Figs A2, A16, A19, App.) it can be seen that the distributions are all very different, the most striking difference being to the distribution of the Golden Weaver, but that may only be a visual impact because this small distribution (Golden Weaver) is easier to grasp by eye.

This may provide an idea of the magnitude of the dissimilarity value beyond which two distributions can be considered to be different from each other. There is still reasonable similarity to the distribution of the Yellow Weaver. Therefore the cutoff point, after which distributions should be considered as different, should be somewhere larger than the respective values observed for the Yellow Weaver for the various methods (see Table 8). For each of the coefficients it is suggested that the largest value for reasonable similarity should be between the respective values shown in Table 8. After this cut-off point distributions are visually very different in appearance.

A first guide would be that dissimilarities larger than 5000, or equivalently 0.5, are definitely not of any interest, a slightly larger cutoff could be used in the case of the Jaccard coefficient. For the two weighted coefficients W5 and W10, a cutoff point of 4000 seems reasonable. Values less than 2500 indicate that the distributions are visually similar.

The weavers with the most similar distributions to that of the Spottedbacked Weaver are the Spectacled, Thickbilled, Forest and Yellow Weavers in order of decreasing similarity (Table 4 and Fig. 4). In the group of weavers and canaries the smallest calculated dissimilarity between two distributions was that between the distributions of the Spottedbacked Weaver and the Spectacled Weaver (see also Figs A16 & A17, App.). Coefficient W5 calculated the smallest value (584). The Jaccard coefficient dissimilarity was 2911 and the value calculated by Coefficient W10 was 643 (Table 4).

CAPE SISKIN

When visually inspecting the atlas maps of the Cape Siskin and the Protea Canary (Figs A6 & A7, App.), one would anticipate that the dissimilarity between these two species was smaller than calculated (Table 1). The value of the Jaccard coefficient was 4920. The method that calculates the smallest dissimilarity between these two species, besides W1, is UW5 (2935) and the value calculated by W10 was 3790.

Both of these species have restricted distributions. It may be that the visual impact of similarity is more striking for small distributions. The outlines of the distributions are approximately the same. The Cape Siskin however appears in more of the grid cells. This is where the sum-component matrices, E to H (equations 13, 14), become useful. They show the components that make up the dissimilarity value (Table 9).

More than half of the difference (68.92%) comes from grid cells where the Protea Canary is absent and the Cape Siskin is present (Table 9). On the other hand, the Protea Canary occurs in few grid cells from which the Cape Siskin is absent, contributing 0.87% to the overall difference between the species. The remaining difference (797 + 348 = 1145, 30.21%) is attributable to differences in reporting rates, or rather transformed reporting rates, in the grid cells where both species were recorded (Table 9).

It can be assumed that in the case of these two species, too few checklists do not cause unreliable reporting rates. There is also not much difference in the number of checklists recorded for each grid cell (Figs A6, A7, App.) and (fig. 5 of Harrison & Underhill 1997).

The calculated dissimilarities indicate that the distributions of the Cape Siskin and the Protea Canary are not that similar after all. This example gives a useful illustration of what the coefficients calculate as opposed to what is perceived when looking at the maps.

In Fig. 1 the order of the two species with the most similar distribution to that of the Cape Siskin are the same for all six coefficients: the Protea Canary is most similar, the second most similar distribution is that of the Bully Canary. After that the ranks do not agree between coefficients. Inspecting the atlas maps, it can be seen that the distributions of Cape, Streakyheaded (Figs A2, A8, A13, App.) and the other canaries, even the Bully Canary, are so different from that of the Cape Siskin (Fig. A6, App.), that it is impossible to say from visual inspection of the distribution maps, which of these should be more similar to the distribution of the Cape Siskin.

MASKED WEAVER

In the case of the Masked Weaver there is no other weaver that has a nearly similar distribution (Fig. 5). The species with the closest distribution to that of the Masked Weaver is the Cape Weaver (Table 5 and Fig. 5). The dissimilarity is 3667 in the case of Coefficient UW1, 4377 for Coefficient W5 and larger than 5000 for the other distance coefficients. The atlas maps (Figs A1 & A3, App.) show that these two distributions are very different and especially that the cores of the distributions do not correspond. The core of the Cape Weaver distribution is in the Western Cape and in patches along the eastern parts of South Africa while the core of the Masked Weaver distribution is a large regular area in the interior of South Africa.

This illustrates again that coefficient values larger than 0.5 (or 5000) should be interpreted as showing that the distributions do not have much in common, i.e. are dissimilar, but also that different coefficients need different cut-off points.

The only difference in the ranking of the first five species is that the Spectacled and the Spottedbacked Weavers appear in this order of similarity in the UW5 and the UW10 coefficients and exchanged for the other coefficients. As can be seen from the distance diagram (Fig. 5), there is almost no difference in the calculated dissimilarities between the distribution of the Masked Weaver and for these two weavers in any of the coefficients, therefore this exchange does not have much meaning.

JACCARD versus UW1 (also Czekanowski)

In the case of the weaver and canary groups (Tables 1-5) these two coefficients have exactly the same orderings. This is because both coefficients 'rank' the distributions according to the degree of overlap. Also the structures of their distance diagrams are proportionally constant, inter-specific distances are proportionally preserved (Figs 1-5). The Jaccard coefficient is more restricted to the right hand side of the distance diagrams, the values are all closer to one. UW1 calculates the smaller dissimilarities, which was stated earlier to be more attractive when the overlap in general is small. Also this coefficient gives a wider spread of the calculated values.

CHANGES IN RANKINGS OF DISSIMILARITIES

Comparing how much changes in the ranking of most similar to least similar distribution when going from one to the other coefficient (Tables 1-5), there is mostly only one pair of differences when switching from W5 to W10 and also when switching from UW5 to UW10. More of the ranks change between UW1 and UW5 and also between UW10 and W10.

SIMILAR versus DISSIMILAR DISTRIBUTIONS

There seems to be a general trend in the ranking of the distances by different coefficients. Vertical bold lines in Tables 1-7 make a distinction between distributions that are considered roughly similar to the one under consideration and the distributions that are too different from the one to which they are compared, the latter distributions are almost complementary. Generally, for dissimilarities smaller than the line, to the right of the line, the rankings are the same. In the cases where this does not hold, the distributions are already dissimilar to the one compared. For example in the comparison to the Yellow Canary (Table 2) different methods rank either the Blackheaded or the Blackthroated Canary into second and third positions (Figs A11 & A12, App.). In the comparison to the Yelloweyed canary (Table 2), the Cape and Forest Canaries (Figs A2 & A15, App.) are in fourth and fifth positions, depending on the coefficient used. All these distributions differ markedly from the one they are compared to in that the area covered by the distribution is almost double or half to that of the distribution they are compared to.

In the case of the Masked Weaver there is no other weaver that has a similar distribution, although in all the coefficients the Cape Weaver is ranked first (Table 5). The four species that have the most similar distribution to that of the Spottedbacked Weaver (Table 4) are ranked in the same order for all six coefficients. After the vertical bold line in Table 4 the ranks change for different methods. This is the same stage at which the distributions start to differ too much from that of the Spottedbacked Weaver.

In Tables 1-7 it appears that for the modified Euclidean distances (W5, W10, UW5 and UW10) a cutoff point at 4000 is a useful guideline to separate roughly similar distributions from those that are too dissimilar for sensible comparison. 4000 should be taken as a maximum, dissimilarities smaller than this show increasingly more obvious similarities between distributions. The cutoff point for the Jaccard coefficient is larger than 4000.

MAGNITUDES OF CALCULATED DISSIMILARITIES

The dissimilarities calculated by the scaled Euclidean dissimilarity coefficients, weighted and unweighted, are mostly smaller than the corresponding value calculated by the Jaccard coefficient. This was to be expected because they are squared distance coefficients. An exception is shown in Fig. 2 (Yellow Canary) where the dissimilarities calculated by the other coefficients are only slightly less than those calculated by the Jaccard coefficient. In Fig. 5 (Masked Weaver) the unweighted

Euclidean coefficients produce larger values. There is no coefficient that consistently calculates the minimum dissimilarity between two distributions.

The magnitudes of the calculated dissimilarities of the modified Euclidean distances shown here (W and UW) are however more comparable to those of the Jaccard coefficient dissimilarities than when the square root was taken. Taking the square root would shift all lines in the distance diagrams to the right, decreasing the range and the spread of the values, and making it more difficult to compare results.

In the following paragraphs two sets of comparisons are discussed with a smaller subset of coefficients.

COLLARED SUNBIRD

For the northeastern part of the distribution (east of 30°E) the distribution of the Olive Sunbird (Fig. A21, App) is more similar to that of the Collared Sunbird (Fig. A20, App) than is the distribution of the Grey Sunbird (Fig. A22, App.), especially in the area 29°S, 30°E, in grid cell 2731CD and also in southern Swaziland. The southern part of the Collared Sunbird distribution resembles more that of the Grey Sunbird, especially in the block 32°S, 26°E where the Olive Sunbird does not occur. This resemblance of the Grey Sunbird with the Collared Sunbird while the Olive Sunbird is absent in the southern part is larger than the resemblance of the Collared Sunbird and Olive Sunbird in the northern parts. It is therefore correct that the Grey Sunbird is ranked closer to the Collared Sunbird than the Olive Sunbird in all of the coefficients (Fig. 6).

The distribution of the Collared Sunbird along the eastern Coast is also the reason why the Black Sunbird is closer in distribution to the Collared Sunbird than either of the Whitebellied or Scarletchested Sunbirds (Figs A24 & A25, App.).

After this it becomes complicated to compare the dissimilarities. The Scarletchested Sunbird has a narrower distribution than the Collared Sunbird, but it has only small reporting rates south of 31°S and does not occur west of 30°E, while the Whitebellied Sunbird has a broader distribution but larger reporting rates south of 29°S resembling more the categories of the Collared Sunbird distribution. The Whitebellied Sunbird is ranked closer to the Collared Sunbird by the weighted coefficients W5 and W10. The Greater Double Collared Sunbird should not be closer in distribution to the Collared Sunbird than either of the Whitebellied or Scarletchested Sunbirds because of its absence east of 31°30'E (Fig. A26, App.). The Coefficient UW1 however ranks the species as follows: Scarletchested, Greater Doublecollared, Whitebellied Sunbird in decreasing similarity (Fig. 6).

The Coefficient SQRTW10 seems to show a good relation of the distances where Grey, Olive Sunbirds are ranked closely, then more dissimilar are the Black Sunbird, Whitebellied and Scarletchested Sunbirds and only then the Greater Doublecollared Sunbird (Fig. 6). The coefficients that do not use weights rank the Scarletchested Sunbird closer in distribution to the Whitebellied Sunbird but it is difficult to judge by visual inspection of the observed distribution maps which of these two has a more similar distribution to that of the Collared Sunbird.

SCARLETCHESTED SUNBIRD

For the Scarletchested Sunbird distribution there is a choice of three sunbird species with the most similar distribution (Fig. 7).

The UW1 Coefficient may give too much weight to the overlap of the distributions. The Olive Sunbird (Fig. A21, App.) distribution has the largest overlap, UW1 = 3370, (Table 7 and Fig. 7) with the distribution of the Scarletchested Sunbird (Fig. A25, App.) but the internal patterns of these two distributions do not correspond. Neither does the pattern of the Whitebellied Sunbird (Fig. A24, App.) distribution match that of the Scarletchested Sunbird. When interested in the internal distribution pattern, so that the cores of the distributions fall into the same grid cells, the distribution of the Purplebanded Sunbird (Fig. 27, App.) is closest to that of the Scarletchested Sunbird. The three coefficients that weight the grid cells show this, W5, W10 and SQRTW10 (Table 7 and Fig. 7).

The overall outline of the Whitebellied Sunbird distribution resembles that of the Scarletchested Sunbird more closely than does the distribution of the Olive Sunbird. This is a very subjective opinion and in this case would rank the species from most similar to least similar as Purplebanded, Whitebellied and Olive Sunbird. The Coefficient SQRTW10 ranks the dissimilarities in this order. Coefficients W5 and W10 calculated the dissimilarity to the Purplebanded Sunbird as smallest followed by the Olive and then the Whitebellied Sunbirds, although there is not much difference in the three magnitudes of the dissimilarities to the Scarletchested Sunbird distribution in these three coefficients. The Coefficient SQRTW10 calculates slightly larger values than the methods which use as weights the number of checklists n_i.

SCATTERPLOTS

In Figs 8 (a) and (b) scatterplots to compare the dissimilarities calculated by different coefficients are shown. The dissimilarities calculated by Coefficients W10 are only slightly larger than those for the five-interval Coefficient W5 (Fig. 8a). This is almost to be expected because there are more possibilities that grid cells fall into different categories and that the values differ by a larger amount. These two coefficients produce almost similar results and it is difficult to choose one of these by comparing the dissimilarities calculated. The Coefficient W10 preserves more of the given information, but 10 categories may not be justified in the case of small distributions where the original observed reporting rates may only have an average of 10%.

When comparing coefficients UW10 and W10 (Figure 8b) the values lie much less along a straight line than in Fig. 8 (a). Most of the UW10 values are larger than the corresponding W10 values. Some values differ by as much as 2000 and more. Only the values in the upper region, larger than 8000, lie along a straight line but these are not the interesting dissimilarities.

Various distance tables of similarity matrices are kept in Appendix C.

AN APPLICATION OF THE DEVELOPED DISSIMILARITY COEFFICIENT

In a group of species from the same genus, the species may be more likely to have similar habitat requirements and may be dependent on the same type of food resources. It is therefore of interest how much they occur in different areas or habitat types to avoid too much competition. For canaries occurring in the Fynbos biome, it was investigated in how far they choose different habitats within this biome (Underhill *et al.* 1998).

DISCUSSION

The Euclidean Distance (eq. 6) and the Coefficient of Divergence (eq. 4) were investigated. The Euclidean distance itself is not useful for the comparison of bird distribution maps because it does not take into account the areas of the two distributions and there is no theoretical maximum value built into the equation. Rather it calculates an average difference between observed values. The Jaccard coefficient and the scaled Euclidean coefficients calculate some sort of overlap; they divide by a function of the total possible area. This seems to be preferable when evaluating the similarity of geographical distributions. These coefficients divide by a sum, as opposed to the Euclidean distance, which takes the average or at the most scales each single sum component when it is in the form of the Taxonomic distance.

The abnormal behaviour of Coefficient W1 is to be expected. Because for this coefficient only one interval for all positive values is used, the transformed reporting rates will all equal zero or one. From this follows that the only possible difference between observed values can be zero or one. The more accurately one wants to estimate the probability that a species occurs in a cell, the more checklists are required. For 10 checklists the reporting rates can only estimate the true probability to an accuracy of 10%. But if the only matter of interest is whether the species occurs in the grid cell or not, a binary response, not many checklists are required. Two or three checklists provide nearly as good an indication of the presence of a species in a grid cell as do 100 checklists. We can therefore assume that in the case of presence-absence data all values are equally accurate and it therefore does not make sense to weight observed reporting rates. If each difference of 'one' is multiplied by the number of checklists, the number of checklists dominate the outcomes too much. This causes the false rankings observed when using the Coefficient W1. Presence/absence data should not be weighted in the case of grid cell data.

A measure of overlap, such as estimated by the Jaccard or UW1 coefficients, was not considered sufficient for comparing the bird distribution maps. These two coefficients however show similar results except that the UW1 dissimilarities are arithmetically smaller. Coefficients only using presence/absence data do not allow comparison of the internal structure, but only of the overall shape.

Between the remaining coefficients it is difficult to decide. The rankings of the dissimilarities are roughly the same for the coefficients W5, W10, UW5 and UW10, especially for distributions that have calculated coefficients of less than 4000, i.e. are fairly similar. For the more dissimilar distributions the ranking of dissimilarities changes more often between coefficients. This can also be explained in that the differences between calculated dissimilarities are in general smaller because usually more species have dissimilar distributions and therefore significantly more dissimilarities will fall into the range 5000 to 10000. If there are small differences between calculated dissimilarities it becomes more likely that these are exchanged in ranking when using other coefficients.

There is almost no difference in the rankings, especially not for the similar distributions of interest when sorting the strictly positive reporting rates into five or

10 categories. There are more differences when changing from an unweighted to a weighted coefficient such as from UW10 to W10. Additionally, those rankings that do change are often borderline cases where it is difficult to judge by visual inspection of the distribution maps, which of the pairs of distributions is more similar.

A concern that arose during the investigation was that the weights, n_i (the number of checklists for grid cell *i*), might be too area dependent, i.e. some regions have large numbers of checklists per grid cell, for example, the South-Western Cape, while other regions have only few checklists per grid cell, for example, the Karoo. A second concern was that this weight function might cause some grid cells to be too dominant when used to weight the squared difference in transformed reporting rates. The reason for this is that the number of checklists collected per grid cell ranges from zero to 1260.

The problem that the number of checklists varies with region only influences the dissimilarity rankings if extensive distributions are compared with distributions that are much smaller, only half their size or less. In such cases, the smaller distribution in the part with few checklists will be found to be more dissimilar than would a smaller distribution in the area with more checklists. The distributions of the Yellow, Blackheaded and Blackthroated Canaries (Figs. A10-A12, App.) provide an example of this. However, we are mainly interested in distributions that roughly cover the same areas and for such pairs of distributions the problem that the number of checklists depends on the region in which the species occurs can almost be ignored. The only other concern is that two distributions in regions with few checklists exhibit more variability between observed reporting rates than two distributions in a region with more checklists. This increases the calculated dissimilarity for such species. It is hoped that by using weights the more reliable grid cells will dominate the calculation.

WEIGHTING BY THE NUMBER OF CHECKLISTS

The second concern was the range of the number of checklists. 30 checklists can be assumed to provide reliable estimates of reporting rates. A larger number of checklists does not make much difference in the accuracy of the estimates and dissimilarity calculations. Therefore it is doubtful whether a grid cell with 200 checklists should obtain more than six times the weight as a grid cell with 30 checklists when 10 categories only allow accuracy of approximately 10%. Also the grid cells with more than 50 checklists may override all other calculated differences in other grid cells. Therefore we changed the weights for each squared difference in a grid cell to the square root of the collected number of checklists for the grid cell. This weight also decreases the problem of different regions of the country being weighted too differently. The only coefficient of this form investigated was SQRTW10 and it showed results at least as good as those of W10.

If more than one category for the strictly positive reporting rates is used, it becomes increasingly more important to weight the grid cells. For example, in a grid cell with only two checklists say both times the species was observed. The observed reporting rate is one, the corresponding percentile when using 10 intervals is 10. This value is a rare event in cells with many checklists, where the observed reporting rate is an

average between observers. Any difference calculated from this value, will be too large. Such reporting rates, originating from grid cells with only a few checklists (especially less than five), should be weighted down so as not to obscure the coefficients. A weighted coefficient becomes particularly important when one or both of the two compared species occur in areas with few checklists such as the Karoo or the Transkei.

NUMBER OF CATEGORIES

Although mainly coefficients that rank the strictly positive reporting rates into 10 categories were investigated (UW10, W10, SQRTW10), there does not seem to be much difference in the dissimilarities when using five categories instead, especially not for similar distributions (dissimilarities less than 4000).

Five categories may be more justified for small distributions and species with low average reporting rates than 10 categories. Therefore we suggest using five categories for the positive reporting rates. The three shade categories used in the atlas distribution maps (Harrison *et al.* 1997 a, b) already provide a good visual description of location of cores and gradients of observed reporting rates.

When only one or two intervals are chosen for the strictly positive reporting rates, it is not possible to compare where the cores of the distributions lie. By this we mean that it is also of interest to determine, if two species occur in the same area whether in that area their exact habitat selections and preferences also coincide. The use of five or 10 categories for the strictly positive reporting rates allows more comparison of the internal structures of the two distributions being compared. But with an increasing number of categories the coefficient may become more dependent on the variability of reporting rates in different regions, i.e. the calculated differences will in general be larger in areas that are more variable, i.e. have fewer checklists collected per grid cell, for example, the Karoo. Using only five categories may partly solve this.

Although we have mainly concentrated on using 10 categories for the strictly positive reporting rates, five categories may be the better option. Firstly there does not seem to be much difference in the estimation of dissimilarity when taking five or 10 categories and five categories may be better when comparing small distributions.

INTERPRETATION OF CALCULATED VALUES

The dissimilarities as they were calculated here and the square roots of these dissimilarities both lie in the range zero to one. The values of the squared dissimilarities compare better in magnitude to dissimilarities calculated by the Jaccard coefficient. If the squared dissimilarities are used, the interesting part of the range (dissimilarities between zero and 0.25) is stretched out. All dissimilarities were left as squared dissimilarities. Also, because the denominator is based on squared values (eq. 11), it is not clear whether taking the square root of the dissimilarities would be justified.

The difficulty in visually comparing the distribution maps is caused to a large extent by the differing areas covered by distributions. Mostly distributions only overlap to some extent and then have another area where the other species does not occur.

For Coefficient SQRTW10, the smallest calculated dissimilarity, the two most similar distributions in the available data set, were those of the Forest Weaver and the Collared Sunbird (see Figs A20 & A32, App.). The calculated dissimilarity was 695. These calculations consider only the distributions south of 27°S. The value of the smallest observed dissimilarity depends to some extend on measurement error, caused by the characteristics of reporting rates. This measurement error can also not be fully removed by weighting the observations.

The dissimilarities calculated here are not to be seen as absolute distances. They are only comparative measures. But the relative magnitudes of the dissimilarities may be seen as meaningful. If for example the distribution of species *j* is compared to the distributions of species *k* and *m*, and if $D_{jk} = 2000$ and $D_{jm} = 4000$, it can be said that the distributions of species *j* and species *m* are considerably more dissimilar than the distributions of species *j* and species *k*. But it is not true to say that species *m* is twice as dissimilar as species *k*, when compared to the distribution of species *j*.

For example, it was established that the Spectacled and the Spottedbacked Weavers (SQRTW10 = 900) have more similar distributions than the Yellow and the Whitethroated Canaries (SQRTW10 = 3173) and these distributions are again more similar than those of Cape Siskin and Protea Canary (SQRTW10 = 3333), (Tables C1 & C2, App. C).

The following guidelines to the interpretation of these dissimilarities are based on the experience obtained in developing them. A dissimilarity of less than 4500 means that the two species occur roughly in the same area but that the one distribution is considerably larger than the other, for example in the case of the Forest and the Streakyheaded Canaries, (Figs A14 & A15), the dissimilarity was 4455 (SQRTW10). In the case of the SQRTW10 coefficient, the dissimilarities less than 4000 seem to include all pairs of distributions where it can be said that they are roughly similar.

Table 1 Dissimilarities of Canary distributions to that of the **Cape Siskin**. Dissimilarities are sorted in descending order, so that the most similar distributions are on the right of the table. The vertical bold line provides a guide, separating distributions that are similar to that of the Cape Siskin from distributions that are considered different. See also Figure 1.

W10	Lemonbr. 10000	Blackthr. 9969	Yelloweye 9947	Blackh. 8947	Streakyh. 7840	Yellow 7634	Forest 7557	Whitethr. 7314	Cape 6581	Bully 6565	Protea 3790
W5	Lemonbr. 10000	Blackthr. 9955	Yelloweye 9920	Blackh. 8719	Streakyh. 7562	Yellow 7338	Forest 7217	Whitethr. 6824	Cape 6430	Bully 6372	Protea 3375
W1	Lemonbr. 10000	Blackthr. 9875	Yelloweye 9414	Blackh. 7792	Streakyh. 6329	Forest 6266	Cape 6265	Yellow 5647	Bully 5635	Whitethr. 5257	Protea 2898
UW10	Lemonbr. 10000	Blackthr. 9961	Yelloweye 9946	Blackh. 9222	Yellow 8962	Whitethr. 8622	Forest 7838	Streakyh. 7379	Cape 7020	Bully 6644	Protea 3084
UW5	Lemonbr. 10000	Blackthr. 9946	Yelloweye 9928	Blackh. 9132	Yellow 8864	Whitethr. 8502	Forest 7688	Streakyh. 7277	Cape 7020	Bully 6534	Protea 2932
UW1	Lemonbr. 10000	Blackthr. 9807	Yelloweye 9585	Blackh. 8614	Yellow 8181	Whitethr. 7959	Cape 7248	Forest 7161	Streakyh. 6850	Bully 6190	Protea 3262
Jaccard	Lemonbr. 10000	Blackthr. 9903	Yelloweye 9789	Blackh. 9256	Yellow 9000	Whitethr. 8864	Cape 8405	Forest 8346	Streakyh. 8132	Bully 7648	Protea 4920
SQRTW10	Lemonbr. 10000	Blackthr. 9961	Yelloweye 9940	Blackh. 8998	Yellow 8379	Wh 7950	Forest 7560	Streakyh. 7497	Cape 6767	Bully 6486	Protea 3333

Table 2 Dissimilarities of Canary distributions from that of the Yellow Canary. Distances are sorted in descending order, so that the most similar distributions are on the right of the table. The vertical bold line provides a guide, separating distributions that are similar to that of the Yellow Canary from those that are considered different. See also Figure 2.

W10	Drksb.Sis 9558	Yelloweye 9374	Forest 8880	Protea 8641	Streakyh. 7736	Cape Sis 7634	Bully 7502	Cape 6384	Blackh. 5524	Blackthr. 4081	Whitethr. 3287
W5	Drksb.Sis 9499	Yelloweye 9122	Forest 8469	Protea 8361	Cape Sis 7338	Streakyh. 7223	Bully 7102	Cape 5902	Blackh. 5327	Blackthr. 4085	Whitethr. 2935
W1	Drksb.Sis 9244	Protea 7329	Yelloweye 7184	Forest 6493	Cape Sis 5647	Bully 5193	Blackh. 4965	Streakyh. 4743	Blackthr. 4497	Cape 3586	Whitethr. 1572
UW10	Forest 9662	Yelloweye 9620	Drksb.Sis 9539	Protea 9436	Cape Sis 8962	Bully 8836	Streakyh. 8449	Cape 7263	Blackthr. 4507	Blackh. 4328	Whitethr. 3007
UW5	Forest 9571	Drksb.Sis 9520	Yelloweye 9504	Protea 9348	Cape Sis 8864	Bully 8681	Streakyh. 8219	Cape 6991	Blackthr. 4463	Blackh. 4247	Whitethr. 2868
UW1	Drksb.Sis 9307	Protea 8967	Forest 8882	Yelloweye 8450	Cape Sis 8181	Bully 7669	Streakyh. 6674	Cape 4976	Blackthr. 3972	Blackh. 3744	Whitethr. 1962
JACCARD	Drksb.Sis 9642	Protea 9456	Forest 9408	Yelloweye 9161	Cape Sis 9000	Bully 8681	Streakyh. 8006	Cape 6646	Blackthr. 5687	Blackh. 5449	Whitethr. 3282
SQRTW10	Drksb.Sis 9522	Yelloweye 9443	Forest 9327	Protea 9091	Cape Sis 8379	Bully 8097	Streakyh. 7965	Cape 6705	Blackh. 4788	Blackthr. 4132	Whitethr. 3173

Table 3 Dissimilarities of Canary distributions to the distribution of the Yelloweyed Canary. The bold vertical line provides a guide, separating distributions that are considered different from that of the Yelloweyed Canary from distributions that are similar. Dissimilaritiesare sorted in descending order, so that the most similar distributions are on the right of the table. See also Figure 3.

W10	Protea 9981	Cape Siskin 9947	Drakensb. S 9879	Blackhead 9683	Yellow 9374	Lemonbr. 9373	Whitethr. 9248	Blackthr. 9164	Cape C. 5497	Forest 5110	Bully 3675	Streakyh. 3578
W5	Protea 9958	Cape Siskin 9920	Drakensb. S 9824	Blackhead 9606	Lemonbr. 9372	Yellow 9122	Whitethr. 9027	Blackthr. 8923	Cape C. 5104	Forest 4853	Bully 3441	Streakyh. 3258
W1	Protea 9575	Cape Siskin 9414	Drakensb. S 9228	Lemonbr. 9179	Blackhead 8752	Whitethr. 7410	Yellow 7184	Blackthr. 7154	Forest 3223	Cape C. 2876	Bully 2565	Streakyh. 2207
UW10	Protea 9978	Cape Siskin 9946	Drakensb. S 9812	Blackhead 9651	Yellow 9620	Whitethr. 9430	Lemonbr. 9381	Blackthr. 9167	Cape C. 5858	Forest 5504	Bully 4167	Streakyh. 4020
UW5	Protea 9962	Cape Siskin 9928	Drakensb. S 9736	Blackhead 9587	Yellow 9504	Lemonbr. 9370	Whitethr. 9343	Blackthr. 8999	Cape C. 5601	Forest 5386	Bully 4004	Streakyh. 3911
UW1	Protea 9767	Cape Siskin 9585	Lemonbr. 9411	Drakensb. S 9176	Blackhead 8880	Whitethr. 8517	Yellow 8450	Blackthr. 7734	Forest 4575	Cape C. 4241	Bully 3617	Streakyh. 3341
Jaccard	Protea 9883	Cape Siskin 9789	Lemonbr. 9697	Drakensb. S 9571	Blackhead 9408	Whitethr. 9200	Yellow 9161	Blackthr. 8723	Forest 6278	Cape C. 5957	Bully 5313	Streakyh. 5009
SQRTW10	Protea 9977	Cape Siskin 9940	Drakensb. S 9835	Blackhead 9603	Yellow 9443	Lemonbr. 9303	Whitethr. 9255	Blackthr. 9070	Cape C. 5663	Forest 5364	Bully 3948	Streakyh. 3769

Table 4 Dissimilarities of weavers to the **Spottedbacked Weaver**, sorted in descending order, so that the most similar distributions to that of the Spottedbacked Weaver are on the right of the table. See also Fig. 4. The bold vertical line is a guide, separating species whose distribution are considered different from that of the Spottedbacked and distributions that are similar.

JACCARD	Redhead. 9868	Masked 8873	Brownthr. 8855	Golden 8496	Lss.Mask. 8238	Cape 7816	Yellow 5957	Forest 4694	Thickb. 4375	Spect. 2911
UW1	Redhead. 9737	Masked 7973	Brownthr. 7944	Golden 7384	Lss.Mask. 7003	Cape 6413	Yellow 4241	Forest 3066	Thickb. 2800	Spect. 1703
UW5	Redhead. 9770	Masked 9119	Brownthr. 8031	Cape 7163	Lss.Mask. 6940	Golden 6931	Yellow 3933	Forest 3304	Thickb. 2190	Spect. 1286
UW10	Redhead. 9779	Masked 9288	Brownthr. 8380	Cape 7264	Lss.Mask. 6997	Golden 6922	Yellow 4119	Forest 3480	Thickb. 2299	Spect. 1322
W5	Redhead. 9872	Masked 7855	Lss.Mask. 7439	Brownthr. 6907	Cape 5899	Golden 4881	Yellow 2234	Forest 1983	Thickb. 866	Spect. 584
W10	Redhead. 9916	Masked 8487	Lss.Mask. 7590	Brownthr. 7526	Cape 6124	Golden 5013	Yellow 2439	Forest 2254	Thickb. 955	Spect. 643
SQRTW10	Redhead. 9858	Masked 8894	Brownthr. 7867	Lss.Mask. 7220	Cape 6655	Golden 5900	Yellow 3177	Forest 2782	Thickb. 1483	Spect. 900

Table 5 Dissimilarities of some weaver distributions to the distribution of the **Masked Weaver**. Dissimilarities are sorted in descending order, so that the most similar distributions are on the right of the table. The vertical line is a guide, separating distributions that are considered different from distributions that are similar to that of the Masked Weaver, in this case none are similar. See also Fig. 5.

W10	Redhead. 9984	Brownthr. 9739	Golden 9544	Lss.Mskd 9538	Yellow 9325	Forest 9148	Thickb. 8816	Spottedb. 8487	Spect. 8348	Cape 5002
W5	Redhead. 9965	Brownthr. 9432	Lss.Mskd 9374	Golden 9230	Yellow 8834	Forest 8606	Thickb. 8215	Spottedb. 7855	Spect. 7756	Cape 4377
UW10	Redhead. 9994	Brownthr. 9957	Golden 9929	Yellow 9870	Lss.Mskd 9862	Forest 9795	Thickb. 9656	Spect. 9329	Spottedb. 9288	Cape 5698
UW5	Redhead. 9988	Brownthr. 9916	Golden 9893	Lss.Mskd 9820	Yellow 9785	Forest 9676	Thickb. 9514	Spect. 9175	Spottedb. 9119	Cape 5406
UW1	Redhead. 9953	Brownthr. 9694	Golden 9605	Lss.Mskd 9471	Yellow 9217	Forest 8785	Thickb. 8618	Spottedb. 7973	Spect. 7915	Cape 3667
Jaccard	Redhead. 9977	Brownthr. 9846	Golden 9799	Lss.Mskd 9729	Yellow 9594	Forest 9354	Thickb. 9258	Spottedb. 8873	Spect. 8837	Cape 5367
SQRTW10	Redhead. 9990	Brownthr. 9889	Golden 9806	Lss.Mskd 9735	Yellow 9676	Forest 9548	Thickb. 9330	Spottedb. 8894	Spect. 8875	Cape 5197

Table 6 The dissimilarities between distributions of the **Collared Sunbird** and other sunbirds. The dissimilarities are sorted in descending order, so that the most similar distributions to that of the Collared Sunbird are on the right of the table. See also Fig. 6.

UW1	Dusky 9898	Orangebr. 9256	Neergaard 9190	Cape Sug. 9022	Marico 8476	Malachite 8029	Gurn.Sugar. 7433	Lss.Dbl.Coll 7021	Purpleb. 6568	Whitebell. 5608	Gr.Dbl.Coll 5048	Scarletch. 4841	Black 4332	Olive 3333	Grey 1903
UW10	Dusky 9984	Orangebr. 9854	Cape Sug. 9785	Neergaard 9538	Malachite 9337	Marico 9122	Lss.Dbl.Coll 8047	Gurn.Sugar. 7959	Purpleb. 6808	Gr.Dbl.Coll 5775	Whitebell. 5431	Scarletch. 5303	Black 4438	Olive 2160	Grey 1681
SQRTW10	Dusky 9967	Orangebr. 9805	Cape Sug. 9686	Neergaard 9581	Malachite 8987	Marico 8902	Gurn.Sugar. 8105	Lss.Dbl.Coll 7564	Purpleb. 6075	Gr.Dbl.Coll 5685	Scarletch. 4468	Whitebell. 4286	Black 3888	Olive 1751	Grey 1030
W5	Dusky 9857	Orangebr. 9658	Cape Sug. 9502	Neergaard 9071	Marico 8825	Malachite 7996	Gurn.Sugar. 7911	Lss.Dbl.Coll 6433	Purpleb. 5427	Gr.Dbl.Coll 4883	Scarletch. 4120	Whitebell. 3548	Black 3026	Olive 1636	Grey 717
W10	Dusky 9931	Orangebr. 9777	Neergaard 9706	Cape Sug. 9637	Marico 8932	Malachite 8521	Gurn.Sugar. 8243	Lss.Dbl.Coll 6923	Purpleb. 5982	Gr.Dbl.Coll 5261	Scarletch. 4291	Whitebell. 3390	Black 3106	Olive 1526	Grey 739

Table 7 Dissimilarities of distributions of sunbirds to the distribution of the **Scarletchested Sunbird**. The dissimilarities are sorted in descending order. See also Fig. 7. The vertical line provides a guide, separating distributions that are similar from distributions that are considered differerent.

UW1	Malachite 9313	Lss.Dbl.Coll 9026	Neergaard's 8709	Gurn.Sugar. 7808	Gr.Dbl.Coll 7711	Marico 6923	Black 6582	Grey 5119	Collared 4841	Purpleband 3870	Whitebellied 3774	Olive 3370
UW10	Malachite 9799	Lss.Dbl.Coll 9754	Neergaard's 9647	Gurn.Sugar. 9456	Gr.Dbl.Coll 8822	Marico 7394	Black 7382	Grey 5800	Collared 5303	Olive 4705	Purpleband 4247	Whitebellied 3464
SQRTW10	Malachite 9745	Lss.Dbl.Coll 9697	Neergaard's 9555	Gurn.Sugar. 9264	Gr.Dbl.Coll 8890	Black 7330	Marico 6921	Grey 5051	Collared 4468	Olive 3776	Whitebellied 3462	Purpleband 3335
w5	Malachite 9448	Lss.Dbl.Coll 9361	Gr.Dbl.Coll 8585	Gurn.Sugar. 8509	Neergaard's 8020	Black 6889	Marico 6511	Grey 4688	Collared 4120	Whitebellied 3318	Olive 2997	Purpleband 2285
w10	Malachite 9652	Lss.Dbl.Coll 9607	Neergaard's 9514	Gurn.Sugar. 9022	Gr.Dbl.Coll 8790	Black 7237	Marico 6629	Grey 4826	Collared 4291	Whitebellied 3671	Olive 3374	Purpleband 2828

Table 8 Dissimilarities between the distributions of Spottedbacked Weaver and the distribution of the Yellow Weaver (first column) and the weaver that is ranked as fifth most similar to the Spottedbacked Weaver (second column).

Coefficient	YELLOW WEAVER	GOLDEN / CAPE WEAVER
JACCARD	5957	7816
UW1	4241	6413
UW5	3933	6931
UW10	4119	6922
W5	2234	4881
W10	2439	5013

Table 9 Components of the dissimilarity between Protea Canary and Cape Siskin. Dissimilarity $D_{jk} = E + F + G + H$, defined in equations 13 and 14. This particular dissimilarity was found using Coefficient W10.

COMPONENT		PERCENTAGE
Siskin = 0, Protea > 0	E = 33	0.87
Siskin > 0, Protea = 0	F = 2612	68.92
Siskin > Protea	G = 797	21.03
Siskin < Protea	H = 348	9.18
Total:	D_{jk} = 3790	100 %

CAPE SISKIN

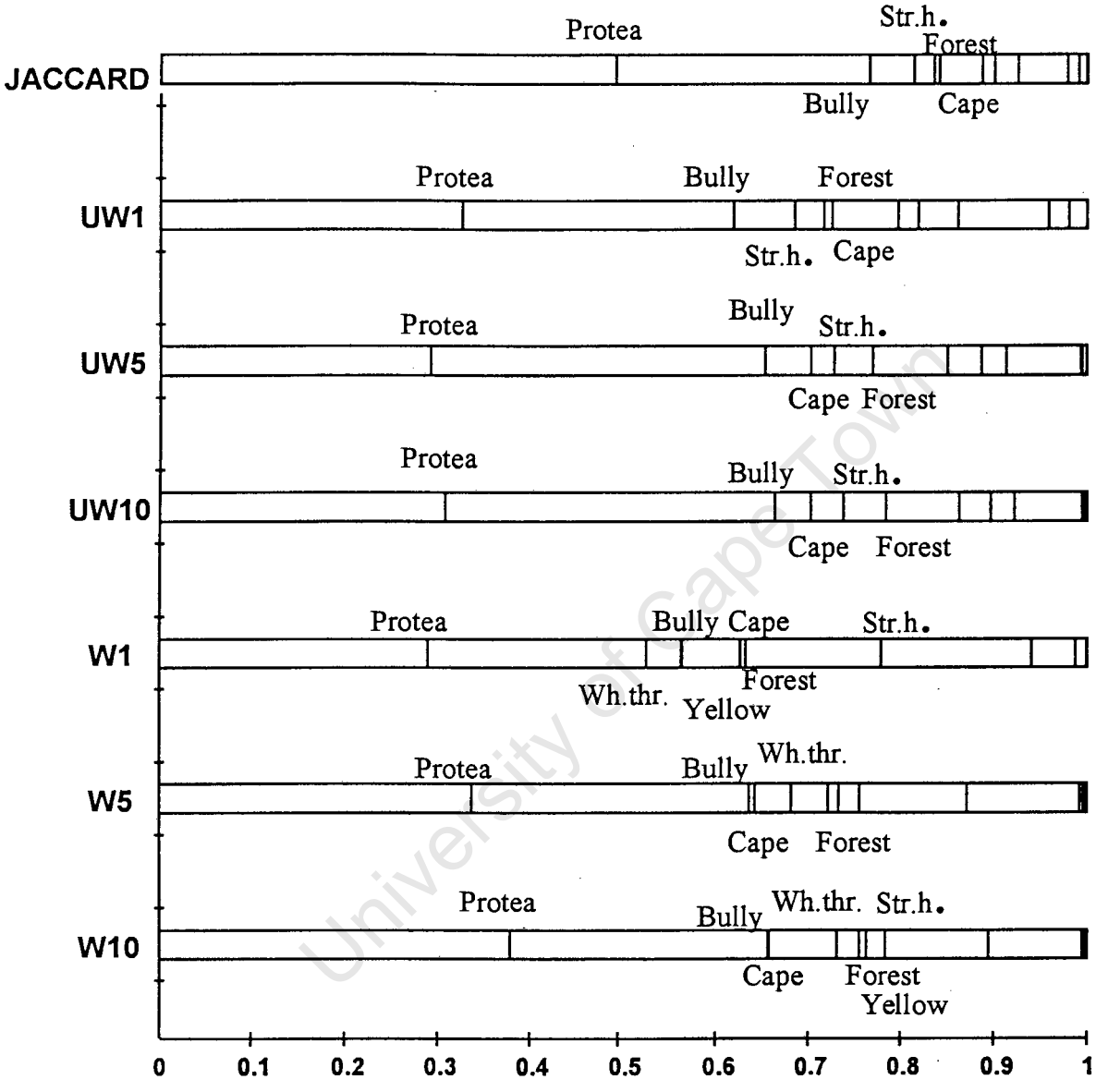


Fig. 1. Distance diagram illustrating the dissimilarities between distributions of the Cape Siskin and other canaries. The dissimilarities were calculated from the dissimilarity coefficient, shown on the left. The species with the most similar distribution to that of the Cape Siskin are closest to the left. The distance scale is shown at the bottom.

YELLOW CANARY

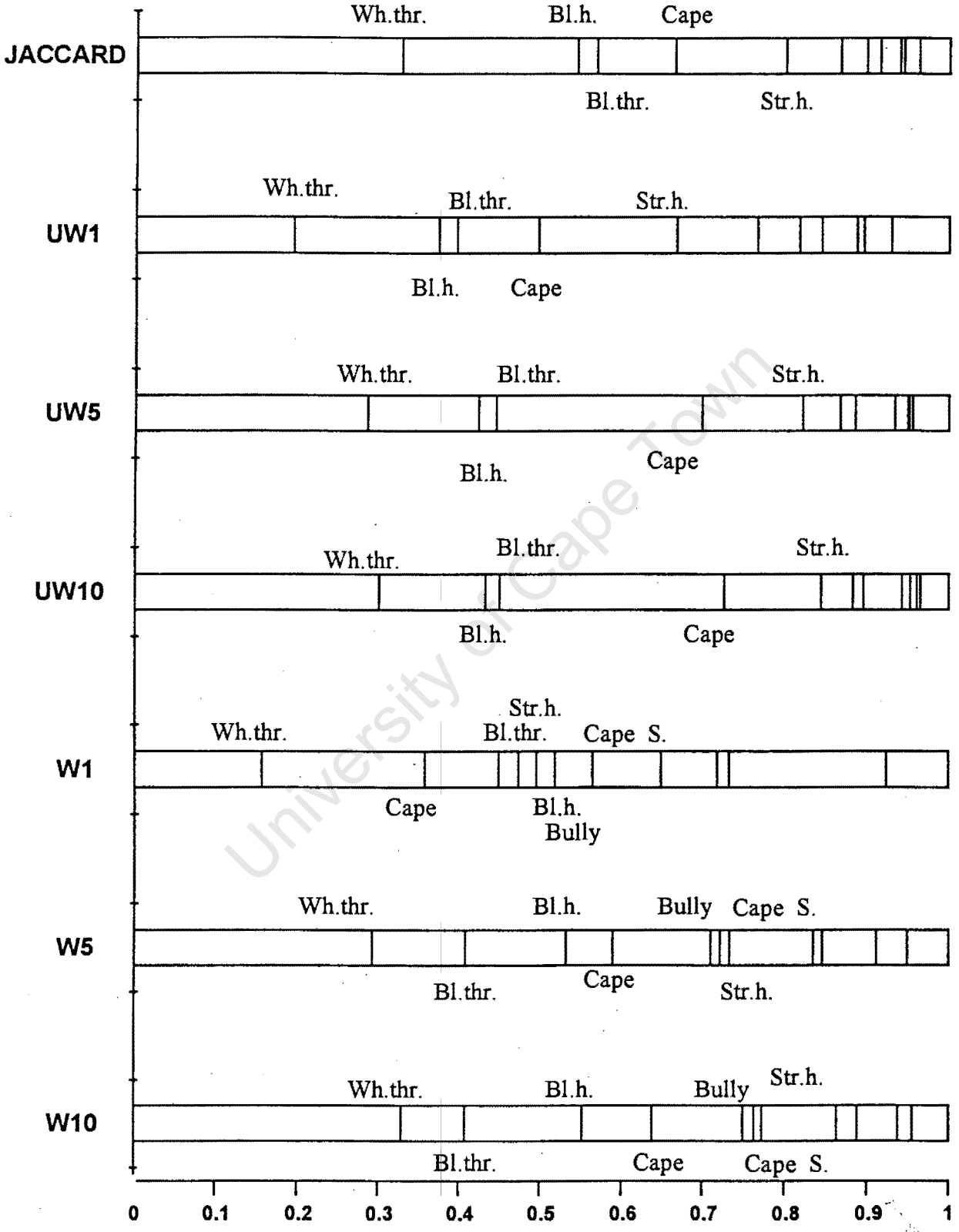


Fig. 2. Distance diagram for the dissimilarities between the distributions of the Yellow Canary and other canaries. The distance scale is shown at the bottom. The species with the most similar distribution to that of the Yellow Canary are closest to the left (smallest distance).

YELLOWEYED CANARY

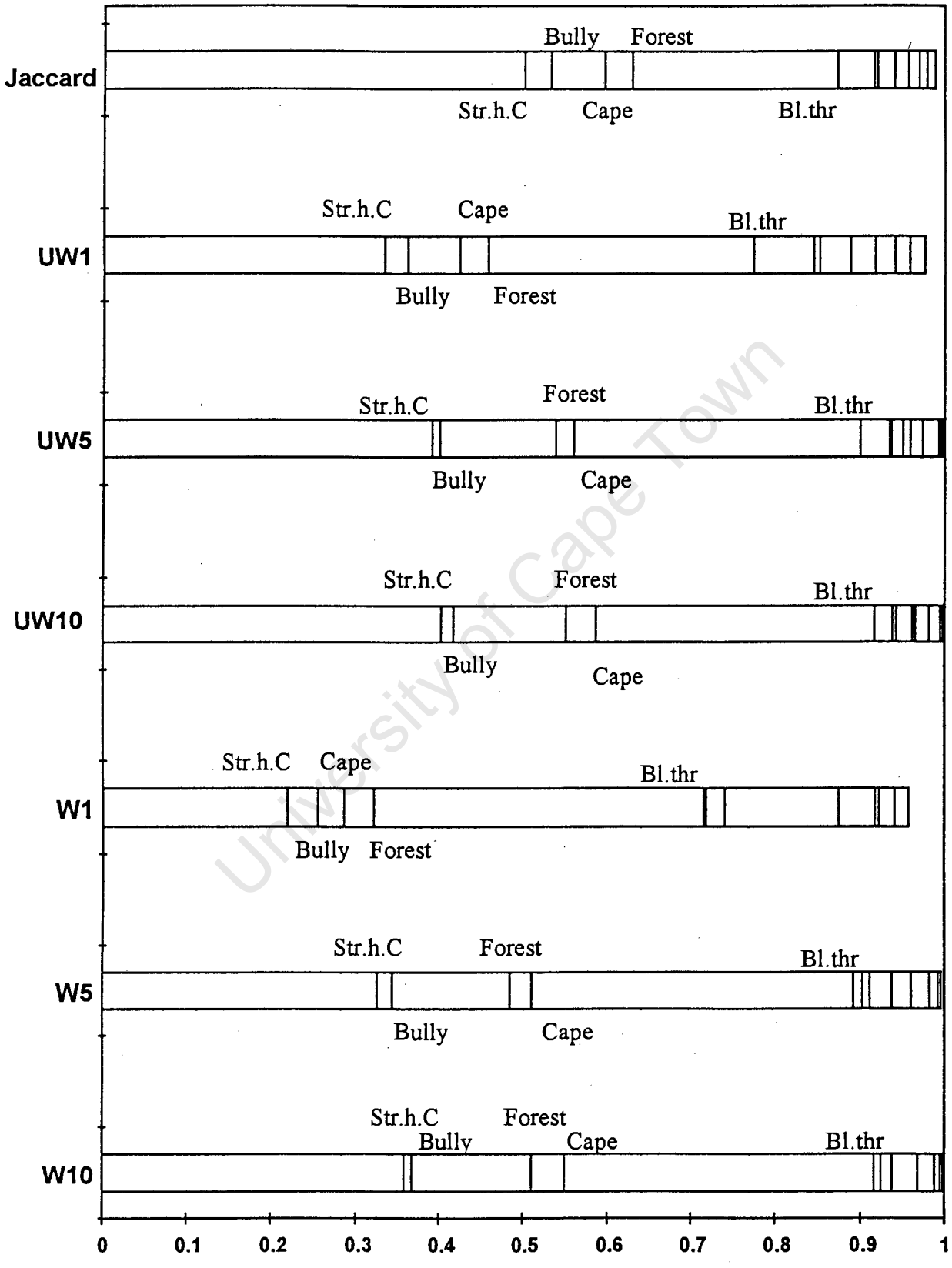


Fig. 3. Distance diagram for dissimilarities between the distributions of the Yelloweyed Canary and other canaries. The coefficient names are shown on the left.

SPOTTEDBACKED WEAVER

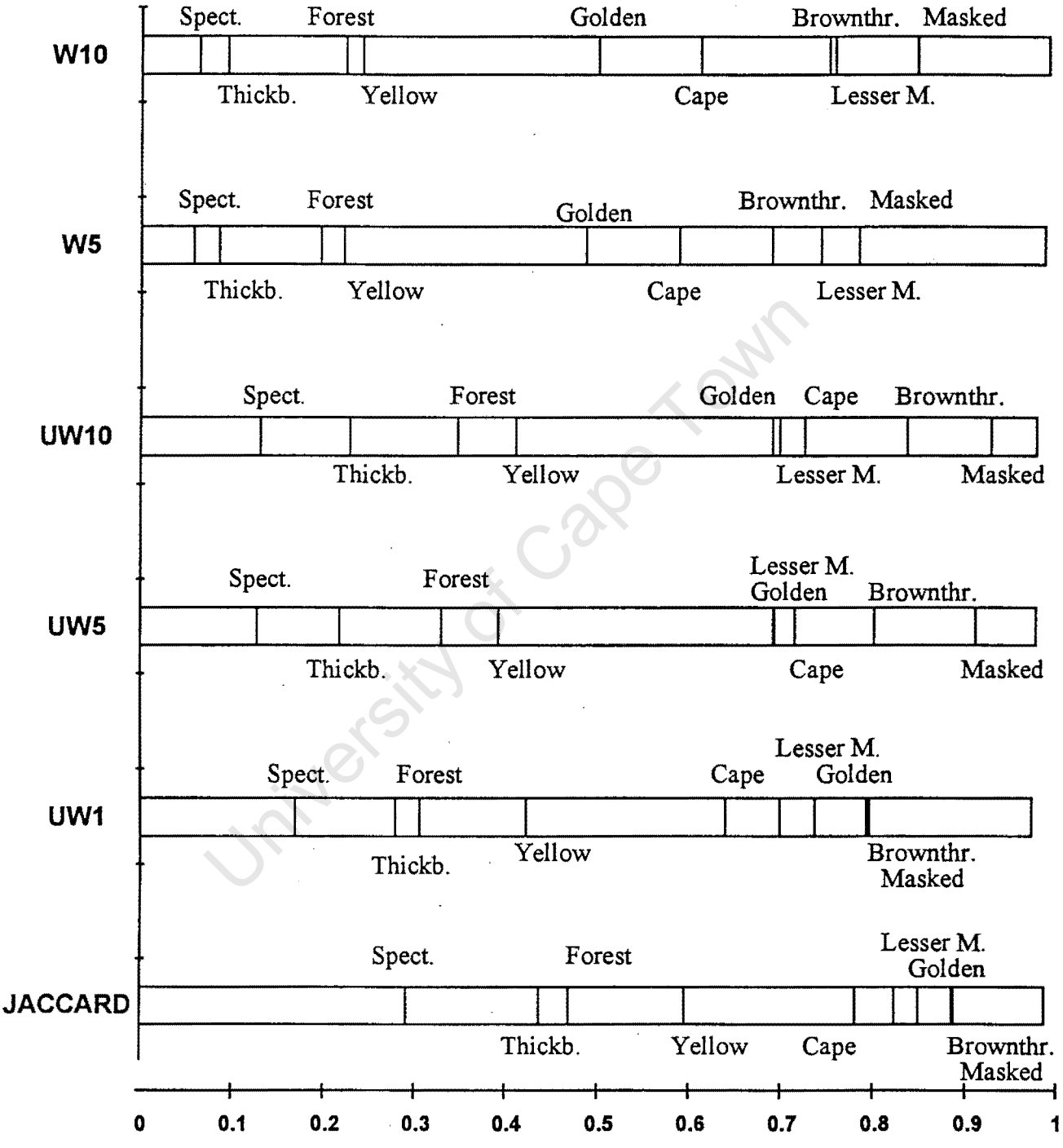


Fig. 4 Distance diagram for the dissimilarities between distributions of the Spottedbacked Weaver and other weavers. The dissimilarities were calculated from the coefficients shown on the left.

MASKED WEAVER

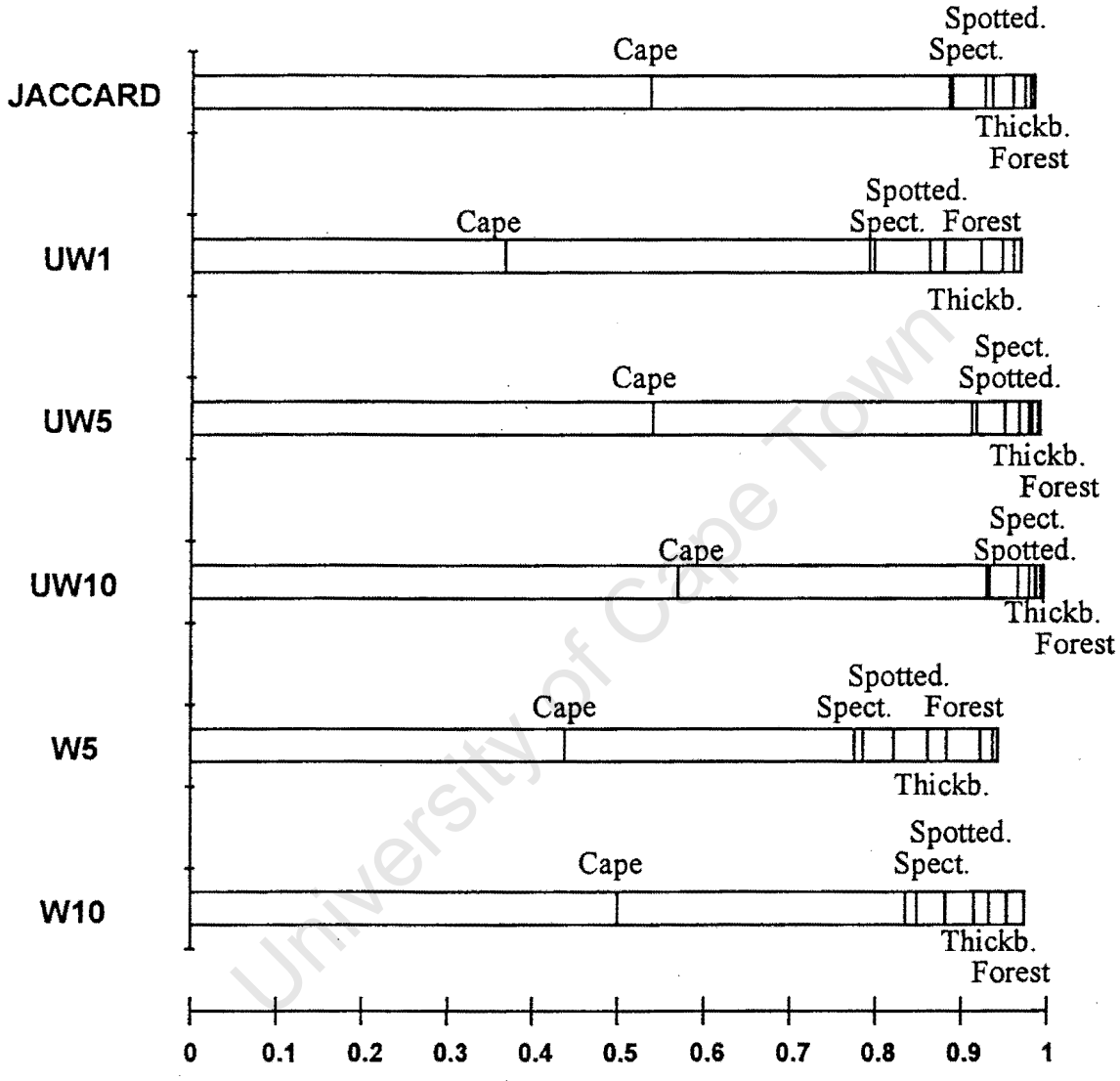


Fig. 5 Distance diagram for dissimilarities between the distributions of the Masked Weaver and other weavers. The dissimilarities were calculated with the coefficients on the left. The distance scale is shown at the bottom. The species with the most similar distribution to that of the Masked Weaver are closest to the left of the bars.

COLLARED SUNBIRD

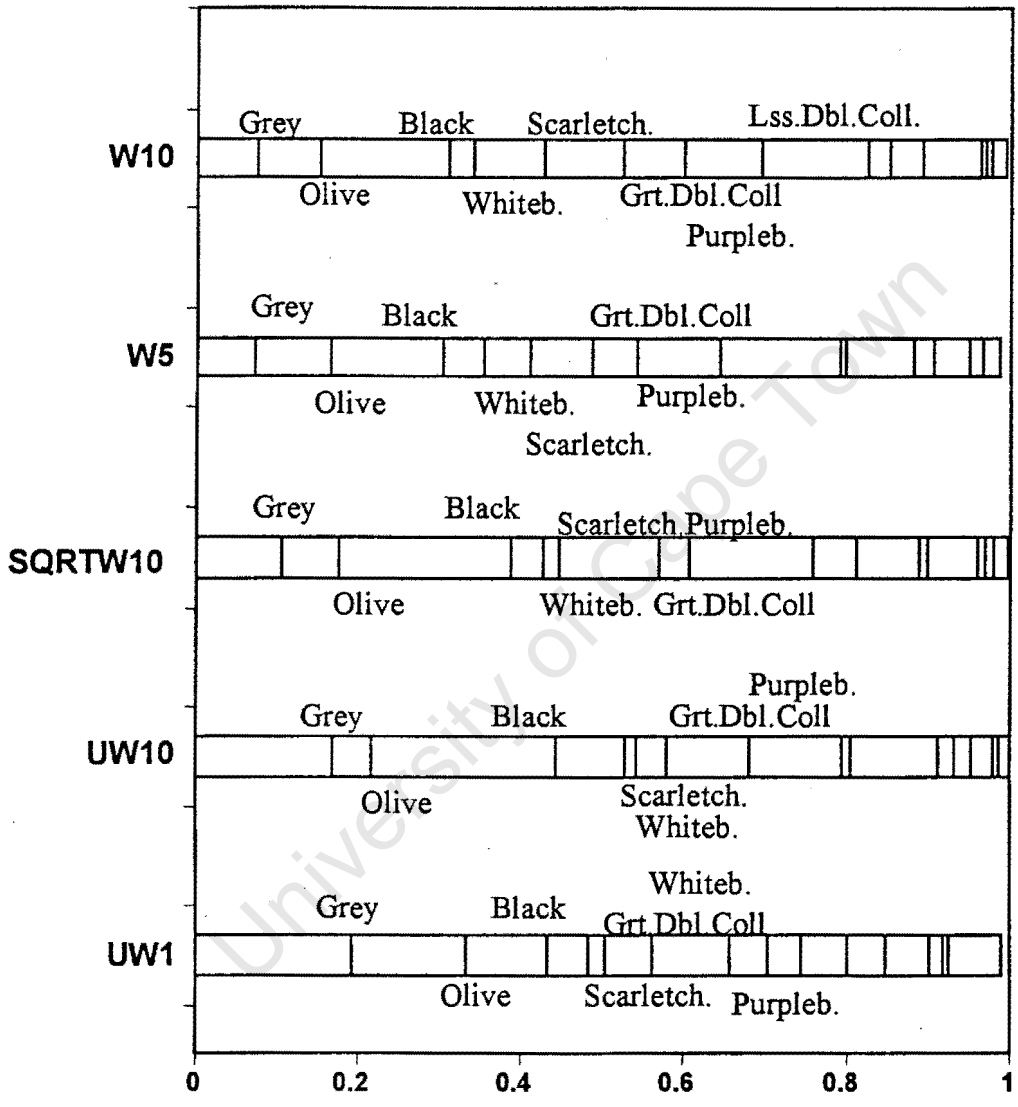


Fig. 6 Distance diagram for the dissimilarities between distributions of the Collared Sunbird and other sunbirds. The coefficients which calculated the dissimilarities are shown on the left.

SCARLETCHESTED SUNBIRD

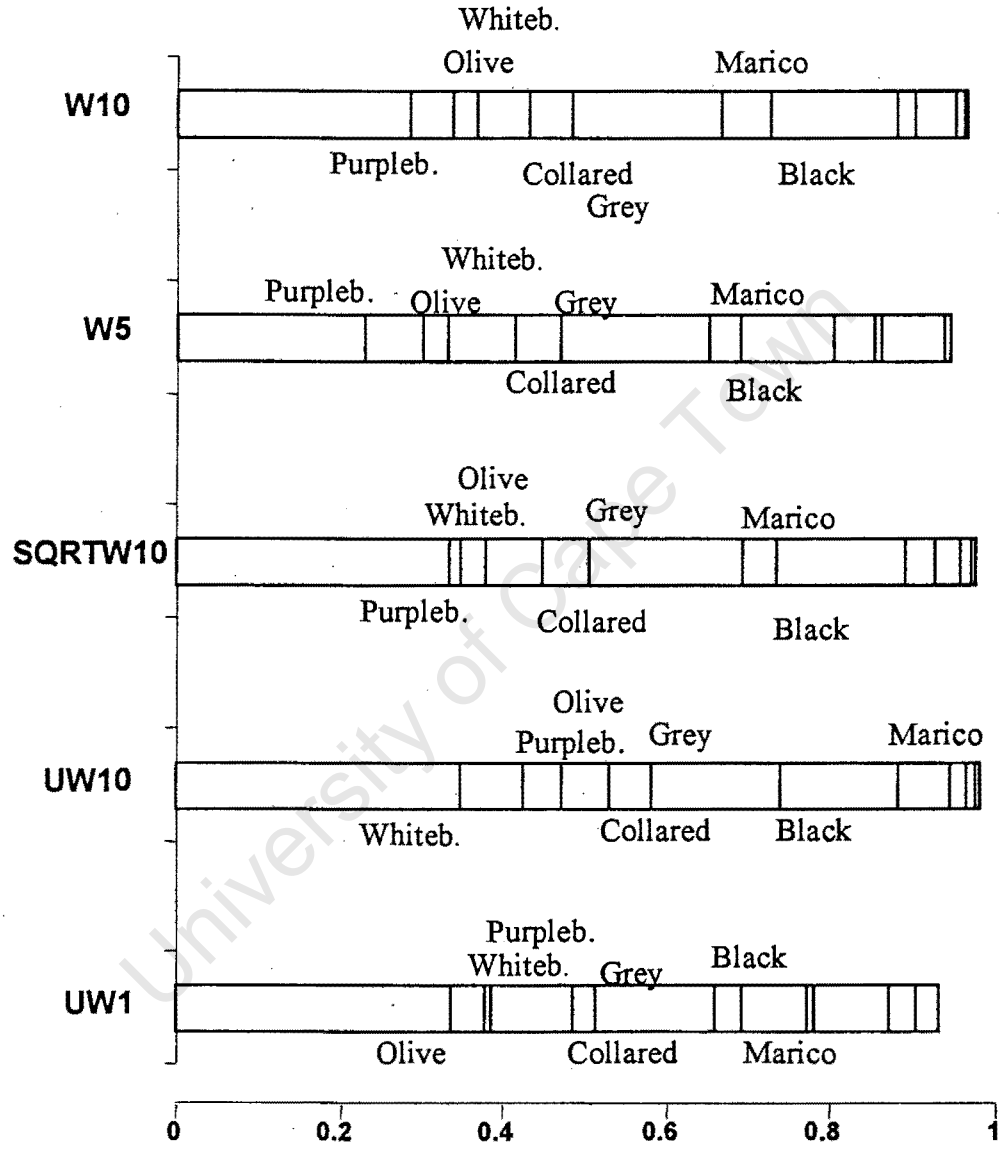


Fig. 7. Distance diagram for the dissimilarities between distributions of the Scarletchested Sunbird and other sunbirds. The coefficients which calculated the respective distances are shown on the left. The species with the most similar distribution to that of the Scarletchested Sunbird is closest to the left in each bar.

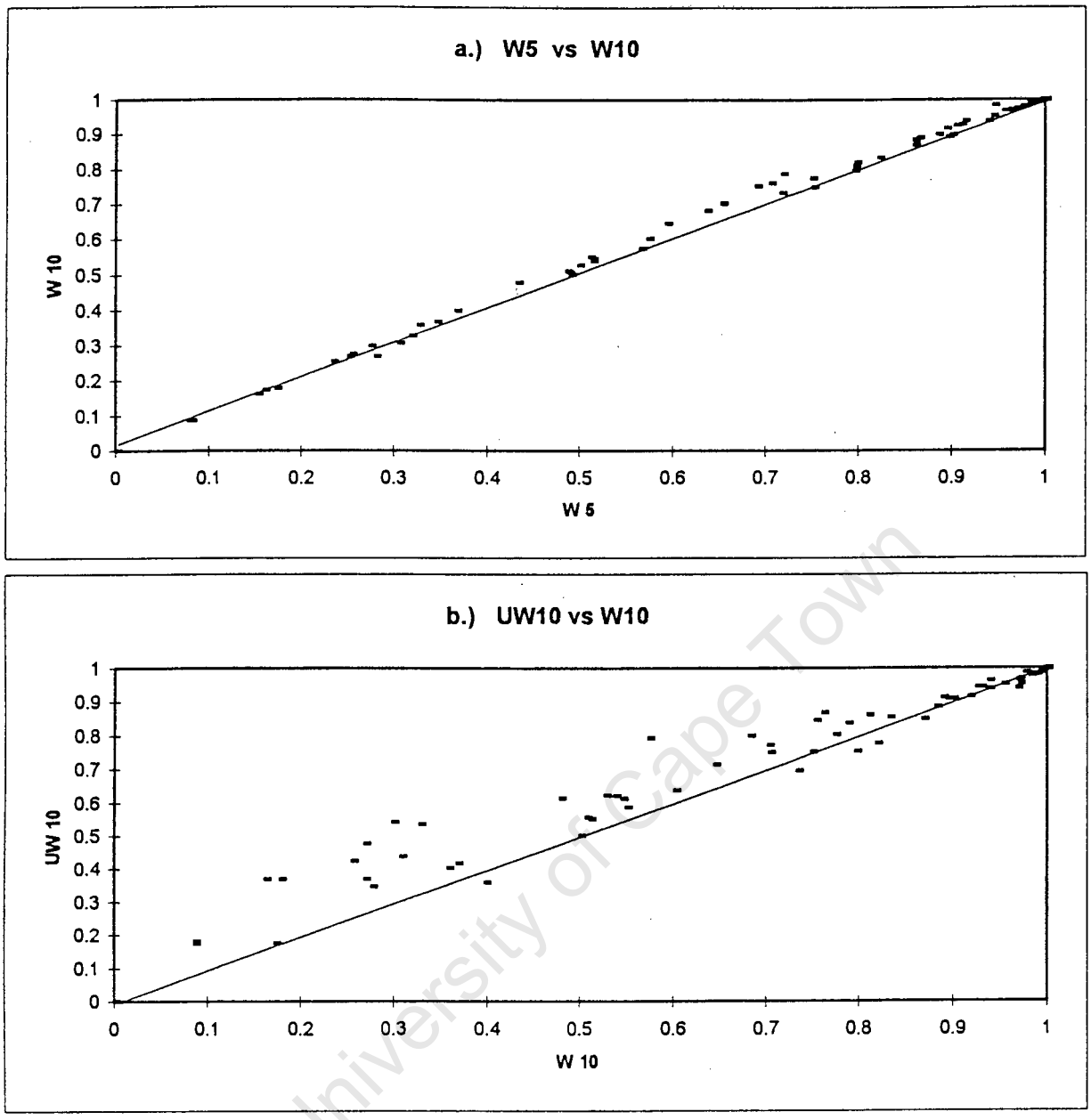


Figure 8 Scatterplots for comparing the dissimilarities calculated by different coefficients. The coefficients being compared here are the Coefficients W5 and W10 which use as weights the number of checklists in the grid cell and sort the positive reporting rates into five and 10 categories respectively (a). In (b) the Coefficients UW10 and W10 which both use ten categories for the positive reporting rates but for UW10 no weights are used and in W10 the weights are the number of checklists in the grid cell being compared.

CHAPTER 2

The Continuity of Bird Distributions

INTRODUCTION

The distributions in *The Atlas of Southern African Birds* (Harrison *et al.* 1997a, b) exhibit different degrees of continuity. Some distributions show almost continuous shading over their entire area of occurrence, for example, the distribution of the Masked Weaver in South Africa (Fig. A3, Appendix) while other distributions are patchy and fragmented, with many holes where the species seems to be absent from isolated grid cells, for example the Dusky Sunbird *Nectarinia bifasciata* (Fig. A28, App.).

How similar the reporting rates of neighbouring grid cells are, depends on how habitat specific the species is, characteristics of the species itself such as how easily it is identified, on the distribution of the habitat on which the species relies and also on the average number of checklists collected in the area. In areas with small numbers of checklists the observed reporting rates exhibit more variability.

In geostatistics the variogram has been developed to describe spatial correlation as this changes with distance between locations. The behaviour at the origin describes the continuity of the distribution at small lags (Cressie 1991). We aimed to use these first variogram values to assess the spatial continuity of a distribution map of the bird atlas and then to compare the degree of continuity between distributions of different species.

METHODS AND THEORY

Spatial data points in close vicinity to a given point are more likely to have more similar values than points which are further away. This feature is called **spatial autocorrelation**. The way in which this is expressed in maps of the data values is that similar values occur in clusters. This means that the observations are not independent. The assumption of independence is an important condition in many statistical approaches to modelling data for prediction and smoothing. Analysis of spatial data is similar to time series analysis, however the additional dimension (for the case of two-dimensional data) introduces new complications of estimating trend and the spatial correlation. Spatial data therefore requires a different approach to the construction of models.

The branch of statistics that has developed from this problem setting is sometimes referred to as 'geostatistics', as it often arises in geographical or geological (mining) contexts (Cressie 1991). Geostatistics provides methods to incorporate both spatial trend and spatial correlation into the modelling process.

The aim in this chapter is to estimate this spatial dependence of reporting rates in the distributions of *The Atlas of Southern African Birds* (Harrison *et al.* 1997a, b). It is not only of interest how fast this dependence in a distribution degenerates with distance but also how the correlation between neighbouring grid cells compares to that of distributions of other species. The latter problem is that of evaluating the continuity of a distribution. The main aim in this chapter will be to find a measure which estimates this continuity, so that different bird distributions can be compared with respect to their continuity of observed reporting rates.

The theory given below was mostly taken from Cressie (1991).

SPATIAL RANDOM PROCESSES

Let $Z(\mathbf{s})$ denote a spatial random process which has been observed at locations $\{ s_1, s_2, ..., s_n \}$. In our case $s_i = (x_i, y_i)$, a two dimensional location variable. If the observations are on a regular grid, such as the grid of the bird atlas, each observation can be written as $s_{k\ell}$, where for the bird atlas data this means the observed reporting rate in the kth row and in the ℓth column.

A value for location s_o is predicted in the following way

$$p(s_o) = \sum_{i=1}^{n} \lambda_i Z(s_i) \qquad (1)$$

where the value at s_o is either of interest because it was not observed or a smoothed value at this location is required. The λ_i's are the weights to determine how much

contribution each of the other observations should make to the prediction of the value at location s_o. Sample values should have a decreasing amount of influence on the prediction the farther they are away from the location for which the value is to be found. The reason for this is, that spatial dependence decreases with increasing distance between any two locations.

The covariance between two locations expresses the degree to which observed values of these locations will differ. In the case of spatial data usually only a single observation has been made at each location, which makes it impossible to estimate the covariance between the values. Instead it is assumed that the covariance of observations depends only on the distance between the locations at which they are taken, in other words it is assumed that the spatial process is stationary.

Rather than using the covariance function to model this spatial dependence, the variogram is used. The variogram requires less strict conditions of stationarity (not second order stationarity but only intrinsic stationarity) and therefore exists in some situations for which no covariance function can be found. The variogram estimates the variance of the differences between observed reporting rates in grid cells a distance *h* apart.

VARIOGRAMS

The term 'variogram' was introduced by Matheron (1963).

The spatial autocorrelation between two locations depends only on their distance apart and not on their exact locations. Expressed more formally, the following two conditions have to hold:

- $E [Z(s + h) - Z(s)] = 0$ (2)
- $Var [Z(s + h) - Z(s)] = 2\gamma(h)$ (3)

where h is the physical distance between locations s and $(s + h)$. Equation 2 means that the mean, $E(s)$, of the observations must be constant over all locations but is not necessarily known, i.e. the data contains no trend. The variance of the difference between observed values only depends on the distance between the respective locations. The quantity $2\gamma(h)$ is called the **variogram** and $\gamma(h)$ is called the semi-variogram. The two above conditions (2 and 3) ensure that the spatial process is **intrinsically stationary**, which is a condition for the prediction equations. If trend is present in the data, methods such as median-polish (Cressie 1991) exist to remove the trend.

The value of the variogram which is approached as $h \rightarrow \infty$ is called the **sill** and should be an asymptote if the data is stationary. The sill of the semi-variogram equals the maximum variation between reporting rates, σ^2. The **range** is the maximum distance at which the process exhibits spatial correlation.

The variogram values of real interest are those at small lags, for our case distances of less than 100 km. Variogram values for lags larger than this distance only indicate that large clusters of similar values are present.

The process of incorporating spatial correlation into prediction was first called 'kriging' by Matheron (1963 in Cressie 1991). For the simplest case, **ordinary kriging**, the situation is as follows:

The process can be modelled by

$$Z(s) = \mu + \delta(s)$$

where μ is unknown but constant and $\delta(s)$ is an intrinsically stationary random process with zero mean.

The variogram for process $\delta(s)$ is $2\gamma(h)$ and is estimated from the data.

The kriging equations are

$$p(\mathbf{Z}) = \sum_{i=1}^{n} \lambda_i Z(s_i) \qquad \text{such that} \quad \sum_{i=1}^{n} \lambda_i = 1.$$

where λ_i is the weight assigned to observation $Z(s_i)$. The aim is to find those weights which minimize the sum of the squared differences between observed and predicted values. In other words, the ordinary kriging predictor minimizes the mean-squared prediction error (Cressie 1991, p.120):

$$\sigma_e^2 = E\left[Z(s_o) - \sum_{i=1}^{n} \lambda_i Z(s_i) \right]^2 = E\left[Z(s_o) - p(Z(s_o)) \right]^2$$

where $Z(s_o)$ are the observed values.

The term inside the brackets can be expressed in terms of the semi-variogram:

$$\left(Z(s_o) - \sum_{i=1}^{n} \lambda_i Z(s_i) \right)^2 = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j (Z(s_i) - Z(s_j))^2 + \sum_{i=1}^{n} \lambda_i (Z(s_o) - Z(s_i))^2$$

$$= -\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j) + 2 \sum_{i=1}^{n} \lambda_i \gamma(s_i - s_j)$$

In matrix notation the weights are found from the variogram in the following way

$$\lambda = \mathbf{C}\, \mathbf{c}$$

where $C_{ij} = \gamma(s_i - s_j)$ and $c_i = \gamma(s_o - s_i)$

This illustrates how the semi-variogram is used in the prediction equations. We will not go further than this here, because we are only interested in the variogram itself

and not its function in the prediction. The remaining theory can be found in Cressie (1991). Our main focus is the difference between the variograms of different species.

VARIOGRAM COMPONENTS

The classical way of estimating the variogram

$$2\gamma(s_1 - s_2) = \mathrm{Var}\,[\,Z(s_1) - Z(s_2)\,] \tag{4}$$

from the observed data is the method-of-moments estimator

$$2\,\hat{\gamma}\,(h) = \frac{1}{|\,N\,(h)\,|} \sum_{N(h)} \left(\,Z\,(s_i\,) - Z\,(s_j)\,\right)^2 \tag{5}$$

where the sum is taken over all distinct pairs of data values s_i and s_j which are a distant h apart.

If the covariance between observations only depends on the absolute distance between their locations and not on the direction, only a single variogram for each data set is required. This assumption can be made for most of the extensive distributions in the bird atlas, but may not hold for smaller distributions, see the discussion for the Grey Sunbird (see also Fig. A22, App.).

It is common to assume that for variogram $2\gamma(h)$ the following is true:

$$E\left[\,Y\,(s+h) - Y\,(s)\right]^2 \to 0 \quad \text{as } h \to 0$$

This means that as the distance between points approaches zero, there will be no difference between the two values at these locations.

NUGGET EFFECT

If the variogram does not equal zero at lag zero, i.e.

$$\gamma\,(\,h\,) \to c_o \quad \text{as } h \to 0$$

c_o is called the nugget effect. This discontinuity at lag zero has two causes. Microscale variation is caused by variation at lags smaller than the minimum distance between points. These are the nuggets referred to and in the case of the bird distributions may originate from strong environmental changes within a single grid cell, so that within a grid cell the probability of observing a species is not constant. The other contribution made to the value c_o is measurement error. Measurement error is frequently assumed to be absent in existing applications. However in the case of binomial data such as in the bird atlas the binomial sampling variance cannot be ignored. The binomial variance depends on the underlying probabilities of occurrence

of the bird species and more importantly on the number of checklists that have been collected for the two grid cells.

The nugget effect c_o can be split up into two components

$$c_0 = c_{MS} + c_{ME}$$

where c_{ME} is variance due to measurement error and c_{MS} is the variance of the microscale process. In the case of the bird atlas data it is assumed that both microscale variation and measurement error are present. The problem is that the two components are not separable because they are observed as only one value c_o. The only possibility is that the measurement error may be estimated. In the bird atlas data the measurement error can be assumed to be the usual binomial variance.

THE BIRD ATLAS DISTRIBUTION AND BINOMIAL OBSERVATIONS

Let X_i denote the number of successes (records of the species) out of the n_i checklists collected for grid cell *i*. $R_i = X_i / n_i$ is the observed reporting rate in grid cell *i*, the observed proportion of successes out of the n_i checklists. X_i here is taken to have a binomial distribution $B(n_i, \pi_i)$, where π_i is the true underlying probability of observing the species in grid cell *i*. Xi has the following mean and variance

$$E(X_i) = n_i \pi_i$$

$$Var(X_i) = ni\ \pi_i (1 - \pi_i)$$

From this follows that the mean and variance of the reporting rate R_i is

$$E\left(\frac{X_i}{n_i}\right) = \pi_i \qquad (6)$$

$$Var\left(\frac{X_i}{n_i}\right) = \frac{\pi_i (1 - \pi_i)}{n_i} \qquad (7)$$

McNeill (1991, 1994) derives a different set of means and variances assuming that the distribution of X_i, conditional on π_i, is binomial with parameters n_i and π_i. If $i \neq j$ then X_i and X_j are conditionally independent given π_i and π_j (McNeill 1991, p.132). The problem introduced by the data having a binomial distribution, is that measurement error is present, and that this is not constant over all observations but depends on the underlying probability of success and on the number of checklists collected for the grid cells.

McNeill (1991, 1994) and Cressie (1991) suggested that the random process should be split up into the following components:

$$Z(s) = \mu(s) + \tau(s) + \eta(s) + \varepsilon(s)$$

where

- $\mu(s)$ is the mean structure, including the trend (large-scale variation)
- $\tau(s)$ is an intrinsically stationary process (small-scale variation). The spatial dependence of this process should be captured in the variogram.
- $\eta(s)$ is the micro-scale variation
- $\varepsilon(s)$ is the measurement error

If all the above processes act independently, the respective variances of these components are also additive (Cressie 1991, McNeill 1994). If values are to be predicted, the components are estimated as one unit, so that it is not necessary to estimate each of these components separately. However the error component should be excluded from the predictions, because the smoothing process aims to exclude errors.

THE PROBLEM OF TREND

Cressie (1991, p.74) warned that if the trend is not constant, $2\hat{\gamma}(h)$, the method-of-moments estimator, is a poor estimate for the variogram and should not be used until the data is detrended. This assumption is particularly important if the kriging predictor is to be used and if a global approach is taken. If in contrast a local window for the kriging is to be used, the variogram values at small lags are not as strongly influenced by trend (McNeill 1994). The bird atlas data are not stationary, trend is present. Trend distorts the estimation of the variation. The effect of trend on the estimated variogram is that it may not reach a maximum (McNeill 1994). Cressie (1991) suggested using the **median-polish kriging** method if trend is present. This method estimates column, row and overall effects first, detrends the data and then uses the residuals to find the variogram. This approach was taken with the Sudden-Infant-Death-Syndrome data.

Problems with this method are that it assumes symmetric data. The bird atlas data is not symmetric because it contains many small reporting rates. McNeill developed a different approach (1991 and 1994). She did not predetermine any trend function but incorporated trend into the kriging equations so that they essentially only smooth the data by removing measurement error, but the trend does not explicitly have to be estimated and separated from the data. The justification given for this was that because local kriging is used (McNeill used a 7x7 window) trend at small lags does not influence these values to such a large extent. Secondly McNeill remarked that, by removing trend, the variation in the variogram often underestimates the true variation (1994, p. 90).

BIAS

McNeill showed (1991, 1994) that the error due to binomial sampling variance contributing to the estimation of the variogram is as follows:

$$\sum_{r=i,j} \frac{\mu(1-\mu)}{n_r} - \sum_{r=i,j} \frac{\sigma^2}{n_r} \tag{8}$$

where μ and σ^2 are the mean and variance of the underlying π_i's, the average probability of encountering the species in grid cell *i*. McNeill (1991, p. 135 & 1994) suggested to remove the bias in the estimation of each variogram term as follows

$$2\tilde{\gamma}(h) = \left(\sum_{i,j}^{N(h)} \left((R_i - R_j)^2 - \sum_{r=i,j} \frac{\hat{\mu}(1-\hat{\mu})}{n_r} + \sum_{r=i,j} \frac{\hat{\sigma}^2}{n_r} \right) \right) / N(h) \tag{9}$$

and called this the 'modified' variogram. McNeill used as estimate for the mean and variance

$$\hat{\mu} = \frac{\sum_{x>0} n_i R_i}{\sum_{x>0} n_i} = \frac{\sum_{x>0} X_i}{\sum_{x>0} n_i} \tag{10}$$

and

$$\hat{\sigma}2 = \frac{\left(\sum (R_i - \hat{\mu})^2\right) - \left(\hat{\mu}(1-\hat{\mu}) \sum \frac{1}{n_i}\right)}{N - \sum \frac{1}{n_i}} \tag{11}$$

where the summations are taken over all *N* grid cells where the observed reporting rates R_i were larger than zero. This estimator aims to remove the binomial measurement error from the estimation of the variogram.

We will not use the variogram for predictions, therefore it may not be necessary to remove the measurement error in the estimation of the variogram. Then the variogram will not be an indication of the continuity of the true underlying probabilities but of the variability of the observed reporting rates.

A disadvantage of removing the bias from each term is that estimates for the overall mean and variance of π_i are required. For this, these values have to be assumed to be constant over the entire area. McNeill remarked (1991) that this may not be a reasonable assumption even if trend is removed because different areas of the country will have different mean reporting rates especially if the distribution covers an extensive area. McNeill suggested that therefore it may be better to use local estimates of μ and σ^2. Unremoved trend would not allow good estimates of μ and σ^2 from the data.

ESTIMATING THE CONTINUITY OF GEOGRAPHICAL DISTRIBUTIONS

The variogram represents the spatial dependence of the data. Values near the origin (at smallest lags) can therefore be used as indicators of the continuity of the random process, (Cressie 1991, p. 60). The magnitude of $2\gamma(1)$, the variogram value at lag one, when compared to the other values of the same variogram, is an indication of how much spatial correlation is present in the distribution. Our main intention in this chapter is to find such a measure of continuity, which allows us to compare the continuity between distributions.

The magnitudes of the variogram values however depend on the magnitude of the average reporting rate of the particular distribution. The variogram values are therefore not comparable between distributions.

This is the same problem as when comparing sample standard deviations between populations. A standard deviation of 10 has a different meaning if the mean of the observations is 100 or 1000. The standard deviation is a better indication of the variability of values if it is compared to the mean, for example as in the coefficient of variation ($CV = \sigma / \mu$). The coefficient of variation expresses the standard deviation as a percentage of the mean.

SCALING FACTOR

A scaling factor for each variogram should be used so that the variograms become comparable between species and are not a function of the mean reporting rate anymore. This scaling factor should transform the variogram values so that they become comparable values expressing the degree of continuity and ranging between a minimum and a maximum value. These limits were here taken to be ideally equal to zero and one. A value of one would then mean that the species has reached the maximal possible variation between reporting rates at a certain lag, meaning that there is no correlation between these values.

The unscaled variogram only gives an indication of relative spatial dependence at different lags, it is never an indication of absolute spatial correlation. The correlation of reporting rates in adjacent grid cells, $\gamma(1)$, can only be compared to the overall variation in the distribution, $\gamma(\infty)$.

Variogram values of zero at lag one are unlikely because measurement error and microscale variation cause a nugget effect (see earlier) and both of these are present in the bird atlas data. Scaled variograms will not always reach a maximum value of one. This may occur when the ratio of the overall variation of the underlying reporting rates, σ^2, to the mean is smaller than that for other species. It may also be possible that a species has too large a variation of observed reporting rates, so that the scaled variogram maximum value will exceed one.

The average reporting rate of a distribution has an influence on the size of the differences between reporting rates. If the mean reporting rate is small (e.g. 0.3 or 0.2

and less), then large differences such as 0.9 - 0.1 are less likely. Whereas with average reporting rates of 0.5 and larger, increasingly larger differences between reporting rates in different grid cells are possible. The average reporting rate does not influence the continuity of a distribution, although there may be some relationship. The chosen scaling factor should conserve the differentiation between differing continuities but remove other effects which influence the magnitudes of the variogram values.

The variogram is formed from squared differences. This suggests that the scaling factor should also be a squared value. Two options are the squared mean observed reporting rate and the average squared reporting rate. Because the numerator is a sum of squares, this suggests that the denominator should be the average squared reporting rate. Because the scaling factors we will be considering are all constants, they will preserve the exact forms of the variograms and only influence the range, maximum to minimum. Using the average of the squared reporting rates for scaling, the variogram becomes

$$2\gamma(h) = \frac{\left(\sum_{i,j}^{N(h)} (R_i - R_j)^2 \right) / N(h)}{\frac{1}{n} \sum_{k=1}^{n} (R_k)^2} \qquad (12)$$

where R_i is the observed reporting rate in grid cell *i*, and N(h) is the number of distinct pairs of grid cells that are a distant *h* apart. The denominator is the average squared observed reporting rate and is calculated only over those R_k which are strictly positive ($R_k > 0$). Grid cells outside of the distributions should not have an influence on the average reporting rate. This value may overestimate the true squared probability for distributions with many zero reporting rate grid cells within the distribution, because true zeroes are left out of the calculation but this effect should not be serious as these effects also increase the value of the numerator.

The following equation is a variogram estimator in which the binomial measurement error is removed and which is scaled by the mean of the squared reporting rates:

$$2\gamma(h) = \frac{\left(\sum_{i,j}^{N(h)} \left((R_i - R_j)^2 - \sum_{r=i,j} \frac{\hat{\mu}(1-\hat{\mu})}{n_r} + \sum_{r=i,j} \frac{\hat{\sigma}^2}{n_r} \right) \right) / N(h)}{\frac{1}{n} \sum_{k=1}^{n} (R_k)^2} \qquad (13)$$

The estimate of the variance itself, σ^2, should not be used to scale the variogram. The sample variance is partly what we want to estimate. Scaling by this value would reduce all sills to equal one but this would destroy the comparison of different variances relative to the mean between distributions.

WEIGHTS

If for a grid cell only few checklists have been collected (less than five) the observed reporting rate is unreliable (see eq. 7). In the estimator of the variogram the sum components for the numerator are the squared differences of two observed reporting rates from grid cells a distance *h* apart. If for one of these two grid cells the number of checklists was small, then not only the corresponding reporting rate is unreliable but also the calculated difference will be unreliable. It does not much improve the situation if a value is subtracted from this trying to reduce the estimated variance. If the difference is not correct in the first place than a subtraction of estimated error does not improve it.

An alternative option to estimate the variogram if the removal of bias (eq. 9) is not satisfactory, is to let reliable differences contribute more to the estimation than unreliable differences. This will not remove the bias but may reduce it.

For each sum component there are two values observed in different grid cells with different numbers of checklists. These are two possible weights, which of the two n_i should be used? One of the reporting rates may be reliable, the other not. In this case the calculated difference between the reporting rates is also unreliable. Therefore the weight given to a calculated difference should only be large if both reporting rates are reliable estimates of the true probabilities. The accuracy of the answer depends on the smaller number of checklists. If the smaller of the two n_i is taken to weight the variogram components, the variogram estimator becomes

$$2\gamma(h) = \frac{\left(\sum_{i,j}^{N(h)} \min(n_i, n_j)(R_i - R_j)^2 \right) / \left(\sum_{i,j}^{N(h)} \min(n_i, n_j) \right)}{\frac{1}{n} \sum_{k=1}^{n} (R_k)^2} \qquad (14)$$

where $\{\min(n_i, n_j)\}$ is the weight assigned to each squared difference of reporting rates and is equal to the smaller of the two numbers of checklists collected for the two grid cells. The *n* in the denominator is only increased when R_k is strictly larger than zero. An alternative is to use the square roots of the n_i as weights. This was considered because the range of the number of checklists is large (in our subset of the data from zero to 1260). For large numbers of checklists (larger than 200) the improvement of the estimation of the true probability does not increase at a level which concerns us.

$$2\gamma(h) = \frac{\left(\sum_{i,j}^{N(h)} \min(\sqrt{n_i}, \sqrt{n_j})(R_i - R_j)^2 \right) / \left(\sum_{i,j}^{N(h)} \min(\sqrt{n_i}, \sqrt{n_j}) \right)}{\frac{1}{n} \sum_{k=1}^{n} (R_k)^2} \qquad (15)$$

This system of calculating the variogram also has the advantage of considering the reliability of the original values. The removal of bias method (eq. 9) ignores this. The squared difference in reporting rates may be very large compared to the estimated bias and then will still have a large effect on the estimation of the variogram.

Reporting rates originating from grid cells with single checklists are more harmful in the estimation than useful. Any difference calculated including such a reporting rate will be wrong, if it is not a true zero.

RESULTS

COMPARISON OF VARIOGRAM METHODS

In Figs 1, 2 and 3 the different methods for estimating the variogram are illustrated for eight selected bird species. The variogram values at large lags (distances larger than 300 km) were included to illustrate the behaviour of variograms at large lags for different species. These variogram values are not meaningful for species where the entire 'width' of their distribution does not cover this distance. The values that are of particular interest to us are the first three or four but especially the very first variogram value. This first value represents the correlation of reporting rates in immediately neighbouring grid cells. The different variograms presented for each of the species were calculated using different scaling factors and different weights.

The variogram of the Blackheaded Canary (Fig. 3c) does not reach an upper limit even at a distance of more than 500 km. This is most likely caused by trend which was not removed from the data. The variograms of the Cape Weaver (Fig. 2a) seem to reach their sill at 500 km. In the case of the Protea Canary (Fig. 3d), the last values of the variogram seem to be larger than those below lags of 400 km. This is probably caused by only few values being available for the estimation of the variogram at these lags and may originate from differences between reporting rates where one of the grid cells falls into the distribution and the other is outside (zero reporting rate) of the distribution of the Protea Canary. The range, the maximum lag at which spatial dependence is present, should not be interpreted here. Firstly, it is not clear how much the trend contributes to the shape of the variogram and to the spatial correlation at large lags. Secondly, is it likely that the shape of the bird distribution (elongated or approximately round), has an effect on the variogram values at larger lags. This is because for elongated, narrow distributions more differences of reporting rates with empty grid cells are included. The maximum values reached by the variograms are therefore not further investigated here and also not the variogram values at larger lags than approximately 100 km.

The species for which the variograms reach a maximum after only a few lags are the Protea Canary at lag 2 (Fig. 3d), the Cape Siskin at lag 2 (Fig. 3b) and the Forest Canary somewhere between lag 4 and lag 6 (Fig. 3a). When visually inspecting the original atlas maps (Harrison *et al.* 1997a, b) corresponding to these species, all of these distributions show a high variability in neighbouring grid cells between observed reporting rates.

From these full variograms, which were calculated over the entire distributions of the species, it is difficult to establish which of the variogram methods gives better results. The variogram is a measure of the continuity of the maps and its values should reflect this, especially the initial values at lags one and two, which represent variability between neighbouring cells.

COMPARISON BETWEEN SPECIES

Fig. 4 shows the first variogram values, $2\gamma(1)$, for some sparrows, canaries and weavers. These variograms were generated using the distributions of the respective species south of 27° S. The species were chosen so that their distribution maps showed different variabilities between the observed reporting rates.

For pairs of species we tried to establish, by investigating the atlas maps, which of the two species has the more variable distribution and should therefore have a larger $2\gamma(1)$ value. In Table 1 these observations are summarised. The subjective opinion was formed from visual inspection of the atlas maps.

When comparing the group of Forest Canary, Protea Canary and the Cape Siskin, the $2\gamma(1)$ value of the Cape Siskin should be smaller than that of the Protea Canary, because the distribution of the Cape Siskin appears to be more continuous. This is the result for all methods except those that use no scaling factor (Bias unscaled and NoWeight unscaled). For these methods the variogram values increase as the average squared mean reporting rate increases, as was expected.

Of all the species considered here, the Cape Sparrow definitely has the smoothest map and should have a clear minimum variogram value. This is not the case for the methods which do not scale the variogram (Fig. 4). This illustrates that if the variogram values need to be comparative, a scaling factor, which removes the dependence of the variogram values on the average observed reporting rate, should be used.

It is difficult to decide visually which of the above three species (Cape Siskin, Forest Canary and Protea Canary) should have the maximum and which the minimum variogram value (Figs. A6, A7, A15, App.). This difficulty is caused by the different shapes and the different sizes of the distributions. The Cape Siskin occurs only in the South Western Cape, covering a small area. The Forest Canary is spread out along the coast. Firstly, its distribution has a larger total area. Secondly, it is narrow. Narrow distributions have a larger circumference than roughly spherical distributions. This may cause edges to make larger contributions to sums in the estimation of the variogram values.

Method W=n (each squared difference is weighted by the smaller number of checklists) produces a larger $2\gamma(1)$ value for the Forest Canary than for the distributions of the Protea Canary and the Cape Siskin (Fig. 4 and Table 1). The reason for this could be that in the areas where many checklists were collected the Forest Canary occurs inconsistently, especially along the southeastern coast (Fig. A15). That would mean that the large differences in this region are heavily weighted.

There is some difference in the magnitude of the estimated bias and therefore in how much the NoWeight variogram changes when bias is removed. For example in the case of the Blackheaded Canary the bias is 0.271 for the $2\gamma(1)$ value. For the House Sparrow the bias is only 0.132 (Table 3). The magnitude of the bias influences how the different methods will rank the $2\gamma(1)$ values.

The NoWeight method (unweighted, no bias removed) is the only method which calculates a lower variogram value for the House Sparrow than for the Blackheaded Canary (Fig. 4, Table 1, Table 3). But from subjective inspection of the maps, it is very difficult to say which of the two distributions is smoother.

The Cape Weaver has a smaller range than the Blackheaded Canary (Figs A1 & A11, App.). In those areas in which it occurs, it appears to be smoother. This is reflected by all methods, but the Bias method does not make a large distinction between the variogram values of these two species.

The rough relationship between the mean (squared) reporting rate and the magnitude of the variogram values is a decrease in $2\gamma(1)$ as the average squared mean increases (Fig.4). One species that forms a striking exception to this trend is the House Sparrow (801). Its variogram values are higher than expected.

COMPARISON OF VARIABILITY IN A FIXED AREA

It is difficult to compare distributions that cover different areas and also have different shapes and sizes, visually. It is easier to find expected rankings of variogram values and compare these with the observed rankings of variogram values when the area is of fixed size, so that impacts on the visual decision by size and shape of the distribution are excluded. We therefore selected a block of 16x16 = 256 grid cells (30° to 33° S, 20° to 23° E, inclusively). Some species whose distributions cover most of this block were chosen to compare the variogram methods (Table 4).

Again we firstly tried to rank pairs of species according to the magnitude of the $2\gamma(1)$ value that we would expect by visually inspecting the distribution maps. These ranks were then compared to the ranks produced by the various variogram methods. The results are summarized in Table 2, the first variogram values $2\gamma(1)$ are plotted in Fig. 5.

There appears to be a relationship between the variogram value $2\gamma(1)$ and the mean reporting rate of the species' distribution. As the mean increases, the variogram value $2\gamma(1)$ in general decreases. The four methods considered here approach each other in their results as the mean reporting rate increases. A striking exception is, as before, the House Sparrow. The House Sparrow, and to a lesser extent the Thickbilled Lark (512), have a greater variogram value between adjoining grid cells than would be expected from the usual relation to the mean, i.e. less correlation occurs than would be expected if the relation would hold.

The two weighted methods generally produce values that are smaller than the value from the NoWeight method. In the case of the Cape Canary and the Lesser Doublecollared Sunbird the weighted variogram values however are larger.

Of the three species Whitethroated Canary, Masked Weaver and Cape Sparrow, the map of the Whitethroated Canary appears to be the smoothest and the distribution definitely appears to be smoother than that of the Masked Weaver. Comparing the variogram values in Fig. 5, however, the value calculated for the Whitethroated

Canary is higher than even that of the Masked Weaver. Instead these values appear in inverse order of their mean reporting rates.

Does this suggest that the scaling factor that was used, the average squared mean, has too large an effect on the outcome of the variogram? Do larger means scale down the variogram values to smaller magnitudes? On the other hand, as discussed above, the results of at least the NoWeight method seem good. The exceptions, House Sparrow and Thickbilled Lark, also oppose this suggestion of the too strong effect of the scaling factor.

From these results it appears that the NoWeight method is as least as good as any of the other methods (Table 2), mostly better, as far as this can be judged from our way of comparing the distribution maps in the atlas.

The Grey Sunbird (Fig. A22) provides a good example of a species with a continuous distribution but where the form or shape of the distribution contributes much to an increased first variogram value. This species has a very narrow distribution along the east coast, only one or two grid cells wide. A large amount of the differences will be taken between a positive reporting rate and zero. The average reporting rate for this species was 15.5% (Harrison *et al.* 1997b). The magnitude of the difference is adjusted by the denominator but the many edge effects contributing to the estimation will cause the estimated variogram value to be larger than expected. This is an example of a species where the spatial correlation is not independent of the direction but is larger in the southwest, northeast direction than in the northwest, southeast direction. For such cases separate variograms are generally necessary, where the spatial correlation does not only depend on the absolute distance h between locations but also on the direction of the distance vector.

BIAS

Between species there is considerable difference in how far apart the results from the Bias and the NoWeight methods are. For this compare the Blackheaded Canary and the Forest Canary (Table 3). For the Blackheaded Canary the difference is 0.271, for the Forest Canary only 0.104 and for the Cape Sparrow the difference between the NoWeight and the Bias value is 0.044 (Table 3 and Fig. 4). This will make a difference in how the two methods rank the first values of the variograms. This provides one way to select the more correct method. In general the bias is constant over distance, but it will be of different magnitudes for different species.

Table 5 shows a list of biases calculated from a distribution with an average squared mean of 4384 (square root) and an estimated standard deviation of 5630, also see equations (8) & (9) for the estimation of the bias.

If in both grid cells the number of checklists was very large, more than 130 checklists as in Table 5 (g) and (h), the estimated bias is small. In (g) 0.0011 is more than 10 times smaller than the squared difference, in (h) the bias is 0.0009 which is 2.5 times smaller than the squared difference in reporting rates. In (f) the difference between reporting rates in the two grid cells was zero. The corresponding numbers of checklists were 12 and 6. The estimated bias was 0.0177. If this is subtracted from

the squared difference, the contribution of this comparison will be negative. The same problem occurs if both numbers of checklists were small, as in Table 5 (n) with four and five checklists for the grid cells. The difference in observed reporting rates is not large, 0.15, and the squared difference is 0.0225. The estimated value for the bias is 0.0318. Again the contribution of this case to the estimation of the variogram would be negative. The contribution of such cases, with both numbers of checklists small, to the variogram estimation should be small but it should not decrease the estimated values by subtracting from the total sum. If the observed reporting rate in one grid cell is considered reliable while the other reporting rate is unreliable such as in Table 5 (i), where the numbers of checklists were 131 and three, the difference between the observed reporting rates is unreliable. The observed squared difference here was 0.5943 and the estimated bias was only 0.0005. This bias for such an unreliable value is too small.

Tables 6 and 7 show the estimated first variogram values calculated by Method NoWeight for the available species. These values were calculated for the part of the distributions south of 27°S. The variogram values are sorted in ascending order, so that the most continuous distributions appear at the top of the table (smallest $2\gamma(1)$, Var(1)) and the most patchy or fragmented distributions at the end of the table, with the largest Var(1) values. In Table 6 the species are listed in the conventional order, by Roberts number. In Table 7 the variogram values are sorted by $2\gamma(1)$ values.

Tables 8 and 9 show the variogram values found by Method SQRTW, Table 8 in the conventional order, Table 9 in sorted order.

Fig. 6 shows a scatterplot of these first variogram values for the two methods NoWeight and SQRTW. All values calculated with Method SQRTW are less than one. The largest Var(1) value for this method was 0.94 for the Redbilled Francolin. The smallest variogram value was calculated for the distribution of the Masked Weaver, 0.1867. The variogram values for the SQRTW method are smaller than those calculated by Method NoWeight with a few exceptions. For species with very small distributions, where the species was only observed in a few isolated cells, the SQRTW method produces outliers. This is the case for the Blackrumped Buttonquail (206), the Dusky Lark (505) and to a lesser extent the Kurrichane Buttonquail (205) and the Monotonous Lark (493). These estimates for the variogram values were based on a small number of observations and are therefore not reliable. The configuration of the weights may be such that small differences are weighted stronger than the larger differences. The values calculated by Method NoWeight are more accurate for these particular distributions. Overall, however, these two methods produce values that are comparable.

Figs 7 (a) and (b) show the contributions to the variogram estimates made by differences between reporting rates. These figures were established using the NoWeight method. The corresponding $2\gamma(1)$ value for the House Sparrow was 0.467 and for the Cape Sparrow 0.09 (Table 3). For the species with the small variogram value, Table 3 (b) more than 25% of all observed differences between reporting rates were less than 0.05 and only few cases where the observed difference between reporting rates was larger than 0.4. For the species with the larger variogram value 12% of all differences were less than 0.05 but larger differences were observed more

frequently than for the species with the small variogram value. The variograms for these two species are shown in Fig. 9, calculated by Method NoWeight.

DISCUSSION

A problem with a single estimate of spatial correlation or a single variogram for each bird distribution, is that the distributions are not uniformly structured all over, but are often smooth in their cores, of which there may be more than one, and which may be small compared to the total area of the distribution. The distributions become more variable towards the edges and these variable parts may be a large or a small percentage of the total area, depending on how large the core of the distribution is. The variogram value can then only be an average estimate of spatial dependence, although in reality this varies much over the distribution.

The concern we had with the bias method developed by McNeill (1991, 1994) was that this approach to removing bias still accepts the absolute difference in reporting rates and aims to correct these by removing a bias term which only depends on the number of checklists observed in the grid cell but not on the observed reporting rate in the grid cell but only on an overall estimate of the mean and variance of π_i, the underlying probability of encountering the species in grid cell *i*. If the observed difference between reporting rates is zero, this method will still aim to remove bias and the contribution to the variogram sum will in such a case be negative. Weighting the contributions according to the smaller number of checklists collected for the two grid cells being compared, also only uses the number of checklists but this method does not have as many potential problems.

The way of comparing the variogram values that we have been using, by inspection of the maps and from that subjectively deciding on what the first variogram value should be, may not be entirely objective. However the size of the clusters in the distributions give a good impression of the overall variability. If clusters are not present, i.e. only isolated grid cells with records occur, this is distinguishable from clusters of very small size, three or four grid cells.

The concept of a probability π_i, of encountering the species, existing for each grid cell is very abstract. The reporting rate variogram is a biased estimator for the variogram of the π_i's because it does not only incorporate the true differences between values but these differences are influenced by the variation in the observed reporting rates. Therefore the variogram calculated from observed reporting rates overestimates the variance of the π_i's (McNeill 1991).

McNeill's method aimed to capture the spatial dependence of the π_i's in the variogram as opposed to the spatial dependence of the observed reporting rates. The π_i's will for almost all species be smoother than what has been observed in the reporting rates, but these are not directly observable.

It is not clear whether it is justified to 'decrease' each observed difference $(R_{ij} - R_{ik})^2$ by subtracting the estimated variance of the measurement error. The term $(R_{ij} - R_{ik})^2$ only becomes a variance term when summed over a large number of cases or when appearing in an expectation. It may be better to estimate a single bias value for each

lag of the variogram or alternatively to let the more reliable values have more influence in the estimation of the variogram.

If the variogram values are used for smoothing in a kriging predictor it may be better to predict a reporting rate for each grid cell than aiming to predict the true probability of encountering the species in the grid cell π_i. In this case one can also estimate the spatial dependence between observed reporting rates instead of between the probabilities of occurrence in grid cells a distance h apart.

We only need a comparative measure for the estimation of continuity between species. If the bias would be approximately constant over all species and if the bias removal would not do much difference in which distributions are considered more continuous and which more patchy, then for our intention this would not be necessary. The bias removal is only necessary when this measurement error is considered larger for some species than others and when this does not really show in the observations already.

The estimated measurement error will be smaller if the variance σ^2 of the species is larger (eq. 12). The estimation of the variance is however not good if trend is present in the data even if an underlying constant mean occurrence of the species is assumed, which is not a reasonable assumption for most bird distributions. The estimation of the measurement error variance is too heavily influenced by the estimates for μ and σ^2.

For the above reasons, the removal of bias method is not recommended. It depends too much on the assumption of an underlying constant mean and on the estimation of μ and σ^2, which are estimated over the entire distribution.

It is however still desirable not to let unreliable reporting rate differences influence the estimation too much. Therefore we chose to use a weighting system that weights each squared difference between reporting rates in two grid cells by the square root of the smaller number of checklists collected for the two grid cells.

The weighted variogram methods have the same purpose as the bias removal suggested by McNeill (1994). This is to reduce the effects produced by the binomial sampling error on the outcome. If we can assume that a large number of checklists produces a better estimate of the underlying true reporting rate, π_i , then these grid cells should contribute more to the estimate of the variability in reporting rates than grid cells with very few cells, where the observed reporting rates are not so reliable.

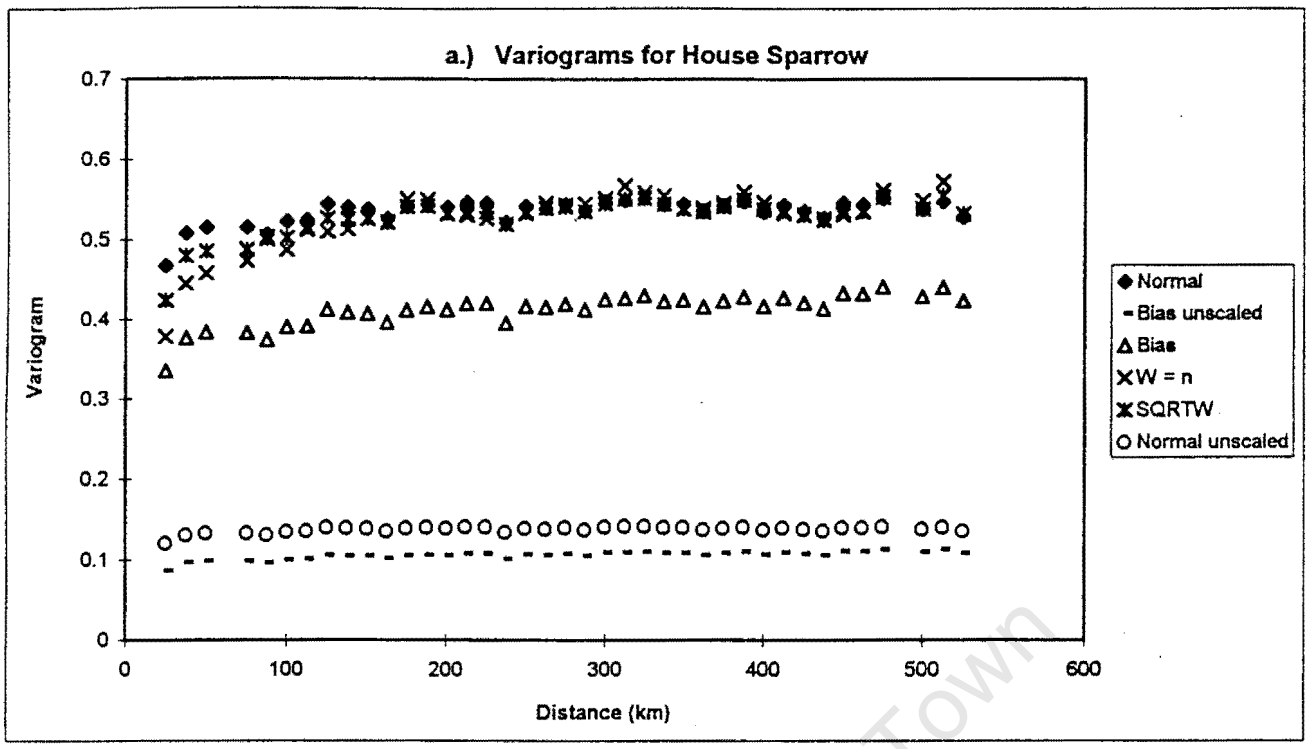
The texture seen on the atlas maps, is the variability in observed reporting rates. This includes the variability of spatial trend, the stationary autocorrelated spatial process and measurement error. For predictions a variogram without measurement errors is required because the true existing values are to be predicted using the true spatial correlation not influenced by errors. The true correlation between values at certain distances can only be estimated if measurement errors are removed.

The magnitude of the variability depends largely on the area in which the species occurs, the area influences how many checklists per grid cell were recorded. For an

example compare the variability of the distribution of the Whitethroated Canary (Fig. A9, App.) in Namibia with that along the west coast of South Africa. It is also dependent on the average commonness of the bird. If the average observed reporting rate is close to 0.5 the variability in the observed rates will be higher than those where the average reporting rate is near 0.1 or 0.9. This may on top still be increased by the errors caused by observers.

Method W = n may prove to be a problem for species that occur both in areas with few checklists (approximately 10 and less) and areas with many checklists (more than 50). If the species has the same variability over the whole area of its occurrence there is no problem but if this variability is different in different regions of the country, the area with more checklists is dominant in the estimation of the variogram. The number of checklists collected and the true variability of a species often depend on the same factors, such as characteristics of the environment.

We recommend using a weighted variogram, rather than removing the bias from single differences. Because the range of the number of checklists in the bird atlas data was large (zero to 1260) the square root of the number of checklists showed better results than using as weights the number of checklists.



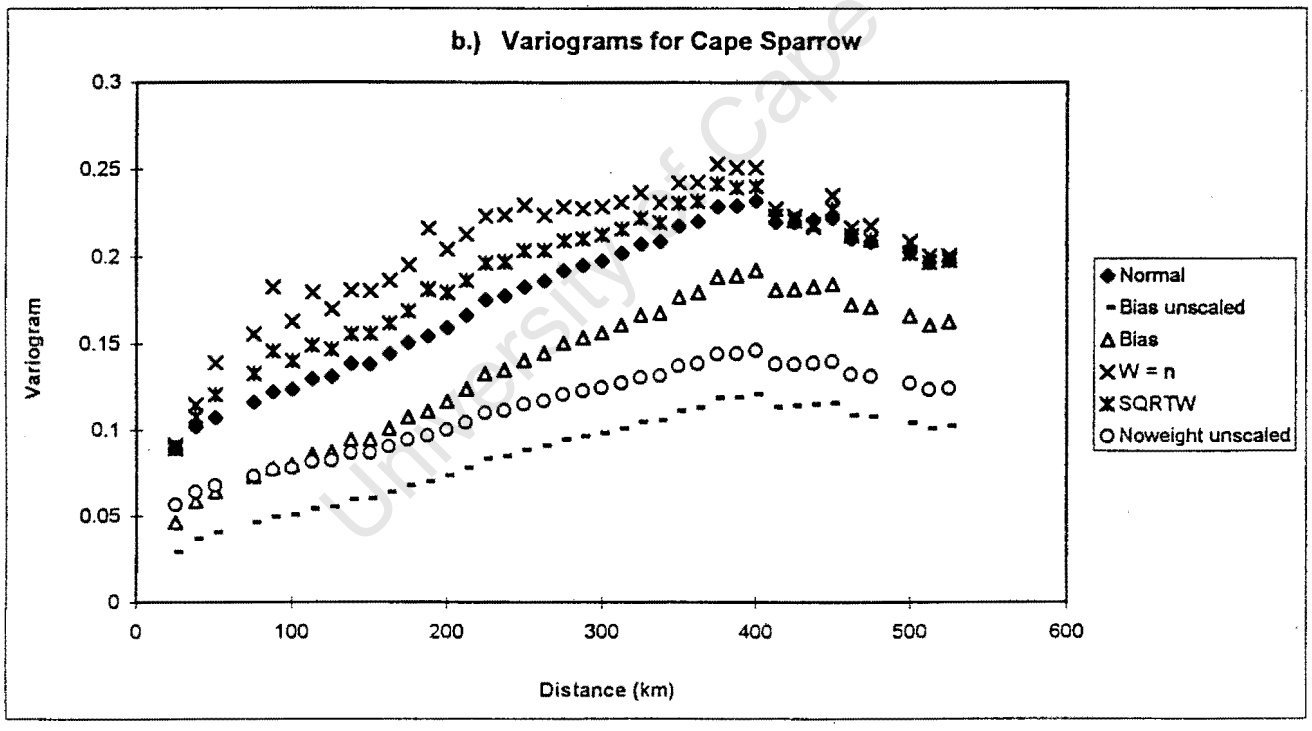
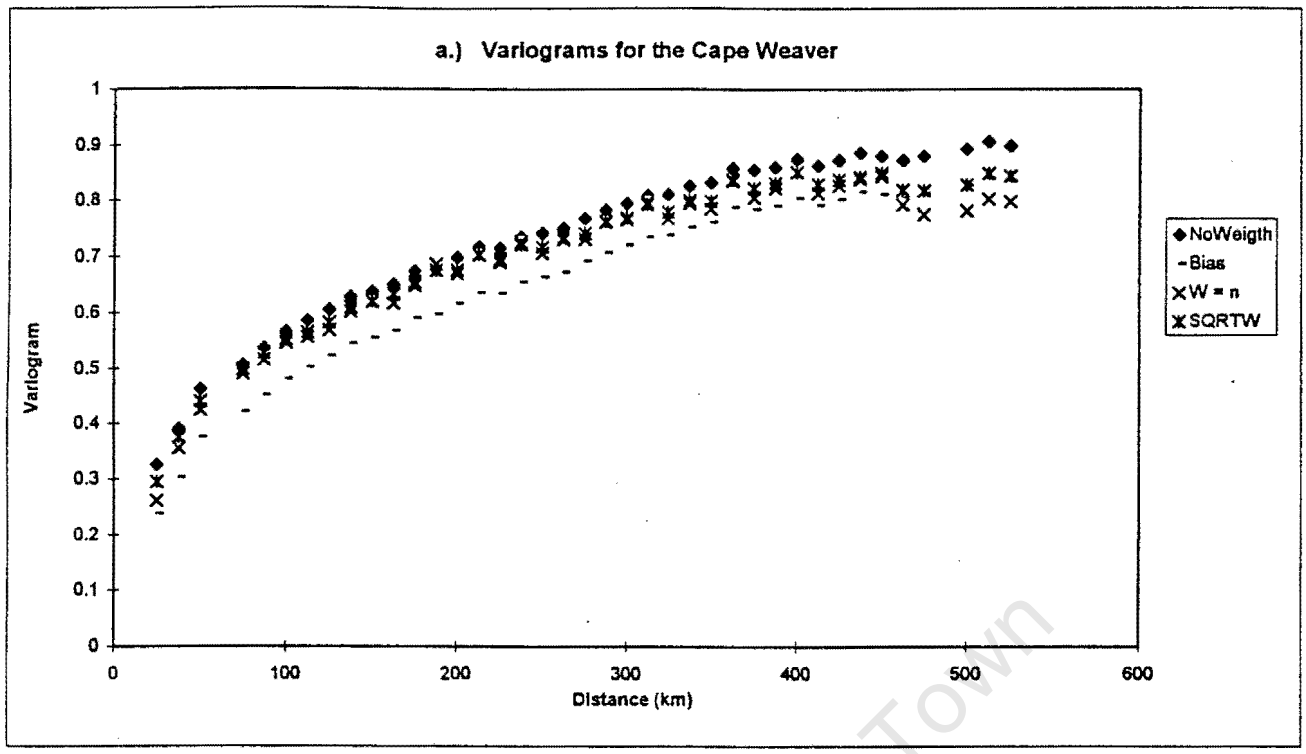


Figure 1 Variograms for the distributions of the (a) House and (b) Cape Sparrows. The methods by which these variograms were calculated are given in the legend.



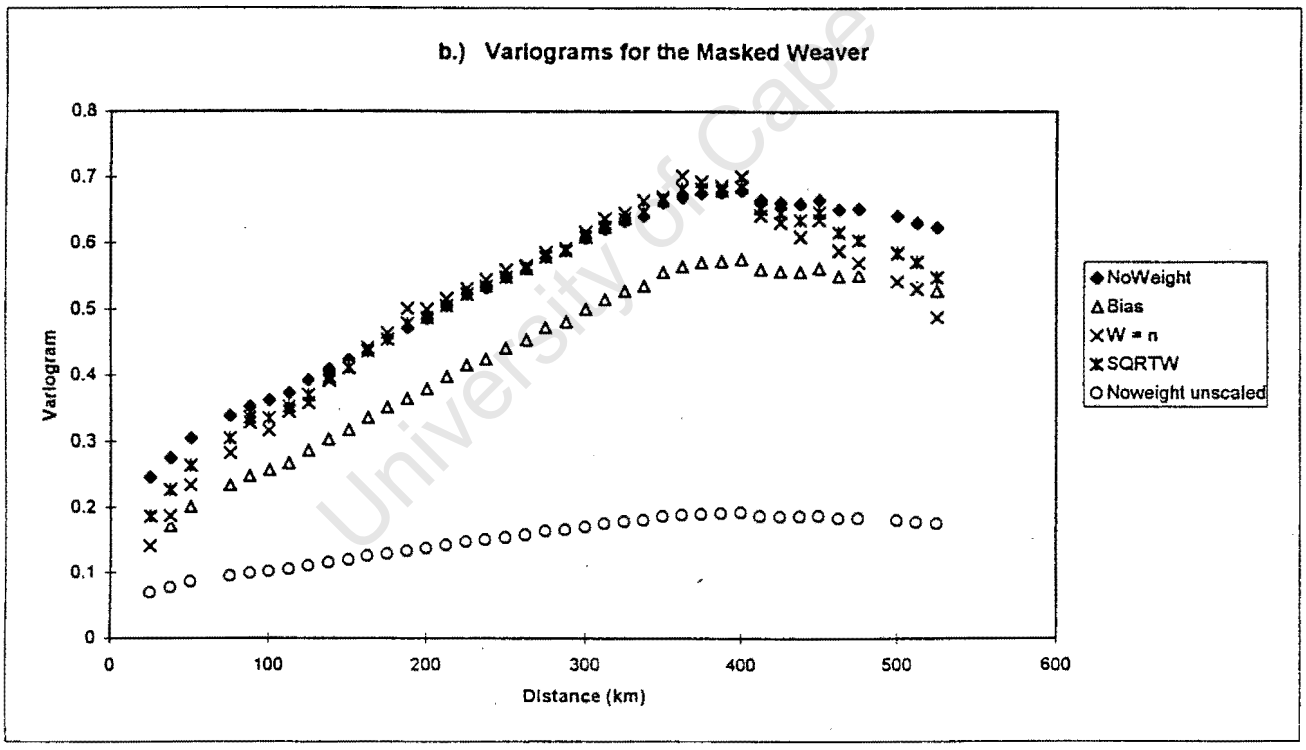


Figure 2 Variograms for (a) the Cape and (b) the Masked Weavers. Variograms calculated by different methods are shown and named in the legend.

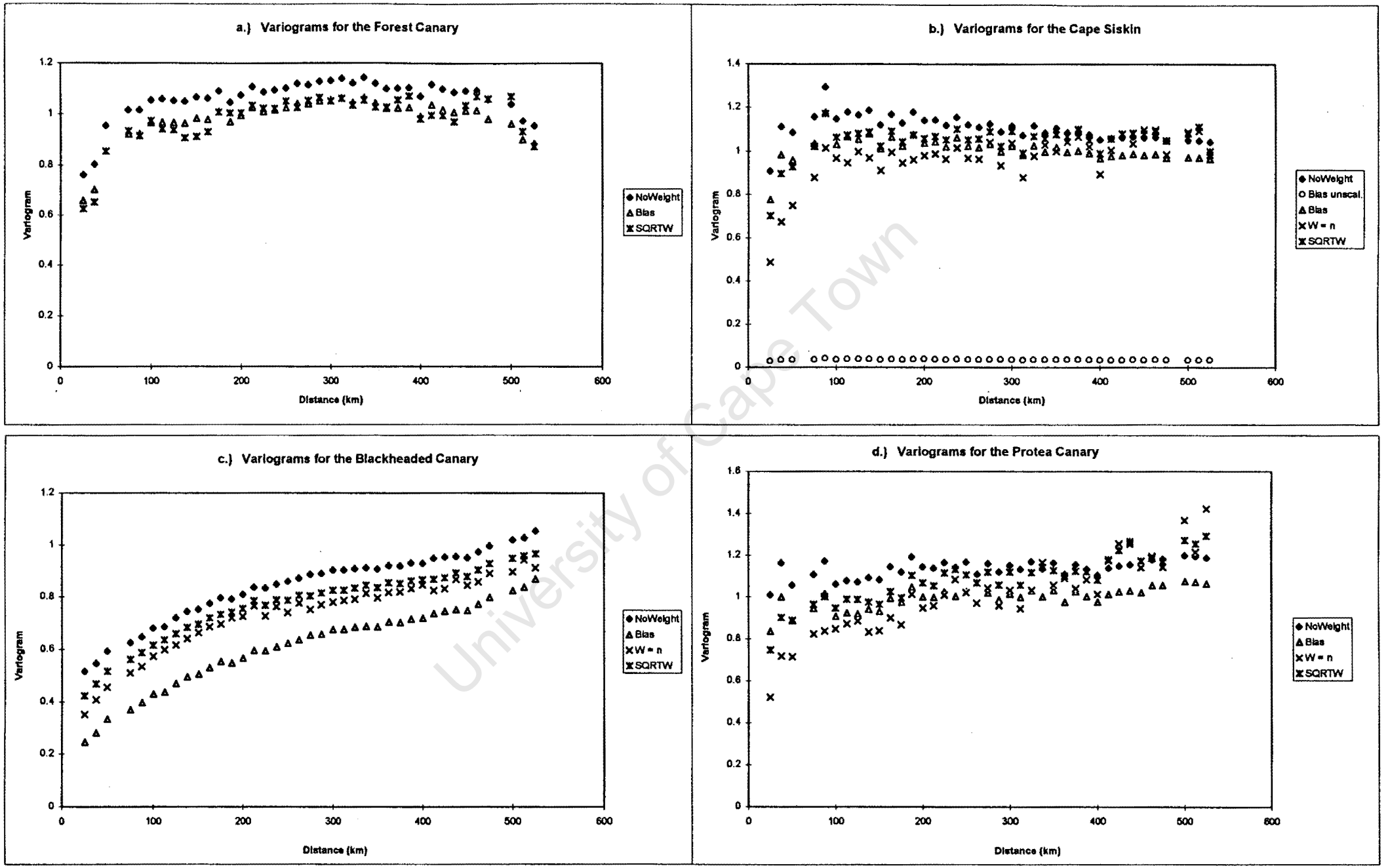


Figure 3 Different variogram methods are illustrated for four canaries. The methods used are shown in the legend.

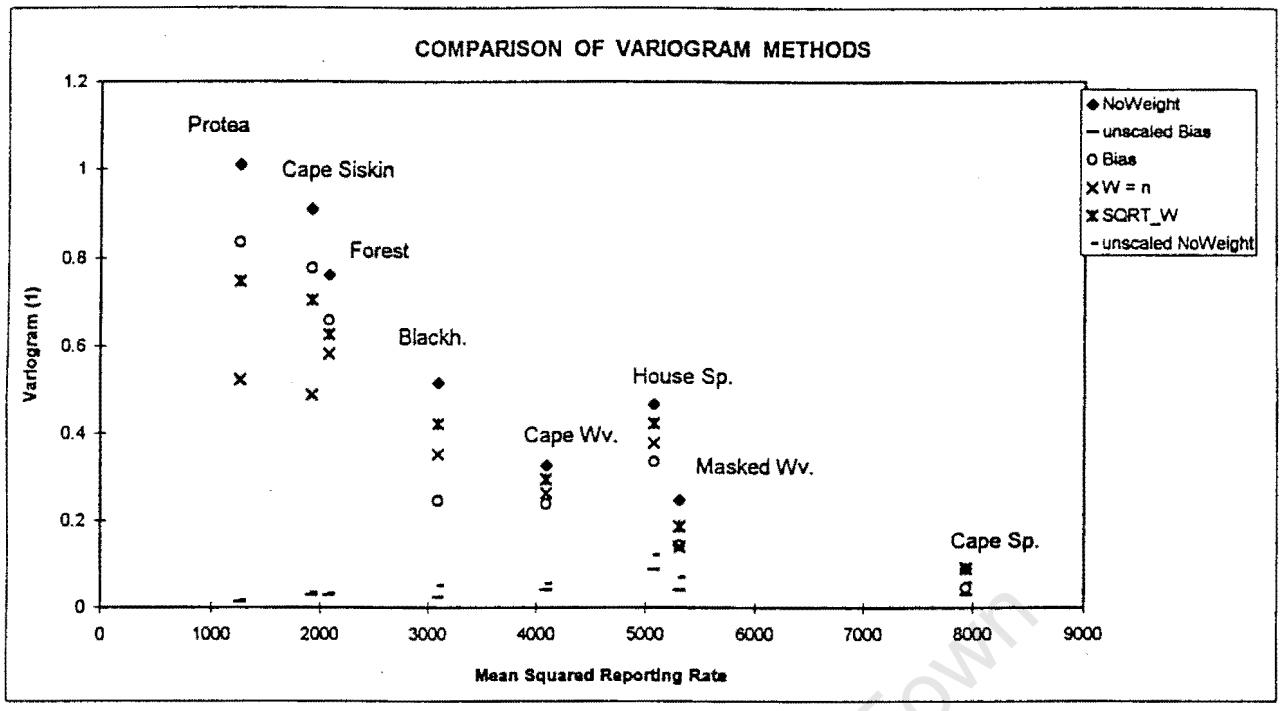


Figure 4. The first variogram values plotted against the mean squared reporting rates of the observed distributions of the species. The variogram methods that were used are specified in the legend. See also Table 1.

Table 1. Comparison of the Var(1) values calculated by different variogram methods. The atlas distribution maps were visually inspected to form a subjective opinion for pairs of species, on which of the two should have the smaller Var(1) value. The methods are compared against this opinion in the last column. See also Fig. 4.

SPECIES A	*SPECIES B*	*SUBJECTIVE OPINION*	*RESULTS*
880 Protea Canary	873 Forest Canary	approximately equal	most methods have Var(1) of 880 larger except W=n and unscaled methods
880 Protea Canary	874 Cape Siskin	874 < 880	all methods but the unscaled methods have 874 < 880
873 Forest Canary	874 Cape Siskin	874 < 873	all but W = n have 874< 873
876 Blackheaded Can.	801 House Sparrow	approximately equal	for the Bias method the difference appears to be too large
876 Blackheaded Can.	813 Cape Weaver	876 > 813	In the Bias method there appears to be too little difference

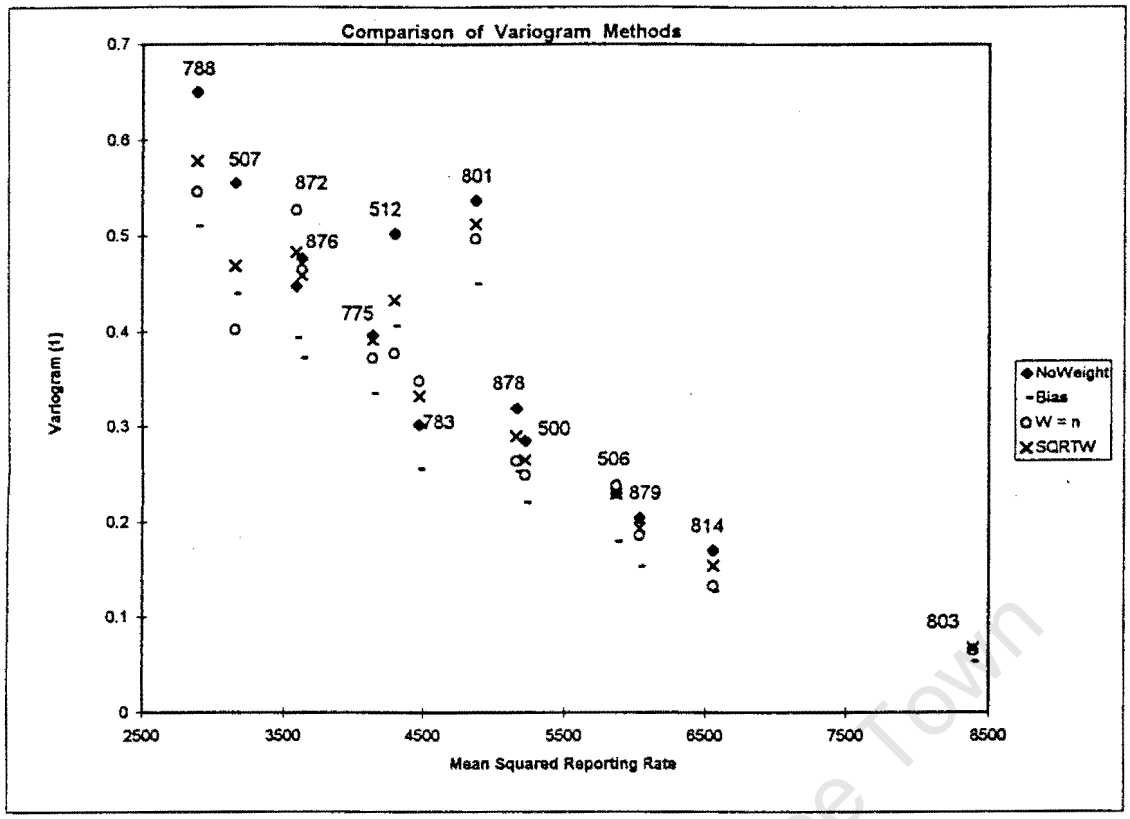


Figure 5. Variogram (1) values calculated by different variogram methods for different species. The species numbers are shown next to the plotted values and the names are given in Table 2. The variogram methods used are shown in the legend.

Table 2. Pairs of species are compared with respect to what their Variogram (1) values should be. An a priori opinion is formed by visual inspection of the atlas distribution maps. This is compared with the outcomes of the different variogram methods. The corresponding Variogram (1) values for the different methods and the various species are plotted in Figure 5.

SPECIES A	*SPECIES B*	*SUBJECTIVE OPINION*	*SUCCESSFUL METHODS*	*COMMENTS*
507 Redcapped Lark	872 Cape Canary	507 > 872	Normal Bias	
872 Cape Canary	876 Blackheaded Can	876 > 872	Normal	872 distribution is smaller
876 Blackheaded Can	512 Thickbilled Lark	512 > 876	Normal	Var(1) values should be almost equal
783 L Dbl.coll. Sunbird	878 Yellow Canary	783 > 878 ?? but not clear cut	Wn, Wrtn	Wn: the difference is too large
783 L Dbl.coll. Sunbird	500 Longbilled Lark	783 > 500	ALL	
872 Cape Canary	512 Thickbilled Lark	512 > 872	Normal Bias	
507 Redcapped Lark	876 Blackheaded Can	507 > 876	all but Wn	
500 Longbilled Lark	506 Spikeheeled Lark	506 > 500	NONE	
775 Malachite Sunbird	512 Thickbilled Lark	512 > 775	ALL	Wn: the difference is too large
801 House Sparrow	512 Thickbilled Lark	801 > 512	ALL	
507 Redcapped Lark	801 House Sparrow	507 > 801	Normal	
872 Cape Canary	801 House Sparrow	801 > 872	ALL but Wn	
879 Wh.throated Can.	814 Masked Weaver	814 > 879	NONE	

Table 3. First lag variogram values, calculated by various methods. These values were calculated over the distributions south of 27°S. See also Table 1 and Fig. 4.

SPECIES				VARIOGRAM METHOD					
		Mean (square)	Sigma	NoWeight	Bias unscal.	Bias	W = n	SQRTW	NoWeight unscal.
House Sparrow	801	5076	2175	0.467	0.086	0.335	0.378	0.424	0.120
Cape Sparrow	803	7930	1792	0.090	0.029	0.046	0.091	0.089	0.057
Cape Weaver	813	4087	2211	0.326	0.0397	0.238	0.262	0.295	0.054
Masked Weaver	814	5310	2383	0.247	0.0401	0.142	0.140	0.187	0.0696
Cape Siskin	874	1915	1277	0.911	0.0285	0.777	0.489	0.704	0.0334
Blackheaded Canary	876	3087	1353	0.516	0.0234	0.245	0.352	0.423	0.0492
Protea Canary	880	1259	806	1.011	0.0133	0.836	0.523	0.747	0.016
Forest Canary	873	2078	1449	0.762	0.028	0.658	0.583	0.626	0.0329

Table 4. Summary of results of variogram calculations when considering only a fixed block of 16x16 grid cells (30° - 33°S, 20°- 23°E). The variogram values shown in the right part of the table are the Variogram (1) values calculated by the various methods for the species. These variogram values are plotted in Fig. 5.

SPECIES		Mean Rep.Rate	Weighted Mean Rep.Rate	Estimated Sigma	Mean squared Rep.Rate	Variogram (1) NoWeight	Bias	W = n	SQRTW
Cape Canary	872	2914	3611	1844	3589	0.448	0.394	0.527	0.484
Blackheaded Canary	876	3129	3031	1095	3630	0.476	0.372	0.465	0.460
Yellow Canary	878	4368	2900	2324	5159	0.320	0.253	0.264	0.290
Whitethroated Canary	879	5605	4735	1612	6036	0.205	0.153	0.186	0.194
Masked Weaver	814	6106	5434	1844	6554	0.171	0.127	0.133	0.155
House Sparrow	801	4177	4404	2025	4870	0.538	0.450	0.497	0.513
Cape Sparrow	803	8140	6995	1673	8391	0.067	0.054	0.065	0.068
Malachite Sunbird	775	3579	3857	1703	4136	0.397	0.335	0.372	0.392
Lesser Doublecollared Sunbird	783	3727	4904	2145	4466	0.302	0.256	0.348	0.333
Dusky Sunbird	788	2270	2086	1183	2880	0.650	0.510	0.546	0.578
Longbilled Lark	500	4425	3545	2345	5222	0.286	0.221	0.250	0.265
Spikeheeled Lark	506	5102	4250	2490	5865	0.231	0.179	0.239	0.230
Redcapped Lark	507	2553	1963	1265	3152	0.556	0.440	0.402	0.469
Thickbilled Lark	512	3524	2420	1975	4288	0.503	0.406	0.377	0.433

Table 5. The estimated bias is shown for various cases of number of checklists and observed reporting rates. These values were calculated for a distribution with estimated mean 4384 and estimated standard deviation 5630. See also eq. 9.

Reporting Rate 1	2	Number of Checklists n1	n2	squared difference	estimated bias
0.167	0.143	12	14	0.0006	0.0109
0.167	0.125	12	8	0.0017	0.0147
0.167	0.000	12	15	0.0278	0.0059
0.167	0.667	12	6	0.2500	0.0177
0.167	0.222	12	9	0.0031	0.0138
0.167	0.167	12	6	0.0000	0.0177
0.771	0.580	131	138	0.0366	0.0011
0.771	0.723	131	184	0.0023	0.0009
0.771	0.000	131	3	0.5943	0.0005
0.813	0.800	80	50	0.0002	0.0023
0.813	0.412	80	17	0.1606	0.0050
0.813	0.773	80	22	0.0016	0.0041
0.750	0.571	4	7	0.0319	0.0278
0.750	0.600	4	5	0.0225	0.0318
0.750	0.286	4	7	0.2156	0.0278

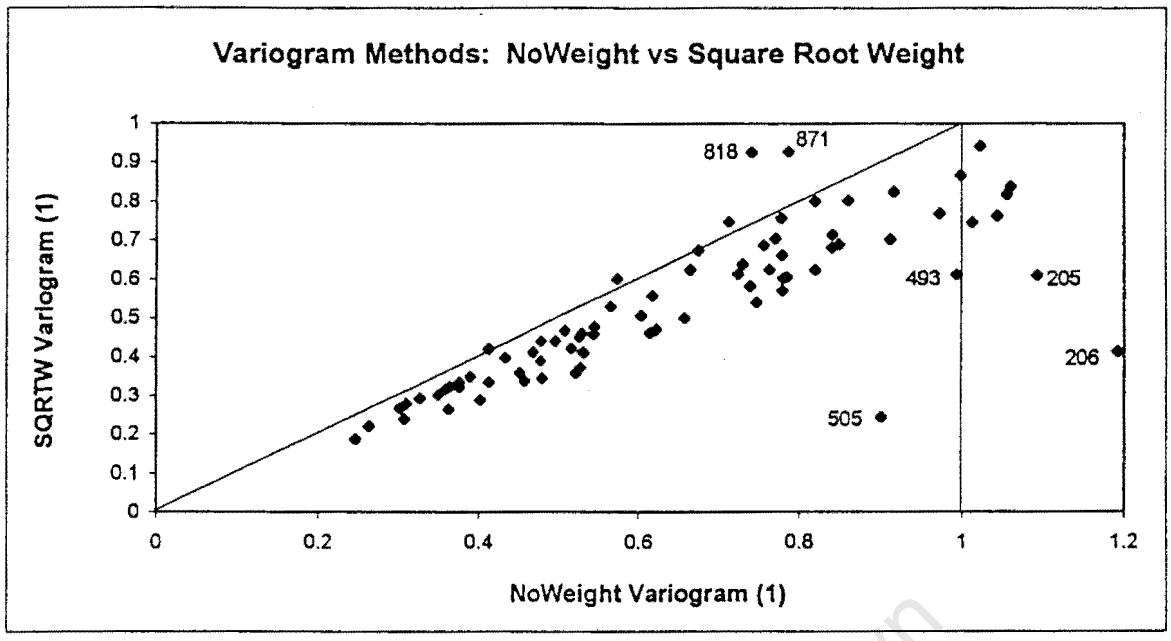
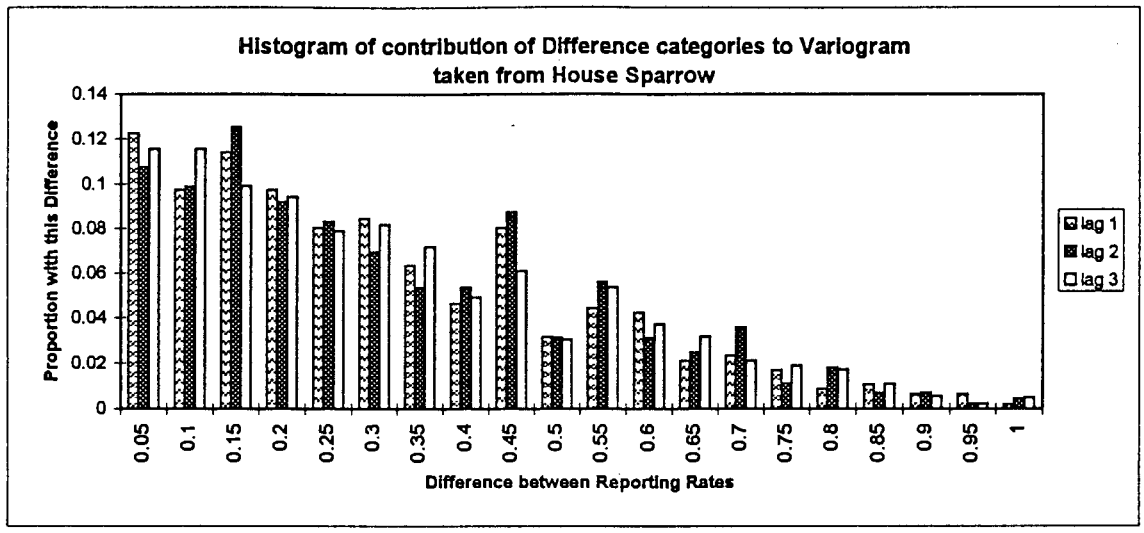


Figure 6. A scatterplot of the first variogram values calculated by the Methods NoWeight, which does not weight the differences, and SQRTW, which weights each calculated difference by the square root of the smaller number of checklists collected for the two grid cells. For the outlying observations the corresponding species numbers are shown, see also Tables 6 and 7.



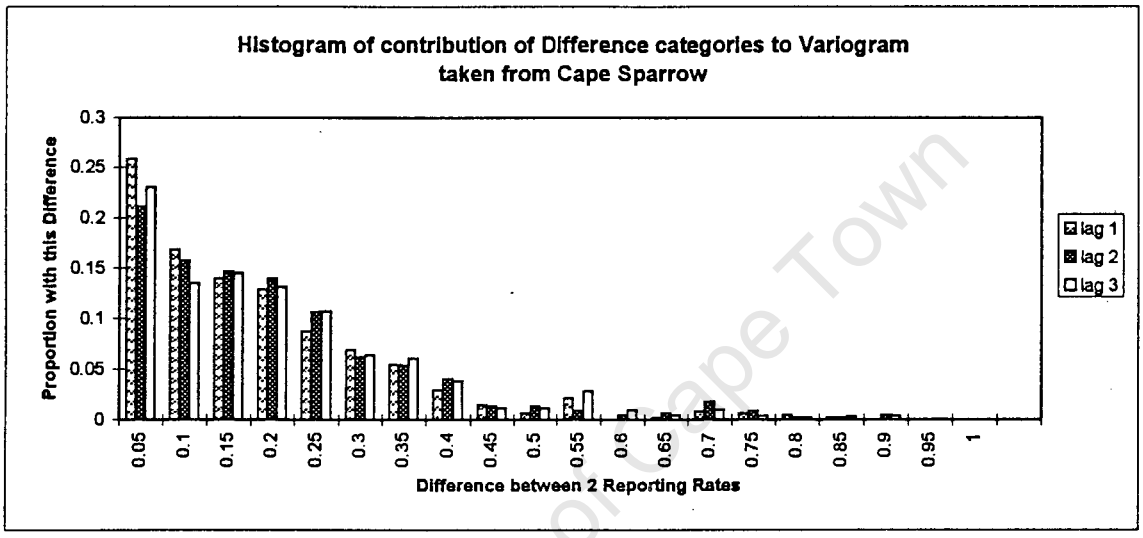


Figure 7 a, b Histograms showing which magnitudes of differences between reporting rates contributed to the estimation of the variogram values at lags 1, 2 and 3. See Figure 9 for the corresponding variograms of these two species.

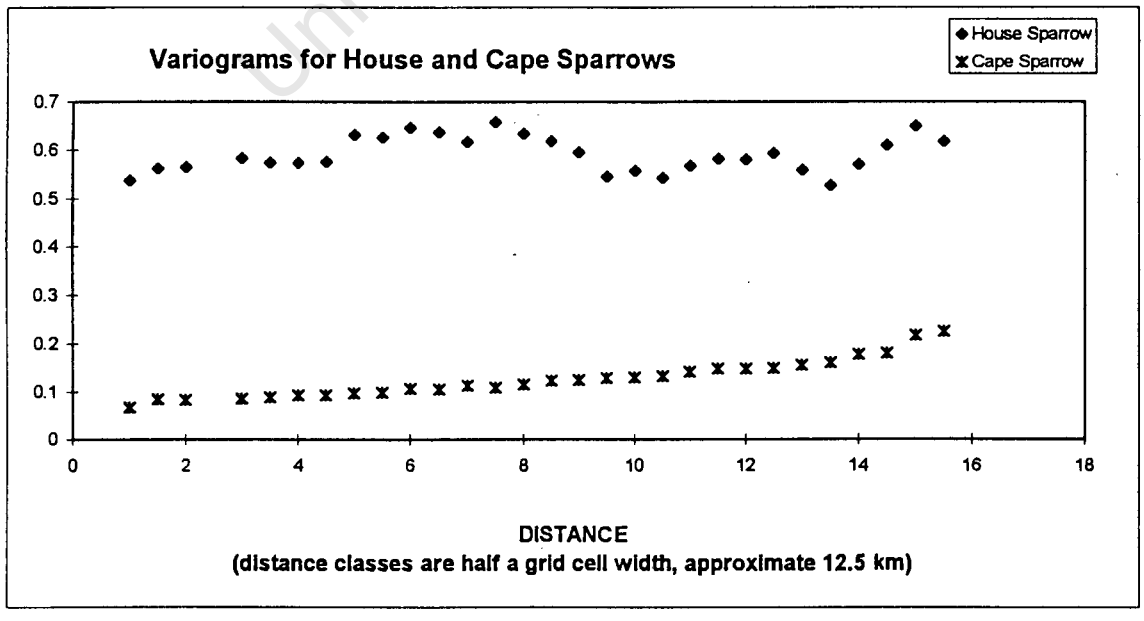


Figure 9. Variograms for the House and the Cape Sparrow, calculated by Method NoWeight on a 16x16 grid cell block (30° - 33°S, 20° - 23°E). See also Table 4.AA23

Table 6. The variogram (1) values calculated for all of the available species given with bird number and name. The variogram method that was used here is the NoWeight estimate, with no weight used, only a scaling factor, the average squared reporting rate.

	SPECIES	Var (1)
188	Coqui Francolin	0.779
189	Crested Francolin	0.572
190	Greywing Francolin	0.779
191	Shelley's Francolin	0.839
192	Redwing Francolin	0.818
193	Orange River Francolin	0.770
194	Redbilled Francolin	1.021
195	Cape Francolin	0.308
196	Natal Francolin	0.777
198	Rednecked Francolin	0.666
199	Swainson's Francolin	0.478
200	Common Quail	0.846
201	Harlequin Quail	1.042
203	Helmeted Guineafowl	0.357
204	Crested Guineafowl	0.780
205	Kurrichane Buttonquail	1.092
206	Blackrumped Buttonquail	1.192
492	Melodious Lark	0.729
493	Monotonous Lark	0.992
494	Rufousnaped Lark	0.412
495	Clapper Lark	0.544
496	Flappet Lark	0.712
497	Fawncoloured Lark	0.457
498	Sabota Lark	0.530
499	Rudd's Lark	1.059
500	Longbilled Lark	0.400
501	Shortclawed Lark	0.840
502	Karoo Lark	0.659
504	Red Lark	0.859
505	Dusky Lark	0.900
506	Spikeheeled Lark	0.375
507	Redcapped Lark	0.622
508	Pinkbilled Lark	0.820
509	Botha's Lark	0.915
510	Sclater's Lark	0.747
511	Stark's Lark	0.740
512	Thickbilled Lark	0.477
773	Cape Sugarbird	0.617
774	Gurney's Sugarbird	0.756

	SPECIES	Var (1)
775	Malachite Sunbird	0.433
777	Orangebreasted Sunbird	0.674
779	Marico Sunbird	0.784
780	Purplebanded Sunbird	0.564
782	Neergaard's Sunbird	0.997
783	Lesser Doublecollared Sunbird	0.390
785	Greater Doublecollared Sunbird	0.508
787	Whitebellied Sunbird	0.412
788	Dusky Sunbird	0.591
789	Grey Sunbird	0.544
790	Olive Sunbird	0.451
791	Scarletchested Sunbird	0.480
792	Black Sunbird	0.377
793	Collared Sunbird	0.529
807	Thickbilled Weaver	0.468
808	Forest Weaver	0.532
810	Spectacled Weaver	0.302
811	Spottedbacked Weaver	0.351
813	Cape Weaver	0.326
814	Masked Weaver	0.246
815	Lesser Masked Weaver	0.602
816	Golden Weaver	1.054
817	Yellow Weaver	0.497
818	Brownthroated Weaver	0.740
819	Redheaded Weaver	0.971
869	Yelloweyed Canary	0.263
870	Blackthroated Canary	0.526
871	Lemonbreasted Canary	0.786
872	Cape Canary	0.362
873	Forest Canary	0.757
874	Cape Siskin	0.911
875	Drakensberg Siskin	0.522
876	Blackheaded Canary	0.517
877	Bully Canary	0.530
878	Yellow Canary	0.307
879	Whitethroated Canary	0.351
880	Protea Canary	1.011
881	Steakyheaded Canary	0.725

Table 7. The variogram (1) values calculated for all of the available species sorted in ascending order are given with bird number and name. The variogram method that was used here is the NoWeight estimate, with no weight used, only a scaling factor, the average squared reporting rate. These calculations only consider the distributions south of 27°S. The smoothest distributions appear at the top of the table (left) while those that are least smooth are at the end (right) of the table. The smoothness' of the distributions decreases with increasing Var(1) value.

	SPECIES	Var (1)
814	Masked Weaver	0.246
869	Yelloweyed Canary	0.263
810	Spectacled Weaver	0.302
878	Yellow Canary	0.307
195	Cape Francolin	0.308
813	Cape Weaver	0.326
811	Spottedbacked Weaver	0.351
879	Whitethroated Canary	0.351
203	Helmeted Guineafowl	0.357
872	Cape Canary	0.362
506	Spikeheeled Lark	0.375
792	Black Sunbird	0.377
783	Lesser Doublecollared Sunbird	0.390
500	Longbilled Lark	0.400
787	Whitebellied Sunbird	0.412
494	Rufousnaped Lark	0.412
775	Malachite Sunbird	0.433
790	Olive Sunbird	0.451
497	Fawncoloured Lark	0.457
807	Thickbilled Weaver	0.468
512	Thickbilled Lark	0.477
199	Swainson's Francolin	0.478
791	Scarletchested Sunbird	0.480
817	Yellow Weaver	0.497
785	Greater Doublecollared Sunbird	0.508
876	Blackheaded Canary	0.517
875	Drakensberg Siskin	0.522
870	Blackthroated Canary	0.526
793	Collared Sunbird	0.529
877	Bully Canary	0.530
498	Sabota Lark	0.530
808	Forest Weaver	0.532
495	Clapper Lark	0.544
789	Grey Sunbird	0.544
780	Purplebanded Sunbird	0.564
189	Crested Francolin	0.572
788	Dusky Sunbird	0.591
815	Lesser Masked Weaver	0.602
773	Cape Sugarbird	0.617

	SPECIES	Var (1)
507	Redcapped Lark	0.622
502	Karoo Lark	0.659
198	Rednecked Francolin	0.666
777	Orangebreasted Sunbird	0.674
496	Flappet Lark	0.712
881	Steakyheaded Canary	0.725
492	Melodious Lark	0.729
818	Brownthroated Weaver	0.740
511	Stark's Lark	0.740
510	Sclater's Lark	0.747
774	Gurney's Sugarbird	0.756
873	Forest Canary	0.757
193	Orange River Francolin	0.770
196	Natal Francolin	0.777
190	Greywing Francolin	0.779
188	Coqui Francolin	0.779
204	Crested Guineafowl	0.780
779	Marico Sunbird	0.784
871	Lemonbreasted Canary	0.786
192	Redwing Francolin	0.818
508	Pinkbilled Lark	0.820
191	Shelley's Francolin	0.839
501	Shortclawed Lark	0.840
200	Common Quail	0.846
504	Red Lark	0.859
505	Dusky Lark	0.900
874	Cape Siskin	0.911
509	Botha's Lark	0.915
819	Redheaded Weaver	0.971
493	Monotonous Lark	0.992
782	Neergaard's Sunbird	0.997
880	Protea Canary	1.011
194	Redbilled Francolin	1.021
201	Harlequin Quail	1.042
816	Golden Weaver	1.054
499	Rudd's Lark	1.059
205	Kurrichane Buttonquail	1.092
206	Blackrumped Buttonquail	1.192

Table 8. The variogram (1) values calculated for all of the available species given with bird number and name. The variogram method used is SQRTW, where the weights are the square roots of the smaller of the number of checklists in the two grid cells. These variogram values were calculated over the distributions south of 27°S. The scaling factor was the average squared reporting rate.

	SPECIES	Var (1)
188	Coqui Francolin	0.571
189	Crested Francolin	0.602
190	Greywing Francolin	0.662
191	Shelley's Francolin	0.682
192	Redwing Francolin	0.801
193	Orange River Francolin	0.705
194	Redbilled Francolin	0.942
195	Cape Francolin	0.280
196	Natal Francolin	0.759
198	Rednecked Francolin	0.624
199	Swainson's Francolin	0.442
200	Common Quail	0.690
201	Harlequin Quail	0.763
203	Helmeted Guineafowl	0.319
204	Crested Guineafowl	0.603
205	Kurrichane Buttonquail	0.611
206	Blackrumped Buttonquail	0.415
492	Melodious Lark	0.638
493	Monotonous Lark	0.613
494	Rufousnaped Lark	0.337
495	Clapper Lark	0.460
496	Flappet Lark	0.748
497	Fawncoloured Lark	0.339
498	Sabota Lark	0.412
499	Rudd's Lark	0.840
500	Longbilled Lark	0.289
501	Shortclawed Lark	0.715
502	Karoo Lark	0.500
504	Red Lark	0.803
505	Dusky Lark	0.245
506	Spikeheeled Lark	0.323
507	Redcapped Lark	0.472
508	Pinkbilled Lark	0.625
509	Botha's Lark	0.826
510	Sclater's Lark	0.542
511	Stark's Lark	0.583
512	Thickbilled Lark	0.391
773	Cape Sugarbird	0.558
774	Gurney's Sugarbird	0.688

	SPECIES	Var (1)
775	Malachite Sunbird	0.399
777	Orangebreasted Sunbird	0.675
779	Marico Sunbird	0.606
780	Purplebanded Sunbird	0.532
782	Neergaard's Sunbird	0.868
783	Lesser Doublecollared Sunbird	0.351
785	Greater Doublecollared Sunbird	0.469
787	Whitebellied Sunbird	0.421
788	Dusky Sunbird	0.463
789	Grey Sunbird	0.478
790	Olive Sunbird	0.361
791	Scarletchested Sunbird	0.345
792	Black Sunbird	0.337
793	Collared Sunbird	0.375
807	Thickbilled Weaver	0.414
808	Forest Weaver	0.414
810	Spectacled Weaver	0.269
811	Spottedbacked Weaver	0.303
813	Cape Weaver	0.295
814	Masked Weaver	0.187
815	Lesser Masked Weaver	0.507
816	Golden Weaver	0.820
817	Yellow Weaver	0.445
818	Brownthroated Weaver	0.927
819	Redheaded Weaver	0.769
869	Yelloweyed Canary	0.222
870	Blackthroated Canary	0.453
871	Lemonbreasted Canary	0.929
872	Cape Canary	0.324
873	Forest Canary	0.626
874	Cape Siskin	0.704
875	Drakensberg Siskin	0.360
876	Blackheaded Canary	0.423
877	Bully Canary	0.462
878	Yellow Canary	0.240
879	Whitethroated Canary	0.264
880	Protea Canary	0.747
881	Steakyheaded Canary	0.614

Table 9. The variogram (1) values calculated for all of the available species sorted in ascending order are given with bird number and name. The variogram method that was used here is SQRTW which weights each difference by the square root of the smaller number of checklists. The scaling factor was the average squared reporting rate. These calculations only consider the distributions south of 27°S. The smoothest distributions appears at the top of the table (left) while those that are least smooth are at the end (right) of the table. The 'smoothness' of the distributions decreases with increasing Var(1) value.

	SPECIES	VAR(1)
814	Masked Weaver	0.187
869	Yelloweyed Canary	0.222
878	Yellow Canary	0.240
505	Dusky Lark	0.245
879	Whitethroated Canary	0.264
810	Spectacled Weaver	0.269
195	Cape Francolin	0.280
500	Longbilled Lark	0.289
813	Cape Weaver	0.295
811	Spottedbacked Weaver	0.303
203	Helmeted Guineafowl	0.319
506	Spikeheeled Lark	0.323
872	Cape Canary	0.324
792	Black Sunbird	0.337
494	Rufousnaped Lark	0.337
497	Fawncoloured Lark	0.339
791	Scarletchested Sunbird	0.345
783	Lesser Doublecollared Sun	0.351
875	Drakensberg Siskin	0.360
790	Olive Sunbird	0.361
793	Collared Sunbird	0.375
512	Thickbilled Lark	0.391
775	Malachite Sunbird	0.399
498	Sabota Lark	0.412
807	Thickbilled Weaver	0.414
808	Forest Weaver	0.414
206	Blackrumped Buttonquail	0.415
787	Whitebellied Sunbird	0.421
876	Blackheaded Canary	0.423
199	Swainson's Francolin	0.442
817	Yellow Weaver	0.445
870	Blackthroated Canary	0.453
495	Clapper Lark	0.460
877	Bully Canary	0.462
788	Dusky Sunbird	0.463
785	Greater Doublecollared Su	0.469
507	Redcapped Lark	0.472
789	Grey Sunbird	0.478
502	Karoo Lark	0.500

	SPECIES	VAR (1)
815	Lesser Masked Weaver	0.507
780	Purplebanded Sunbird	0.532
510	Sclater's Lark	0.542
773	Cape Sugarbird	0.558
188	Coqui Francolin	0.571
511	Stark's Lark	0.583
189	Crested Francolin	0.602
204	Crested Guineafowl	0.603
779	Marico Sunbird	0.606
205	Kurrichane Buttonquail	0.611
493	Monotonous Lark	0.613
881	Steakyheaded Canary	0.614
198	Rednecked Francolin	0.624
508	Pinkbilled Lark	0.625
873	Forest Canary	0.626
492	Melodious Lark	0.638
190	Greywing Francolin	0.662
777	Orangebreasted Sunbird	0.675
191	Shelley's Francolin	0.682
774	Gurney's Sugarbird	0.688
200	Common Quail	0.690
874	Cape Siskin	0.704
193	Orange River Francolin	0.705
501	Shortclawed Lark	0.715
880	Protea Canary	0.747
496	Flappet Lark	0.748
196	Natal Francolin	0.759
201	Harlequin Quail	0.763
819	Redheaded Weaver	0.769
192	Redwing Francolin	0.801
504	Red Lark	0.803
816	Golden Weaver	0.820
509	Botha's Lark	0.826
499	Rudd's Lark	0.840
782	Neergaard's Sunbird	0.868
818	Brownthroated Weaver	0.927
871	Lemonbreasted Canary	0.929
194	Redbilled Francolin	0.942

CHAPTER 3

Smoothing of Bird Atlas Distribution Maps, Based on Reporting Rates

INTRODUCTION

In *The Atlas of Southern African Birds*, the distribution maps show the observed data, with no interpolation or smoothing.

Interpolation was a minor problem, because only 2% of 4537 grid cells were not surveyed, but could be a major problem in atlases for other taxa. Smoothing has however the potential to make a substantial contribution to the improvement of the distribution maps for the bird atlas. This is particularly true in those areas, such as the Karoo, where relatively small numbers of checklists were obtained for each grid cell, and the reporting rates are therefore unreliable, due to sampling vagaries. The intuitive feeling is that some form of judicious averaging over adjoining grid cells with similar habitats would result in improved distribution maps.

The statistical problem is that smoothing and interpolation of binomial data is a poorly developed area, apart form the special 'bernoulli' case of presence/absence data.

Interpolation of atlas distribution maps was a controversial issue within the Atlas Publication Committee (L.G. Underhill & J.A. Harrison pers. comm.). The statisticians within the committee were keen to 'improve' the distribution maps by employing a smoothing technique; the biologists on the committee considered that the maps should display the 'truth', in their view, the observed data. The committee decision, implemented in the atlas, was to show the observed reporting rates, apart from some special cases (described by Harrison & Underhill 1997, p.lviii) involving grid cells for which only one or two checklists were obtained.

Inconsistency and irregularity of data occurs particularly in areas where there are only a few checklists or where some of the grid cells do not have checklists at all. We hypothesised that a smoothed version of the existing reporting rate maps would be a more accurate representation of the true distribution of bird species.

There are many factors that influence observed reporting rates. One important factor is the number of checklists collected for any particular grid cell. Other factors are observer effort, the variety of habitats in the grid cell, the accessability of the terrain, the conspicuousness of the bird, its overall commonness and seasonal effects such as breeding plumage and migration.

With little data the influence of these above factors becomes stronger. The smaller the number of checklists collected, the more fluctuation will there be in the resulting reporting rate. The reporting rate is the proportion of successes out of the total number of checklists. With only one checklist, reporting rates of one and zero are the only possibilities, none of which is a good estimate of the true underlying reporting rate. But as the number of checklists increases, the reporting rate will gradually approach some constant value inherent to that particular grid cell.

Since the reporting rates are the only values that are shown on the final distribution maps the above explains why there is so much fluctuation in the shades between adjoining grid cells in these maps. This gives the motivation for finding methods to smooth these observed reporting rates.

In this chapter we will be looking at a variety of methods that find a smoothed value through regression, more particular through generalized linear models. These models predict a value for a grid cell using the information of nearby cells.

METHODS

PROBLEM STATEMENT

Distribution maps in *The Atlas of Southern African Birds* (Harrison et al. 1997a, b) are based on data that may be treated as binomial. The number of checklists in individual grid cells varies between zero and 1260. As a result of this the reporting rates have different accuracies. This is reflected in the variability of the shades in the atlas maps in areas where, a priori, uniform shading would be anticipated. This is especially true in areas where few checklists were obtained.

The smoothing of binomial spatial data has not been extensively researched. Spatial correlation also exists so that the individual observations in adjoining gird cells are not independent. A factor that complicates the smoothing process is the fact that the spatial correlation is not constant for all species but varies between species, which suggests that different degrees of smoothing should be used for each species.

The reporting rates presented in the distribution maps are only estimates for the true probabilities of the species occurring in each of the grid cells. The accuracy of the observed reporting rates depend most importantly on the number of checklists from which they were calculated ($r_i = S_i / n_i$). There is therefore a much higher degree of variability in the illustrated maps than is present in the true situation.

There are several reasons why these maps may require smoothed versions. For one, the reporting rates show more variability than is actually present, due to observer effort and characteristics of the species. This measurement error should be removed. Secondly, for more general reference books, the maps should be smoother than the direct observations. Thirdly checklists were not collected for all grid cells. For such cases the smoothing would at the same time find an interpolated value for the cell.

Our aim was to smooth the observed reporting rates so that they become more consistent with the surrounding information. The degree of adjustment should be based on the accuracy of the information in the surrounding grid cells and also on the spatial correlation between the cell itself and its surrounding cells. Above all else, the smoothing method is required to be rapid, because of the large area covered (4537 grid cells) and because of the large number of species (more than 900).

ASSUMPTIONS

If surrounding values are used for predicting the probability of occurrence of the species in a cell, one must assume that there is some form of relationship between the probabilities of occurrence in these cells. The first assumption that has to be made is that the species does occur in patches that are larger than a single grid cell (approximately 25km x 25km), i.e. that there is some consistency in occurrence between neighbouring grid cells. This assumption will not be reasonable for all

species; for example, habitat specific species, such as wetland birds and forest birds are patchily distributed in discrete habitats. We developed a measure of continuity of distribution in the previous chapter. This measure can here be used as an indicator whether smoothing is valid or not, and the extent to which smoothing should be used.

We must also assume that the observed reporting rates are a valid estimate for the probability of occurrence of a species in any grid cell. This implies that it is valid to smooth these reporting rate maps in order to get a more general overview of the true distribution of the species. This problem is discussed in more detail in a later section. The observed reporting rates for a species in a grid cell are a measure of the average probability of encountering the species in the grid cell (Underhill et al. 1992). The distribution maps in *The Atlas of Southern African Birds* aim to present the probability that a species occurs in a grid cell.

LITERATURE REVIEW

Cressie (1991) provided an extensive reference of existing methods and theory on the analysis of spatial data and with it gives many application examples, specialising however on geostatistics and kriging.

SMOOTHING SPATIAL DATA

Discussions of smoothing methods for spatial data can be found in Cressie (1991), Ripley (1981) and McNeill (1994).

Trend Surface Analysis, where a polynomial is fitted to the data by the method of least squares, has the problem that it does not consider spatial correlation. This may cause that polynomials of too high orders are fitted so that the clusters of similar data values can be described. If each observation is not equally reliable, as is the case with binomial data where the variance of the observation depends on the total number of trials, a weighted least squares approach should be taken (Draper & Smith 1981). Trend Surface Analysis is a global approach and some of the problems are that the form of the trend has to be determined and that the relationship to these trend parameters may change in different regions. A local approach would reduce the number of parameters required and would ensure more similar characteristics of the area. The most serious problem however is that fitting polynomials to the data ignores the spatial correlation which is present and assumes independence between observations.

Kriging smoothing or interpolation methods especially solve the problem of spatial autocorrelation. A covariance function is established which assumes that the association of data values depends only on the distance between the respective locations, i.e. that the data is stationary.

The predicted or smoothed value for location s_0 has the following form

$$\hat{Z}_o = \sum_{i=1}^{n} w_i Z_i$$

The weights minimise the expected squared error

$$E\left[\left(\sum_{i=1}^{n} w_i Z_i - Z_o\right)^2\right]$$

In the case of simple kriging the weights are found by solving

$$\mathbf{Cw} = \mathbf{c}$$

where **C** is the covariance matrix with elements c_{ij} the covariance of Z_i and Z_j and **c** is a vector of covariances with elements c_i the covariance between Z_i and Z_o.

Then the solution is

$$\hat{Z}_o = \mathbf{c}^T\mathbf{C}^{-1}\mathbf{Z}$$

This method of simple kriging can be extended to include other covariates (co-kriging) or to model trend as a polynomial in x and y (universal kriging). Usually the trend has to be removed so that the remaining process is stationary. An advantage of kriging is that the covariance function determines the appropriate degree of smoothing (McNeill 1994).

Kriging where the observations include measurement error has not been researched extensively. In the binomial data of the bird atlas each observation is subject to the binomial variance and is therefore only an estimate of the true underlying situation. If π_i is the true probability of observing a bird species in grid cell *i* then the observed reporting rate R_i is only an estimate for π_i,

$$\pi_i = R_i + \varepsilon_i$$

where ε_i is the observation error. In addition this error is not constant over all observations but varies, depending directly on the number of checklists collected for the grid cell. McNeill (1991, 1994) developed a form of kriging which smooths data with measurement error, concentrating on binomial and poisson data. The problems encountered with this approach were that of trend estimation and removal and the estimation of the measurement error contributing to the covariance function. The final model for the binomial atlas data assumed that trend in a local window (7x7) does not have a large effect on the variogram values. The trend was therefore modelled implicitly in the kriging equations, aiming only to remove the measurement error.

Kernel smoothing uses a weighted average of nearby observations. This is a local approach. The weights are a function of the inverse of the distance between the observation location and the point to be estimated. This function, the kernel function, can be adjusted to smooth the data to different degrees. Kernel smoothing ignores spatial correlation between data points.

'SUDDEN - INFANT - DEATH - SYNDROME' DATA

A data set of 'Sudden-Infant-Death-Syndrome', (SIDS), from North Carolina is widely used in the literature to illustrate the geostatistical theory for the case of counts data (Cressie 1991, 1994). It is similar to the bird atlas data in that it counts successes out of a total number of counts. For the SIDS data, the number of sudden-infant-deaths out of a total of n_i life births in a county are counted. The counties are not on a regular grid but their locations were transformed to grid coordinates. The difference of the SIDS data to the bird atlas data is that all of the n_i are extremely large compared to the numbers of checklists for single grid cells in the bird atlas data, not causing the same degree of unreliability of true probabilities as in the bird atlas data, where some of the n_i equal only zero, one or two.

The problem that the SIDS and the bird atlas data have in common, is that the variance of the observations is not constant but does depend on the mean and on the total number of items for the grid cell. To make the variance independent of the mean the *Freeman-Tukey (square root) transformation* was used (Cressie 1991 p. 245).

$$Z_i = \left(\frac{1000\, S_i}{n_i}\right)^{1/2} + \left(\frac{1000\,(S_i + 1)}{n_i}\right)^{1/2} \qquad (1)$$

Then the variance of the Z_i will be

$$\text{Var}\,(Z_i) \cong \frac{\tau^2}{n_i}$$

where τ is a constant. The Freeman-Tukey transformation removes the dependence on the mean but the variance still depends on n_i.

Cressie stated that this transformation symmetrizes the data. Symmetric data is more likely to give rise to additive models and also increases the chance of equal variances (Cressie 1985).

The next step for the SIDS data was to remove the trend so that the residuals are intrinsically stationary, using the median-polish algorithm. **Median-polish** is easy to use if the data points lie on a regular grid. It assumes that the mean has as additive components an overall effect, a row effect and a column effect ($\mu(s) = a + r_k + c_l$). Interaction terms are also possible but are more complicated to estimate. The residual process, $\delta(s)$, is assumed to be intrinsically stationary and is used to estimate the variogram. To predict a value for location s_o, the following equation is used:

$$\tilde{Z}(s_o) = \tilde{\mu}(s_o) + \hat{R}(s_o)$$

where $\tilde{\mu}(s_o)$ is the median polish estimate and $\hat{R}(s_o)$ is the estimate of the residual found by the kriging equations which make use of the variogram estimated from the residual data. The median-polish algorithm requires that the data are approximately symmetrical. It is robust if outliers are present.

Alternatively, if the variances differ greatly between observations, as is the case with binomial data when the total n_i in each observation differ, a weighted median-polish algorithm should be used instead. This was used on the SIDS data (Cressie 1991, p.396-402). Weighted median-polish weights each data point by $\sqrt{n_i}$. To find intrinsically stationary residuals each of the residuals is multiplied by the corresponding $\sqrt{n_i}$ to give a set of normalized residuals from which the spatial dependence can be estimated. The intrinsically stationary process of normalized residuals is

$$\delta(s) = \sqrt{n_i}\ R(x_i, y_i)$$

This approach based the variogram on the number of successes, therefore the data was rather modelled with a Poisson distribution than with a binomial distribution. We however are only concerned about the spatial dependence of the observed reporting rates, not the number of 'successes' per 10000 say. It is believed that even if a weighted median-polish algorithm for the removal of trend was used, the variance of the residual reporting rates would still heavily depend on the number of checklists n_i.

Cressie (1991, p.183) stated that for spatial two-dimensional data, or higher dimensions, trend, $\mu(.)$, will usually decompose additively into directional components.

We will not use the kriging approach but will develop a smoothing approach based on regression concepts. Roughly speaking this approach aims to 'predict' the reporting rate of a grid cell from a small set of surrounding observed reporting rates. The kriging and generalized linear model approaches are compared at a later stage.

DISCUSSION OF THEORY

LEAST SQUARES LINEAR REGRESSION

Linear regression assumes that the variable to be modelled, **Y**, is related to the explanatory variables, matrix **X**, as shown in eq. 2. The variable of interest, **Y**, is also called the **response** variable.

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \qquad (2)$$

The observed value Y_i is a linear function of the explanatory variables $\boldsymbol{x}$, and observed error ε_i. For this relationship to hold the following conditions have to be met (Cox & Snell 1989):

- The distribution of the response Y_i is normal
- The errors ε_i have a normal distribution (the error distribution) with
- $E(\varepsilon_i) = 0$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (i.e. the variance is constant over all *i* observations) (3)

The least squares estimate for β in this situation is

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{Y}$$

b is an unbiased estimate for β, provided the model is correct (Draper & Smith 1981).

Often the above conditions do not hold. It may be the case that the response **Y** does not have a normal distribution. Then a transformation of the response may convert it to have a normal distribution but this still does not guarantee that the individual observations have constant variance.

EXPLANATORY VARIABLES

a.)

(x,y)		x		
		-1	**0**	**1**
y	**-1**	(-1,-1)	(0,-1)	(1,-1)
	0	(-1, 0)	(0, 0)	(1, 0)
	1	(-1, 1)	(0, 1)	(1, 1)

b.)

(x,y)		x				
		-2	**-1**	**0**	**1**	**2**
y	**-2**	(-2,-2)	(-1,-2)	(0,-2)	(1,-2)	(2,-2)
	-1	(-2,-1)	(-1,-1)	(0,-1)	(1,-1)	(2,-1)
	0	(-2, 0)	(-1, 0)	(0, 0)	(1, 0)	(2, 0)
	1	(-2, 1)	(-1, 1)	(0, 1)	(1, 1)	(2, 1)
	2	(-2, 2)	(-1, 2)	(0, 2)	(1, 2)	(2, 2)

Figure 1 The (x,y) coordinates assigned to grid cells relative to the central grid cell are shown here a.) for blocks of nine grid cells and b.) for blocks of 25 grid cells. These coordinates will be used as explanatory variables in the regression models.

Environmental factors that influence the suitability of a habitat for a bird species change more or less gradually so that nearby grid cells can be expected to have more similar reporting rates than cells that are far apart. This means that the reporting rates are not independent but spatially autocorrelated. This spatial autocorrelation can be modelled if it is incorporated into the explanatory variables. We can choose as explanatory variables the coordinates on a north-south and east-west set of axes. This provides a description of the spatial relation among the grid cells and allows us to assume that the residuals from this model are approximately independent.

The explanatory variables are the north-south and west-east coordinates relative to the central grid cell, which will have coordinates (x,y) = (0,0) (Figure 1). The y-coordinate represents the north-south axis and increases from north to south (using the conventions of the southern hemisphere). The x-coordinate represents the east-west axis and increases from west to east. The dependent variables (response variables) are the reporting rates in each of the grid cells in the block.

With this setup it is possible to fit a curved surface to the observed reporting rates in each block of cells and from that predict a smoothed reporting rate for the central grid cell.

THE ATLAS DATA

Let us introduce the bird atlas data (Harrison *et al.* 1997a, b) at this point. For the quarter degree grid cells of southern Africa checklists of species were collected. In each of these, each species was either recorded as present or as absent. The observed response is the proportion of total checklists on which the species was recorded as present, this is the reporting rate. Our aim is to model the probability of success, the probability that the species is recorded in any given grid cell.

S_i is the **number of successes** out of the n_i checklists collected for cell *i*. $r_i = S_i / n_i$ is the observed **reporting rate** for cell *i*. The number of successes, S_i, can be modelled, approximately, by a binomial distribution (Underhill *et al.* 1992) with the following expected value and variance:

$$E(S_i) = n_i \pi_i \tag{4}$$

$$\text{Var}(S_i) = n_i \pi_i (1 - \pi_i) \tag{5}$$

The expected value and variance of the reporting rate r_i are given by

$$E\left(\frac{S_i}{n_i}\right) = \pi_i \tag{6}$$

$$\text{Var}\left(\frac{S_i}{n_i}\right) = \frac{\pi_i (1 - \pi_i)}{n_i} \tag{7}$$

where π_i is the true probability of recording the species in cell *i*.

From the form of eq. 7 it can be seen that the variance of the reporting rates depends on the number of checklists n_i. This results in some observations, those originating from a large number of checklists, being more reliable, i.e. these have smaller variances. This means that the variance of the observed reporting rates is not constant over all grid cells but inversely proportional to n_i.

Regression aims to model the response on the vector of explanatory variables $\boldsymbol{x}$. This is done by relating the response to $\boldsymbol{x}$ through a function. The simplest form would be a linear relationship:

$$r_i = \beta_o + x_i \beta + \varepsilon_i \qquad (8)$$

where r_i is the observed reporting rate, β_o is a constant and ε_i is the observed error. Then the model for π_i, the *true* probability of observing the species in cell *i* has the following form:

$$\pi_i = \beta_o + x_i \beta \qquad (9)$$

But the conditions for the linear relationship do not hold (eq. 3). The error distribution is not normal but binomial, the linear function does not restrict π_i to lie in the interval [0, 1] and the variance of the errors is not constant but depends on π_i and on n_i.

For the binomial distribution however, the following holds (Cox & Snell 1989, p.15, 16)

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + x_i\beta \qquad (10)$$

This is the **logistic transformation** of π_i which is linearly related to the explanatory variables. The type of model is called the linear logistic model (Cox & Snell 1989).

Rewriting this to make π_i the subject of the equation gives

$$\pi_i = \frac{\exp(\beta_o + x_i\beta)}{1+\exp(\beta_o + x_i\beta)} \qquad (11)$$

so that $0 \leq \pi_i \leq 1$, and the values for π_i are held within the possible range.

The explanatory variables chosen were the coordinates of the north-south and the east-west axes, relative to the central grid cell, which is assigned explanatory variables (0, 0), see Figure 1. Adding an interaction term, xy, enables the model to fit

a curved surface, not just a plane to the data. For the 25-cell blocks, the terms, x^2 and y^2, were also included in some of the models.

The reporting rate surface for blocks of nine grid cells has, for example, the following possibilities:

- It can increase towards the south if the coefficient of the y-term is positive, and vice versa.
- It can increase towards the east if the coefficient of the x-term is positive, and vice versa.
- It can increase from the north-east to the south-west if the coefficients of both the x-term and the y-term are positive (with other combinations of coefficients possible).
- If, in addition, the coefficient of the xy-term has a non-zero value, curvature is added to this plane.

The possible forms of the surface in blocks of 25 grid cells are more general, especially when the square terms x^2 and y^2 are added. These surfaces are likely to be sufficiently flexible for our purposes.

The linear logistic model for the prediction of π_i in the case of nine-cell blocks will have the form

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 x + \beta_2 y + \beta_3 xy \qquad (12)$$

and for 25-cell blocks will be

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 x + \beta_2 y + \beta_3 xy + \beta_4 x^2 + \beta_5 y^2 \qquad (13)$$

where β_o is a constant term. The four and six β's are the parameters to be estimated from the data for the nine-cell block and the 25-cell block models respectively.

WEIGHTED LEAST SQUARES REGRESSION

It is often possible and necessary to transform the response variable **Y** to another set of observed variables **Z** that do follow the conditions in (3). (Draper & Smith 1981, p.108), so that the transformed variables do have a normal distribution and the errors have constant variance.

If, after the transformation, the error variance is still not constant but known, the regression equations can be further adjusted. Expressed in matrix notation, this is

- $E(\varepsilon) = 0$
- $Var(\varepsilon) = V\sigma^2$
- $\varepsilon \sim N(0, V\sigma^2)$

where **V** is the variance-covariance matrix of the errors. If the entries of **V** are known then a weighted least squares regression can be used so that the maximum likelihood estimate for β becomes:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} \qquad (14)$$

Y must be linearly related to the explanatory variables and in the case of the binomial distribution must be replaced by **Z**, the logistically transformed response

$$Z_i = \ln\left(\frac{r_i}{1 - r_i}\right) \qquad (15)$$

Equation 14 is a single set of matrix multiplications. It provides a possible way to simplify the regression equations usually used in the iterative generalized linear model process (see later section) for binomial data. For the binomial distribution the variance of the observations depend on n_i, the number of checklists collected. Therefore weighted regression is used. The elements of **V** are known or can at least be directly estimated from the observed values. **V** is diagonal because we have assumed that the observations between grid cells are independent.

$$V_{ii} = \frac{\pi_i(1 - \pi_i)}{n_i} \qquad (16)$$

This expression is the variance of the reporting rate in grid cell *i*. The i'th weight is indirectly proportional to the variance. The larger the variance the more unreliable is the data value in the grid cell and the less weight should be attributed to the observation. The matrix of weights, **W**, is diagonal and is the inverse of V

$$\mathbf{W} = \mathbf{V}^{-1} \qquad (17)$$

with

$$w_{ii} = \frac{n_i}{\pi_i(1 - \pi_i)} \qquad (18)$$

These are the weights associated with the observed reporting rates, r_i. Grid cells with a larger number of checklists obtain a larger weight. The π_i's in these equations are not known but can be estimated from the observed reporting rates r_i.

PROBLEMS WITH ZERO AND ONE REPORTING RATES

The above method of weighted least squares regression introduces a problem. In the atlas data, 'zero' and 'one' reporting rates are frequently observed. This causes the weights to be undefined (eq. 18) because 'zero' or 'one' reporting rates cause a division by zero. If these grid cells are excluded from the regression, which effectively means that the weights corresponding to zero or one reporting rates are set to zero, this may cause the matrix ($\mathbf{X}^T\mathbf{W}\mathbf{X}$) to be singular, not invertible, in many cases. 'Zero' and 'one' reporting rates are often observed in areas where few

checklists have been collected (for example in the Karoo and Namibia) and these are also the areas where smoothed reporting rates are most needed.

In Cox & Snell (1989) and Cox (1970) methods to solve this problem are given. These are the topic of the following section.

EMPIRICAL LOGISTIC TRANSFORM

The empirical logistic (logit) transform is presented here in the context of the bird atlas data:

- n_i = the total number of checklists that have been collected for grid cell *i*
- S_i = the number of successes out of the total number of checklists
- r_i = the reporting rate, the observed proportion of successes, S_i / n_i
- π_i = the true but unknown probability of spotting the bird species in grid cell *i*

We wanted to model the probability of success in grid cell *i*. The observed reporting rate r_i can be modelled by a binomial distribution with the following mean and variance:

$$E\left(\frac{S_i}{n_i}\right) = \pi_i$$

$$Var\left(\frac{S_i}{n_i}\right) = \frac{\pi_i(1-\pi_i)}{n_i}$$

In the previous section we said that the logistic transformation

$$\lambda_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \qquad (19)$$

is linearly related to the explanatory variables. Instead

$$Z_i = \ln\left(\frac{S_i}{n_i - S_i}\right) \qquad (20)$$

can be used to estimate the logistic transformation of π (eq. 19), provided the number of successes and the number of failures are not too small (Cox & Snell 1989). Z_i is called the **Empirical Logistic Transform** of π_i. Its distribution is approximately normal with mean

$$E(Z_i) = \lambda = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

and variance

$$\text{Var}(Z_i) = \frac{1}{n_i \pi_i (1 - \pi_i)}$$

Var (Z_i) can be estimated by

$$V = \left(\frac{n_i}{S_i (n_i - S_i)} \right)$$

The estimate Z is biased especially when the number of successes or the number of failures are small. Reduction in bias can be obtained by using a Taylor series expansion on λ (Cox & Snell 1989).

The new estimate for the logistic transform λ will then be

$$Z_i = \ln \left(\frac{S_i + 0.5}{n_i - S_i + 0.5} \right) \qquad (21)$$

with corresponding variance

$$V = \frac{(n_i + 1)(n_i + 2)}{n_i (S_i + 1)(n_i - S_i + 1)} \qquad (22)$$

For our purposes there is one important advantage in the use of these equations (21, 22), namely that undefined weights caused when $S_i = 0$ or $S_i = n_i$ are avoided.

Cox & Snell (1989, p. 32) state that "the above modification of the empirical logistic transform and the associated variance are appropriate if unweighted linear combinations of the transforms are to be used. It can be shown that if a weighted least squares analysis of the transforms is to be used it is preferable to take

$$Z = \ln \left(\frac{S_i - 0.5}{n_i - S_i - 0.5} \right) \qquad (23)$$

with variance

$$V = \frac{n_i - 1}{S_i (n_i - S_i)} \qquad (24)$$

but they give no further reasons. The matrix **W** is the inverse of the diagonal variance matrix **V** such that

$$w_{ii} = \frac{S_i (n_i - S_i)}{n_i - 1} \qquad (25)$$

This would again lead to grid cells with $S_i = 0$ or $S_i = n_i$ having zero weight and therefore these data values would not contribute to the regression. As stated previously, this feature is not what we want. Grid cells with zero and one reporting

rates contain as much information as grid cells with other reporting rates. 'Zero' and 'one' reporting rates should definitely influence the outcome of the regression. Another disadvantage caused by deleting the 'zero' and 'one' reporting rates is that less data are left to estimate the regression parameters.

It is more important that the number of checklists for a grid cell determine which information is important rather than to discard values. It is quite possible to observe a zero reporting rate in a cell for which more than 20 checklists have been collected, even if all surrounding cells have observed reporting rates larger than zero.

Empirical logistic transformations may be used for weighted least squares regression. They have the advantage of not requiring iterative calculations. However, to save computation time is no longer as important as in the early days of computing. Cox & Snell (1989) remarked on this in their later edition. They suggested to rather use the full iterative process which will result in more accurate regression results. But they also commented that the idea of the empirical logistic transform may become useful in some non-standard problems (Cox & Snell 1989).

Initially we thought that even a slight reduction in time and calculation effort could overall save much time. This is because the regression would have to be run over all the grid cells of Southern Africa and this process would have to be repeated for every one of the 900 species. For a species with an extensive distribution this would involve more than 4000 regressions, for example the Cape Turtle Dove *Streptopelia capicola* was recorded in 4111 of the grid cells (Harrison *et al.* 1997a, p.510).

To summarize, the two methods investigated from this section are:

1.) $$Z_i = \ln\left(\frac{S_i + 0.5}{n_i - S_i + 0.5}\right)$$

$$V_i = \frac{(n_i + 1)(n_i + 2)}{n_i (S_i + 1)(n_i - S_i + 1)}$$

2.) $$Z_i = \ln\left(\frac{S_i - 0.5}{n_i - S_i - 0.5}\right)$$

$$V_i = \frac{n_i - 1}{S_i (n_i - S_i)}$$

These are used to find estimates for the β parameters in a single set of matrix multiplication equations

$$\mathbf{b} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Z}$$

The diagonal matrix of weights is $\mathbf{W} = \mathbf{V}^{-1}$.

GENERALIZED LINEAR MODELS

If the distribution of the response variables is not normal but some other distribution of the exponential form, instead of linear regression, **generalized linear models** are used (McCullagh & Nelder 1989). Some transformation of the response is related to a linear function of the explanatory variables.

In the case of the binomial distribution this transformation is the logit of the probabilities (eq.10) as given above. The model aims to predict the probability of success as the explanatory variables change.

We used this theory to investigate whether a few iterations would not improve on the results of the one-step regression methods. Statistical packages, such as GENSTAT, provide functions to perform this operation. The software provided by GENSTAT was too computationally time consuming and it was considered too difficult to manipulate the programs for special cases and modifications.

Generalized linear models are most commonly used with poisson, binomial, exponential, gamma and beta responses. Solving the regression equations involves an iterative process, called 'Iteratively Reweighted Least Squares' (IRWLS) regression (McCullagh & Nelder 1989). This process is the topic of the following section.

ITERATIVELY REWEIGHTED LEAST SQUARES REGRESSION

If the error distribution is not normal but of some other exponential form, such as of the poisson, binomial or exponential distribution, then iteratively reweighted least squares regression should be used instead of the linear least squares regression approach. The theory discussed here can be found in more detail in McCullagh & Nelder (1989). The number of successes out of the total number of checklists for a grid cell in the bird atlas data, meets the requirements for a binomial distribution at least approximately.

The aim is to model the probability of success in grid cell i, π_i, on a set of explanatory variables. The explanatory variables are the north-south and the east-west coordinates relative to the central grid cell.

For linear regression we need some function of the mean π_i (the expected reporting rate), to be linearly related to the explanatory variables. The linear predictor

$$\eta_i = x^t\beta \qquad (26)$$

is the linear function of explanatory variables that will be used in the prediction of π_i.

If the error distribution is binomial it can be shown that the function of the mean, relating it to the linear predictor, is

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = x'\beta \tag{27}$$

This is the canonical **logistic link** and is the most common link function used for the binomial distribution. It is called a link function because it links the mean to the explanatory variables.

Rewriting this, π_i is related to the vector of explanatory variables, in the following way.

$$\pi_i = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \tag{28}$$

The maximum likelihood equations to estimate β are:

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{W} \mathbf{z}$$

$$\mathbf{b} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z} \tag{29}$$

where

- **X** is the design matrix with elements the explanatory variables
- $\mathbf{X}^T$ is the transpose of **X**
- **b** is the maximum likelihood estimate of β.
- **W** is the diagonal matrix of weights

$$w_{ii} = \frac{n_i}{\pi_i(1 - \pi_i)}\left(\frac{\partial \pi_i}{\partial \eta_i}\right)^2 \tag{30}$$

- The z values are the modified dependent variables

$$z_i = \eta_i + \frac{y_i - n_i\pi_i}{n_i}\left(\frac{\partial \eta_i}{\partial \pi_i}\right) \tag{31}$$

where y_i is the response, the observed number of successes Si

The equations (29-31) form an iterative process. This means that w_{ii} and z_i must be recalculated after each iteration using the β estimates from the previous iteration. For these calculations the main step is to find the new estimated π_i values with equation (28) substituting the β's estimated by the previous iteration. This process is called **Iteratively Reweighted Least Squares** Regression.

Because the data has a binomial distribution, we can develop equations (30) and (31) further by finding the exact derivatives.

If

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = g(\pi_i) = \eta_i$$

then

$$\frac{\partial \eta_i}{\partial \pi_i} = \frac{1-\pi_i}{\pi_i}\frac{\partial}{\partial \pi_i}\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$= \frac{1-\pi_i}{\pi_i}\frac{(1-\pi_i)-\pi_i(-1)}{(1-\pi_i)^2}$$

$$= \frac{1}{\pi_i(1-\pi_i)}$$

and

$$\frac{\partial \pi_i}{\partial \eta_i} = \pi_i(1-\pi_i)$$

Then the diagonal elements of **W** will become

$$w_{ii} = n_i \pi_i (1-\pi_i) \qquad (32)$$

All the off-diagonal elements equal zero because we have assumed that there is no correlation between grid cells. The modified dependent variables will become

$$z_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \frac{y_i - n_i \pi_i}{n_i \pi_i (1-\pi_i)} \qquad (33)$$

When the iterative process has converged, the predicted reporting rate for grid cell *i* can be found by

$$\hat{\pi}_i = \frac{e^{x_i^t b}}{1 + e^{x_i^t b}} \qquad (34)$$

To start the regression process, an initial set of weights and dependent variables are required:

$$z_i = \ln\left(\frac{r_i}{1-r_i}\right) \qquad (35)$$

$$w_{ii} = n_i r_i (1-r_i) \qquad (36)$$

are generally used and are found by substituting as initial π_i estimates the observed reporting rates r_i in equations (32) and (33).

Note that if the observed reporting rate in a cell equals 'zero' or 'one' the corresponding weight will equal zero (eq. 30) and the dependent variable *z* will be undefined because of either a division by zero or the logarithm of zero is taken (eq.

35). Earlier we stated that it is desirable to keep 'zero' and 'one' reporting rates as data values instead of discarding this information.

McCullagh & Nelder (1989) suggested that by using a trick for the first iteration of the iterative regression process, the necessary number of iterations until convergence, can be reduced. This involves modifying the dependent variable z. At the same time this helps to solve the problem introduced by 'zero' and 'one' reporting rates. The empirical logistic transform (Cox & Snell 1989) discussed earlier (eq. 21) is ideal for this purpose.

The trick is to allow the modified dependent variables (instead of eq. 35) for the first iteration to be

$$Z = \ln\left(\frac{S+0.5}{n-S+0.5}\right) \tag{37}$$

with corresponding weights the inverse of the variance

$$V = \frac{(n+1)(n+2)}{n(S+1)(n-S+1)} \tag{38}$$

For all the remaining iterations, equations (29) and (32 – 34) are taken.

DECIDING BETWEEN MODELLED AND OBSERVED DATA

If the observed reporting rate in a grid cell differs from the otherwise constant reporting rates observed in neighbouring grid cells, this will be suspicious only if there are few checklists for this grid cell. If there are many checklists for the cell this is strong support that the observed reporting rate is correct. The more checklists there are for any grid cell, the more likely it is that the observed reporting rate is correct, and that the observed value approaches the true reporting rate for that grid cell.

We want a method that ensures that reporting rates which are accurate and reflect the situation in a given grid cell correctly, are not smoothed. For example, a grid cell may have a different habitat to its surrounding cells. This may occur if its area contains a wetland or a forest patch that does not overlap into neighbouring cells. The problem arises not only for a single cell but will also occur if two or three grid cells differ from the rest. If we smooth these true values we may obtain a poorer map than the observed data in the sense that it is farther from the truth.

When the number of records for a cell was large (more than say 20 cards), the regression outcome should have little effect on the final smoothed predicted value. In such a case the regression should not alter the observed reporting rate much. Therefore we considered methods that decide on which of the two values to choose, based on the number of checklists. A weighted average between the observed and the predicted reporting rate is a sensible approach.

The weight function should have the following properties:

- the weights depend on n_i, the number of checklists collected for grid cell *i*. Let f (n_i) denote this function, the weight for the model-predicted reporting rate
- $f(n_i)$ should range between zero and one
- f (n_i) must be defined for all possible values of n_i, theoretically 0, 1, ... , ∞
- if $n_i = 0$ then $f(n_i) = 1$, so that all weight is given to the model
- as $n_i \rightarrow \infty$ more weight should be given to the observations, so that $f(n_i) \rightarrow 0$; for infinitely large n_i, all weight is given to the data.

The following exponential function satisfies these conditions:

$$f(n_i) = \exp(-\alpha n_i) \qquad (39)$$

where α is some value greater than or equal to zero, depending on how strongly the number of checklists collected for the grid cell should influence the weights. α controls the 'steepness' of function $f(n_i)$, see Fig. 4 (p.B7) and Table 6 (p.B6) for the behaviour of $f(n_i)$ for a selection of different α's.

The final smoothed reporting rate is then found by:

$$R_{sm} = f(n_i) * R_{mod} + (1 - f(n_i)) * R_{obs} \qquad (40)$$

where

- R_{sm} is the final smoothed reporting rate
- R_{mod} is the predicted reporting rate found by the regression model
- R_{obs} is the original observed reporting rate

The weights $f(n_i)$ and $(1 - f(n_i))$ sum to one. In the extreme case of α equal to zero ($f(n_i) = 1$ for all n_i) the final smoothed value P_{sm} would equal the reporting rate predicted by the model, regardless of the number of checklists. If α equals one ($f(n_i) = \exp(-n_i)$) the influence of the predicted values from the model decreases rapidly as n_i increases (Fig. 4 & Table 6): in this case, if there is one checklist for a grid cell, the observed reporting rate is given weight 0.63; for two checklists, the weight given to the observed reporting rate is 0.86. For values of α greater than one, there will be little difference between the smoothed distribution and the observed distributions, apart from grid cells with no checklists or one checklist.

The chosen value of α should be species specific. If a species shows continuity of distribution, the value of α should be close to zero, whereas if a species has a highly fragmented distribution, a larger value of α (close to one, say) should be chosen. A range of α values needs to be explored, to decide on the appropriate magnitude for a particular species.

ZERO CUT-OFF POINTS

When drawing the smoothed maps there will always be reporting rates with extremely small values. In the atlas, reporting rates less than 0.02 were represented by a cross. No lower cutoff was needed at which grid cells would be left blank. However, with the predicted reporting rates, a lower cutoff is required, otherwise a species would be recorded as present in grid cells, whenever the predicted reporting rate is strictly positive. Grid cells with very small reporting rates should clearly not be counted as part of the distribution of the species and these cells should rather be left blank.

If the cutoff value below which reporting rates are set to zero is ε, then this is equivalent to setting the number of checklists for the grid cell to $1 / \varepsilon$. We chose $\varepsilon = 0.005$, which is equivalent to 200 checklists. In broad terms, this means that we have shown a species as absent if the predicted reporting rate is so small that the species is expected to be recorded only once in more than 200 checklists.

CHOOSING A REGRESSION METHOD

Initially we considered only regression methods that involve a single step of matrix multiplications, i.e. no iterations. Intuitively these should be considerably faster than using an iterative process. An approximation seemed to be enough, since the reporting rates are presented on maps in shading, each of which represents an interval of reporting rates. At this stage we were only concerned with accuracy to about 5%. We would have chosen the one approach that would have consistently given the closest results to those obtained by the generalized linear model approach (here implemented in GENSTAT). As a comparison we included a smoothing approach which for each central grid cell predicts a new value which is the mean reporting rate in the local area, calculated as a weighted average of the reporting rates

$$\hat{\mu} = \frac{\sum_{R_i>0} R_i n_i}{\sum_{R_i>0} n_i} \qquad (41)$$

The Iteratively Reweighted Least Squares regression procedure, a method which requires several iterations, was also investigated. The methods considered are the following:

1. **NR+0.5**: Normal regression using the first form of the empirical logistic transform, which adds 0.5 to the number of successes for the modified dependent variable.

2. **NR-0.5**: Normal regression using the second form of the empirical logistic transform, which subtracts 0.5 from the number of successes in the modified dependent variables. For this case zero and one observed reporting rates have to be deleted as data points.

3. **WR+0.5**: Weighted regression using the first form of the empirical logistic transform, which adds 0.5 to the number of successes in the weights and the modified dependent variables.

4. **WR-0.5**: Weighted regression using the second form to the empirical logistic transform, which subtracts 0.5 from the number of successes in the weights and the modified dependent variables. Again zero and one reporting rates are excluded from the regression.

5. **W-AVG**: Weighted Average. This is the average reporting rate in the square block of grid cells, where each of the observed reporting rates was weighted by the square root of the number of checklists recorded for the grid cell.

6. **W=n**: This is a variation on WR+0.5. A different set of weights is used, each reporting rate is weighted by the number of checklists from which it is calculated. To include zero and one reporting rates in the regression, 0.5 is added to the number of successes to find the modified dependent variables.

7. **IRWLS**: Iteratively Reweighted Least Squares regression was derived from the generalized linear model theory. A fixed number (five) of iterations were used. For the first iteration the first form of the empirical logistic transform is used, adding +0.5 to the number of successes.

The single step methods WR+0.5 and NR+0.5 were examined on blocks of sizes nine and 25 grid cells. The Method IRWLS was only examined on blocks of nine grid cells.

Results obtained from the GENSTAT software were used as a guideline to what were correct predictions and estimated coefficients. GENSTAT results were obtained for blocks of size nine and for blocks of size 25. For the 25 block case, two different parameter models were investigated and compared: a.) The model included the additional square terms x^2 and y^2 and b.) Only coefficients for the terms x, y and xy were estimated.

- MODEL CHECKING

An approximate goodness-of-fit test for the model in the case of generalized linear models is the **scaled deviance** (GENSTAT 5 Committee 1993). This is defined to be twice the difference between the maximal possible likelihood and the likelihood of the fitted model (McCullagh & Nelder 1989) and (GENSTAT 5 Committee 1993). The scaled deviance has approximately a χ^2 distribution with d degrees of freedom where d is the residual degrees of freedom (five in the case of nine-grid cell blocks where terms x, y and xy are fitted). This approximation is only good for a large number of observations and when not many extreme observations occur. In the bird atlas data, for a regression on nine grid cells the number of data points is not large compared to the number of parameters fitted and extreme reporting rates 'zero' and 'one' occur frequently.

The change in deviance when terms are added and subtracted from the model is usually a better indication of whether the model has improved or not.

OVERDISPERSION

Overdispersion is caused by clusters of entities which are more similar to each other than to other units (McCullagh & Nelder 1989). In the case of the bird atlas data, the probability of success in a grid cell is only constant if each observer spends the same effort and time observing and has the same skill of identifying birds and does not communicate with other observers on which species are present, especially rare birds. But the way in which the reporting rates were gathered, the same observer could have collected several checklists for the same grid cell. Checklists collected in one particular month or season are also more likely to have more similar probabilities of success than checklists collected in another season, meaning that the checklists within a grid cell are not entirely independent. All these factors were ignored when assuming that the reporting rate data can be modelled using a binomial distribution. McCullagh & Nelder (1989) remarked however that overdispersion can be expected in many practical situations. The effect of overdispersion is that the variance of the response, the number of successes in the case of the binomial distribution, is higher than would be expected from the theoretical binomial variance $[n\pi(1 - \pi)]$. The variance of the response Y can instead be described by

$$Var(Y) = \sigma^2 n\pi(1 - \pi)$$

where σ^2 is the dispersion parameter. The dispersion parameter can be incorporated into the theory as a constant without much change. If the dispersion parameter does not equal 'one', the scaled deviance equals the residual deviance divided by the dispersion parameter.

The problem is the estimation of the dispersion parameter. A wrong model, missing explanatory variables, outliers, the wrong link function and overdispersion all have the same effect of a large residual mean deviance. The dispersion parameter is therefore estimated from the model that is believed to contain all possible explanatory variables which could be responsible for the changes in response.

GENSTAT estimates the dispersion parameter from the residual mean deviance if it is not fixed at 'one' or some other value. The residual mean deviance is the residual deviance divided by its degrees of freedom. If the dispersion parameter is not fixed at 'one' the standard error of the estimated coefficients are multiplied by the estimated dispersion parameter. This will reduce the 'significance' of the parameters or in the case of underdispersion increase the significance.

In the atlas data, overdispersion is to be expected. It is however not clear to what extent this is caused by a lack of some explanatory terms in the model, for example the square terms x^2 and y^2 in the nine grid cell models. Therefore caution is needed when interpreting the residual mean deviances and the significance of the estimated coefficients, both in the case of the t-value and the change in deviance. The estimated dispersion parameter (if not fixed at 'one') would change from regression to regression on blocks of nine grid cells when for a single species the true dispersion

parameter could be expected to vary only slightly. In some regressions the dispersion parameter will even be smaller than one, suggesting underdispersion. This will be the case if the explanatory variables explain more of the variation than expected for binomial data.

We decided to fix the dispersion parameter at 'one' in each case and rather let a large residual mean deviance be interpreted by as partly overdispersion and partly missing terms if the reporting rates have a specific configuration which may not be completely explainable by the available terms x, y and xy. The residual mean deviance is probably also not a good estimate for σ^2 if little data is available and in the case of many extreme observations. The residual mean deviance for a block of nine grid cells depends strongly on how well the explanatory variables can explain the trend and variation of the observed reporting rates. The residual mean deviance in this case is therefore more an indication of the model fit than of overdispersion. σ^2 would have to be estimated from a model with more explanatory variables than x, y and xy. The estimation of σ^2 is not necessary for our purpose, because the predicted values do not change when the dispersion parameter changes.

The dispersion parameters were however estimated for some selected regressions and are shown in Tables 1 - 4.

We trusted that the generalized linear model output obtained by the GENSTAT statistical package (GENSTAT 5 Committee 1993) is the best possible answer to be achieved for any regression. These generalized linear model results were taken as a standard to compare other regression results against.

The regression process had to be generalized and automized. It is impossible to consider special cases and choose only the significant terms. Therefore the regression output estimates coefficients for x, y, and the interaction term xy, and the constant term; whether they are significant or not.

COMPUTER PROGRAMS

There are good reasons for not using the GENSTAT package to perform the regressions for all cells of a species:

- GENSTAT requires constant user interaction while running. This means that only a single regression can be achieved at any time, although with highly advanced programs the whole process could probably be achieved in a single step. A C++ (Borland International 1992) program will be able to run the whole process in one continuous step, from finding a block of nine or 25 grid cells, to predicting and saving a value and carrying on to the next block of grid cells.

- User written programs are considerably faster.

- It is easier to manipulate output and input formats with C++ than in GENSTAT.

- Problems with the predefined method in GENSTAT were encountered. It is easier to customize the programs for special cases that have to be considered.

RESULTS

Figures and tables for this chapter can be found in Appendix B.

COMPARISON OF REGRESSION METHODS

TABLE 1: MASKED WEAVER *Ploceus velatus*

To compare outcomes between different smoothing methods 14 grid cells were selected from the distribution of the Masked Weaver (see Figure A3, Appendix). The results are shown in Table 1, where the cases are labelled from (a) to (p). To assist in the understanding of the results, the original shades, as they appear in the atlas map, are shown on the left hand side of Table 1. The 'comments' column contains the a priori expectation of the form the estimated coefficients of the explanatory variables should take, if anything can be said at all. These suggestions were formed by inspection of the block of nine cells for the case.

As a first step towards finding a best smoothing method, generalized linear models results, here implemented by the GENSTAT package, were taken as a guideline to the ideal regression results. The shaded rows in each table cell show the predicted reporting rate for each grid cell with each method. The values preceding the shaded row are the estimated coefficients for the constant term and the x, y and xy terms respectively.

The + 0.5 Methods

The methods which involve adding 0.5 to the data values both have advantages and disadvantages. One of the advantages is that these methods produce results even when GENSTAT fails to do so. For the blank area (Table 1, f) the predicted value must be zero but these methods predict a reporting rate of 7 %. Therefore, for all blank blocks the predicted values must be explicitly set to zero for the +0.5 methods. In the last two examples of Table 1, (n and p) the smoothed reporting rates must not necessarily equal zero. The GENSTAT program gives warning messages that the predicted values are not to be trusted. The three Methods WR+0.5, NR+0.5 and W = n, produce results for these two cells. In both cases (n and p), the predictions would produce a light shade (2% to 24.9%) for the central cell. In example (n), reporting rates of 12.6%, 12.3% and 11% are predicted by the three Methods. For example (p), the predicted values are 15.8% and 15.3%. These predictions may be too large, but it is more unlikely that the predicted values should equal zero.

For Method NR+0.5 the predictions are within 3% of the GENSTAT values, except in (e) where the difference between the predictions is 5% and in (i) where the difference is 8%. This method seems to work adequately, but does have faults. The signs of the parameters are not always the same as those for the parameters estimated by GENSTAT. For example in (b) the interaction term changes from 0.00 in GENSTAT

to -0.19 in Method NR+0.5, and in (c) the interaction term changes from 0.69 to -0.12. In the latter cell this difference cannot be caused by the fact that the interaction term should not be in the model, because the GENSTAT-derived 't-value' was larger than two. On the other hand, many of the predictions are within 1% of the GENSTAT predictions, for example (a), (b), (g), (k) and (m).

There are on average larger differences between Method WR+0.5 and the GENSTAT values, but all predictions are within 6% of the GENSTAT predictions. The signs of the coefficients also differ for some of the selected grid cells, for example in (e), (g), and (h). But in all cases the signs of the coefficients that are significant ($t > 2$) in GENSTAT are the same in the +0.5 methods results.

Method W = n produces similar results to Methods NR+0.5 and WR+0.5. The only difference between Method W=n and Method WR+0.5 is that a different set of weights was used. It is difficult to judge which of these three methods produces the best results.

Methods WR-0.5 and NR-0.5 perform little worse than the '+0.5' methods already investigated. In (c) the differences in predicted values to those from GENSTAT are 12.5% and 8%, in (d) there is a difference of 7.5% and in this example the y-term is also changed from -0.84 in GENSTAT to 0.24 and 0.23 in Methods WR-0.5 and NR-0.5, respectively.

With Method WR-0.5 more unusual wrong results occur. For (n) the predicted reporting rate equals 100%, which clearly is wrong. The estimated values for the parameters are also abnormal, the estimated coefficient for the interaction term equals 15.99. For Method WR+0.5 the corresponding estimate was -0.64. The –0.5 methods only have three data values to work with in the case of (n) because all cells with zero and one reporting rates were deleted. This is because the weights for these cells would otherwise be undefined.

Of the five single step regression methods mentioned so far, Method W = n most consistently produces results closest to those of GENSTAT. The signs of the estimated coefficients also stay the same more often, for example (e) and (h).

Provided that enough data (five or more strictly positive values) are present, the IRWLS (iteratively reweighted least squares regression) Method converged at the latest after four iterations for the blocks of the Masked Weaver distribution that were chosen for Table 1. With fewer positive values than five, a final result was always achieved with the seventh iteration. It is interesting to see that for the blocks of cells that go to seven iterations, the final predicted reporting rates are always zero in Table 1. This implies that if the values have not converged after four or five iterations, the predicted value is most likely equal to zero. But there are also cases where only four cells had positive reporting rates and where the final predicted value is larger than zero and the process converged after four iterations (Table 2 a, e, j and k). The value of the smoothed reporting rate most importantly depends on the configuration of the reporting rates in the block of nine grid cells.

For Method IRWLS, after a maximum of seven iterations but in more of the cases after three iterations e.g. Table 1(a), (b), (e) and (j) to (m), the results are exactly the

same as in GENSTAT to two decimal places, for example (c), and usually to more decimal places. One advantage of Method IRWLS is that it calculates results where GENSTAT does not, for example in (n) and (p).

TABLE 2: CAPE WEAVER *Ploceus capensis*

The results obtained by one-step methods in Table 2 for selected cells from the distribution of the Cape Weaver (see Figure A1, Appendix) contain larger deviations from the GENSTAT results than were obtained for the Masked Weaver in Table 1. This may make it easier to choose between them.

In grid cell (b) there are 12.9% and 14% differences between the GENSTAT prediction and the predictions from Methods NR+0.5 and WR+0.5, respectively. In (c) these differences are 9.7% and 10.4%, respectively. The signs of significant (GENSTAT-derived) values remain the same but signs of the non-significant parameters may change, for example in (b), (d) and (e). The two methods that produce the most similar predictions and estimates of coefficients to those of GENSTAT, are Methods W=n and WR-0.5. In cells (c), (d), (e) and (h) the predictions from Method W=n are most similar. In cells (b), (f), (a) and (j) the predictions from Method WR-0.5 are more similar to the GENSTAT predictions. However in the last two cases (a) and (j), there are large differences in the coefficient estimates for the interaction term between GENSTAT and Method WR-0.5.

As was the case in Table 1, Methods WR-0.5 and NR-0.5 produce wrong results. For example in (c) the significant y-term changes from -1.91 in GENSTAT to 0.46 in WR-0.5 and the predicted value of this method is 40.6% while the predicted reporting rate from GENSTAT was 5.1%. For Method NR-0.5 no results could be calculated for this block of grid cells. There is a similar problem in (g) where the predicted value from Method WR-0.5 is 50%, but from the block of grid cells on the left it can be seen that this is wrong. There are five positive data points in this block. Five or fewer positive data points in a block of nine grid cells occur frequently throughout the atlas distributions; it would be an undesirable feature if a method cannot produce a reliable result for such blocks. In block (e) none of the single step methods pick up the significant negative xy term sufficiently. For the –0.5 Methods the interaction terms is even positive and this has the result that these two methods produce a too large predicted reporting rate. The predictions by the other single-step Methods are also too large. The predicted reporting rate from GENSTAT is 0.056. The closest value to this produced by the single-step methods is 0.118 from Method W=n.

For cases (j) and (k) four of the nine grid cells have strictly positive observed reporting rates. This shows that it will not always be the case that few (less than five) positive data values produce a predicted reporting rate of zero. Also if the observed reporting rate of the central grid cell was zero and only four of the other grid cells had positive reporting rates, as is the case in Table 2 (e), a sensible regression can be found. It will not be very accurate in the one-step methods, but the predicted values are still within 5% or 6% of the predictions from GENSTAT, see blocks (a), (c), and (e). For these same cases the GENSTAT predictions are not accurate either. For (a) the prediction with its approximate standard error is 0.047 ± 0.026 for (c) 0.051 ± 0.033 and for (e) 0.056 ± 0.036. This shows that the deviations of values from the

results obtained by GENSTAT should not be taken as absolute answers but should also be directly compared to the original shaded blocks.

There was no single step method that was consistently 'better' than the others. How well the results compare to those obtained by GENSTAT also seems to depend on the magnitude of the predicted value for the particular cell. For larger predicted values, the results from all methods are more similar. This may explain why the results in Table 1 for the Masked Weaver, which has a larger average reporting rate than the Cape Weaver, are more consistent with the results of GENSTAT than the results in Table 2 for the Cape Weaver. For the Cape Weaver the results often differ by a multiple of two or more, e.g. (a), (b), (c) and (e) in Table 2. For the Masked Weaver the results do not differ to this degree from the GENSTAT results. This makes it more difficult to compare between the one step methods. However, if a method from the five single-step methods has to be chosen, we would suggest the W=n method, which most often produces the nearest predicted values and parameter estimates to those from GENSTAT.

TABLE 1 AND TABLE 2: GENSTAT

If by inspecting the shaded blocks of grid cells no possible surface is obvious, which may happen if there is large variation in the shading, then this is reflected well in the GENSTAT outcomes in that the constant term is significant (topmost value in each table cell), for example Table 1 (g), (h), (i) and Table 2 (a), (k). In the blocks where we could predict the significant parameters, the GENSTAT results confirm this, for example Table 2 (b) and Table 1 (j) and (m).

In both Tables 1 and 2 there are no results produced by GENSTAT for which it can be said that the predicted values are unreasonable. In each of the cells the resulting smoothed value can be explained by the surrounding shades. This inspection of the validity of the smoothed values is simplified by that there are only nine grid cells used. The only problem caused by the GENSTAT approach is that the program does not run for certain configurations of observed reporting rates. These cases can however be supplemented with the results obtained by the Method IRWLS.

Of the methods investigated, IRWLS using a fixed number of five iterations is definitely the preferred. To make a final decision on the validity of the model and to decide whether the GENSTAT results are good in the first place, another way of checking the model is necessary.

One way of checking the results is to compare the resulting smoothed maps from each method to the original atlas map.

SIZE OF BLOCKS

It is more difficult to decide what the form of the parameters should be when working with blocks of 25 grid cells. Simple patterns are replaced by increased variability (see the shaded blocks to the left of Tables 3 and 4). These have the same central cells as

were used in Tables 1 and 2. This is to simplify comparisons between these tables. It is also more difficult to assign a priori expectations to coefficients because their meaning in a larger data set is more complex, especially for the square terms x^2 and y^2. It is possible to compare the GENSTAT predictions from nine-cell blocks to those from 25-cell blocks because the immediate local pattern is contained in the larger blocks.

Inspection of the blocks of 25 grid cells reveals that these do not add any more information about what the central value could be, but rather confuse the image and add extra variation, for example in Table 3 (c), (d) and Table 4 (a). It appears to be more sensible to predict the central value by just considering the immediate neighbouring cells.

The outermost 16 cells are more than 25 kms distant from any point of the central cell. This physically makes it doubtful whether they should have any influence on the predicted value of the central grid cell. Because of this decreasing importance with increasing distance the kriging algorithm, touched on briefly in the previous chapter, would become more relevant in this situation, more so if the size of the blocks would increase further.

When using blocks of 25 grid cells the regression process becomes more complex to calculate. A 6x6 matrix has to be inverted, for the nine-cell blocks this matrix was only 4x4. From Tables 3 and 4 it becomes clear that if a fixed number of terms has to be chosen, the square terms x^2 and y^2 have to be left in the model. In Table 4 (a) the nine-cell model predicts a reporting rate of 0.047. The value from the 25-cell model, including square terms is 0.017 and excluding them is 0.244. This last prediction is 14 fold that of the prediction from the nine-cell model. Here both of the square terms were significant. When removing the square terms from the model the approximate standard error for the prediction increases from 0.009 to 0.018. The prediction obtained from the model without square terms would produce a medium shade (14.3% to 39.3%), which is unacceptable. In Table 4, block (c), the predicted values are 0.051, 0.130 and 0.201 respectively. The prediction error decreases although both of the square terms were significant.

A particular example is the one in Table 4 block (e). The blank central patch is a section between two mountain ranges. Therefore one would think that the data for the central cell is correct as it is (zero observations of the Cape Weaver in six checklists). Comparing the results of the three different models, we find that the nine-cell model gives the answer closest to a zero predicted reporting rate: 0.056. The predicted reporting rate for the 25-cell model including square terms is 0.147 which is conspicuously too high and for the model excluding square terms is 0.202. There is too much smoothing in the latter two models. One would anticipate that the interaction term would become more significant when more data confirm a gap in the distribution but this does not happen in this case. It seems that there is more variation in these large blocks than the models can handle and this causes local features to be smoothed out.

In Table 4 (g) the prediction by the IRWLS method was 0.011. The GENSTAT program did not run for this block. The predictions from the 25-cell models were 0.110 and 0.127 for the model with and without square terms respectively. Here it is

difficult to say whether the nine-cell prediction is too small or the 25-cell predictions are too large.

Table 3 (c) is another example where it is difficult to say what the final smoothed reporting rate should be. Both square terms are significant. The observed value was 0.250. The nine-cell model gives the smallest predicted value, 0.209. But in general the additional cells only add confusion. The example of Table 3 (d) is still more variable than in (c) and it is even more difficult to judge on the predicted values.

In the case of the Cape Weaver (Table 4), which has the lower average reporting rate, the difference in the predicted values between the nine-cell and the 25-cell models is often more than 5% (Table 3 c, d and Table 4 c, e, f, g) and sometimes larger than 10% (Table 4 b).

All this seems to suggest that it is not worth including another two parameters and thereby complicating the model only to accommodate the considerable extra variation caused by the additional cells. The better way seems to choose the block size as nine grid cells and leave the model as simple as possible and as physically valid as possible.

If a decision has to made on which terms to include in the model it is not difficult to see that the square terms should be added when the block size is increased to 25. Although at this block size it is more difficult to find clear patterns, except at edges of a distribution (e.g. Table 4 f, g) or where large geographical landmarks are present (e.g. Table 4 e), the differences between the models when including and excluding the square terms becomes larger the more pattern is present. Patterns in the shades means that some of the terms will become significant, otherwise only the constant term is significant.

What does it mean that the prediction errors for the model without the square terms are almost always smaller than for the model with the square term? If fewer terms are included in a model that will often cause the constant term to become more significant. The predictions are only based on this constant term and the approximate standard errors for the prediction are only based on the standard error of the constant coefficient. With increasing significance of the constant term, its standard error often decreases and so with it the standard error of the predicted reporting rate. Not too much importance therefore should be given to these standard errors.

The last two methods considered in Tables 3 and 4 (Methods WR+0.5 and NR+0.5) can only be compared to the 25 size blocks without the square terms (column 6) and were merely included to investigate how they perform with more information. For the Cape Weaver (Table 4) the predictions produced by the Weighted Regression model WR+0.5 are consistently larger than those of the GENSTAT model. The values from the Normal Regression (NR+0.5) are closer to the GENSTAT values but are very poor if the square terms of the model were significant (Table 4 a, d), the difference between predicted values was 9.4% and 15.9%. Again considering the three blocks where it is easy to identify a trend (Table 4 e, f, g) these two models (WR+0.5 and NR+0.5) show the worst performance of all. These models change zero reporting rates too much, caused by the addition of 0.5 to the number of successes and the number of failures. This negative effect can only be adjusted with further iterations.

The aim is to fit the best possible surface, independent of whether there are too many terms in relation to available data points or not.

In summary this seems to indicate that blocks with only nine grid cells are the better choice for predicting the immediate trend in the area. They are not excessively influenced by further surrounding information that takes away the concentration from the centre, the grid cell of primary interest.

THE SMOOTHED DISTRIBUTION MAPS

COMPARISON OF THE CAPE WEAVER MAPS FROM DIFFERENT SMOOTHING METHODS

The atlas map of the observed reporting rates is shown in Fig. A1 (App.). NR+0.5 appears to smooth the original distribution most heavily, followed by WR+0.5 and then W-AVG, (see Figs. 4, 5, 7). Maps in which larger areas with the same shade occur have a more smoothed appearance. When comparing the smoothed maps from WR+0.5 and NR+0.5 (Figs. 4, 5) the stronger smoothing in the NR+0.5 map can be observed in the following areas: 31°S 28°E, where NR+0.5 has a larger area of the lightest shade, 3226DD where NR+0.5 has a more continuous dark area with only one cell left that is of a lighter shade, 2820CB where the original observed reporting rate was 1 / 108 checklists.

The observed reporting rate in grid cell 2820CB was one success out of 108 checklists. This clearly is a stray observation and any smoothing method should eliminate this value. In Method NR+0.5, however, this observed reporting rate and that of all the surrounding eight cells is increased so that the smoothed reporting rates are larger than two percent. WR+0.5 has the smoothed values of surrounding cells increased to more than two percent, the predicted reporting rate in the central cell is still less than two percent.

IRWLS is the Method that does the least amount of smoothing. Here the features of the original data are best preserved (Fig. 8), especially in the block of grid cells in the Karoo, the area defined by 30° S to 33° S and 20° E to 24° E.

One area where IRWLS Method seems to smooth the data too heavily is in the area 31°S and 28°E (Fig. 8 and Fig. A1, App.). In the original data zero reporting rates occur in what seems not a random pattern. It seems as if these should have been preserved in the modelling. In the area of 28°S, 30°E the original data shows some zero reporting rates, but these are also smoothed in the IRWLS method. Using a weighted average between the observed reporting rate and the model-predicted reporting rate improves these situations slightly but removes the smoothed appearance of the rest of the distribution of the Cape Weaver (Figs 9-14). From these Figures it also appears that the value of α should have a value between 0 and 0.05 (Figs 8, 9), so that only the reliable observations have an impact in this adjustment of the smoothed maps.

Method IRWLS will delete isolated single values, e.g. the data of cells 2820CB and 2819CC (Fig. 8). The respective reporting rates in these cells were one out of 108 and one out of 19. Smoothing methods, that take into account the information of surrounding cells, should eliminate these isolated small values. It is likely that the species generally does not occur in this cell, except as a vagrant, out of its range. Exceptions to this are isolated cells with special environmental, habitat features. The IRWLS process can not do much about this situation, neither can any other smoothing method. The weighted average between the model-predicted and the observed reporting rate will insert isolated observations back into the distribution maps almost immediately (compare grid cells 2820CB and 2819CC in Figs 8 & 9).

The IRWLS is certainly more appealing to the eye than the original atlas map. The discontinuity has been removed. This has the effect that individual features are more striking for example the decline in the reporting rates in the area 3226DA, and the edges of the distribution are more defined.

There are some peculiarities that are due to the setup of the model. In 2828AD the shade of this particular cell was not originally dark but because of the surrounding the predicted reporting rate is larger than the observed, while the predicted reporting rates of the two cells to the north and east of the central cell are smaller than the observed reporting rates. This is not a fault in the calculation of this predicted value and may not be a fault at all, but is caused by the choice of the explanatory variables the x and y coordinates relative to the central grid cell.

DISTRIBUTION MAPS OF THE BLACKHEADED CANARY *Serinus alario*

The methods, appearing in the order of their degree of smoothing are: NR+0.5, WR+0.5, W=n (nearly the same as WR+0.5), WAVG, IRWLS (Figs 15-18)..

The three single step regression methods (WR+0.5, NR+0.5, W = n) produce bad side effects in the Namibian part of these smoothed distribution maps (Figs 15-17). These are caused by that only few of these grid cells have non-zero reporting rates. The one-step regression methods spread out any observed reporting rate, mainly caused by adding a 0.5 to the dependent variables. The empirical logistic transformation is useful to overcome the problem of zero and one reporting rates, but only if further iterations remove disadvantages such as the wrong predictions in southern Namibia in Figs 15 and 16.

The W+0.5 regression method is at the same time the first step in the iterative process of the generalized linear models approach. With further iterations this negative aspect of the single step methods which add 0.5, disappears. In the fifth iteration only the smoothing that can be justified remains. The maps in this area is similar to the original data map, with the addition of some crosses and that the shades are lighter.

This suggests, that if only few cells in an area have checklists, the smoothing process is not of any use and it is questionable if smoothing can be justified at all. This also can be related to the continuity that was discussed in the previous chapter, which influences the dependence of values of nearby cells. If this spatial correlation falls

away, any smoothing method that does not preserve this patchy appearance is not valid.

Comparing the smoothed map from the IRWLS process (Fig. 19), with the atlas map (Fig. A11) it appears that the core parts of the Blackheaded Canary distribution are better defined than in the original map.

The two patches of zero reporting rate grid cells, 3022BD to 3023AC and 2918DB to 2919CC in the distribution of the Blackheaded Canary (Fig. A11) are not kept in the one step methods but reappear in the IRWLS Method, especially in the fifth iteration (Fig. 19) and see Figs 19 & 26-28 for iterations 1-4. The map produced by the IRWLS approach is more similar to the original data values than any of the single-step Methods.

DETAILED DISCUSSION OF A PARTICULAR SCENARIO

Fig. 3 and Table 5 (both on p.B5) describe a situation which was encountered when smoothing the distribution of the Blackheaded Canary (see also Fig. A11, App.). The smoothed map shows two striking gaps in the distributions at 3022BB and 2918DB which are present in the original distribution but more obvious in the smoothed version. The first of these two blocks is discussed here and its surrounding area is shown in Fig. 3.

The number of checklists in the blank area ranges from six to 11. This number is high enough to suggest an underlying environmental feature but not large enough to be certain. If seven iterations would be used, all of the originally blank cells would have predicted values of zero. This is a reasonable conclusion considering the size of the blank block and that the central cell is always surrounded by at least three more cells with zero observed reporting rates. A trend can be found in each of the blocks of nine grid cells.

The C++ results have not converged after five iterations but this process is more controlled than the GENSTAT results. If all the estimated coefficients are of approximately equal magnitudes this is usually an indication of a problem in the regression calculation process.

Are the estimated predictions wrong? GENSTAT predicts a smoothed value of zero for each of the four cells. After five iterations of the Method IRWLS all predictions were smaller than 2% (illustrated as crosses on the maps) and after six iterations the value for 3023AC was zero, 3022BB and 3022BD both had a predicted value of 0.006 and 3023AA had a predicted value of 0.004 which would be set to zero (after inserting the zero cutoff limit of $\varepsilon = 0.005$). For 3022BD the interaction term should be positive (see Fig. 3) because the top left and bottom right cells have larger reporting rates than their neighbours, y should be positive as the reporting rates increase towards the south and the outcome of the x term is probably dominated by the cells 3022BA and 3022BC on the left and the zero reporting rates to the east of these. The signs of the estimated coefficients are correct although the magnitudes are suspicious. The nonconvergence pointed at in the GENSTAT results after two iterations also means that after a fixed number of iterations the process has not converged to zero, but would if more iterations were allowed which is shown in the

last column. These final GENSTAT results however also warn that the "iterative weights have become zero or have been held at a limit".

We conclude that the predicted values in all four of these cells should finally equal zero. There is however no harm if the number of iterations is fixed and the prediction has not approached zero yet. For our purposes this might rather be an additional advantage because there is a chance that the species is present but has not been observed, also because the number of checklists is not large and if one wanted to expect a more gradual change over from zero to positive reporting rates. This however partly shows that certain patterns can be explained well by this choice of spatial explanatory variables.

NUMBER OF ITERATIONS REQUIRED

In the IRWLS approach the level of smoothing decreases with further iterations after WR+0.5, which is equivalent to the first iteration we have been using for the IRWLS approach (Figs 16, 19, 26-28). Fig. 19 shows the smoothed map produced by Method IRWLS after five iterations ($\alpha = 0$ means that the values shown are the pure model-predicted reporting rates). The core of the distribution is established with approximately the third or fourth iteration (Figs 27 & 28), but along the edges of the distribution too heavy smoothing still occurrs. The original distribution (Fig. A11) is enlarged by many crosses around the edges (predicted reporting rates less than 2%) and this smoothed map after the fourth iteration still has an unpolished appearance (Fig. 28). This effect disappears during the fifth iteration (Fig. 19). More of the predicted reporting rates that were less than 2% in the fourth iteration equal zero in the fifth. The predicted values which are larger than 2% are established in the fifth iteration and do not change with another iteration (Fig. 19), only a few of the small predicted values (less than 2%, shown as crosses) became less than the zero cutoff point of 0.005 used here. Predictions less than 0.005 are set to zero in the smoothed maps. The predictions for isolated reporting rates with an increasing number of iterations match more those of the raw reporting rates (Fig. A11), except that they may be of lighter shades (2716BD, 2719BC / BD, 2722DB, 2729CD), because more of the small reporting rates are set to zero in the fifth iteration. Some of the remaining crosses (smaller than 0.02 predictions) may disappear with a sixth or seventh iteration but this was not considered necessary, no changes will be made with more than five iterations to the predictions which are larger than 2%.

WEIGHTING OF MODELLED AND OBSERVED REPORTING RATES

When using $\alpha = 0.05$ (see eqs 39 & 40) the smoothed distribution maps (Fig. 20) still are 'smooth' compared to the raw data (Fig. A11). Blank grid cells in the core of the distributions remain to be absent. There are, however, more patches, the map has a more separated appearance. Single grid cells are less connected by equal shades. The choice of α has an effect on the spatial autocorrelation of the resulting map. With

larger choices of α (closer to one) the maps will increasingly more resemble the original atlas map. With a choice of $\alpha = 0.1$ and larger, the smoothed distribution maps increasingly resemble the raw data maps (Fig. A11). There is a rapid decline in the smoothness of the distribution maps if a weigthed average between the model-predicted and the observed reporting rates is used. For this also see the distribution maps for the Cape Weaver (Figs 6-12 and Fig. A1, App.), for the Cape Siskin (Figs 30 to 36 and Fig. A6) and for the Dusky Sunbird (Figs 37-43 and Fig. A28). The Cape Siskin was included to provide an example of a distribution which does not exhibit much spatial correlation and which is fairly small compared to the other distribution maps shown here.

ZERO CUT-OFF POINTS

Fig. 29 shows the smoothed distribution map of the Blackheaded Canary if instead of 0.005 a cutoff limit of 0.001 is used below which the predicted reporting rates are set to zero (compare to Fig. 19, where the 0.005 cutoff limit was used).

The model fits a curved surface to the data, very often at the edges of the distribution there will be a smoothing out effect, so that all around the edges previously zero values will have predicted values that are still not high but not zero. The way the maps are drawn is that for very small reporting rates (< 2%) crosses are drawn. But if the data can be trusted, as we assume, the values outside of the observed distribution should be kept at zero and not extend the distribution. Often these predicted values, if the model worked well, are so small that they are not meaningful as reporting rates. For example a reporting rate of 0.0001 means that on average once in a 10 000 tries will this species be spotted. In terms of presence and absence of a species this means that the species does not occur in this gridcell and such a reporting rate should be read as zero.

DISCUSSION

KRIGING versus GENERALIZED LINEAR MODELS

In the case of the bird atlas distribution maps there are two reasons for smoothing distributions. The one is to remove measurement error which is caused by binomial sampling variance and depends mainly on the number of checklists that have been collected for each of the grid cells. The effect of the measurement error is larger in areas with grid cells for which few checklists have been collected. These areas will appear to have a less smooth distribution of the species than areas with generally a large number of checklists per grid cell. The other reason for smoothing is to provide more general maps for reference.

All of the bird distribution maps contain measurement error, but not all distributions show a strong correlation between the reporting rates of adjacent grid cells. The degree of smoothing should therefore primarily depend on an estimated value of the degree of dependence between values present. This value was obtained from the variogram developed in the previous chapter, $2\gamma(1)$, which is the variogram value at the smallest lag and represents the correlation between immediately adjacent grid cells.

The weights developed in the kriging prediction equations do not determine the degree of smoothing but only the relative weights that should be given to each of the observations according to their correlation to the location for which a value is to be found. In the case of no autocorrelation in the distribution, all values will be given equal weight to predict the location of interest.

McNeill (1991, 1994) developed a method where the number of checklists influence these weights, otherwise no kriging method exists which considers the reliability of individual observations.

In a kriging approach to smoothing, the main problem is that of trend removal. The median polish method is not local enough for our purposes. The data show so many local minima and maxima, that this approach would be likely to smooth too heavily. Median-polish trend removal also requires the data to be symmetric. The bird atlas data contains too many zero and small reporting rates to be symmetric. A transformation, such as the Freeman-Tukey square root transformation (Cressie 1991) would have to be used. This again introduces the problem of which of the zero reporting rates to keep and which of them to leave out of the analysis. The Freeman-Tukey transformation also does not remove the dependence of the variance of the observations on the n_i, the number of checklists collected for grid cell *i*.

There is no doubt that a local approach is more justified, for several reasons, even if it takes more computational time. There is much underlying geographical and environmental change so that any kind of global approach, aiming to fit a global surface would results in too much averaging and too little consideration of the true causes in the changing values. The structures of the bird distributions are in general

too different in different areas, so that the relationship to the covariates is not constant but changes and the data should rather be split up into subregions, which in itself is a complicated process.

For a local approach fewer assumptions need to be made, for example, that the process is stationary in the mean and variance over the entire area. The matrices that have to be inverted are smaller, although more matrices have to be inverted. If the spatial autocorrelation is strongest only at small lags there is no gain in including more than the necessary surrounding values. With a local approach there is more guarantee that existing features are not smoothed out but specifically contribute to the estimation.

It is more likely than not that the environmental variables cause the trend. Because the set of explanatory variables would be far too many to include in a 3x3 grid cell block regression, it is assumed that a combination of directional explanatory variables has the same effect as all the environmental variables combined.

The previously known methods of kriging did not pay much attention to measurement error in the data and if they did they did not consider cases where this measurement error is not constant but different in each observation. McNeill developed an approach which aims to subtract the measurement error during the estimation of the variogram. The kriging equations then aim to predict or smooth the reporting rates without these measurement errors.

In the generalized linear models approach, to control the degree of smoothing, either different block sizes can be used or the number of explanatory variables can be increased or decreased. We suggest however that blocks for smoothing the bird atlas data should not contain more than nine grid cells and no other explanatory variables beside the x, y and xy terms. The $2\gamma(1)$ value could however be used as a judgement of how much a final smoothed value should consist of the model-smoothed value and how much of the original data. With little spatial correlation in a particular distribution, too heavily smoothed reporting should be prevented.

For the generalized linear models approach no assumptions about the normality, stationarity and symmetry of the data have to be made, in contrast the only assumption is that the data follow a binomial distribution. With more assumptions being made, there is more possibility that specific data sets do not following these assumptions. Each data set would therefore have to be specifically inspected and adjusted prior to smoothing.

McNeill (1994) discussed available methods to smooth and interpolate spatial data. A section in Cressie (1991) is also dedicated to such methods. There are many complications with the case of binomial data in the kriging approach. Many assumptions on the form of the data have to be made, for example that the data are symmetric and stationary. If a trend is fitted parameters must often be determined. Global approaches keep these parameters constant once estimated.

The next problem is that of removing trend. McNeill (1991, 1994) solved this problem by assuming that in a local neighbourhood (7x7 grid cell blocks) the trend does not have a large influence on the variation of the observed reporting rates.

McNeill's kriging equations implicitly model trend. In these kriging equations the smoothing only aims to remove measurement error from the observations and to predict the other part of the random process as one without separating it into components in the estimation process. This effect can also be observed on the resulting smoothed map for the Pied Crow, the example which was used for illustration by McNeill (1991). In areas with many checklists the smoothed reporting rates fit the original data well but in areas with few checklists heavily smoothed reporting rates are produced, especially in the Karoo near 31° S and 20° E (figs. 1 & 3 of McNeill 1991). This degree of smoothing in this area is also caused by the size of the blocks that was chosen.

The variogram values do not directly determine the degree to which the data should be smoothed. This can only be decided by restricting the size of the window. For our purposes the variogram method would work the wrong way around. If there is not much correlation present in a distribution, the variogram levels off almost immediately. This would imply that the weights for all data points are similar, which in turn would cause heavily smoothed distributions. In contrast, if close grid cells have a strong autocorrelation, the nearby cells would obtain larger weights and cells at a distance would obtain smaller weights. This is good for large spatial correlation values but not if spatial correlation is almost absent.

The required matrix inversion in the kriging approach is of the same size as the number of data points in the block considered for the local window and as many weights have to be estimated. If, as in the case of McNeill a 7x7 grid cell block is used, a 49x49 covariance matrix would have to be inverted and 49 weights would have to be estimated for each grid cells to be smoothed. For a 3x3 window a 9x9 matrix would have to be inverted and nine parameters would have to be estimated. Computational effort is however not the main concern.

For each species a variogram model has to be fitted to the observed variogram. This may involve an iterative procedure if the sill does not equal the estimated σ^2. However this sill is not required if the local kriging window is small enough. Then only the initial values of the variogram are of interest.

There may be concern that the reporting rates are not linearly related to the kriging predictor (McNeill 1991). Kriging methods were originally developed using normal data.

The reason why generalized linear models are in general not used for spatial data is that these methods ignore spatial autocorrelation and instead assume the observations to be independent. But if the explanatory variables are chosen to explain the correlation in the data this disadvantage falls away. In addition, generalized linear models provide all the relevant theory for the case of binomial data. Kriging was used primarily for the interpolation of values at locations that were not sampled. Therefore many of these methods provide perfect interpolations of the values and therefore use global fitting procedures.

REPORTING RATES AND THE BINOMIAL DISTRIBUTION

Do the reporting rates as observed in the bird atlas data, have a binomial distribution? The assumptions that have to hold for a random variable to have a binomial distribution are that the observations are independent (the visits to a single grid cell are independent) and that for a given grid cell the probability of success (in this case, encountering the species) is constant.

The first assumption is that the observations in different checklists are independent. The ideal situation would be that each checklist was collected by a different observer, where observers did not communicate with each other, and made their observations in non-overlapping time periods. Communication between observers would alter the probabilities of encountering a species for an observer who is told where to look for the species. But such communication would mainly involve rarities, which do not concern us much here. We believe that the assumption of independence of checklists is a reasonable one.

The second assumption, that of a constant probability of success in encountering a species during the compilation of a checklist can be less well justified. The ideal situation would be for the same number of individual birds of the species to be present in the grid cell throughout the year and to remain uniformly conspicuous in terms of behaviour and plumage. Furthermore, the identification skills of the observers need to be uniform and the time spent compiling each checklist needs to be standardized. All these issues are discussed by Harrison & Underhill (1997). At best, the theoretical motivation for there being a fixed success probability for each species in each grid cell is a weak one.

Departures form this assumption generally reveal themselves as large values for the estimated overdispersion parameter in the generalized linear modelling performed by GENSTAT, which for the binomial distribution should be approximately unity. However, more than 50% of the overdispersion estimates were less than two (Tables 1 & 2). Thus we feel confident that in spite of all the caveats, the assumption of constant probability of success is sufficiently adhered to, so that the binomial distribution provides a useful tool for modelling the reporting rates.

During the field collection period of the atlas, observers were encouraged not to send in records of single unusual observations. This was to keep the data in the form of an overall summary and aimed not to distort the number of checklists for other species that were present but not recorded (Harrison & Underhill 1997, p. xlv). A checklist should estimate the complete list of all species in the cell, with the purpose of using these reporting rates as binomial data and as indices for relative species abundance in grid cells.

It is not justifiable to use more than the immediate adjacent cells for smoothing. The environment in some regions changes so fast that for smoothing one would have to include other variables in addition to those which only describe the spatial relationship, if cells are too far from the value to be smoothed. This would complicate the process unnecessarily if the same effect can be achieved with smaller blocks and fewer parameters in the model.

TERMS IN THE MODEL

With more terms in the model, the fitted surface becomes more complicated. It is not clear what the additional effect of the square terms on the form of the surface is. It may also be that these terms would explain too much of the variation which was required to be smoothed and should not be explained by a trend model.

For many areas and for many species it will be the case that the occurrence does not necessarily stretch over more than a single cell. The area of one cell covers approximately 25 km². This area is large enough to provide living space for a species. Habitats can change considerably within 50 km (two grid cells farther). Isolated cells of occurrence in distributions may occur if forest patches are present in isolated cells or in the case of isolated wetlands. In blocks of nine grid cells it is more likely that this situation will be captured as it is. In larger blocks occasional zero reporting rates are more likely to be smoothed out.

The number of terms that are kept in the model does not matter as much as in the usual modelling process where the final model is of interest as a description of the situation and where the terms in the model are restricted to those that are significant. This ensures that models are more general and hence better for a general understanding of the situation and for predictions. In the case of the smoothed atlas maps we are not interested in the model terms themselves but only in a single predicted value, trying to find this from the best possible fit to the data, but at the same time restricting the surface to have a fixed number of parameters. Leaving insignificant terms in the model does not do much harm, it only slightly instead of significantly increases the fit to the data.

If a best possible fit to the data is the intention, one could ask, why not add more terms to the model? Firstly a limited amount of data is present, a maximum of nine data points for each regression. This would in many cases not be sufficient for the estimation of more parameters especially if grid cells without checklists and many zero reporting rates are present. By adding more terms to the model the resulting smoothed map would look more like the original data, which we were trying to smooth. Square terms x^2 and y^2 are not justified in a block of nine grid cells and might be able to explain much more of the variation even though they do not have a true physical interpretation. The other terms x, y and xy are enough to explain any trend that may be present in a block of nine grid cells.

With six (for 25-grid cell blocks, including square terms) instead of four parameters to estimate in each iteration for each regression the calculation process would be much complicated. A 6x6 matrix would have to be inverted in each iteration and the

It is not justifiable to use more than the immediate adjacent cells for smoothing. The environment in some regions changes so fast that for smoothing one would have to include other variables in addition to those which only describe the spatial relationship, if cells are too far from the value to be smoothed. This would complicate the process unnecessarily if the same effect can be achieved with smaller blocks and fewer parameters in the model.

TERMS IN THE MODEL

With more terms in the model, the fitted surface becomes more complicated. It is not clear what the additional effect of the square terms on the form of the surface is. It may also be that these terms would to explain too much of the variation which was required to be smoothed and should not be explained by a trend model.

For many areas and for many species it will be the case that the occurrence does not necessarily stretch over more than a single cell. The area of one cell covers approximately 25 km². This area is large enough to provide living space for a species. Habitats can change considerably within 50 km (two grid cells farther). Isolated cells of occurrence in distributions may occur if forest patches are present in isolated cells or in the case of isolated wetlands. In blocks of nine grid cells it is more likely that this situation will be captured as it is. In larger blocks occasional zero reporting rates are more likely to be smoothed out.

The amount of terms that are kept in the model does not matter as much as in the usual modelling process where the final model is of interest as a description of the situation and where the terms in the model are restricted to those that are significant. This ensures that models are more general and hence better for a general understanding of the situation and for predictions. In the case of the smoothed atlas maps we are not interested in the model terms themselves but only in a single predicted value, trying to find this from the best possible fit to the data, but at the same time restricting the surface to have a fixed number of parameters. Leaving insignificant terms in the model does not do much harm, it only slightly instead of significantly increases the fit to the data.

If a best possible fit to the data is the intention, one could ask, why not add more terms to the model? Firstly a limited amount of data is present, a maximum of nine data points for each regression. This would in many cases not be sufficient for the estimation of more parameters especially if grid cells without checklists and many zero reporting rates are present. By adding more terms to the model the resulting smoothed map would look more like the original data, which we were trying to smooth. Square terms x^2 and y^2 are not justified in a block of nine grid cells and might be able to explain much more of the variation even though they do not have a true physical interpretation. The other terms x, y and xy are enough to explain any trend that may be present in a block of nine grid cells.

With six (for 25-grid cell blocks, including square terms) instead of four parameters to estimate in each iteration for each regression the calculation process would be much complicated. A 6x6 matrix would have to be inverted in each iteration and the

understanding of significant terms would decrease. With the only terms in the model being x, y and xy it is easier to check whether the estimated coefficients are correct.

FINAL MODEL

The regression method with the best results was the **iteratively reweighted least squares** model using a fixed number of five iterations on blocks of nine grid cells. A surface is fitted to the reporting rates of these nine cells and this is used to predict a reporting rate for the central grid cell. The IRWLS method is not perfect. There are areas where it appears to smooth too heavily, but there are more advantages:

- When compared to GENSTAT results it produces numerically equal estimated coefficients and predictions, to at least two decimal places. This should however not be surprising, because GENSTAT uses the same algorithm and theory (generalized linear models). The only difference is the initial choice of weights and modified dependent variables.
- The IRWLS regression process does not take significantly longer than single-step methods.
- The IRWLS regression has advantages over the GENSTAT method, because it takes significantly less time to run, the programs can be manipulated more easily and the results are calculated for some cases which the GENSTAT program cannot handle.
- The results are more reliable than those of the single-step methods, the effects of the +0.5 for small number of successes or failures can be readjusted with more iterations.
- The predicted values nearly preserve the size of the distribution, the area of the original observed distribution is not increased by too heavy smoothing. This would happen if the explanatory variables could not sufficiently explain small scale trend.
- By carefully choosing explanatory variables, it is possible to incorporate spatial dependence into the estimation of the coefficients.
- Generalized linear models weight each observation according to its reliability or variance.
- With data not on a regular grid, it would not be a problem to name the explanatory variables for each data point.

REFERENCES

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, **36**, 192 - 225.

Borland International. (1992). *Turbo C++, Version 3.0, User's guide*. California: Borland International.

Carstensen Jr., L.W. (1987). A Measure of Similarity for Cellular Maps. *The American Cartographer*, **14**, 345 - 358.

Clifford, H.T. and Stephenson, W. (1975). *An Introduction to Numerical Classification*. New York: Academic Press.

Collett, D. (1991). *Modelling Binary Data*. London: Chapman & Hall.

Cox, D.R. (1970). *The Analysis of Binary Data*. London: Chapman & Hall.

Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data, 2nd ed.* London: Chapman & Hall.

Cox, T.F. & Cox, M.A.A. (1994). *Multidimensional Scaling*. London: Chapman & Hall.

Cressie, N. (1985). When are relative variograms useful in geostatistics? *Journal of the International Association for Mathematical Geology*, **17**, 693 - 702.

Cressie, N.A.C. (1991) *Statistics for Spatial Data*. New York , Wiley.

Cressie, N. & Chan, N.H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association*, **84**, 393 - 401.

Cressie, N. & Hawkins, D.M. (1980). Robust estimation of the variogram, I. *Journal of the International Association for Mathematical Geology*, **12**, 115 - 125.

Cressie, N. & Read, T.R.C. (1985) Do sudden infant deaths come in clusters? *Statistics and Decisions*, Supplement Issue No. 2, **3**, 333 - 349.

Cressie, N. & Read, T.R.C. (1989) Spatial data analysis of regional counts. *Biometrical Journal*, **31**, 699 - 719.

Cressie, N. (1986) Kriging nonstationary data. *Journal of the American Statistical Association*, **81**, 625 - 634.

Digby, P.G.N. & Kempton, R.A. (1987). *Multivariate Analysis of Ecological Communities*. London: Chapman & Hall.

Draper, N.R. & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). New York: Wiley.

GENSTAT 5 Committee of the Statistics Department Rothamsted Experimental Station. (1993). *GENSTAT 5 Release 3 Reference Manual*. Oxford: Oxford University Press.

Gower, J.C. (1985). Measures of Similarity, Dissimilarity and Distance. In: Kotz, S., Johnson, N.L. (eds) *Encyclopedia of Statistical Sciences*, Vol. 5. New York: Wiley.

Harrison, J.A., Allan, D G, Underhill, L.G., Herremans, M., Tree, A.J., Parker, V. & Brown, C.J. (eds). (1997a) *The Atlas of Southern African Birds.* Vol. 1: Non-passerines. Johannesburg: BirdLife South Africa.

Harrison, J.A., Allan, D G, Underhill, L.G., Herremans, M., Tree, A.J., Parker, V. & Brown, C.J. (eds). (1997b) *The Atlas of Southern African Birds.* Vol. 2: Passerines. Johannesburg: BirdLife South Africa.

Harrison, J.A. & Underhill, L.G., (1997). Introduction and methods. In: Harrison, J.A., Allan, D G, Underhill, L.G., Herremans, M., Tree, A.J., Parker, V. & Brown, C.J. (eds) (1997a) *The Atlas of Southern African Birds,* Vol. 1. Johannesburg: BirdLife South Africa. pp. xliii - lxiv.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246 – 1266.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

McNeill, L. (1991). Interpolation and smoothing of binomial data for the Southern African Bird Atlas Project. *South African Statistical Journal.* **25**, 129-136.

McNeill, L. (1994). *Topics in Interpolation and Smoothing of Spatial Data.* PhD Thesis. University of Cape Town.

Rice, J.A. (1988). *Mathematical Statistics and Data Analysis.* San Fransisco: Wadsworth.

Ripley, B.D. (1981). *Spatial Statistics.* New York: John Wiley.

Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy.* San Francisco: Freeman.

Sokal, R.R. & Sneath, P.H.A. (1973). *Numerical Taxonomy.* San Francisco: Freeman.

Underhill, L.G., Erni, B. & Mashinini, F.X. (1998). Ecological differentiation of canaries (Fringillidae) in the Western Cape. *Durban Museum Novitates*, **23**, 56 – 60.

Underhill, L.G., Prys-Jones, R.F., Harrison, J.A. & Martinez P. (1992). Seasonal patterns of occurrence of Palaearctic migrants in southern Africa using atlas data. *Ibis.* **134**, suppl. 1, 99 – 108.

APPENDIX A

The distribution maps in this appendix were taken from *The Atlas of Southern African Birds* (Harrison *et al.* 1997 a, b).

Figure A1. Cape Weaver	*Ploceus capensis*
Figure A2. Cape Canary	*Serinus canicollis*
Figure A3. Masked Weaver	*Ploceus velatus*
Figure A4. European Bee-eater	*Merops apiaster*
Figure A5. Wattled Starling	*Creatophora cinerea*
Figure A6. Cape Siskin	*Pseudochloroptila totta*
Figure A7. Protea Canary	*Serinus leucopterus*
Figure A8. Bully Canary	*Serinus sulphuratus*
Figure A9. Whitethroated Canary	*Serinus albogularis*
Figure A10. Yellow Canary	*Serinus flaviventris*
Figure A11. Blackheaded Canary	*Serinus alario*
Figure A12. Blackthroated Canary	*Serinus atrogularis*
Figure A13. Sreakyheaded Canary	*Serinus gularis*
Figure A14. Yelloweyed Canary	*Serinus mozambicus*
Figure A15. Forest Canary	*Serinus scotops*
Figure A16. Spottedbacked Weaver	*Ploceus cucullatus*
Figure A17. Spectacled Weaver	*Ploceus ocularis*
Figure A18. Yellow Weaver	*Ploceus subaureus*
Figure A19. Golden Weaver	*Ploceus xanthops*
Figure A20. Collared Sunbird	*Anthreptes collaris*
Figure A21. Olive Sunbird	*Nectarinia olivacea*
Figure A22. Grey Sunbird	*Nectarinia veroscii*
Figure A23. Black Sunbird	*Nectarinia amethystina*
Figure A24. Whitebellied Sunbird	*Nectarinia talatala*
Figure A25. Scarletchested Sunbird	*Nectarinia senegalensis*

Figure A26. Greater Doublecollared Sunbird *Nectarinia afra*

Figure A27. Purplebanded Sunbird *Nectarinia bifasciata*

Figure A28. Dusky Sunbird *Nectarinia fusca*

Figure A29. Cape Sugarbird *Promerops cafer*

Figure A30. House Sparrow *Passer domesticus*

Figure A31. Cape Sparrow *Passer melanurus*

Figure A32. Forest Weaver *Ploceus bicolor*

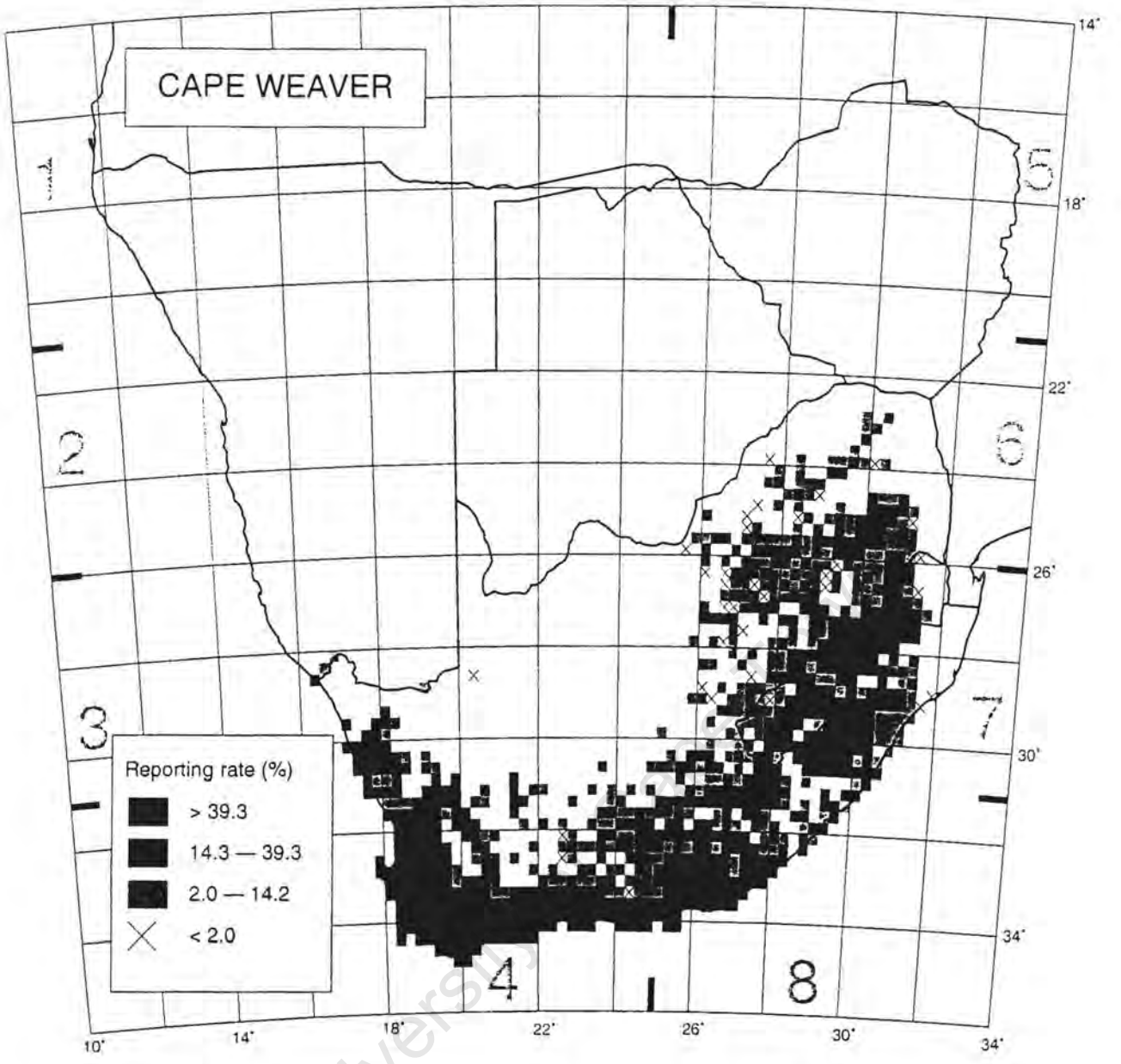


Figure A1. The atlas distribution of the Cape Weaver *Ploceus capensis*.

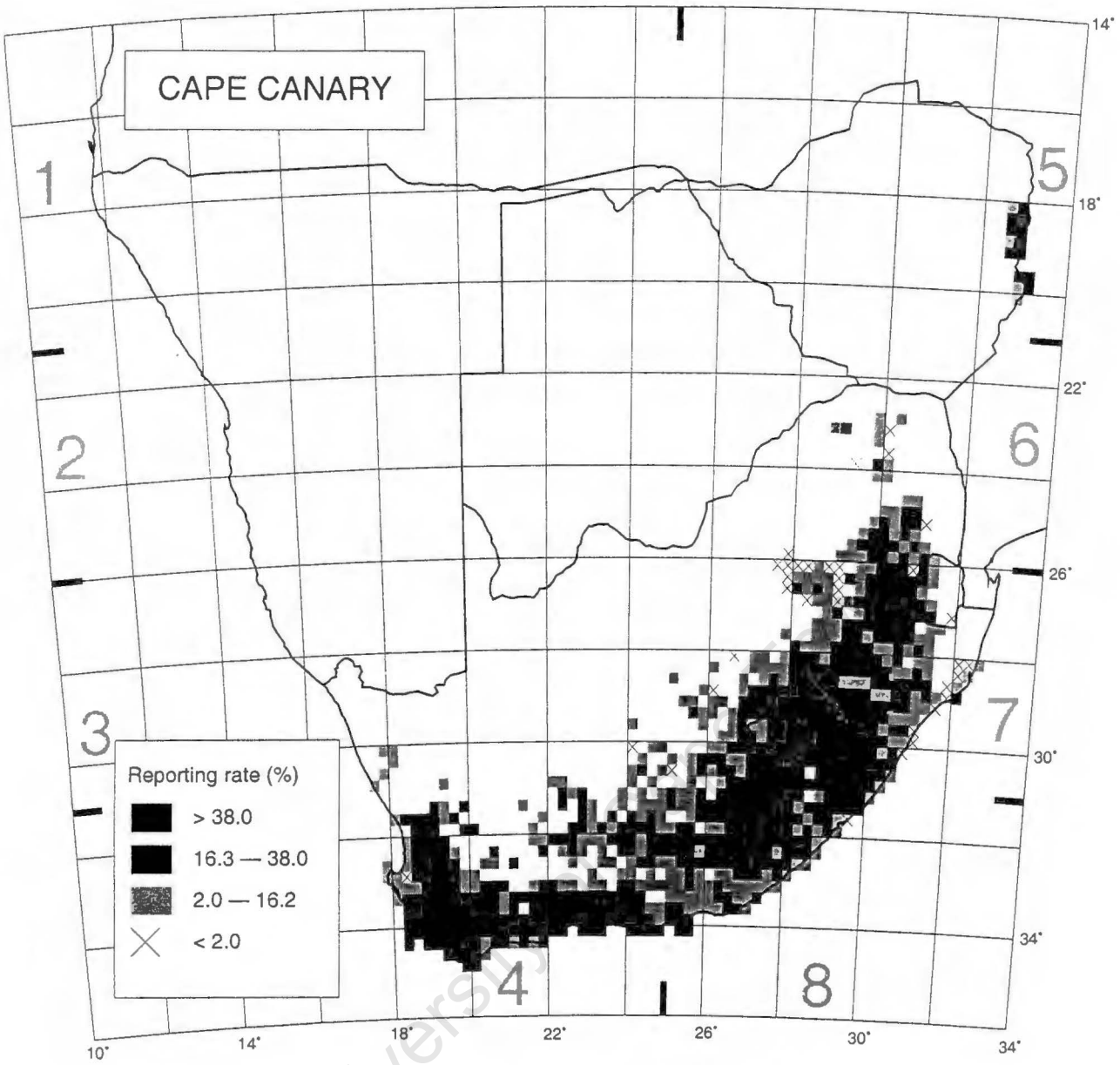


Figure A2. The atlas distribution of the Cape Canary *Serinus canicollis*.

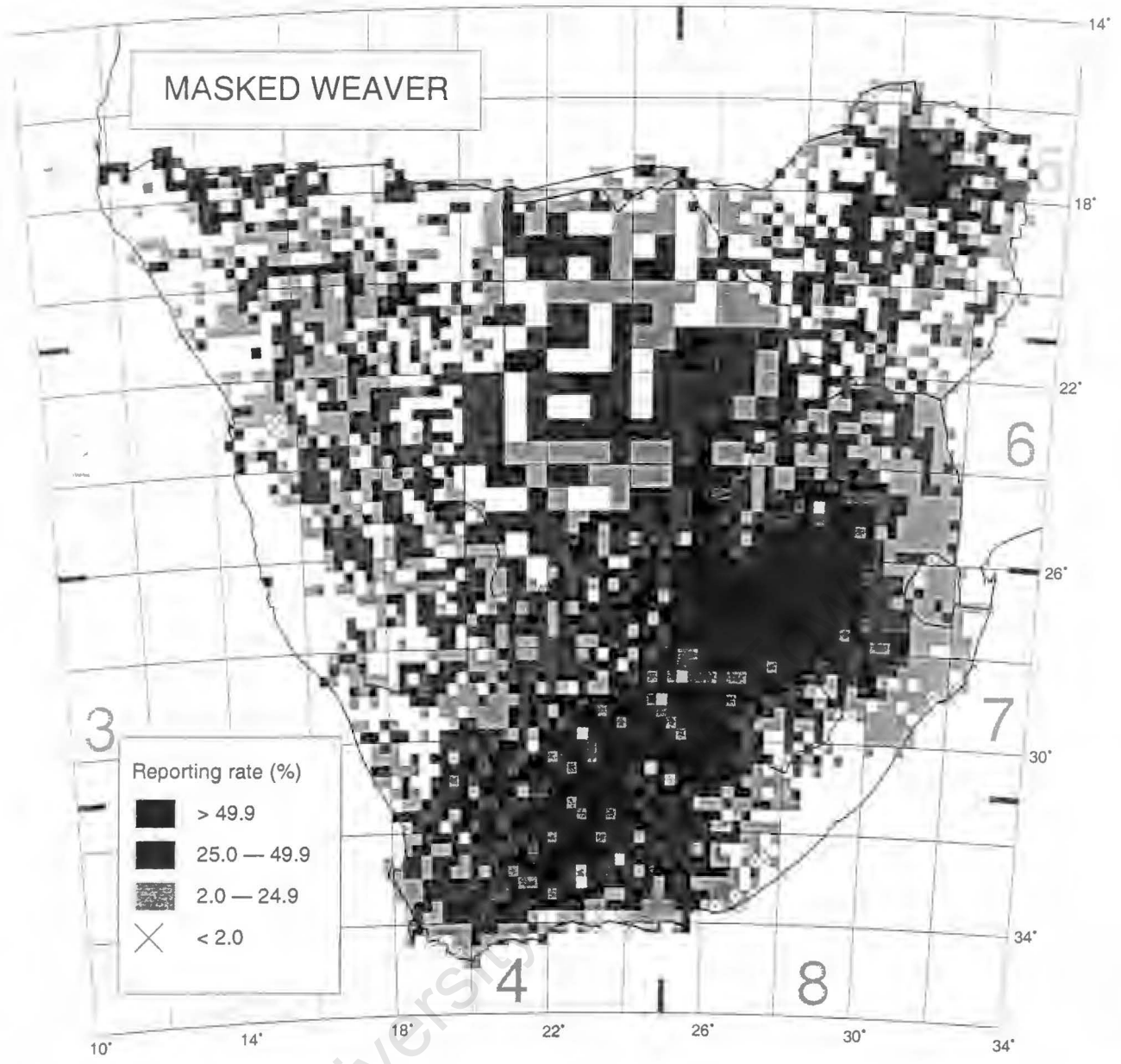


Figure A3. The atlas distribution of the Masked Weaver *Ploceus velatus*.

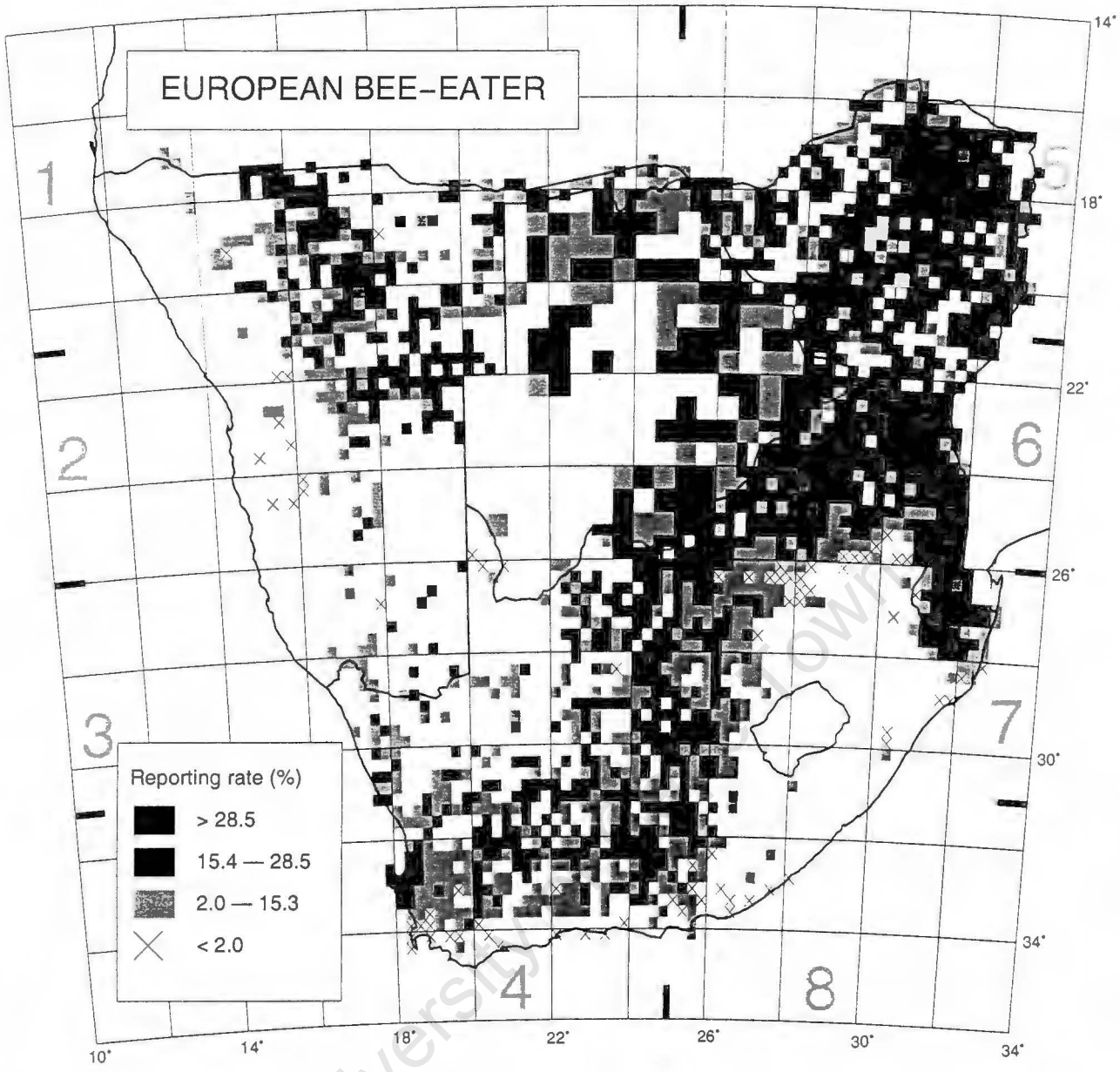


Figure A4. The atlas distribution of the European Bee-eater *Merops apiaster*.

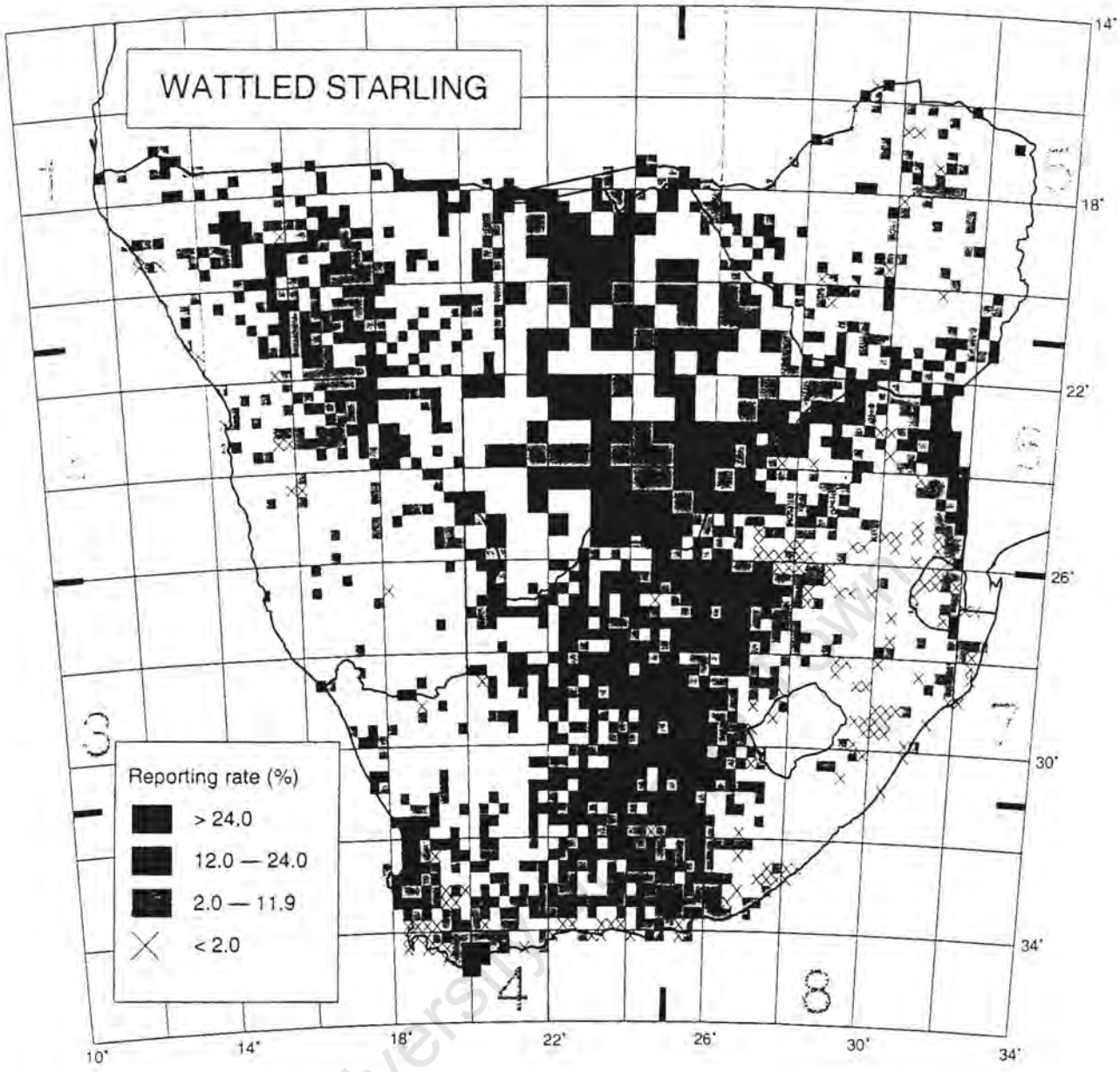


Figure A5. The atlas distribution of the Wattled Starling *Creatophora cinerea*.

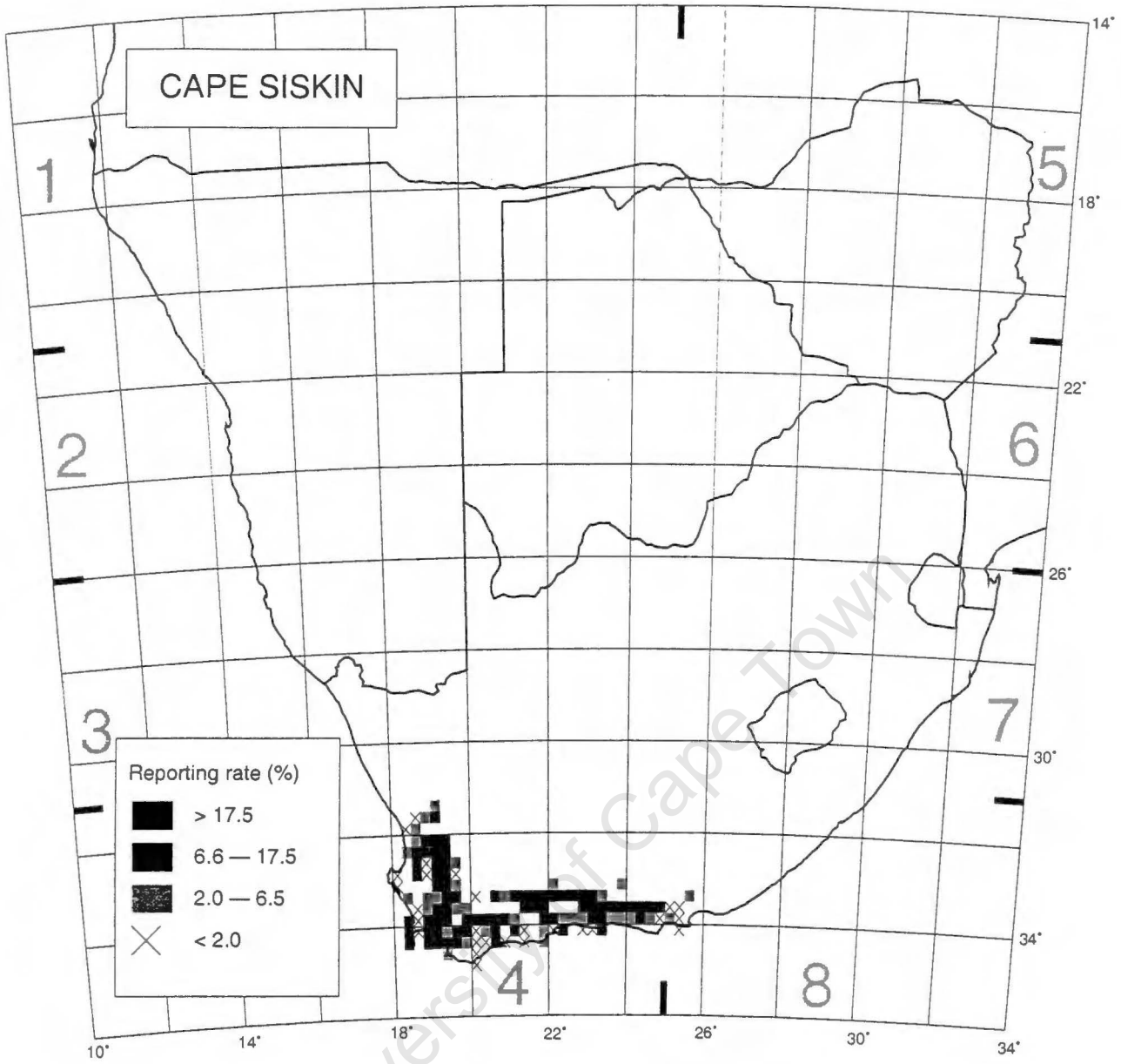


Figure A6. The atlas distribution of the Cape Siskin *Pseudochloroptila totta*.

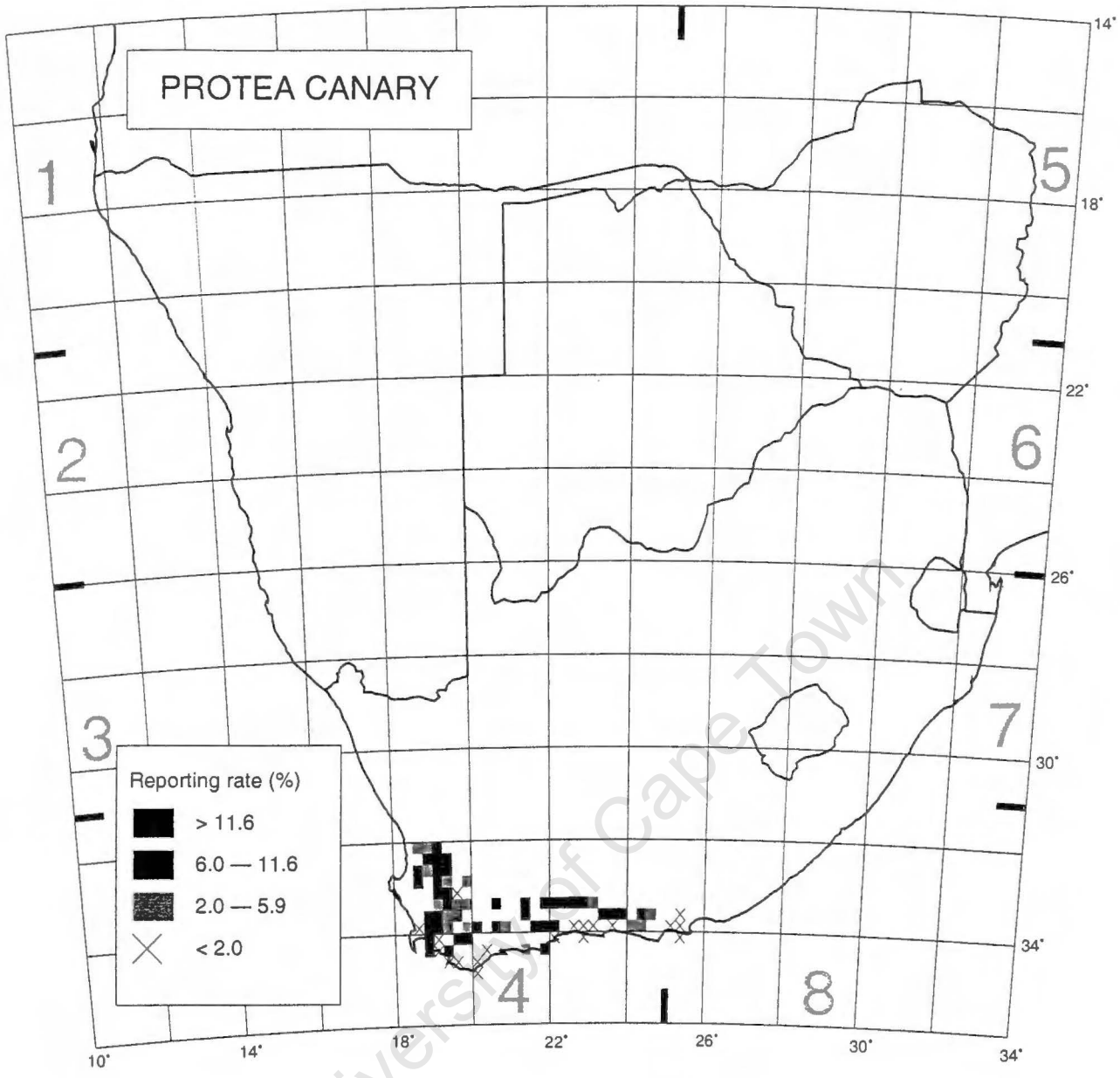


Figure A7. The atlas distribution of the Protea Canary *Serinus leucopterus*.

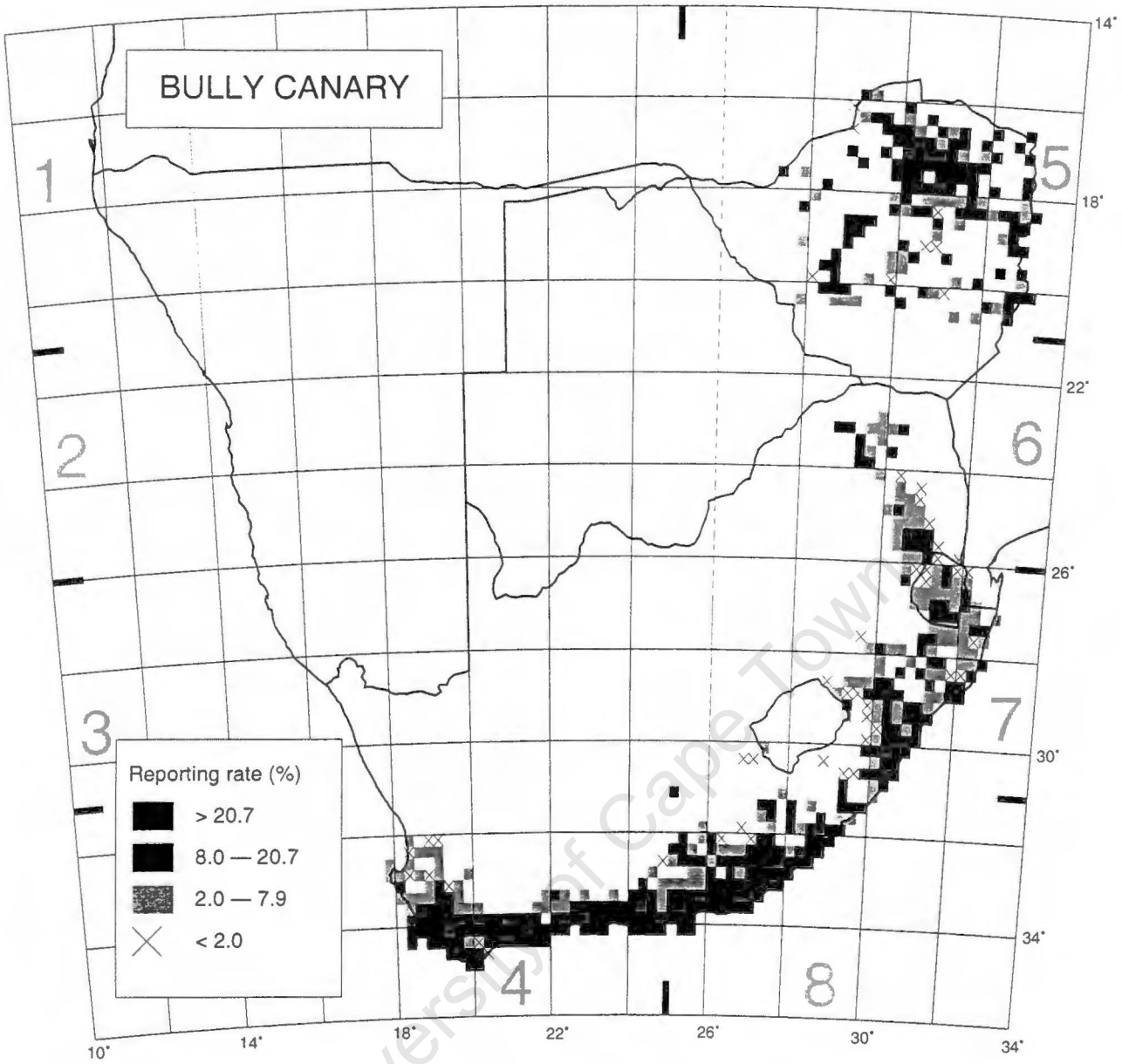


Figure A8. The atlas distribution of the Bully Canary *Serinus sulphuratus*.

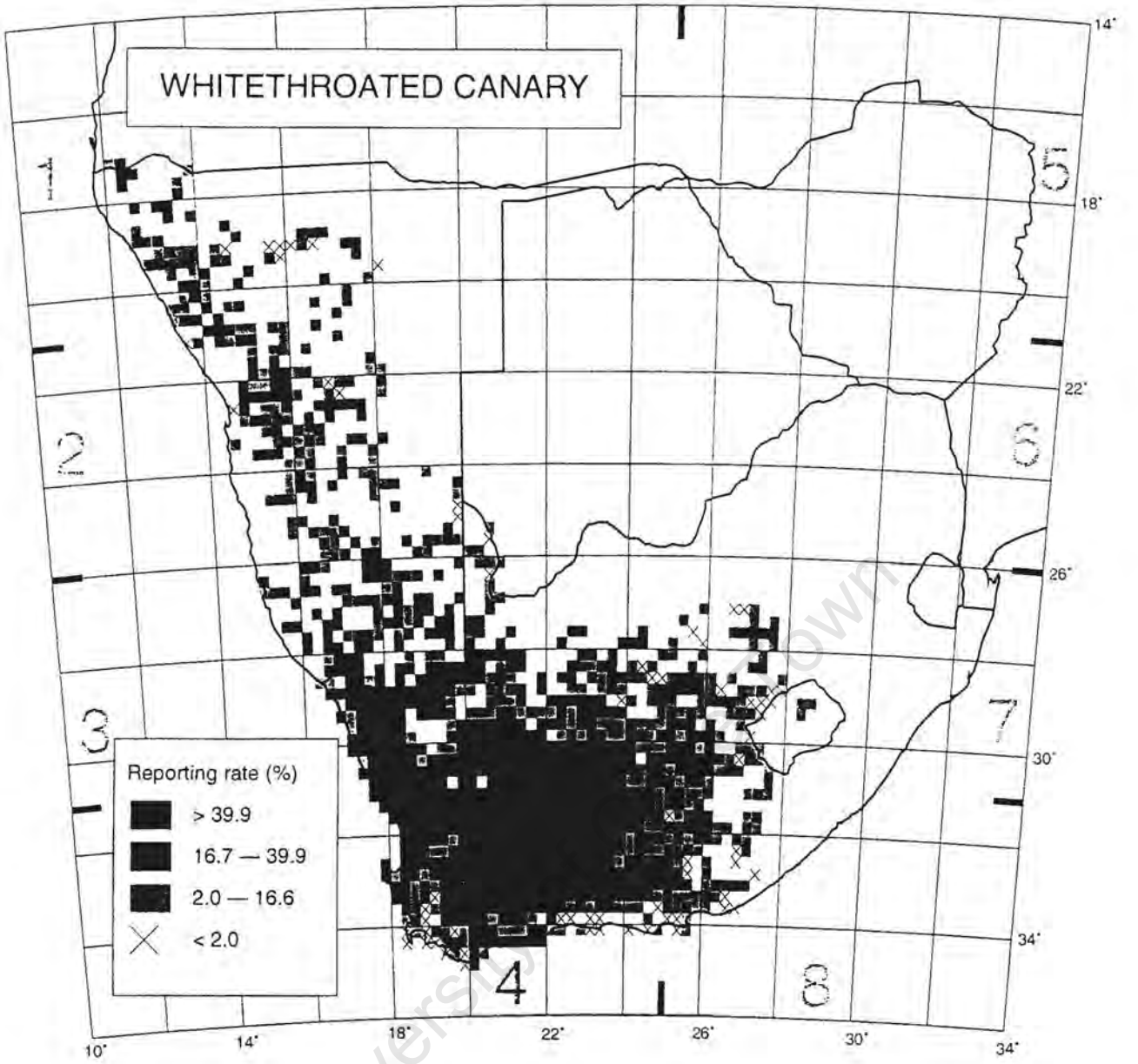


Figure A9. The atlas distribution of the Whitethroated Canary *Serinus albogularis*.

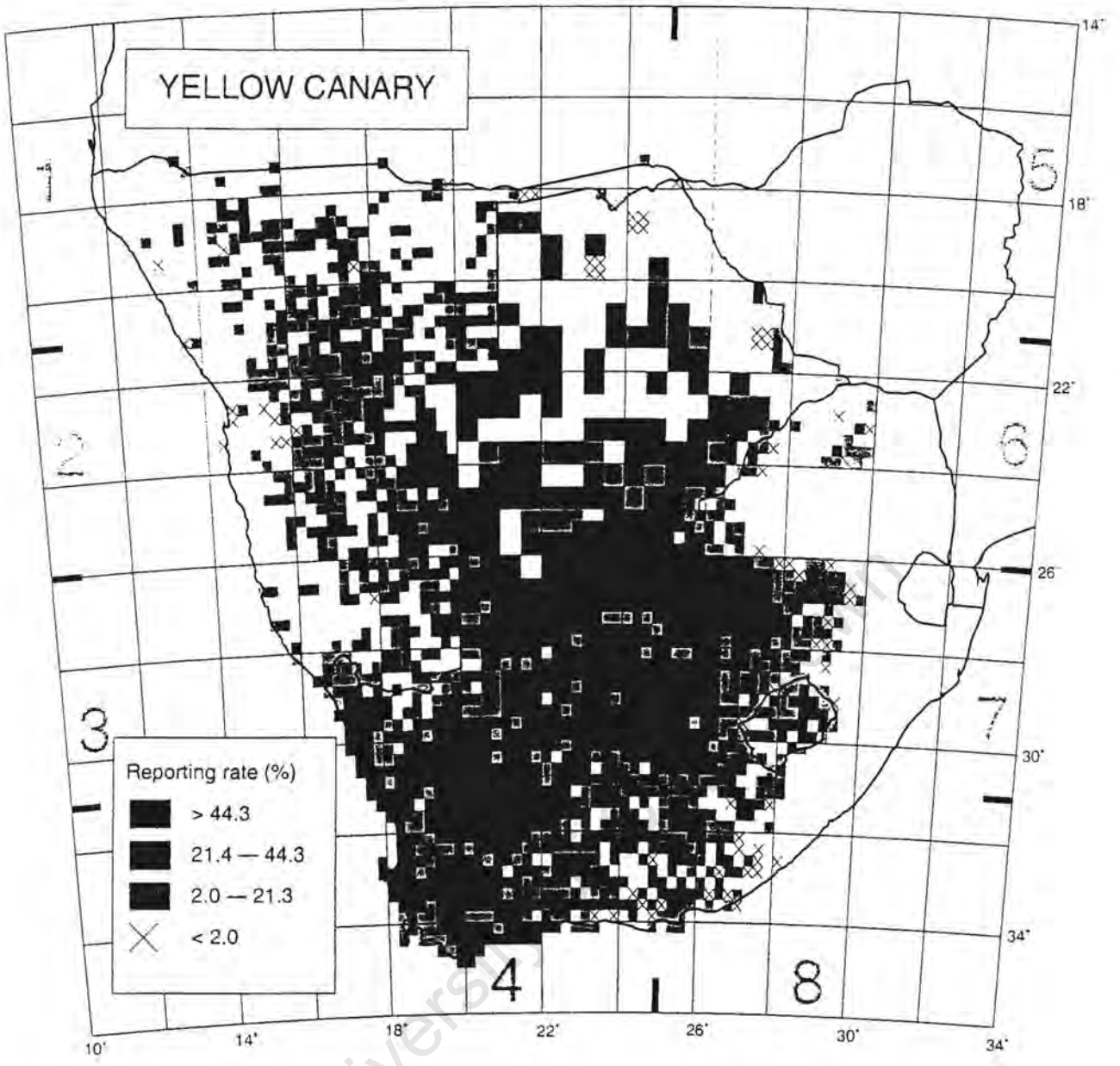


Figure A10. The atlas distribution of the Yellow Canary *Serinus flaviventris*.

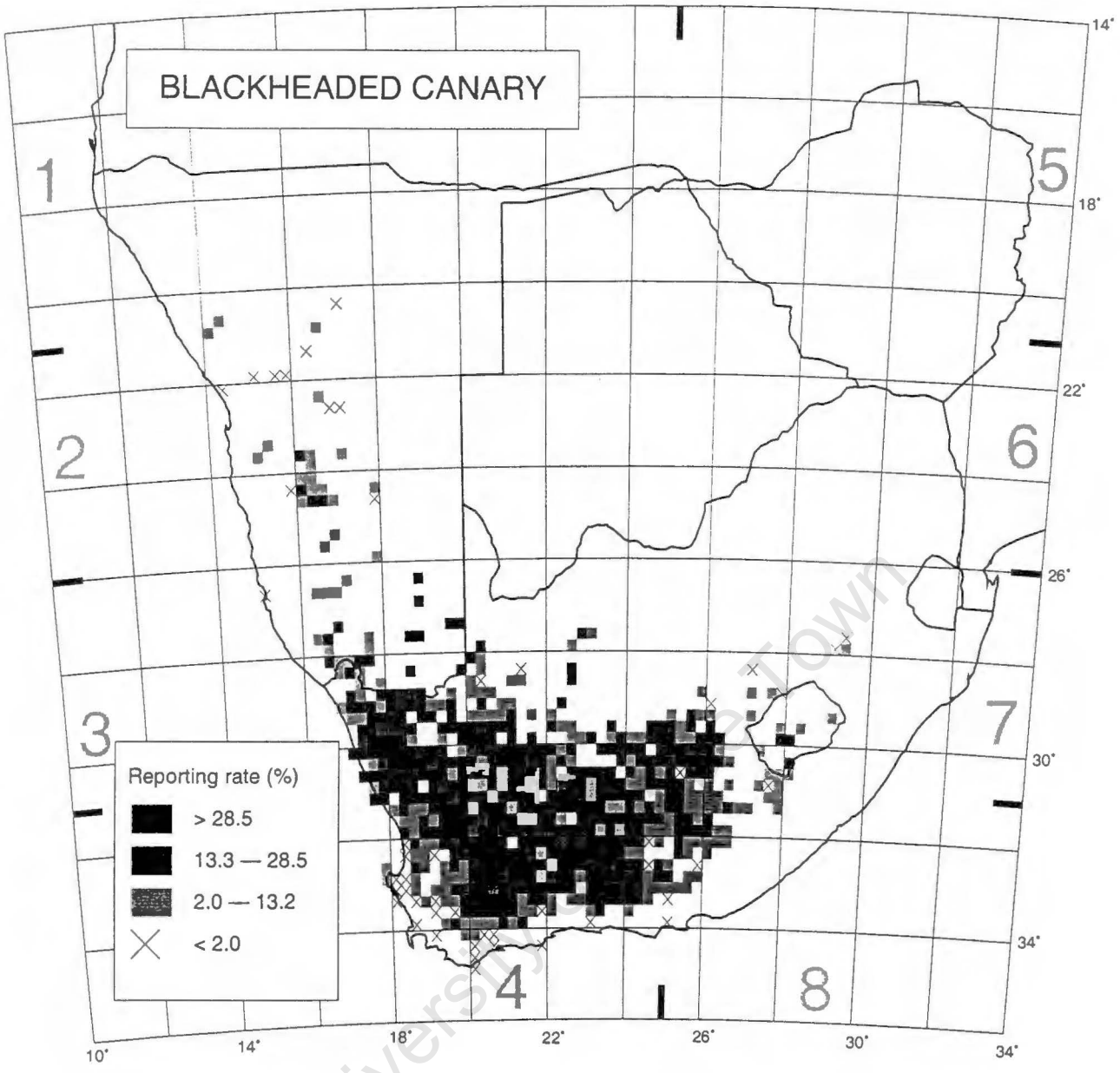


Figure A11. The atlas distribution of the Blackheaded Canary *Serinus alario*.

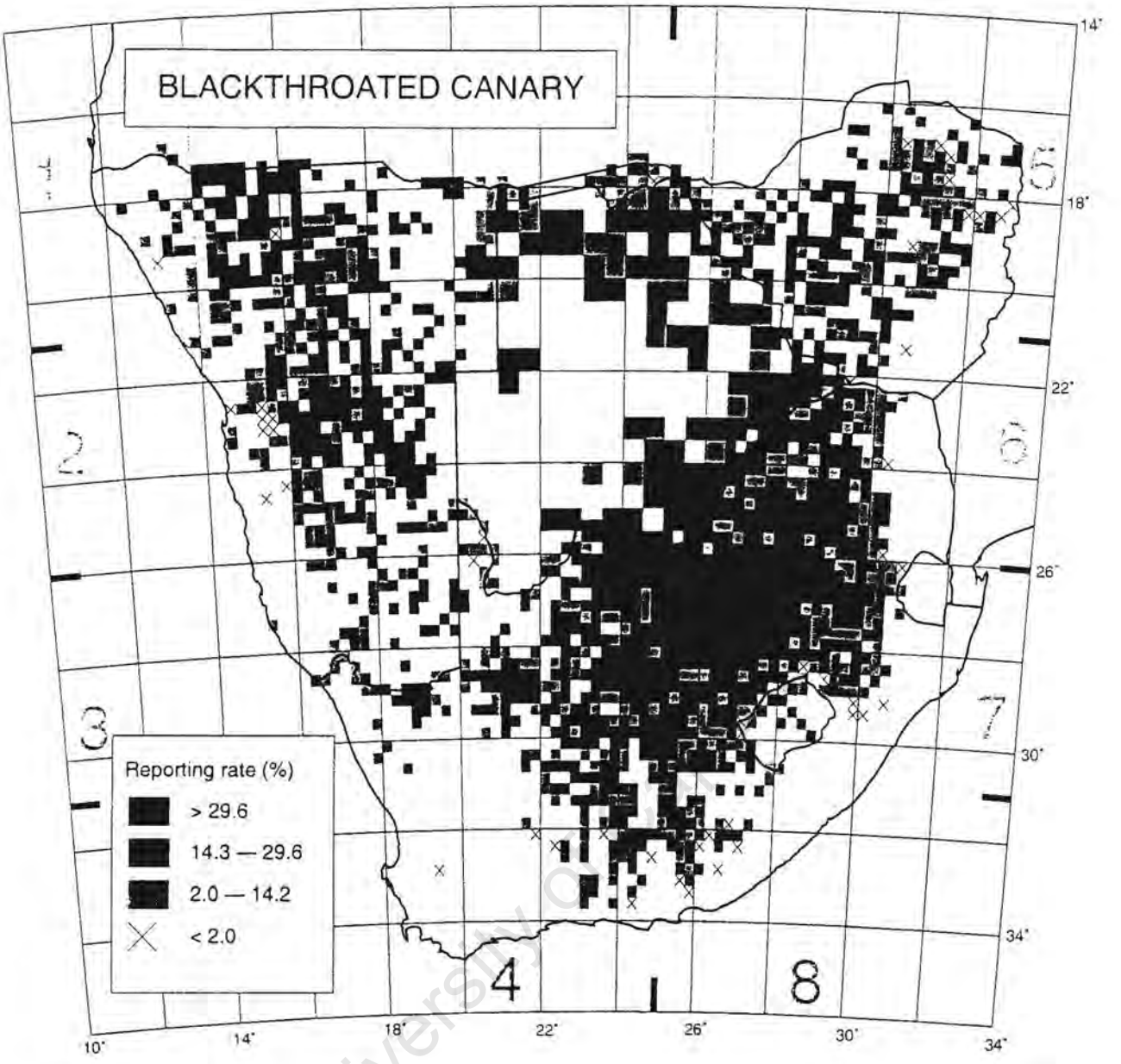


Figure A12. The atlas distribution of the Blackthroated Canary *Serinus atrogularis*.

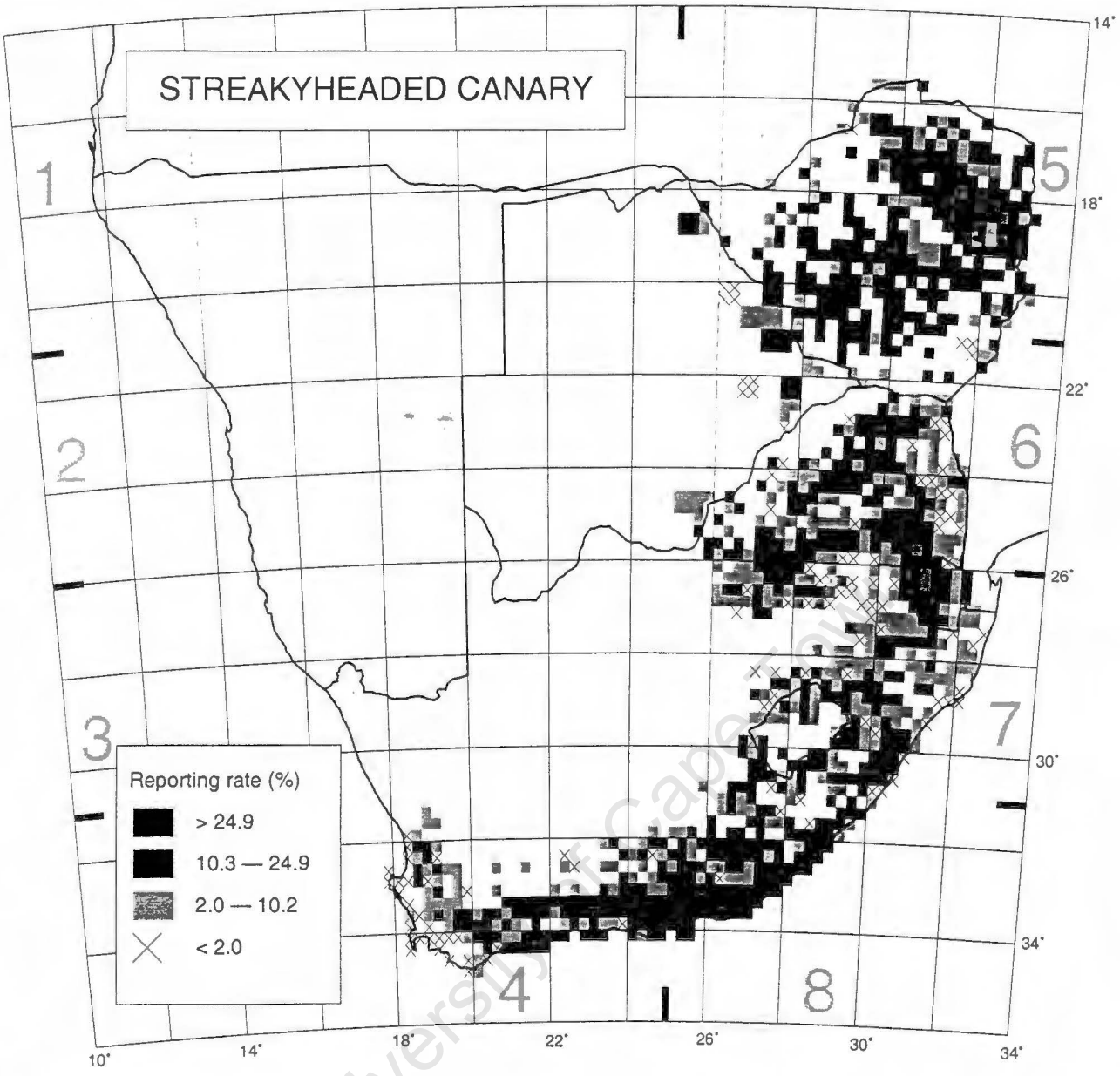


Figure A13. The atlas distribution of the Streakyheaded Canary *Serinus gularis*.

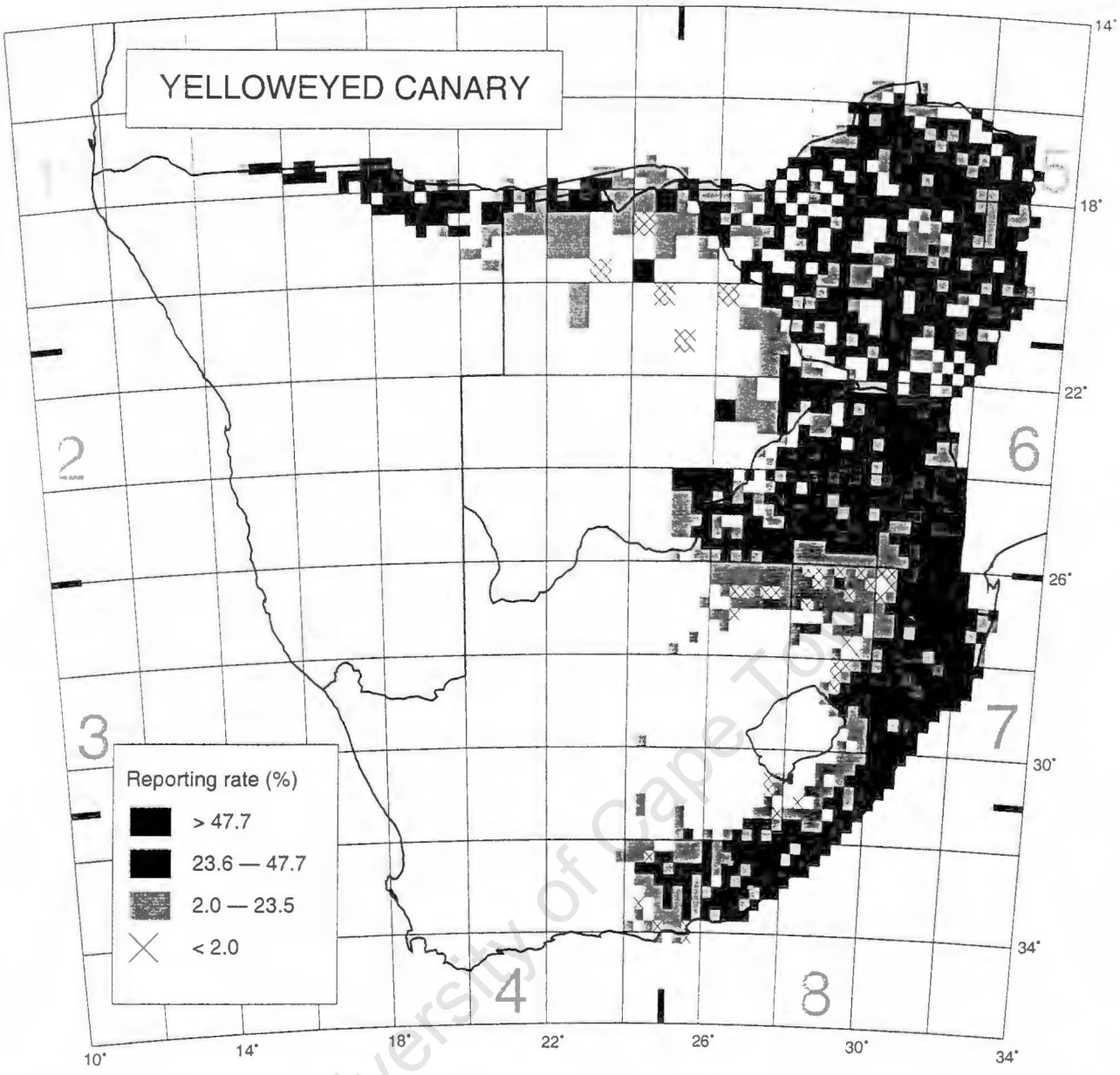


Figure A14. The atlas distribution of the Yelloweyed Canary *Serinus mozambicus*.

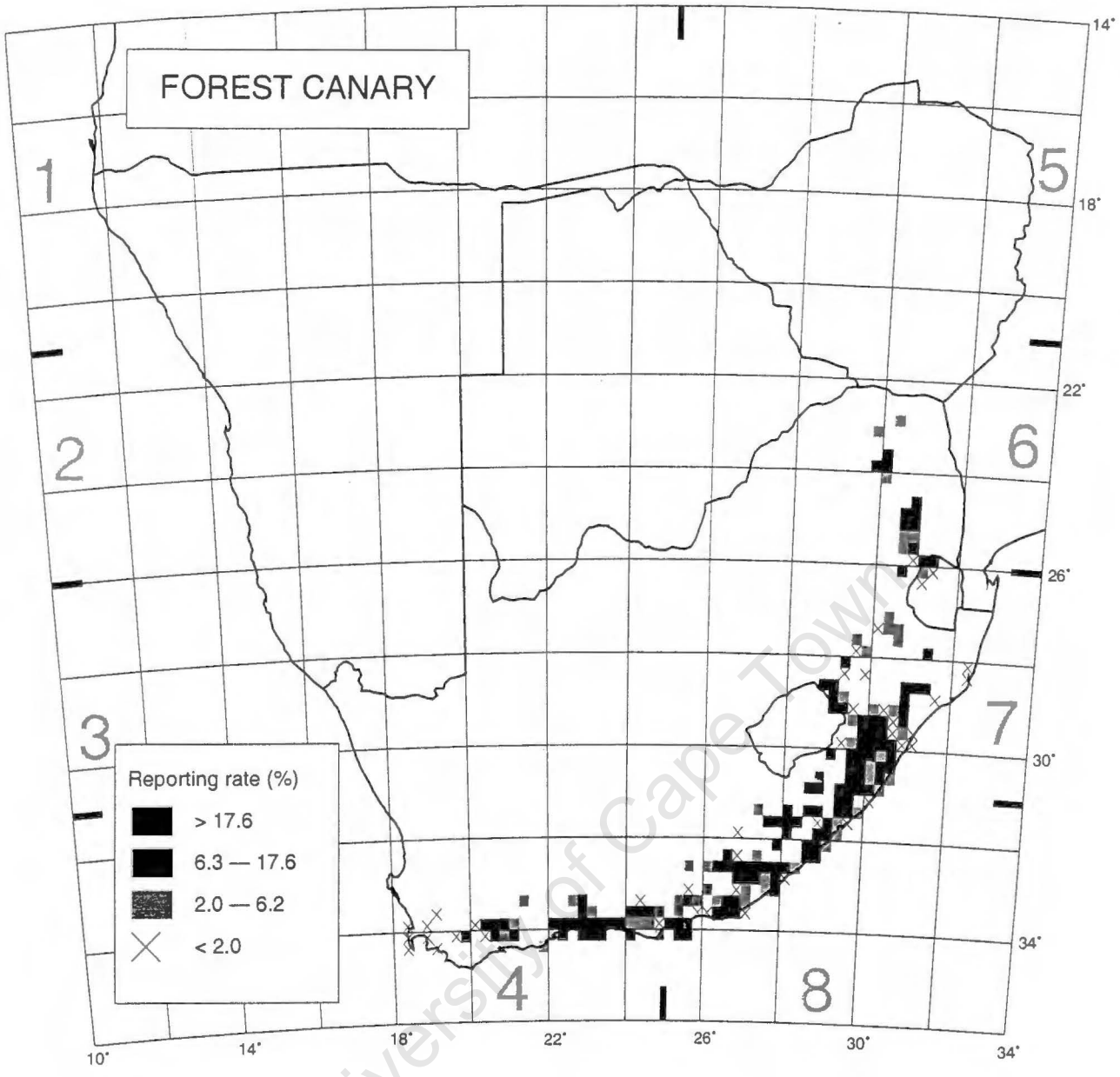


Figure A15. The atlas distribution of the Forest Canary *Serinus scotops*.

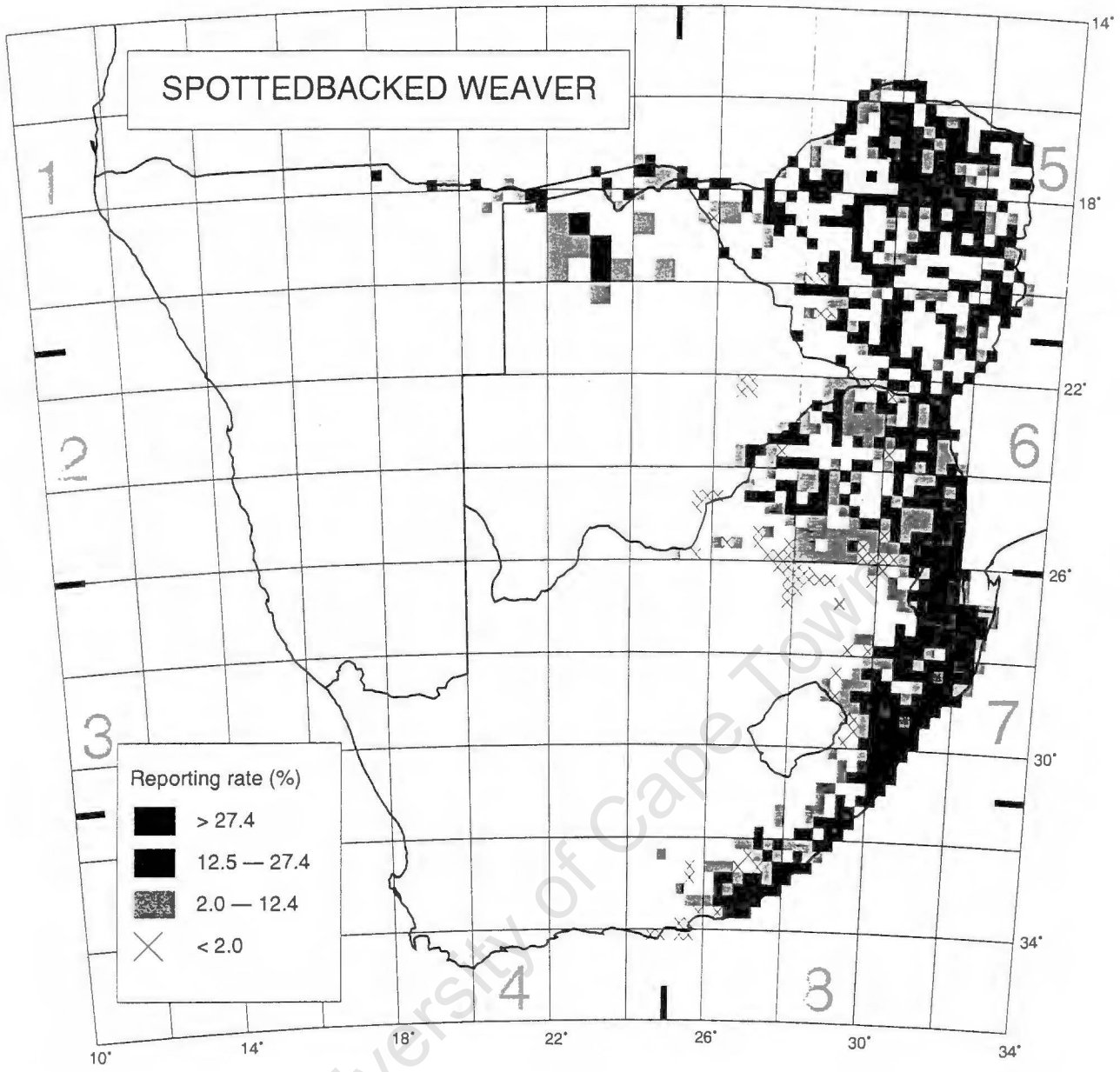


Figure A16. The atlas distribution of the Spottedbacked Weaver *Ploceus cucullatus*.

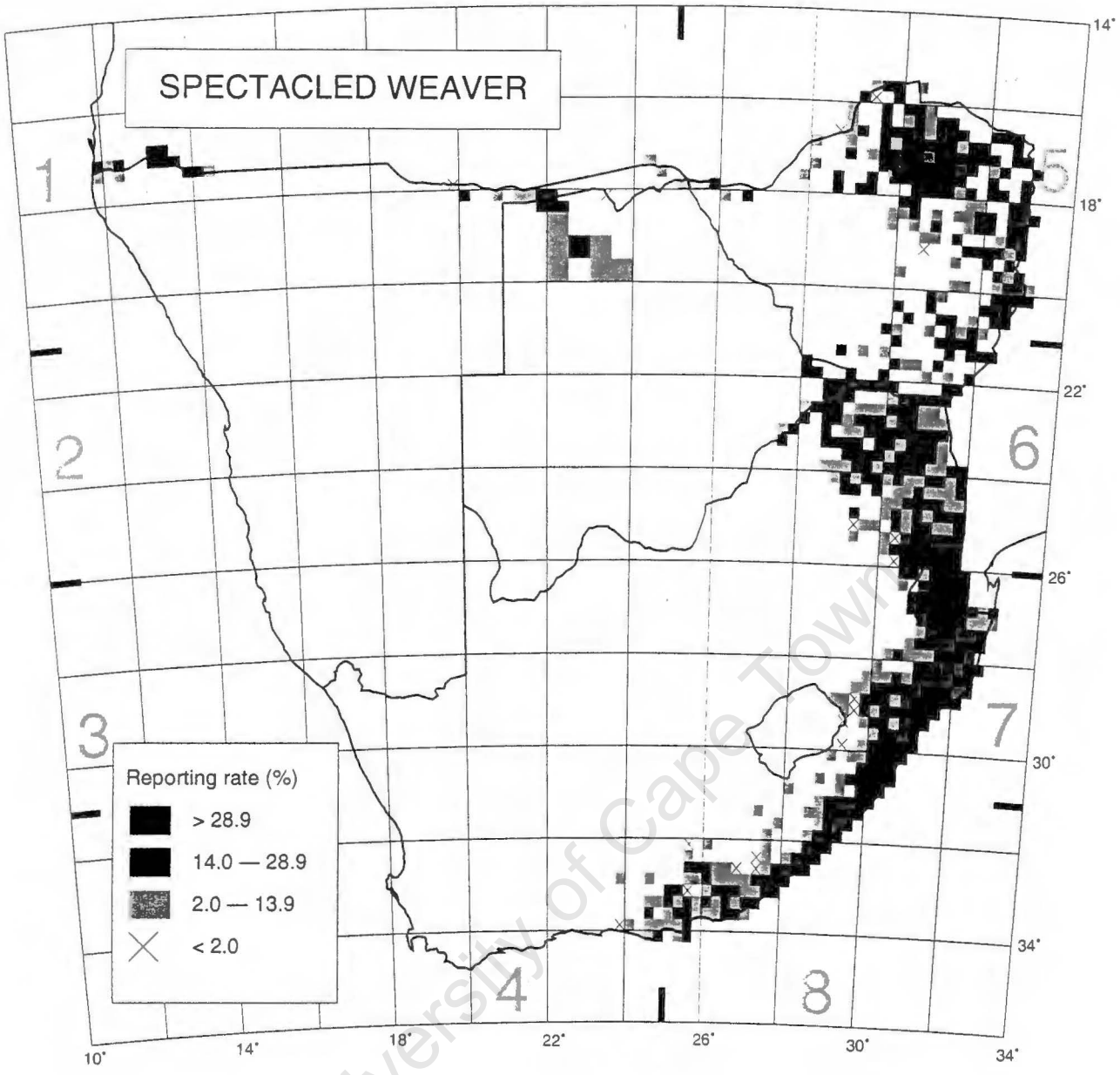


Figure A17. The atlas distribution of the Spectacled Weaver *Ploceus ocularis*.

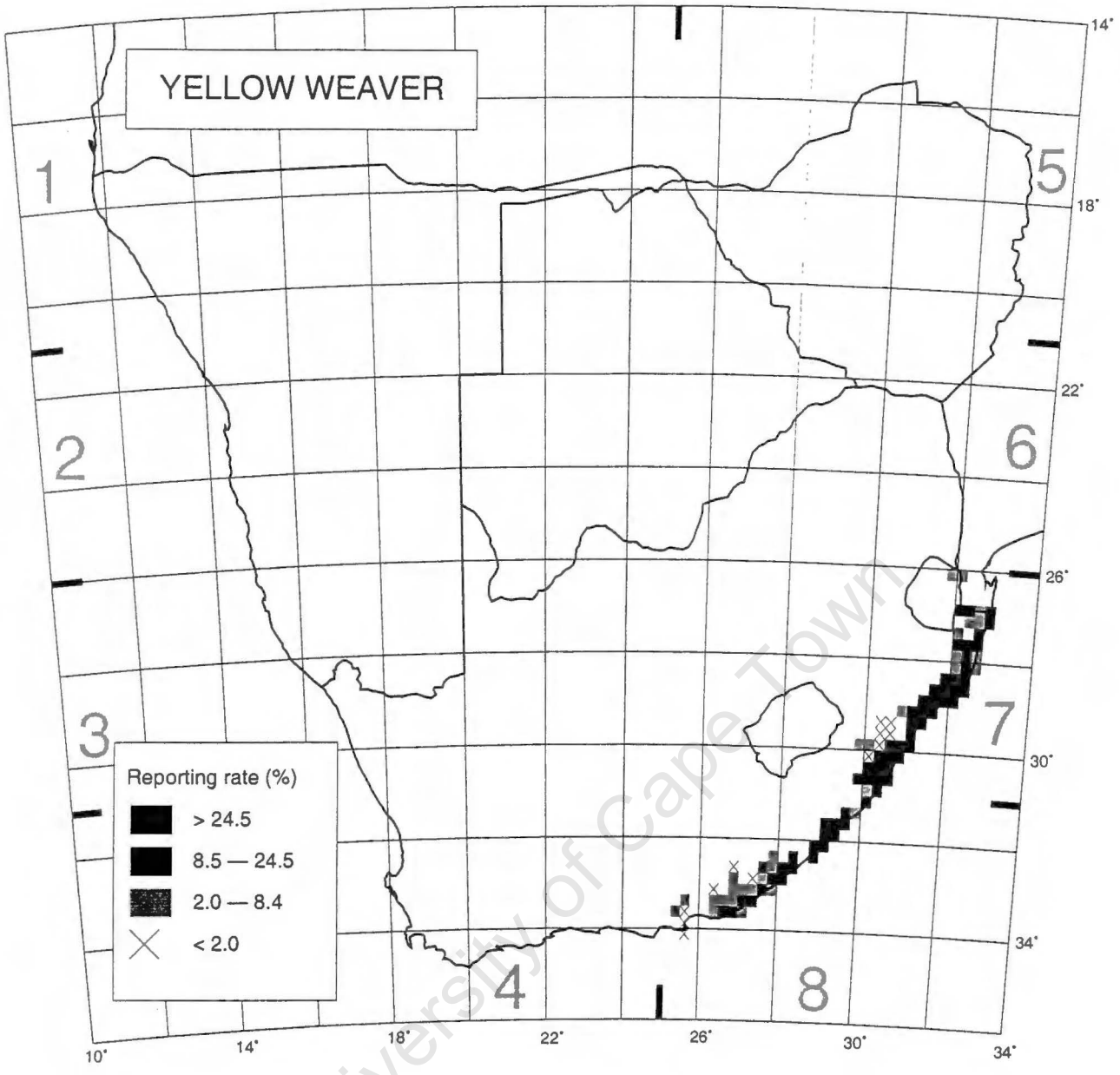


Figure A18. The atlas distribution of the Yellow Weaver *Ploceus subaureus*.

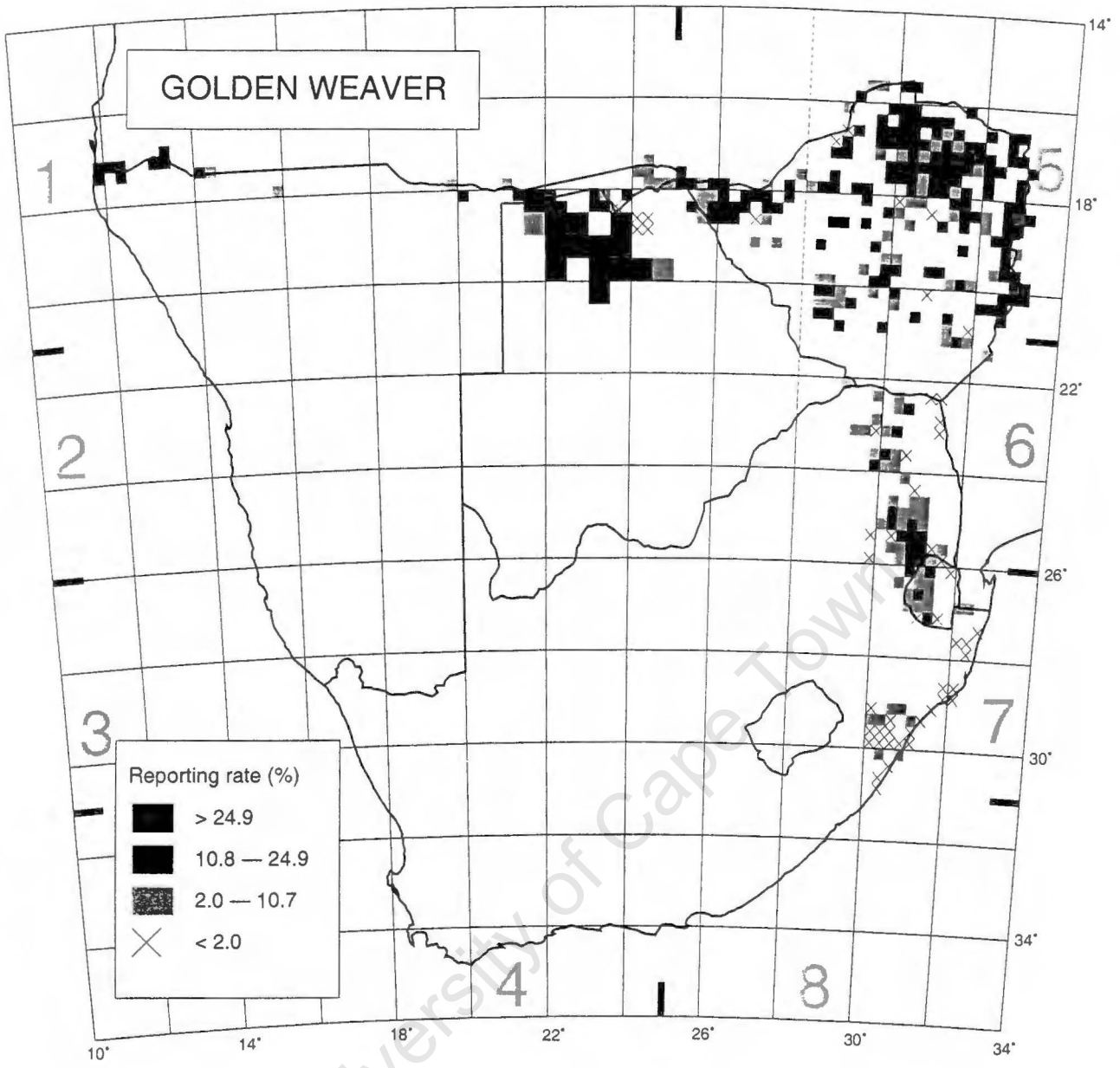


Figure A19. The atlas distribution of the Golden Weaver *Ploceus xanthops*.

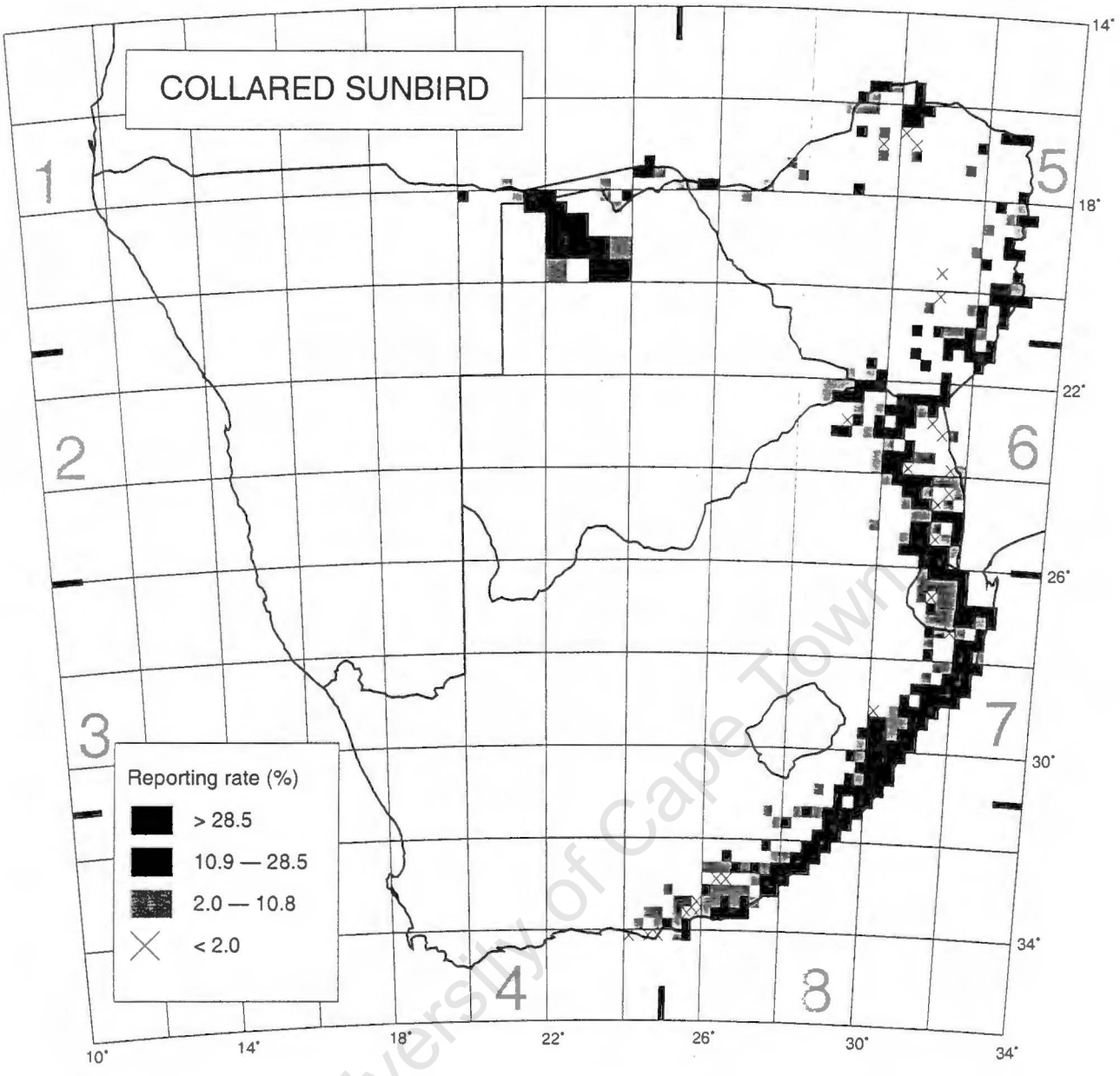


Figure A20. The atlas distribution of the Collared Sunbird *Anthreptes collaris*.

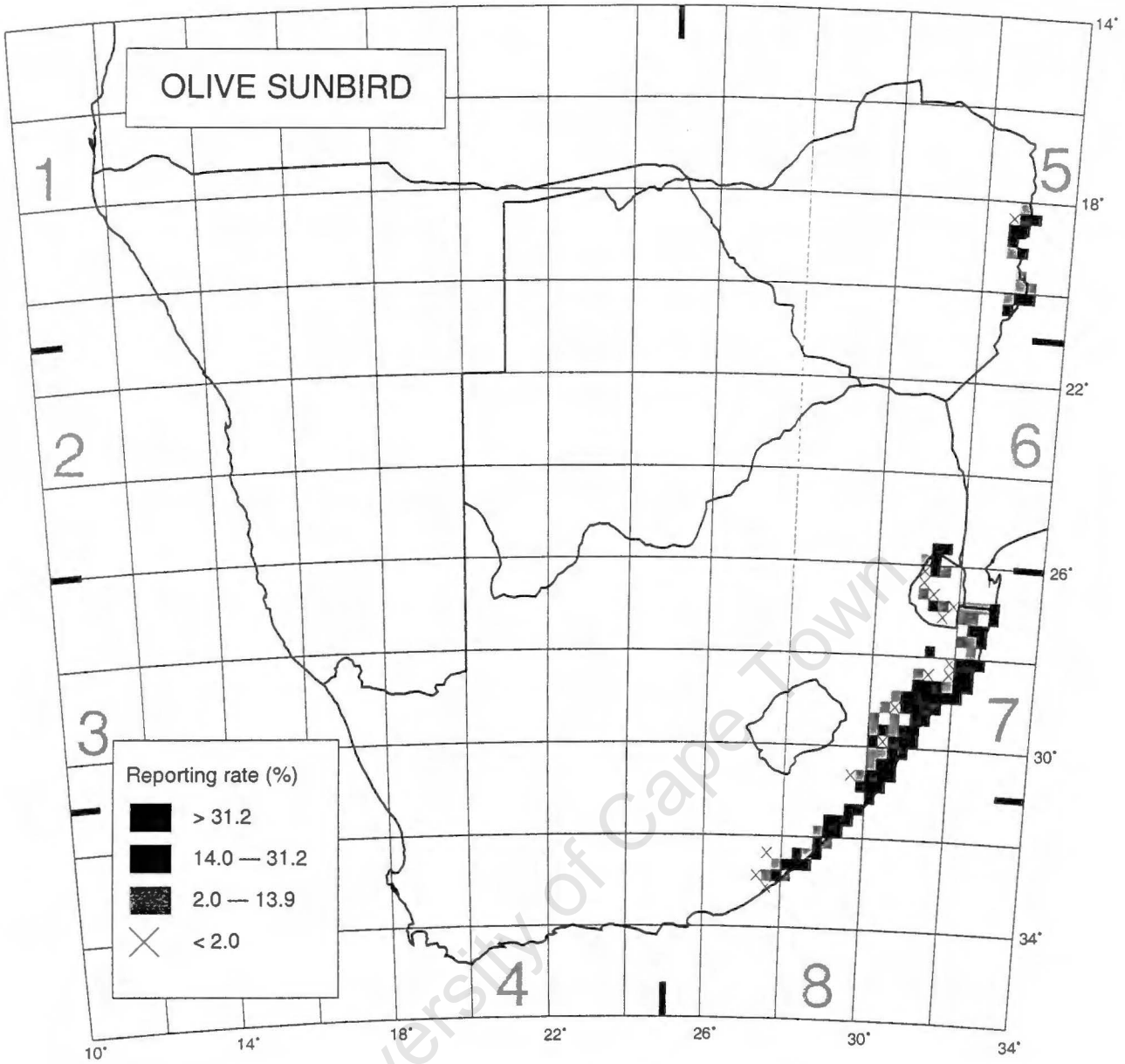


Figure A21. The atlas distribution of the Olive Sunbird *Nectarinia olivacea*.

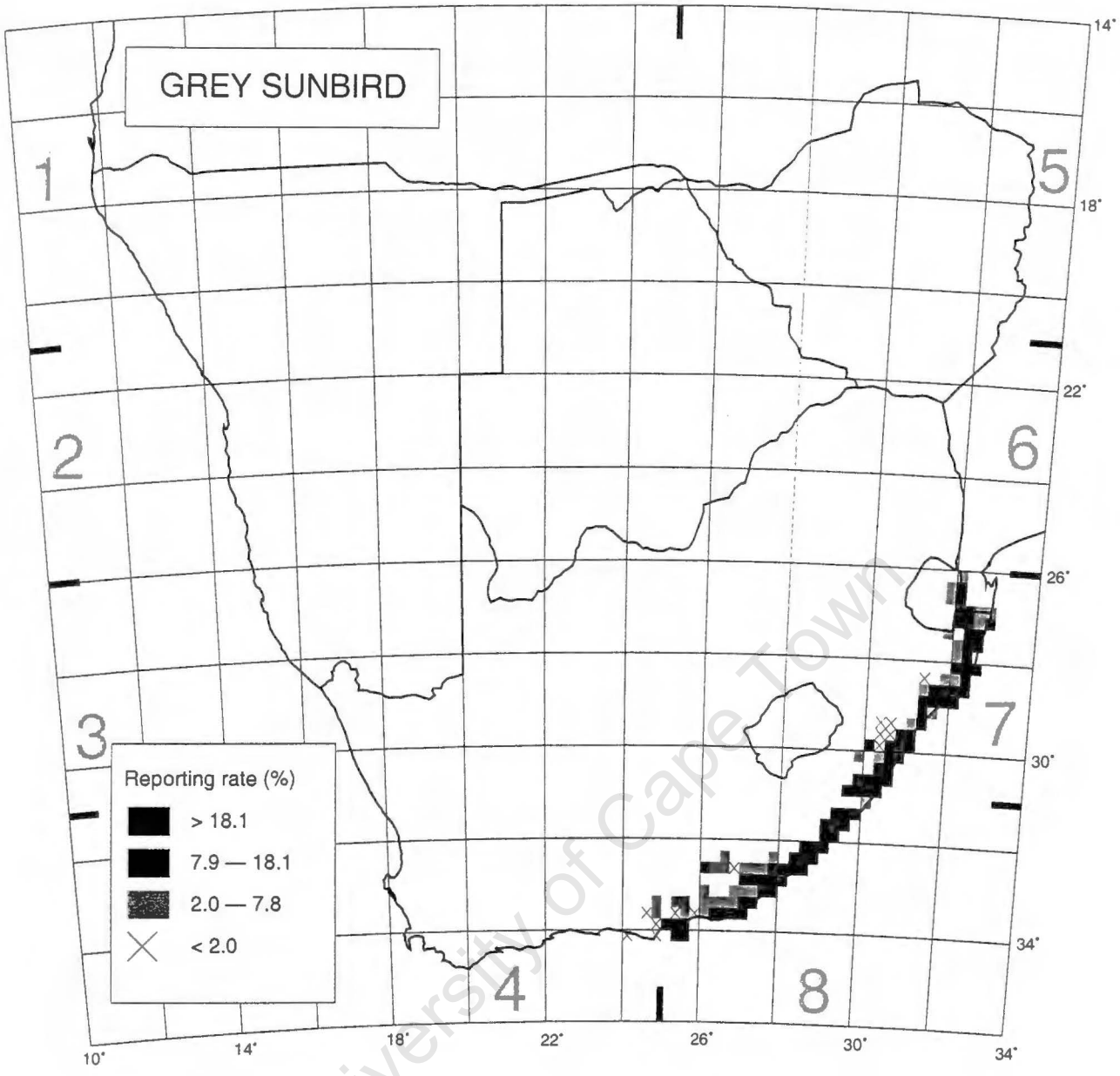


Figure A22. The atlas distribution of the Grey Sunbird *Nectarinia veroscii*.

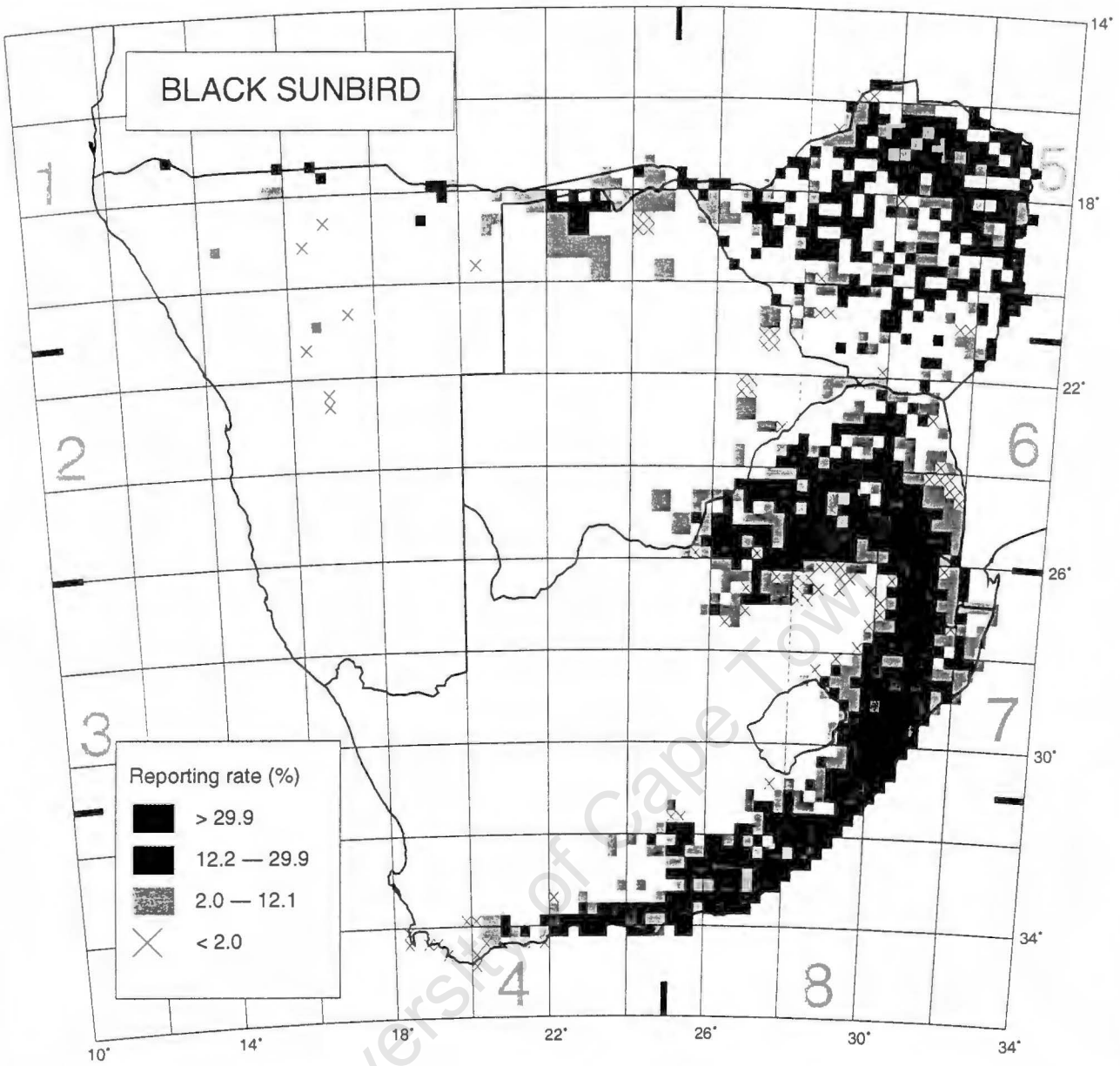


Figure A23. The atlas distribution of the Black Sunbird *Nectarinia amethystina*.

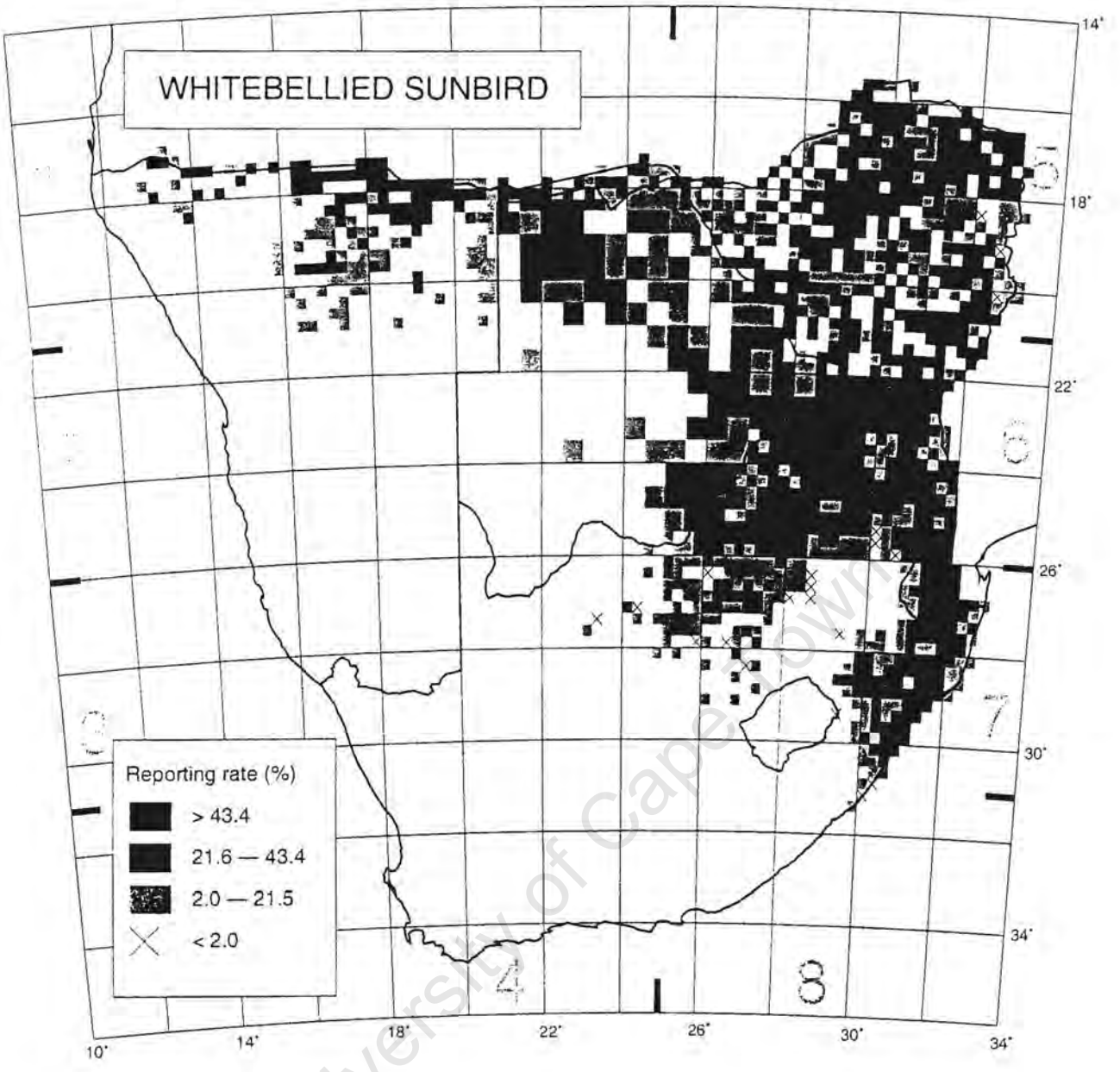


Figure A24. The atlas distribution of the Whitebellied Sunbird *Nectarinia talatala*.

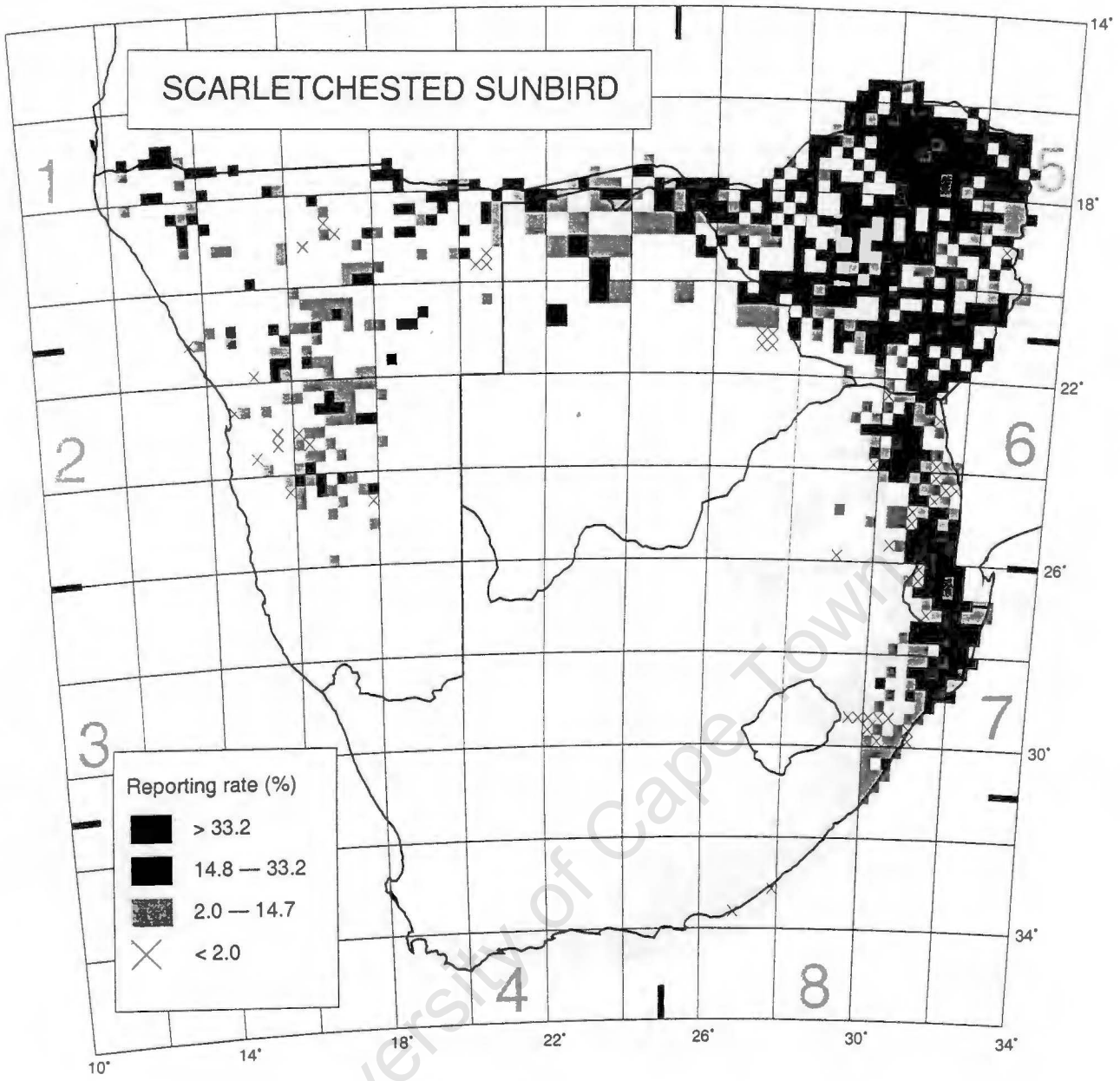


Figure A25. The atlas distribution of the Scarletchested Sunbird *Nectarinia senegalensis*.

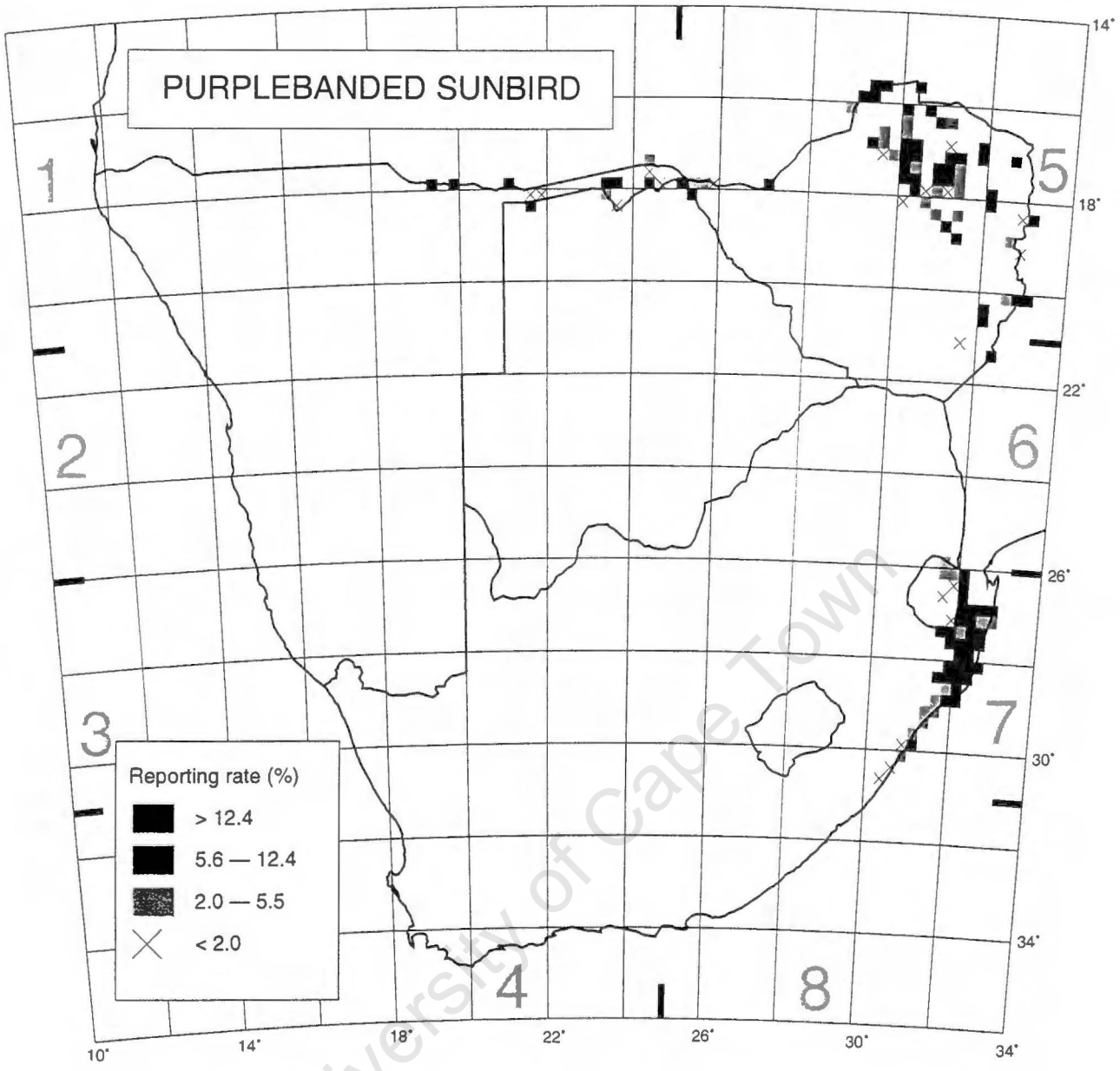


Figure A27. The atlas distribution of the Purplebanded Sunbird *Nectarinia bifasciata*.

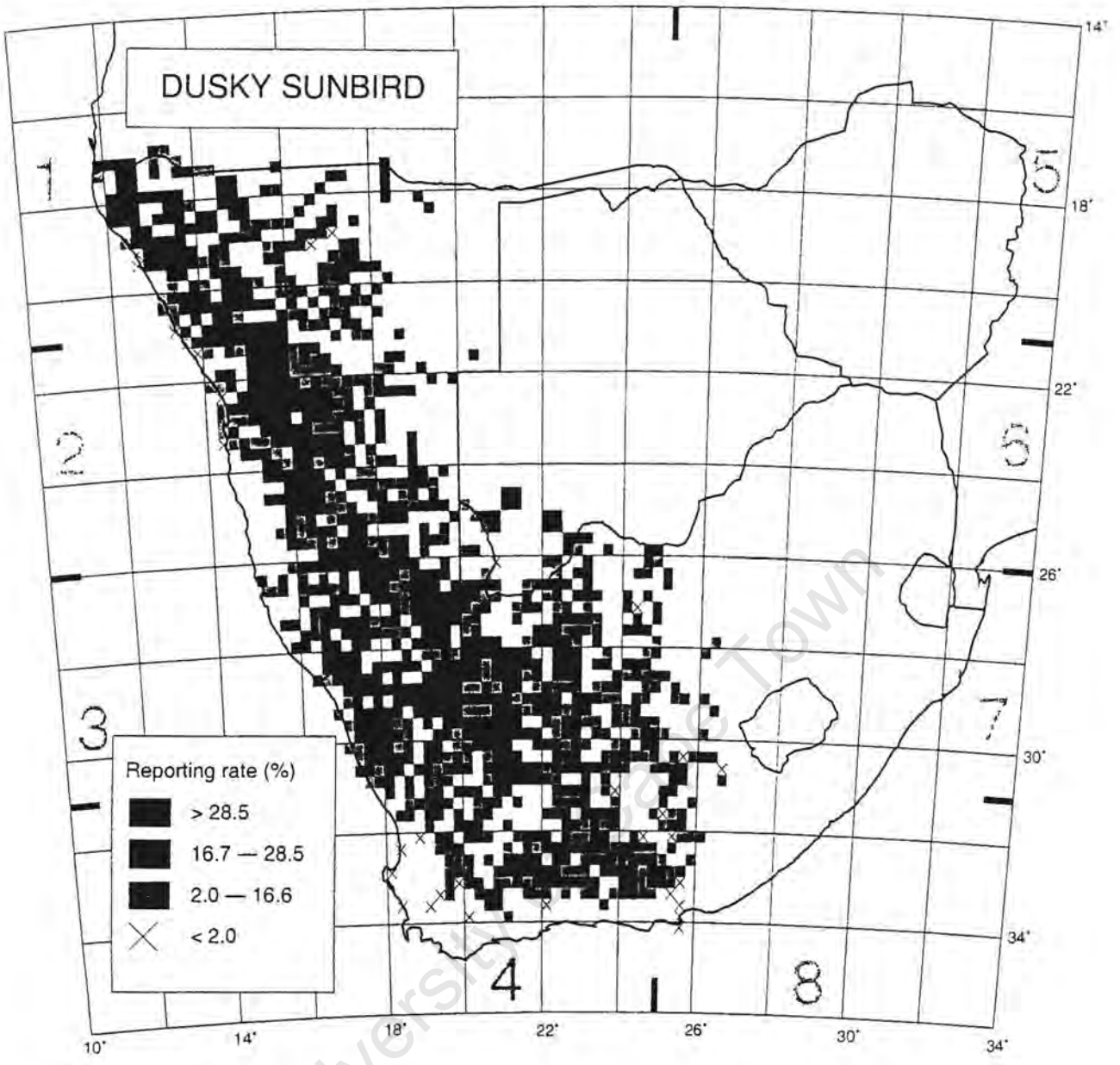


Figure A28. The atlas distribution of the Dusky Sunbird *Nectarinia fusca*.

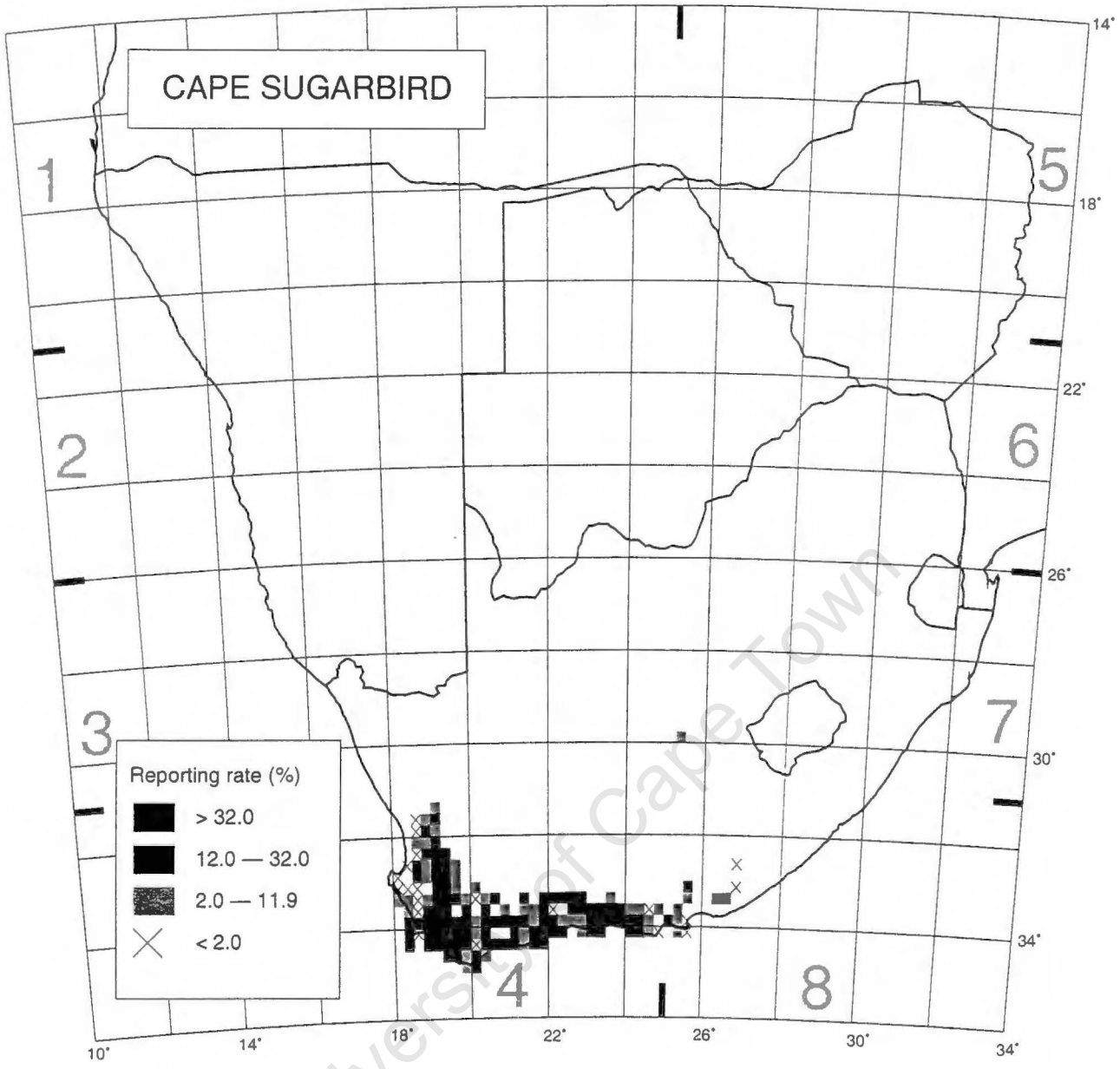


Figure A29. The atlas distribution of the Cape Sugarbird *Promerops cafer*.

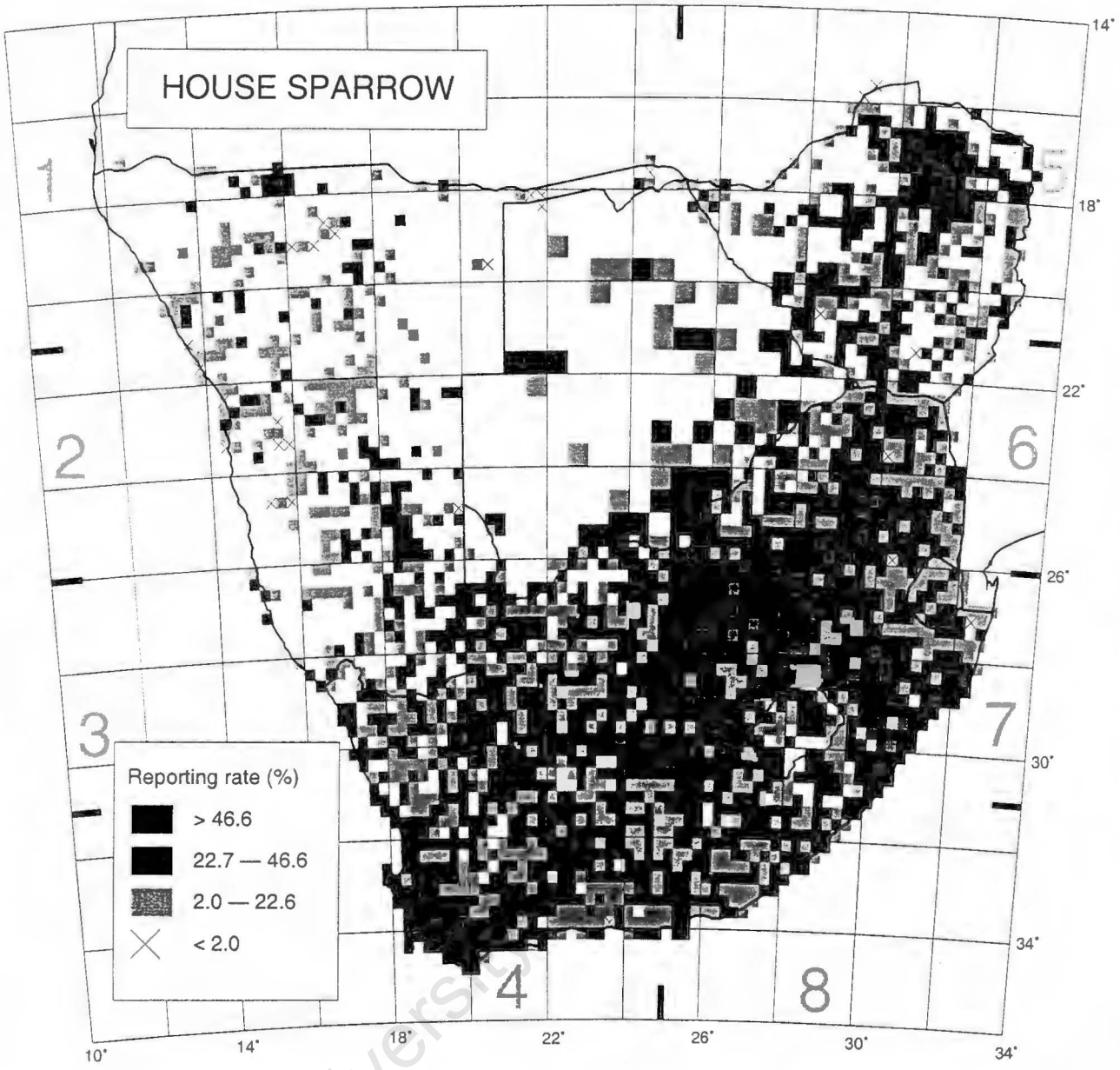


Figure A30. The atlas distribution of the House Sparrow *Passer domesticus*.

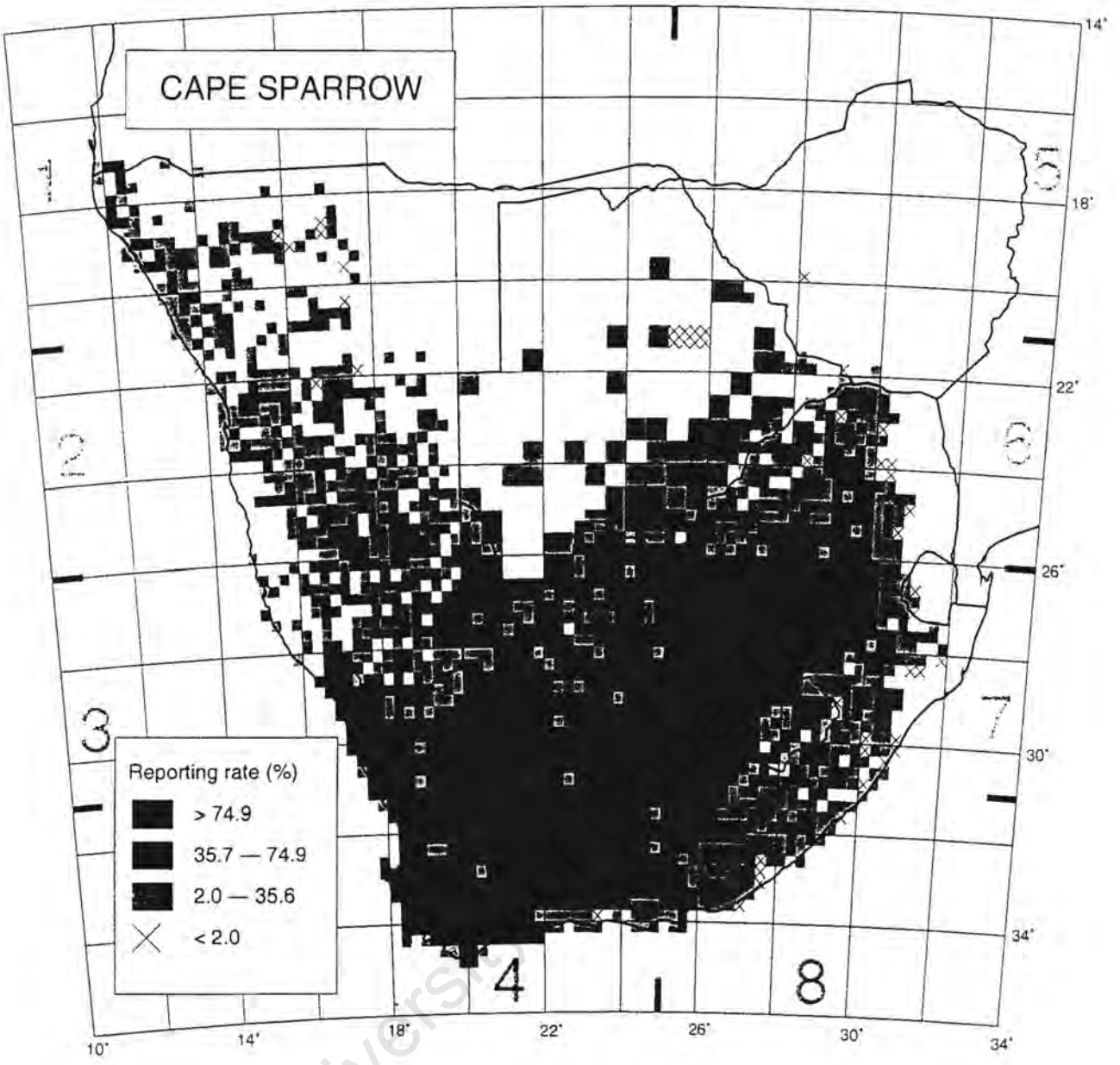


Figure A31. The atlas distribution of the Cape Sparrow *Passer melanurus*.

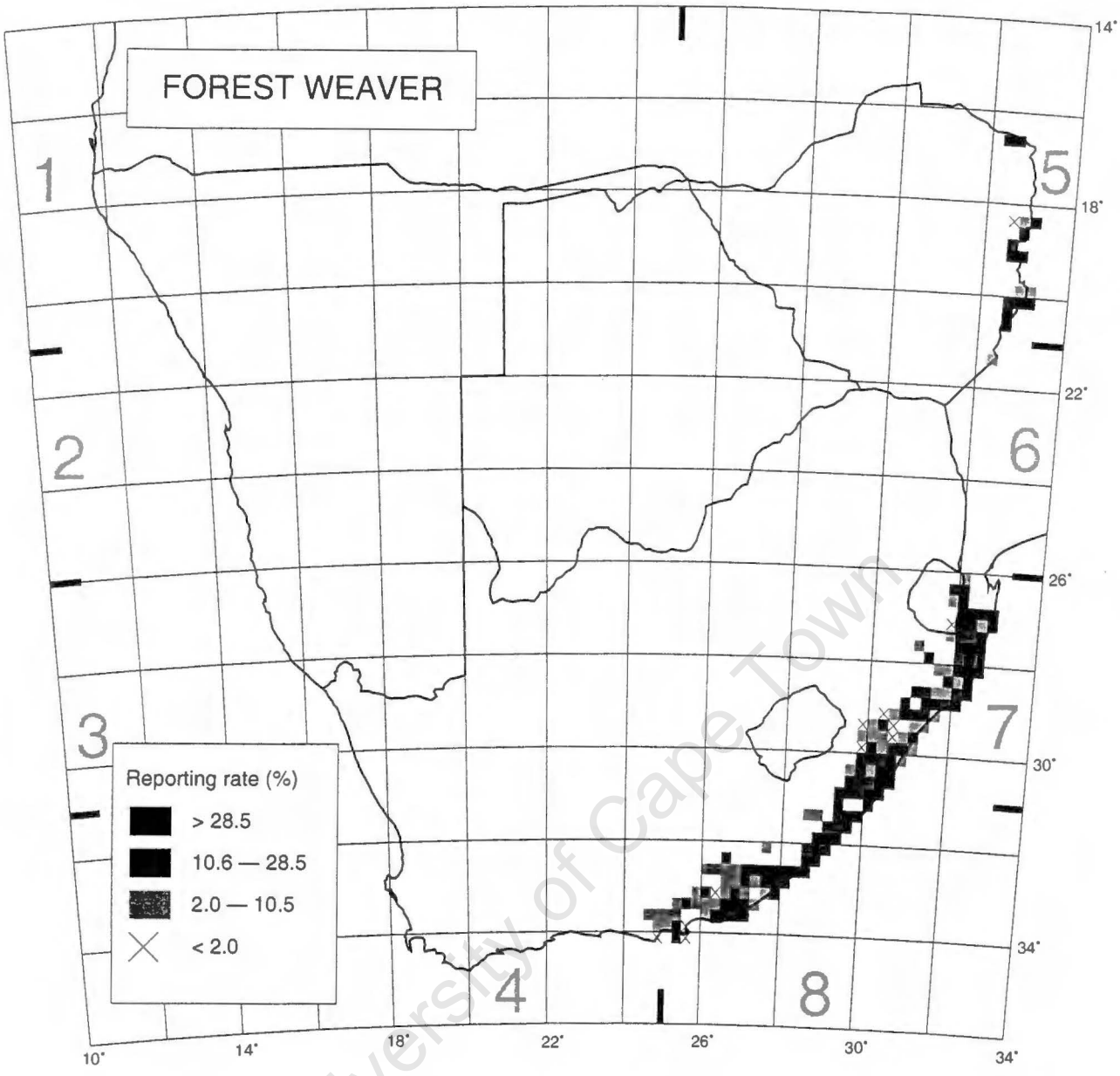


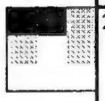
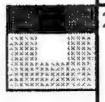
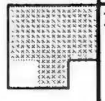
Figure A32. The atlas distribution of the Forest Weaver *Ploceus bicolor*.

APPENDIX B

The Tables and Figures for Chapter 3 are kept in this appendix.

TABLE 1 Comparison of different regression methods for data from the **Masked Weaver** for 14 selected quarter degree grid cells. To the left of the table are shown the original shades for the grid cells. Values shown in the table are the estimated coefficients for the following explanatory variables: constant, x, y, interaction xy. The shaded values are the predicted reporting rates for the central grid cell. The dispersion parameter is shown in the sixth column for each block of cells.

?	t > 2	significant
*	t > 3	highly significant
**	t > 10	very highly significant

	GRID CELL	no. of checklists	COMMENTS	obs. RepR central cell	GENSTAT	significant terms and disp. par.	NR +0.5	WR + 0.5	W AVG	W = n	NR - 0.5	WR - 0.5	ITERATIVELY REWEIGHTED LEAST SQUARES REGRESSION 2nd	3rd	4th	5th
	2925CB	18, 12, 12	no trend	13 / 19	const: 0.57	*	0.59	0.50		0.56	0.71	0.57	0.57	0.57	same	same
		11, 19, 14			x: - 0.06		-0.09	-0.02		-0.09	-0.13	-0.01	-0.06	-0.06		
	(a)	13, 25, 17			y: - 0.23		-0.10	-0.24		-0.19	-0.09	-0.30	-0.23	-0.23		
					xy: - 0.03		-0.11	0.02		-0.06	-0.19	0.03	-0.03	-0.03		
				= 0.68	pred: 0.638	2.18	0.644	0.623	0.639	0.636	0.671	0.640	0.638	0.638		
	3024DA	39, 8, 13	no trend ?	0 / 2	0.43		0.41	0.39		0.47	0.30	0.35	0.42	0.43	same	same
		16, 2, 69			0.23		0.59	0.20		0.27	0.09	0.13	0.23	0.23		
	(b)	23, 13, 6			-0.95	*	-0.93	-0.79		-1.16	-0.41	0.48	-0.94	-0.95		
					0.00		-0.19	0.13		-0.29	0.72	0.61	0.01	0.00		
				= 0	0.605	2.39	0.6	0.596	0.599	0.616	0.575	0.586	0.605	0.605		
	2927CD	58, 18, 19		2 / 8	-1.33	*	-1.44	-1.02		-1.44	-0.91	-0.69	-1.27	-1.33	-1.33	same
		49, 8, 21	x negative		-0.66	?	-0.94	-0.35		-0.82	0.11	0.12	-0.60	-0.66	-0.66	
	(c)	39, 23, 4			0.95	*	0.50	0.74		0.53	1.38	1.26	0.88	0.95	0.95	
					0.69	?	-0.12	0.50		0.18	0.92	1.12	0.62	0.68	0.68	
				= 0.25	0.209	3.91	0.192	0.264	0.255	0.192	0.287	0.334	0.220	0.210	0.210	
	2919AC	78, 8, 27		0 / 7	-1.99	*	-1.87	-1.67		-1.84	-1.41	-1.41	-1.93	-1.98	-1.99	same
		69, 7, 17			-0.75	?	-0.67	-0.53		-0.70	-0.27	-0.27	-0.70	-0.75	-0.75	
	(d)	7, 4, 12	y negative		-0.84	?	-0.77	-0.66		-0.72	0.23	0.24	-0.80	-0.84	-0.84	
					0.73		0.67	0.79		0.74	1.85	1.71	0.77	0.74	0.73	
				= 0	0.121	1.56	0.133	0.159	0.219	0.137	0.196	0.196	0.127	0.121	0.121	
	2930AA	192, 120, 35	Note: number of check-	0 / 50	-1.56	**	-1.93	-1.63		-1.63	-1.63	-1.49	-1.55	-1.56	same	same
		258, 50, 157	lists for central cell = 50		-0.17		-0.33	0.08		-0.19	-0.35	-0.15	-0.19	-0.17		
	(e)	221, 416, 333	y negative ?		-1.31	**	-1.16	-1.34		-1.26	-1.18	-1.32	-1.30	-1.31		
					-0.51	*	-0.35	-0.54		-0.58	-0.36	-0.57	-0.52	-0.51		
				= 0	0.174	9.76	0.127	0.164	0.177	0.163	0.163	0.184	0.176	0.174		
	3129AB	15, 12, 9	Blank area	0 / 4	does not		-2.63	-2.67		-2.76			-3.77	-4.80	-5.81	-6.81
		2, 4, 5			run (as		-0.55	-0.56		-0.61	does not	does not	-0.49	-0.47	-0.47	-0.46
	(f)	1, 7, 29	6th iteration: 0.0004		expected)		0.29	0.31		0.31	work	work	0.26	0.24	0.24	0.24
			7th iter: 0				-0.87	-0.91		-0.95			-0.82	-0.79	-0.79	-0.78
				= 0			0.067	0.065	0	0.060			0.023	0.008	0.003	0.001
	2930DD	132, 62, 448	the one missing cell	48 / 1260	-3.13	**	-3.05	-3.18		-3.08	-3.15	-3.10	-3.06	-3.12	-3.09	same
		122,1260, 729	falls in ocean		-0.34		-0.01	-0.52		-0.44	-0.26	-0.48	-0.45	-0.35	-0.34	
	(g)	28, 521, 0			-0.16		-0.23	0.23		-0.05	-0.21	-0.09	-2.05	-0.16	-0.16	
			no sign. trend		-0.07		0.44	-0.87		-0.22	0.01	-0.28	-0.06	-0.06	-0.07	
				= 0.038	0.042	2.30	0.045	0.040	0.043	0.044	0.041	0.043	0.046	0.042	0.042	

	GRID CELL	no. of checklists	COMMENTS	obs. RepR central cell	GENSTAT	significant terms and disp. par.	NR +0.5	WR + 0.5	W AVG	W = n	NR - 0.5	WR - 0.5	ITERATIVELY REWEIGHTED LEAST SQUARES REGRESSION 2nd	3rd	4th	5th
	2726DD	32, 13, 111	constant	32 / 43	1.13	**	1.05	1.05		1.18	1.11	1.08	1.13	1.13	same	same
		162, 43, 30			-0.43	*	-0.39	-0.42		-0.38	-0.41	-0.44	-0.43	-0.43		
	(h)	65, 72, 54			-0.11		0.03	-0.20		-0.02	0.02	-0.23	-0.11	-0.11		
					-0.17		-0.17	-0.13		-0.13	-0.16	-0.12	-0.16	-0.17		
				= 0.744	0.756	5.20	0.740	0.740	0.744	0.765	0.753	0.747	0.755	0.756		
	3126BC	19, 9, 12	no sign. trend	37 / 69	-1.10	*	-1.57	-0.85		-1.33	-1.80	-0.83	-1.09	-1.10	-1.10	same
		36, 69, 54			-0.13		-0.10	-0.26		-0.02	-0.17	-0.34	-0.14	-0.13	-0.13	
	(i)	8, 13, 15			-0.49		-0.57	-0.36		-0.69	-0.42	-0.20	-0.48	-0.49	-0.49	
					-0.31		-0.35	-0.36		-0.26	-0.51	-0.38	-0.32	-0.31	-0.31	
				= 0.5362	0.250	10.10	0.172	0.299	0.220	0.208	0.142	0.304	0.252	0.250	0.250	
	2925CA	11, 18, 12	7th grid cell : RR = 1	5 / 11	0.67	?	0.57	0.56		0.62	0.66	0.61	0.67	0.67	same	same
		9, 11, 19			-0.04		-0.03	0.00		-0.05	0.01	0.06	-0.04	-0.04		
	(j)	4, 13, 25			0.33		0.39	0.26		0.31	0.46	0.27	0.33	0.33		
			negative interaction		-1.29	*	-1.12	-1.10		-1.14	-1.28	-1.25	-1.28	-1.29		
				= 0.4545	0.662	1.26	0.638	0.636	0.612	0.649	0.659	0.648	0.661	0.662		
	3218DD	7, 14, 23		5 / 20	-0.69	*	-0.68	-0.59		-0.61	-0.89	-0.75	-0.68	-0.69	same	same
		7, 20, 15			0.65	?	0.46	0.61		0.61	0.63	0.77	0.64	0.65		
	(k)	67, 16, 55			0.58	?	0.46	0.50		0.53	0.65	0.65	0.58	0.58		
					0.45		0.38	0.46		0.48	0.19	0.35	0.45	0.45		
				= 0.25	0.335	1.73	0.338	0.356	0.365	0.352	0.292	0.320	0.336	0.335		
	3320DD	18, 22, 28		20 / 47	-0.33	?	-0.34	-0.31		-0.34	-0.40	-0.35	0.33	-0.33	same	same
		51, 47, 23			-0.20		-0.19	-0.18		-0.21	-0.26	-0.22	-0.20	-0.20		
	(m)	38, 30, 27	y negative		-1.13	*	-1.11	-1.06		-1.11	-1.26	-1.18	-1.13	-1.13		
					-0.64	?	-0.61	-0.60		-0.62	-0.73	-0.69	-0.64	-0.64		
				= 0.4255	0.419	0.47	0.415	0.423	0.431	0.415	0.400	0.415	0.419	0.419		
	2918BC	5, 4, 78	6th iteration: 0.0035	0 / 12	warning:		-1.94	-1.96		-2.09		8.84	-2.59	-3.14	-3.82	-4.70
		10, 12, 69			iterative		0.80	0.84		0.92	does not	-0.51	1.38	1.94	2.63	3.51
	(n)	8, 2, 7	7th iter: 0		weights have		-0.70	-0.78		-0.79	work	4.85	-0.98	-1.41	-2.22	-3.13
					become		-0.65	-0.64		-0.69	Det = 0	15.99	-0.57	-0.13	0.68	1.60
				= 0	zero	0.29	0.126	0.123	0.219	0.110		0.9999	0.070	0.041	0.021	0.009
	2918AB	7, 7, 4	6th iteration: 0.007	0 / 5	warning:		-1.67	-1.67		-1.71		-1.66	-2.30	-2.83	-3.42	-4.16
		3, 5, 4	7th iter: 0		iterative		0.40	0.41		0.40	does not	-0.81	0.73	1.12	1.75	2.52
	(p)	4, 10, 12			weights have		-1.18	-1.17		-1.19	work	-1.66	-1.67	-2.22	-2.83	-3.58
					become		-0.83	-0.79		-0.77	Det = 0	2.20	-0.81	-0.46	0.18	0.96
				= 0	zero	0.16	0.158	0.158	0.143	0.153		0.160	0.091	0.056	0.032	0.015

REPORTING RATES (%)

2.0 - 24.9

25.0 - 49.9

> 49.9

TABLE 3

Regression results for the **Masked Weaver**, when taking blocks of size 25 grid cells. Same grid cells as in Table 1. As comparison the GENSTAT results for blocks of size 9 were included. Significant parameters are indicated by a '?' if the t-value was larger than 2 and by a '*' if the t-value was larger than 3. '**' means that the parameter is very highly significant for the model. Results are shown when the square terms(x^2 and y^2) are included and when they are excluded. The shades of the blocks to the left of the table are those used in the atlas maps.

	GRID CELL	Parameter	GENSTAT 9 cells	signif	GENSTAT, 25 cells with square terms	signif	disp. par.	GENSTAT, 25 cells only x, y, xy	signif.	disp. par.	WR ÷ 0.5	NR ÷ 0.5
	2925CB	const	0.57	*	0.66			0.75	*		0.65	0.65
		x	-0.06		0.20	?		0.17	?		0.18	0.20
		y	-0.23		0.06			0.05			0.04	0.14
		xy	-0.03		-0.10			-0.08			-0.06	-0.11
	(a)	x^2			-0.10							
		y^2			0.13							
		predict	0.638		0.660	0.047	1.86	0.680	0.023	2.02	0.657	0.657
	3024DA	const	0.43	?	0.29			0.67	*		0.60	0.52
		x	0.23		-0.05			-0.08			-0.07	-0.16
		y	-0.95	*	-0.42	*		-0.42	*		-0.35	-0.35
		xy	0.00		0.01			0.04			0.06	0.02
	(b)	x^2			0.08							
		y^2			0.13	?						
		predict	0.605		0.573	0.048	2.65	0.661	0.021	2.67	0.645	0.627
	2927CD	const	-1.33	*	-0.81	*		-0.43	*		-1.25	-1.04
		x	-0.66	?	-0.41	*		-0.19	?		-0.49	-0.58
		y	0.95	*	-0.17	*		-0.18	?		-0.68	-0.12
		xy	0.69	?	0.05			0.06			-0.13	0.02
	(c)	x^2			-0.22	*						
		y^2			0.21	*						
		predict	0.209		0.307	0.038	5.83	0.395	0.031	7.46	0.223	0.261
	2919AC	const	-1.99	*	-1.30	*		-1.50	*		-1.00	-1.06
		x	-0.75	?	-0.07			-0.07			-0.24	-0.23
		y	-0.84	?	-0.62	*		-0.69	*		-0.39	-0.17
		xy	0.73		0.72	*		0.70	*		0.30	0.20
	(d)	x^2			-0.19							
		y^2			0.05							
		predict	0.121		0.214	0.047	2.36	0.183	0.025	2.27	0.270	0.257
X	2930AA	const	-1.56	**	-1.41	**		-1.65	**		-1.56	-1.90
		x	-0.17		0.17	*		0.19	*		0.19	0.14
		y	-1.31	**	-0.97	**		-0.95	**		-0.84	-0.89
		xy	-0.51	*	0.17	*		0.11	*		0.13	0.27
	(e)	x^2			-0.21	*						
		y^2			-0.03							
		predict	0.174		0.196	0.014	7.02	0.161	0.008	8.06	0.174	0.130
X	2930DD	const	-3.13	**	-2.97	**		-3.34	**		-2.32	-2.83
		x	-0.34		-0.32	?		0.06			0.00	-0.18
		y	-0.16		-0.58	*		-0.15			0.61	0.02
		xy	-0.07		0.00			0.16	?		0.10	-0.01
	(f)	x^2			-0.22	?						
		y^2			-0.27	*						
		predict	0.042		0.049	0.005	3.48	0.034	0.003	5.27	0.089	0.056
	2726DD	const	1.13	**	1.30	*		1.09	*		0.97	0.95
		x	-0.43	*	-0.04			-0.05			-0.06	0.05
		y	-0.11		-0.26	*		-0.26	*		-0.26	-0.21
		xy	-0.17		0.04			-0.02			-0.01	0.03
	(g)	x^2			0.00							
		y^2			-0.14	*						
		predict	0.756		0.786	0.022	4.00	0.749	0.013	4.04	0.725	0.721

REPORTING RATES (%)

X < 2.0

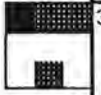
2.0 - 24.9

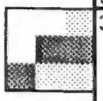
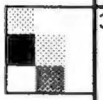
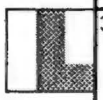
25.0 - 49.9

> 49.9

TABLE 2.

Regression results for 10 grid cells taken from the distribution of the **Cape Weaver** are shown. The values in the table are the estimated parameter coefficients for the constant, x, y and the interaction term xy respectively. The shaded rows contain the **predicted reporting rates for the central grid cell.** The shades blocks to the left of the table are as they appear n the atlas (Harrison *et al.* 1997b). In the 6th column those parameters which were significant in the GENSTAT results are marked by a '*' if the t-value was larger than three and by a '?' if the t-value was larger than two. The values in this column are the approximate errors of the prediction. The dispersion parameter is shown in the sixth column.

	grid cell	number of checklists	COMMENTS	observed RepRate in central cell	GENSTAT	signif. & disp.par.	NR +0.5	WR +0.5	W AVG	W = ni	NR - 0.5	WR - 0.5	ITERATIVELY REWEIGHTED LEAST SQUARES REGRESSION 2nd	3rd	4th	5th
	3129AA	17, 15, 12	constant, no trend?	0 / 2	-3.02	*	-2.19	-2.21		-2.43	-2.45	-2.45	-2.78	-2.99	-3.02	
		17, 2, 4			0.24		0.42	0.33		0.33	0.84	0.84	0.29	0.25	0.24	same
		13, 1, 7			0.78		0.73	0.59		0.64	-0.32	-0.32	0.67	0.76	0.78	
	(a)				0.53		0.37	0.35		0.46	-0.47	-0.47	0.42	0.51	0.53	
				0	0.047	1.0	0.100	0.099	0.053	0.081	0.080	0.080	0.058	0.048	0.047	
	3029CC	34, 4, 78		0 / 15	-2.78	*	-1.47	-1.40		-1.79	-2.34	-2.42	-2.14	-2.59	-2.76	-2.78
		17, 15, 12			0.60		-0.25	-0.04		-0.50	1.09	1.02	0.16	0.42	0.58	0.60
		17, 2, 4	neg. y coeff.		-3.37	*	-1.31	-1.72		-2.01	-3.31	-3.15	-2.65	-3.17	-3.34	-3.37
	(b)		positive interaction		2.44	*	1.52	1.53		1.05	2.95	2.94	1.93	2.25	2.41	2.44
				0	0.058	1.4	0.187	0.198	0.269	0.144	0.088	0.082	0.105	0.070	0.060	0.058
	3029CD	4, 78, 62		0 / 12	-2.93	*	-1.75	-1.70		-2.27		-0.38	-2.37	-2.78	-2.92	-2.93
		15, 12, 9	negative y term		0.18		-0.35	-0.06		0.30	Matrix	2.02	0.05	0.13	0.17	0.18
		2, 4, 5			-1.91	?	-0.59	-0.64		-1.07	not	0.46	-1.33	-1.77	-1.90	-1.91
	(c)				0.29		0.25	-0.03		0.17	invertible	2.38	0.14	0.24	0.28	0.29
				0	0.051	2.7	0.148	0.155	0.179	0.094		0.406	0.086	0.058	0.051	0.051
	3226BA	18, 23, 125		4 / 10	-0.30	?	-0.69	-0.19		-0.38	-0.89	-0.19	-0.30	-0.30	-0.30	
		51, 10, 81			0.45	*	0.45	0.37		0.50	0.50	0.37	0.45	0.45	0.45	same
		40, 15, 160	smoothing valid ?		-0.07		0.09	-0.18		0.01	0.21	-0.21	-0.07	-0.07	-0.07	
	(d)				-0.11		-0.11	-0.02		-0.15	-0.13	0.00	-0.10	-0.10	-0.10	
				= 0.4	0.426	8.4	0.333	0.454	0.426	0.407	0.290	0.454	0.427	0.426	0.426	
	3219BB	7, 3, 6		0 / 6	-2.83	*	-1.98	-1.87		-2.01	-1.39	-1.39	-2.46	-2.76	-2.82	
		8, 6, 6			-0.03		0.21	0.07		0.12	-0.81	-0.81	-0.01	-0.03	-0.03	same
		24, 8, 7			-0.16		-0.34	-0.17		-0.14	-0.29	-0.29	-0.13	-0.15	-0.16	
	(e)		interaction negative		-1.74	?	-0.77	-0.89		-0.95	0.29	0.29	-1.35	-1.66	-1.73	
				0	0.056	1.2	0.122	0.134	0.105	0.118	0.200	0.200	0.078	0.060	0.056	
	3320BB	17, 12, 5	edge of distribution	3 / 26	-2.21	*	-1.76	-1.77		-1.81	-2.38	-2.34	-2.11	-2.20	-2.21	
		24, 26, 8	negative x term		-1.13		-0.66	-0.72		-0.72	-1.15	-1.15	-1.03	-1.12	-1.13	same
		24, 19, 9			0.53	?	0.31	0.35		0.30	0.57	0.51	0.46	0.52	0.53	
	(f)				-0.06		-0.11	-0.19		-0.17	0.02	-0.17	-0.13	-0.06	-0.06	
				= 0.116	0.099	0.45	0.146	0.145	0.156	0.141	0.084	0.088	0.108	0.099	0.099	
	3321AA	12, 5, 11	y, x negative ?	0 / 8	warning:		-2.20	-2.07		-2.12		0.00	-2.65	-3.20	-3.82	-4.51
		26, 8, 13	interaction positive ?		iterative		-0.17	-0.07		-0.12	Matrix	1.05	-0.42	-0.96	-1.59	-2.29
		19, 9, 42	6th it. pred = .005		weights		0.57	0.76		0.75	not	-1.11	1.25	1.81	2.42	3.12
	(g)		7th it: pred = 0		have become zero		0.64	0.82		0.79	invertible	-0.58	1.28	1.83	2.46	3.15
				0	0.000	0.13	0.100	0.112	0.120	0.107		0.500	0.066	0.039	0.022	0.011

	grid cell	number of checklists	COMMENTS	observed RepRate in central cell	GENSTAT	signif. & disp.par.	NR +0.5	WR +0.5	W AVG	W = ni	NR - 0.5	WR - 0.5	ITERATIVELY REWEIGHTED LEAST SQUARES REGRESSION 2nd	3rd	4th	5th	
	3123DA	9, 16, 25		3 / 14	-2.68	*	-2.15	-2.10		-2.34	-1.44	-1.44	-2.5	-2.65	-2.68	-2.68	
		10, 14, 5			0.19		0.07	-0.11		-0.09	-1.41	-1.49	0.04	0.16	0.19	0.19	
		11, 31, 54			1.06	?	0.74	0.61		0.68	-0.21	-0.28	0.88	1.03	1.06	1.06	
	(h)				-0.91		-0.46	-0.55		-0.44	0.76	0.73	-0.74	-0.88	-0.90	-0.91	
				= 0.2142	0.064	1.60	0.104	0.110	0.104	0.088	0.192	0.192	0.076	0.066	0.064	0.064	
	3019CB	8, 5, 9		2 / 17	-2.67	*	-2.16	-1.97		-2.06	-2.27	-2.27	-2.42	-2.61	-2.67		
		7, 17, 9			-1.38	?	-0.74	-0.88		-0.78	-1.93	-1.93	-1.12	-1.32	-1.38	same	
		7, 21, 11			0.51		-0.09	0.08		-0.01	0.32	0.32	0.26	0.45	0.51		
	(j)				0.62		0.23	0.10		0.15	-1.91	-1.91	0.35	0.55	0.62		
				= 0.1176	0.065	1.64	0.103	0.123	0.093	0.113	0.094	0.094	0.082	0.068	0.065		
	3121AB	6, 7, 3		1 / 6	-2.37	*	-1.90	1.71		-1.81			-2.20	-2.35	-2.37		
		7, 6, 5			0.50		0.49	0.49		0.56	Matrix	Matrix	0.51	0.50	0.50	same	
		8, 12, 4			0.68		0.22	0.29		0.28	not	not	0.53	0.66	0.68		
	(k)				0.52		0.34	0.32		0.39	invertible	invertible	0.46	0.52	0.52		
				= 0.1667	0.085	1.27	0.131	0.153	0.096	0.140			0.100	0.087	0.085		

REPORTING RATES %

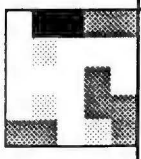
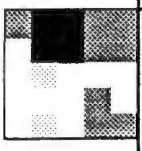
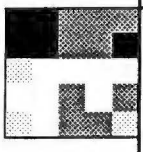
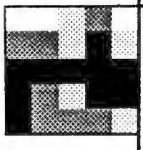
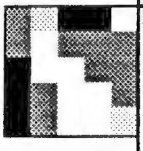
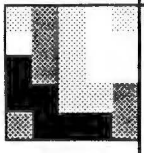
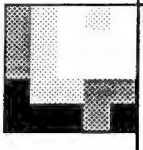
2.0 - 14.2

14.3 - 39.3

> 39.3

TABLE 4

Comparison of different Regression methods for grid cells taken from the Cape Weaver when taking blocks of size 25 grid cells. Significant parameters are marked by a '?' if the t-value was larger than 2 and by a '*' if it was larger than 3 and the corresponding change in deviance was larger than 5. Same central grid cells as in Table 2.

	GRID CELL	Parameter	GENSTAT 9 cells	signif.	GENSTAT, 25 cells with square terms	signif.	disp. par.	GENSTAT, 25 cells only x, y, xy	signif.	disp. par.	WR + 0.5	NR + 0.5
	3129AA	constant	-3.02	*	-4.04	*		-1.13	*		-1.02	-1.81
		x	0.24		-0.34	?		-0.08			-0.04	0.07
		y	0.78		-0.41	*		-0.27	*		-0.16	-0.23
		xy	0.53		0.23	*		0.17	*		0.12	0.08
	(a)	x^2			0.23	*						
		y^2			0.72	*						
		predict	0.047		0.017	0.009	4.66	0.244	0.018	7.40	0.264	0.141
	3029CC	constant	-2.78	*	-1.38	*		-1.49	*		-1.20	-1.52
		x	0.60		0.03			-0.05	?		-0.11	0.03
		y	-3.37	*	-0.99	*		-0.88	*		-0.66	-0.53
		xy	2.44	*	0.48	*		0.43	*		0.34	0.22
	(b)	x^2			-0.02							
		y^2			-0.09							
		predict	0.058		0.201	0.041	5.08	0.184	0.029	4.64	0.231	0.179
	3029CD	constant	-2.93	*	-1.90	*		-1.38	*		-1.15	-1.32
		x	0.18		0.09			-0.07	?		-0.08	-0.16
		y	-1.91	?	-0.85	*		-0.55	*		-0.44	-0.44
		xy	0.29		0.34	*		0.33	*		0.26	0.25
	(c)	x^2			0.36	*						
		y^2			-0.19	?						
		predict	0.051		0.130	0.026	3.58	0.201	0.023	5.60	0.241	0.210
	3226BA	constant	-0.30	?	-0.24			-0.28	*		-0.20	-1.01
		x	0.45	*	0.32	*		0.29	*		0.21	0.04
		y	-0.07		0.33	?		-0.02			-0.06	0.33
		xy	-0.11		-0.25	*		-0.22	*		-0.21	-0.01
	(d)	x^2			0.27	*						
		y^2			-0.50	*						
		predict	0.426		0.441	0.031	7.86	0.429	0.018	15.80	0.450	0.267
	3219BB	constant	-2.83	*	-1.76	*		-1.37	*		-1.09	-1.38
		x	-0.03		-0.24	?		-0.25			-0.17	-0.21
		y	-0.16		-0.09			-0.11			-0.12	-0.19
		xy	-1.74	?	-0.08			-0.07			-0.06	-0.07
	(e)	x^2			0.24	?						
		y^2			-0.06							
		predict	0.056		0.147	0.046	1.84	0.202	0.030	1.91	0.251	0.201
	3320BB	constant	-2.21	*	-1.68	*		-1.56	*		-1.24	-1.49
		x	-1.13		-0.26			-0.26			-0.15	-0.23
		y	0.53	?	0.69	*		0.67	*		0.49	0.47
		xy	-0.06		0.15			0.14			0.08	0.01
	(f)	x^2			0.07							
		y^2			-0.02							
		predict	0.099		0.157	0.033	1.61	0.174	0.024	1.51	0.224	0.184
	3321AA	constant			-2.09	*		-1.93	*		-1.45	-1.66
		x	iterative weigths		-0.33			-0.37			-0.19	-0.27
		y	have become		0.77	*		0.86	*		0.58	0.48
		xy	zero		0.19	?		0.23	?		0.11	0.10
	(g)	x^2	GLM (5 its): 0.011		0.05							
		y^2			0.07							
		predict	0.000		0.110	0.027	1.33	0.127	0.023	1.25	0.190	0.159

REPORTING RATES %

X < 2.0

2.0 - 14.2

14.3 - 39.3

> 39.3

Table 5. Regression results for a particular example from the distribution of the Blackheaded Canary, see also Figure 3. The table shows the different results obtained when using GENSTAT and the C++ program. Significant values from GENSTAT are indicated by a '*'. 'not conv.' is an error message given by GENSTAT which means that the results have not converged.

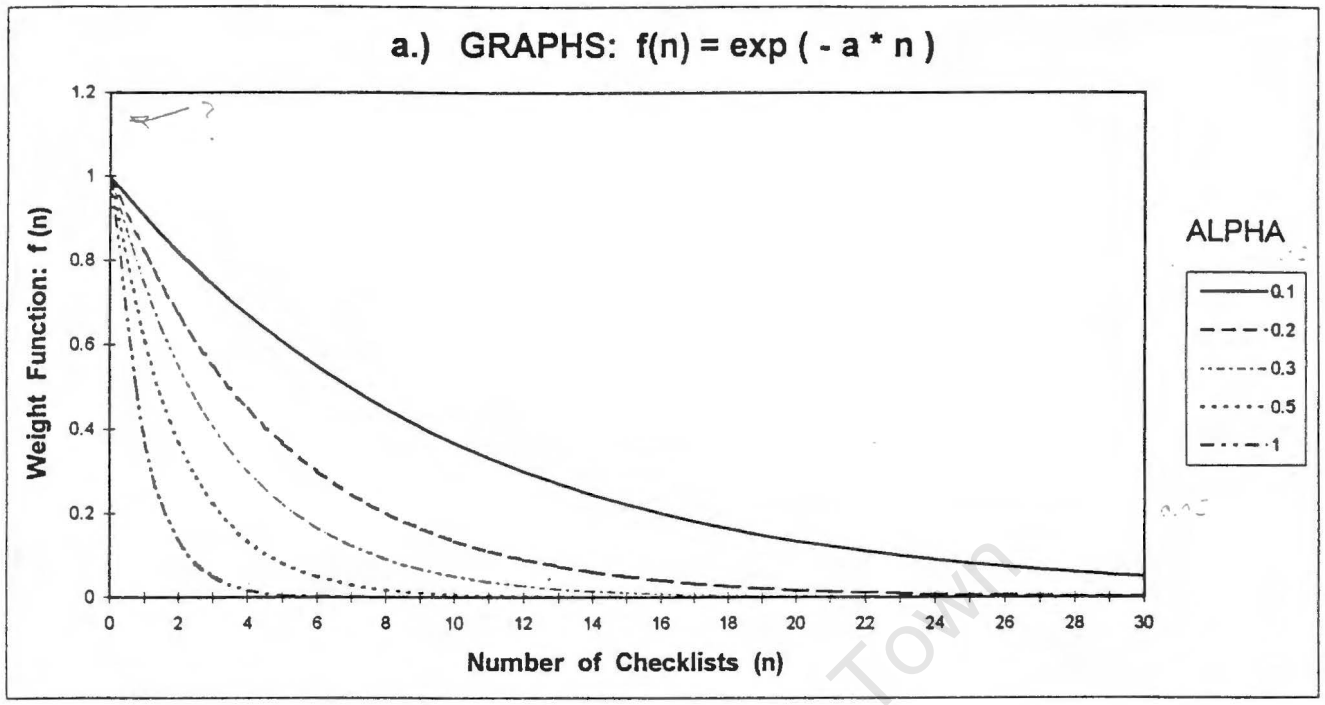
GRID CELL	PARAM.	C++ RESULTS after 4 it's	after 5 it's	6 it's	GENSTAT RESULTS 1 it	2 it's	all it's
3022BB	const	-3.72	-4.40	-5.17	-2.49	-2.83	-13
	x	-1.87	-2.55	-3.32	-0.56	-0.96	-12
	y	-1.93	-2.62	-3.39	-0.49	-0.98	-12
(a)	xy	-2.16	-2.84	-3.61	-0.66	-1.19	-12
	predict	0.024	0.012	0.006	0.076	0.056 not conv.	0
3023AA	const	-3.91	-4.73	-5.63	-2.56	-3.15	-10.6
	x	2.10	2.92	3.83	0.67	1.04	8.5
	y	-2.16	-2.99	-3.89	-0.90	-1.43	-8.9
(b)	xy	2.02	2.85	3.75	0.14	0.87	8.4
	predict	0.020	0.009	0.004	0.072	0.041 not conv.	0
3022BD	const	-3.57	-4.28	-5.10	-2.48	-2.93	-14
	x	-1.89	-2.61	-3.43	-0.63	-1.19	-12
	y	2.18	2.91	3.72	* 0.84	* 1.38	* 12
(c)	xy	2.38	3.10	3.91	* 1.05	* 1.63	* 13
	predict	0.028	0.014	0.006	0.077	0.051 not conv.	0
3023AC	const	-3.60	-4.37	Pie	-2.21	-2.75	-10.9
	x	2.36	3.14	too	* 0.98	* 1.59	9.7
	y	2.25	3.03	small	* 0.77	* 1.39	9.5
(d)	xy	-2.63	-3.41		* -0.93	* -1.62	* -9.8
	predict	0.027	0.012	8.76E-07 0	0.099	not conv. 0.060	0

2922DC (11) 909	2922DD (8) 1250	2923CC (5) 2000	2923CD (7) 0
3022BA (10) 2000	3022BB (6) 0	3023AA (11) 0	3023AB (7) 4285
3022BC (8) 1250	3022BD (6) 0	3023AC (8) 0	3023AD (8) 0
3022DA (7) 1428	3022DB (14) 2142	3023CA (14) 2857	3023CB (16) 1875

Fig. 3 This is a subset of the Blackheaded Canary distribution. The values shown are reporting rates (multiplied by 10000). The top right hand corner of each grid cell gives the number of check-lists that were collected for this grid cell. The shades are as in the original bird atlas maps. See Table 5 for the corresponding regression results

Table 6 This table shows the weights calculated for a number of alpha values for differing numbers of checklists, f(n) = exp (-a * n) (equations 39 and 40 in text). The calculated weights here are those assigned to the model-predicted reporting rates. See also Figure 2.

ALPHA	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.15	0.05
Checklists												
1	0.90	0.82	0.74	0.67	0.61	0.55	0.50	0.45	0.41	0.37	0.86	0.95
2	0.82	0.67	0.55	0.45	0.37	0.30	0.25	0.20	0.17	0.14	0.74	0.90
3	0.74	0.55	0.41	0.30	0.22	0.17	0.12	0.09	0.07	0.05	0.64	0.86
4	0.67	0.45	0.30	0.20	0.14	0.09	0.06	0.04	0.03	0.02	0.55	0.82
5	0.61	0.37	0.22	0.14	0.08	0.05	0.03	0.02	0.01	0.01	0.47	0.78
6	0.55	0.30	0.17	0.09	0.05	0.03	0.01	0.01	0	0	0.41	0.74
7	0.50	0.25	0.12	0.06	0.03	0.01	0.01	0	0	0	0.35	0.70
8	0.45	0.20	0.09	0.04	0.02	0.01	0	0	0	0	0.30	0.67
9	0.41	0.17	0.07	0.03	0.01	0	0	0	0	0	0.26	0.64
10	0.37	0.14	0.05	0.02	0.01	0	0	0	0	0	0.22	0.61
11	0.33	0.11	0.04	0.01	0	0	0	0	0	0	0.19	0.58
12	0.30	0.09	0.03	0.01	0	0	0	0	0	0	0.17	0.55
13	0.27	0.07	0.02	0.01	0	0	0	0	0	0	0.14	0.52
14	0.25	0.06	0.01	0	0	0	0	0	0	0	0.12	0.50
15	0.22	0.05	0.01	0	0	0	0	0	0	0	0.11	0.47
16	0.20	0.04	0.01	0	0	0	0	0	0	0	0.09	0.45
17	0.18	0.03	0.01	0	0	0	0	0	0	0	0.08	0.43
18	0.17	0.03	0	0	0	0	0	0	0	0	0.07	0.41
19	0.15	0.02	0	0	0	0	0	0	0	0	0.06	0.39
20	0.14	0.02	0	0	0	0	0	0	0	0	0.05	0.37
21	0.12	0.01	0	0	0	0	0	0	0	0	0.04	0.35
22	0.11	0.01	0	0	0	0	0	0	0	0	0.04	0.33
23	0.10	0.01	0	0	0	0	0	0	0	0	0.03	0.32
24	0.09	0.01	0	0	0	0	0	0	0	0	0.03	0.30
25	0.08	0.01	0	0	0	0	0	0	0	0	0.02	0.29
26	0.07	0.01	0	0	0	0	0	0	0	0	0.02	0.27
27	0.07	0	0	0	0	0	0	0	0	0	0.02	0.26
28	0.06	0	0	0	0	0	0	0	0	0	0.01	0.25
29	0.06	0	0	0	0	0	0	0	0	0	0.01	0.23
30	0.05	0	0	0	0	0	0	0	0	0	0.01	0.22



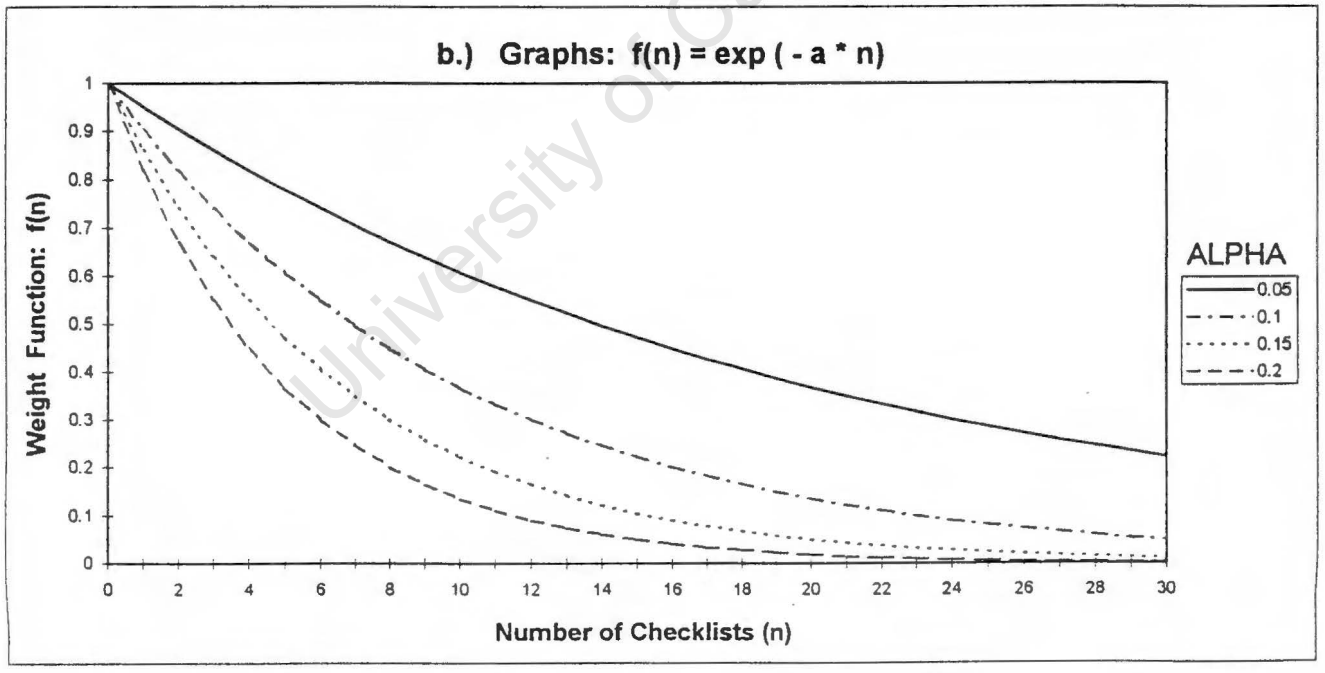


FIGURE 4. These two figures illustrate the behaviour of the weight function f(n). The various functions displayed here, are formed by substituting different values for the alpha value 'a' (see legend). f(n) is the weight that will be assigned to the model-predicted reporting rate, while the weight [1 - f(n)] is assigned to the observed reporting rate. See also the text, equations 39, 40.

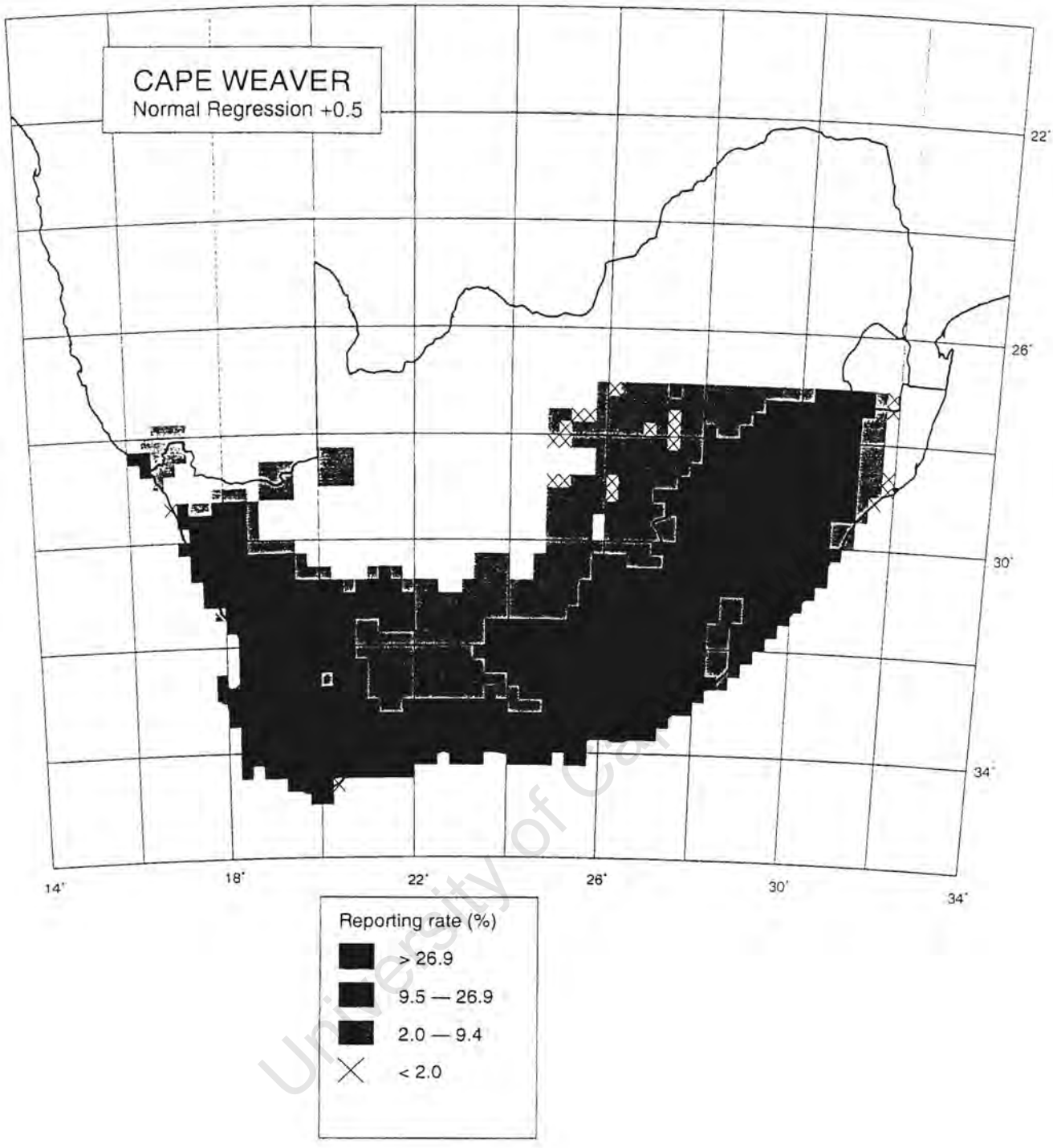


Figure 4. A smoothed distribution map for the Cape Weaver, produced by Method NR+0.5.

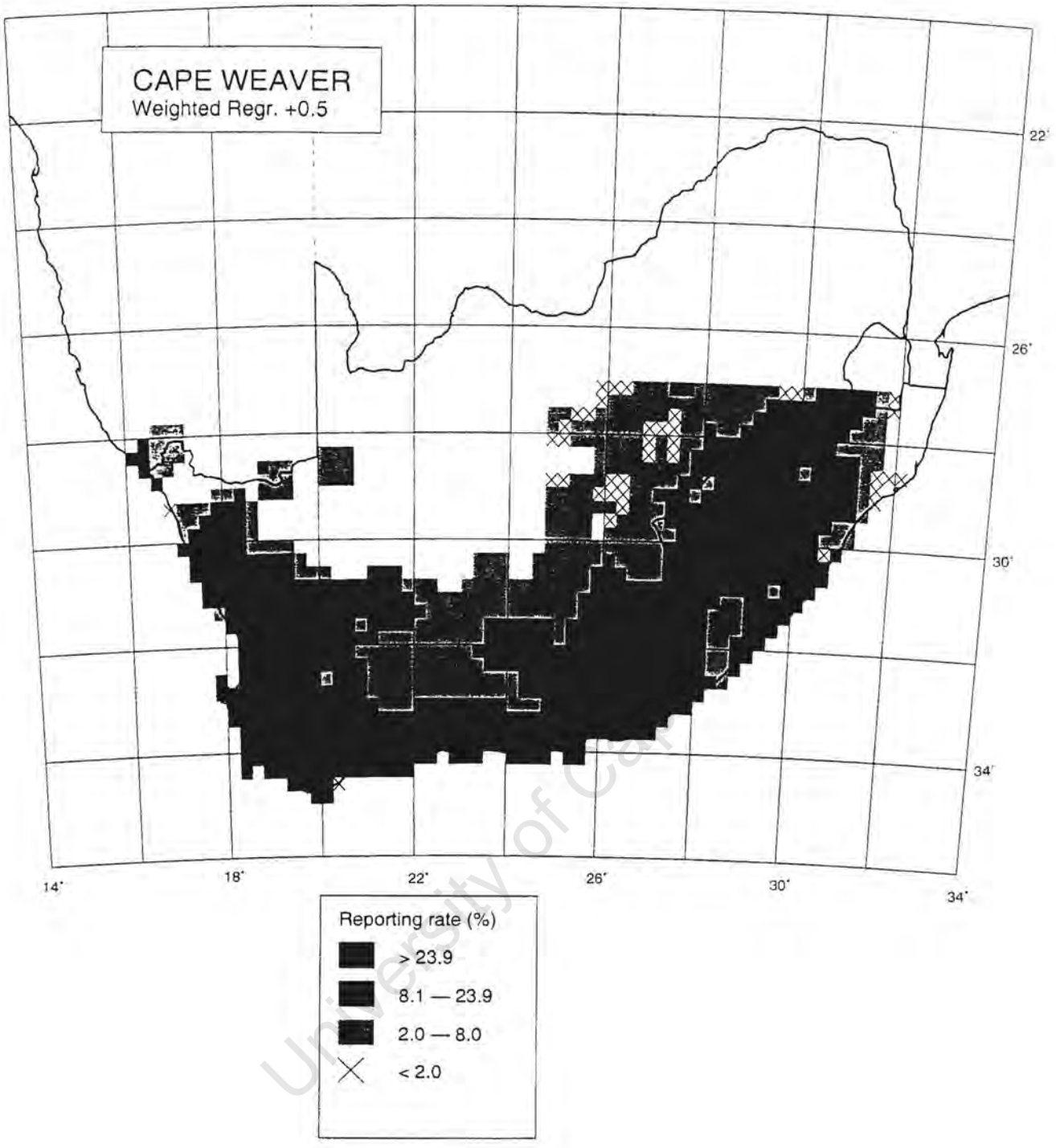


Figure 5. A smoothed distribution map for the Cape Weaver, produced by Method WR+0.5.

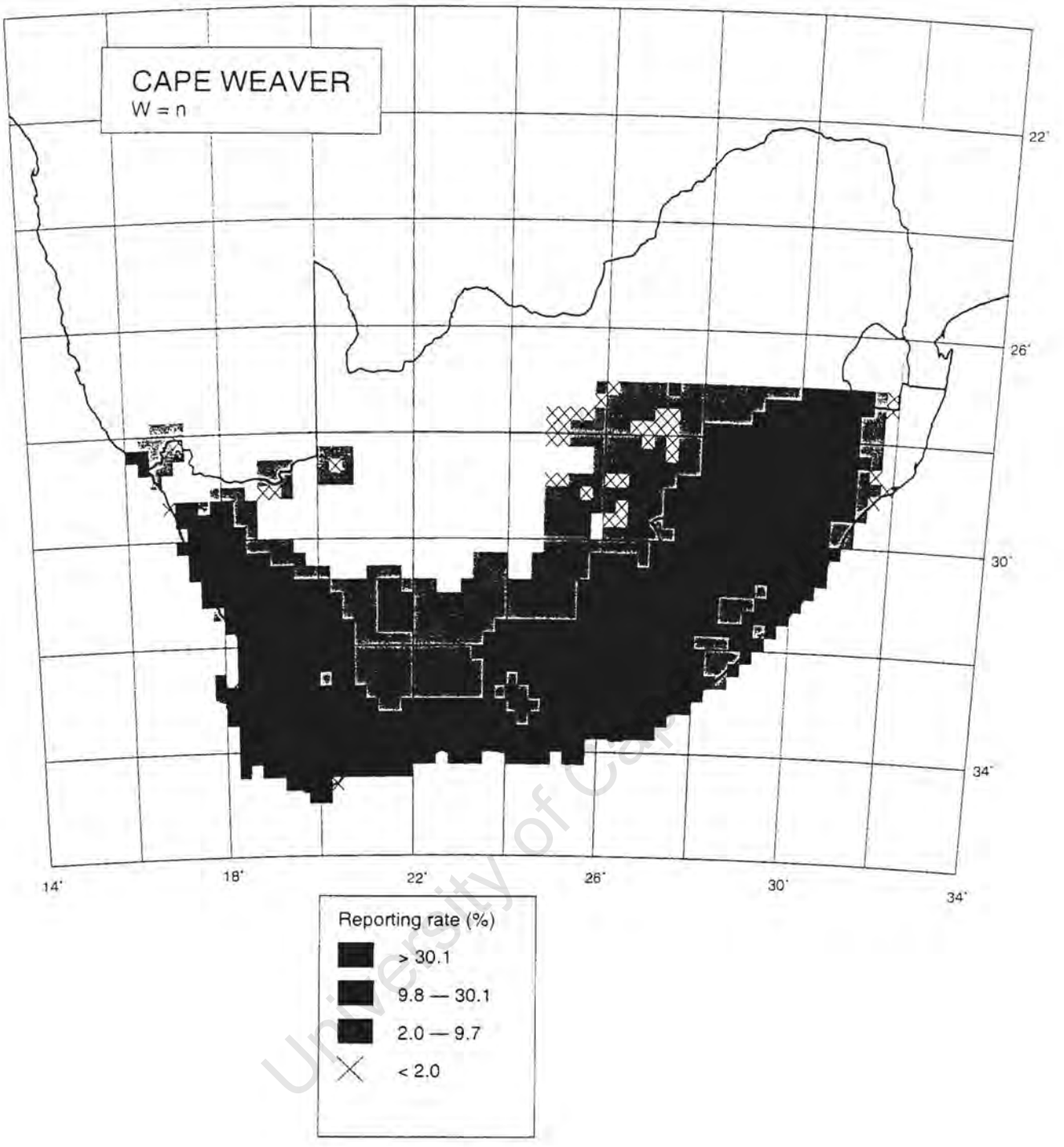


Figure 6. A smoothed distribution map for the Cape Weaver, produced by Method W=n.

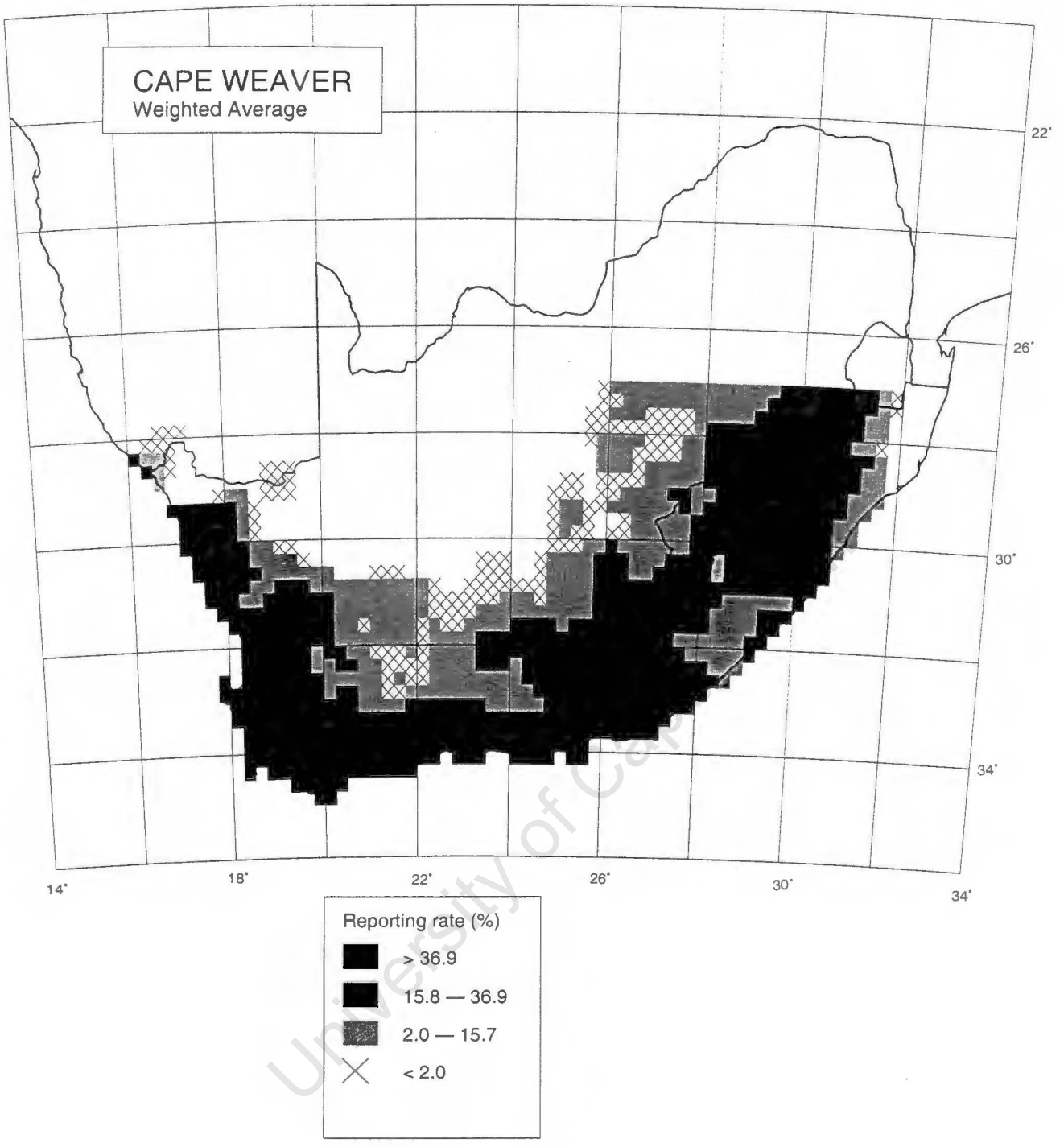


Figure 7. A smoothed distribution map for the Cape Weaver, produced by Method WAVG.

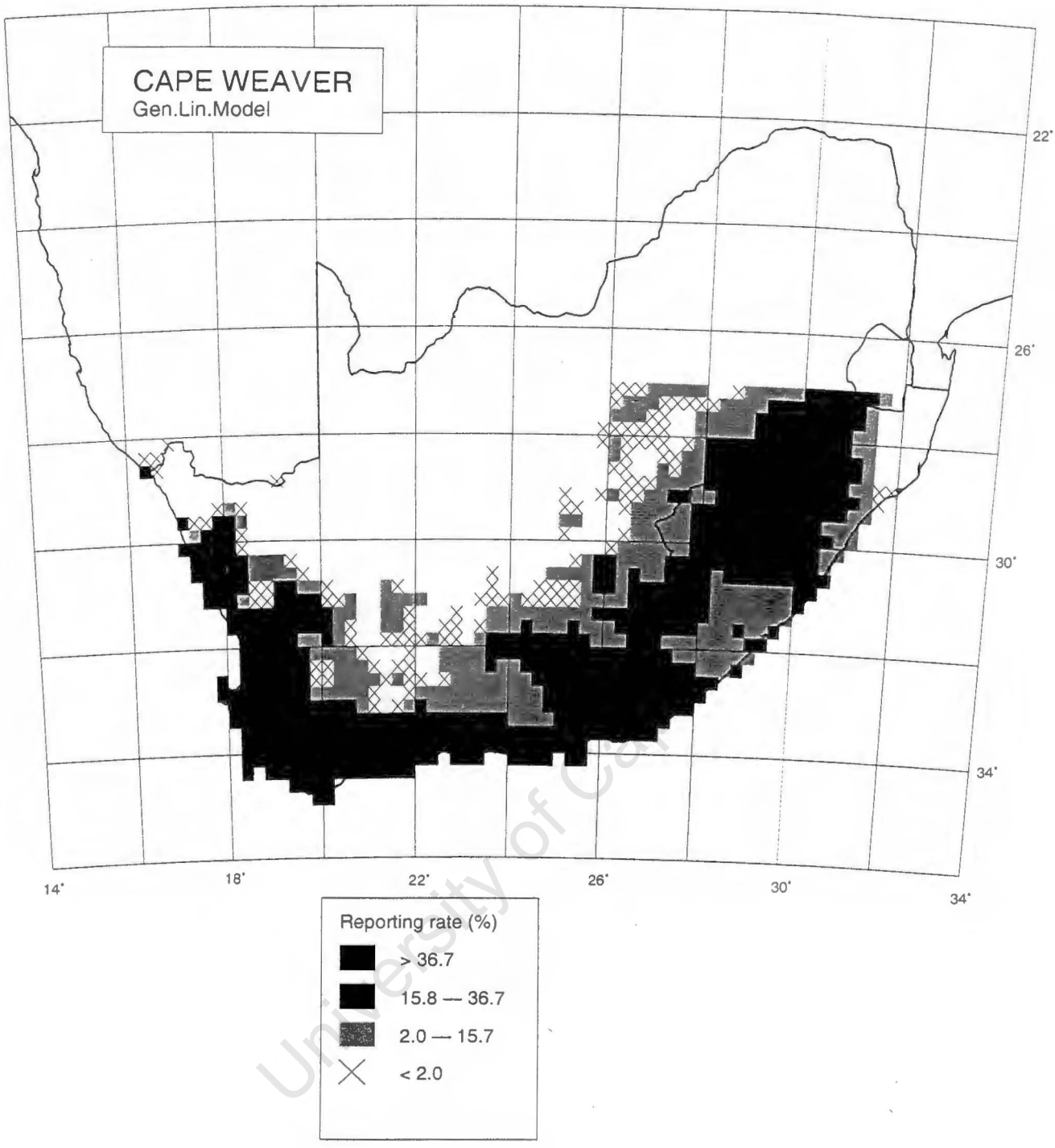


Figure 8. A smoothed distribution map for the Cape Weaver, produced by Method IRWLS, values are pure model-predicted reporting rates.

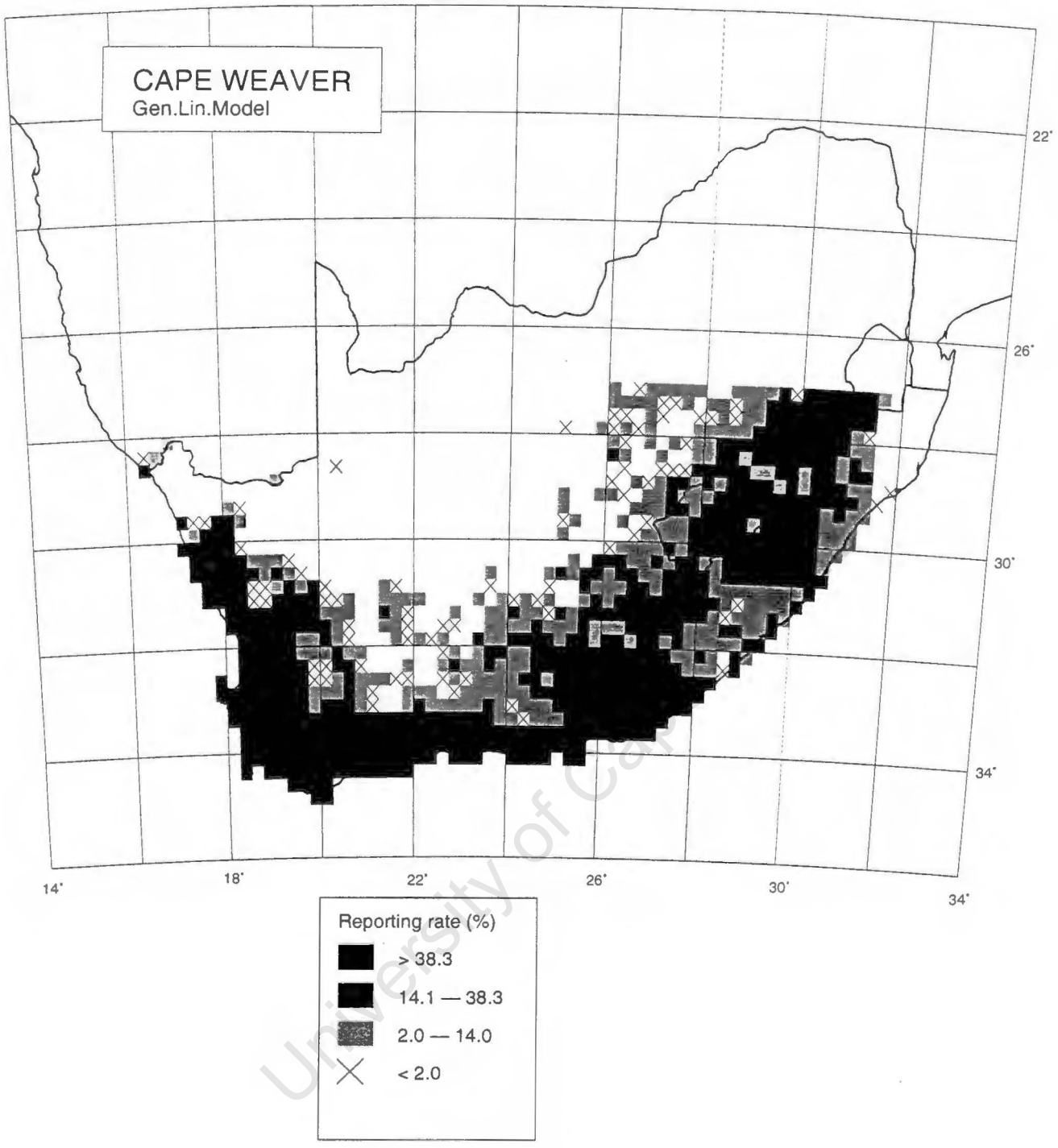


Figure 9. A smoothed distribution map for the Cape Weaver, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.05$.

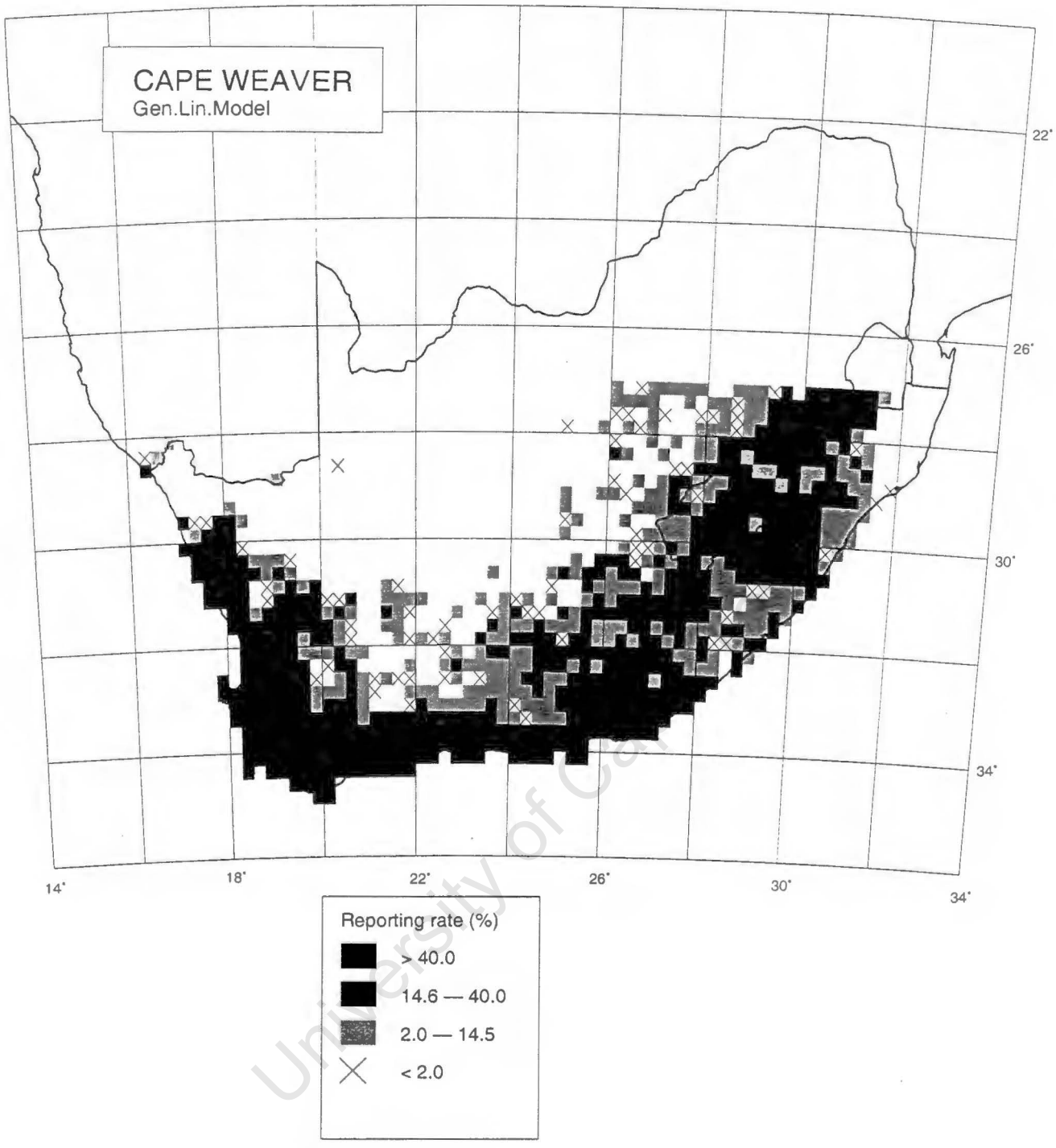


Figure 10. A smoothed distribution map for the Cape Weaver, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.1$.

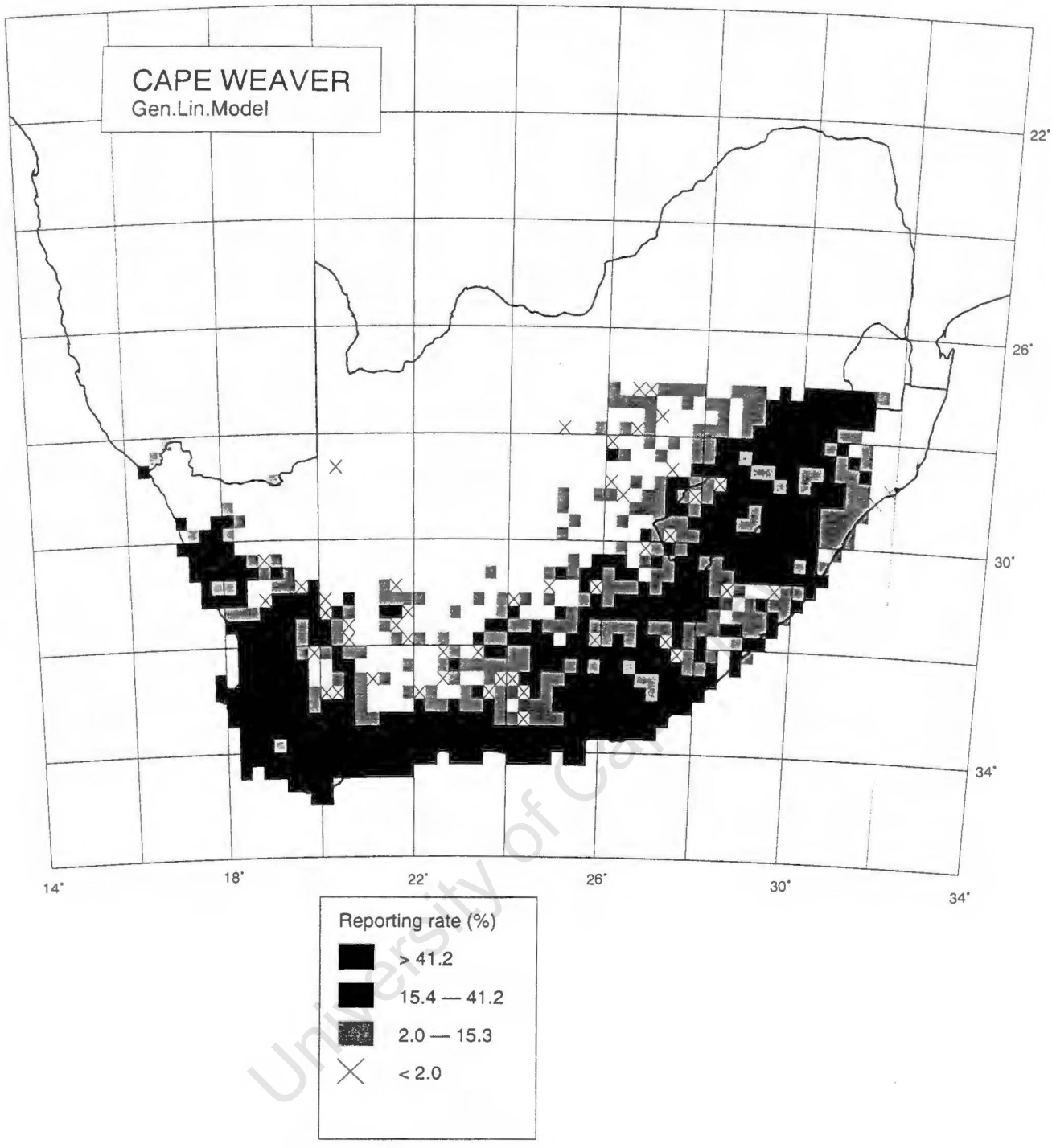


Figure 11. A smoothed distribution map for the Cape Weaver, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.2$.

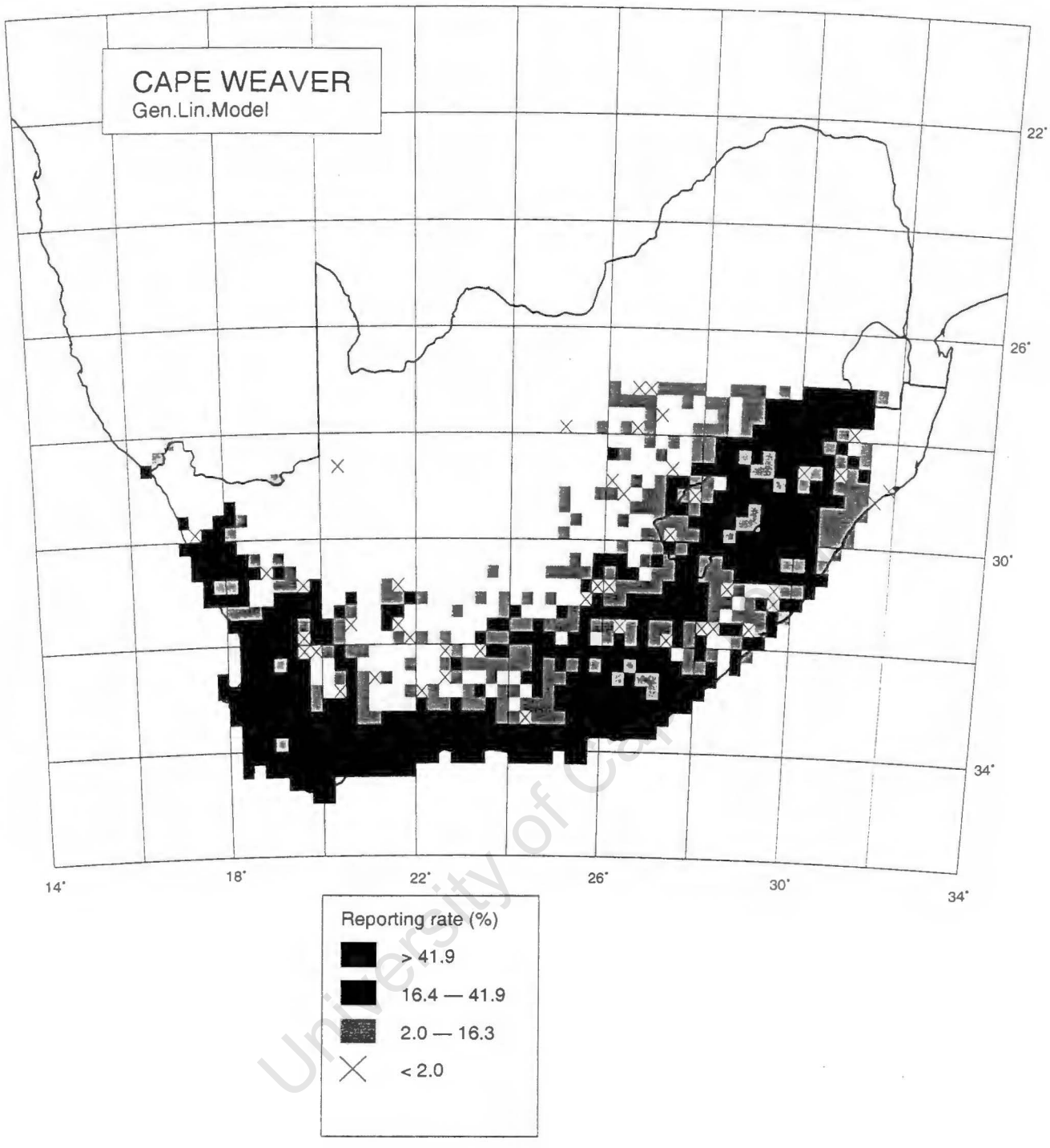


Figure 12. A smoothed distribution map for the Cape Weaver, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.3$.

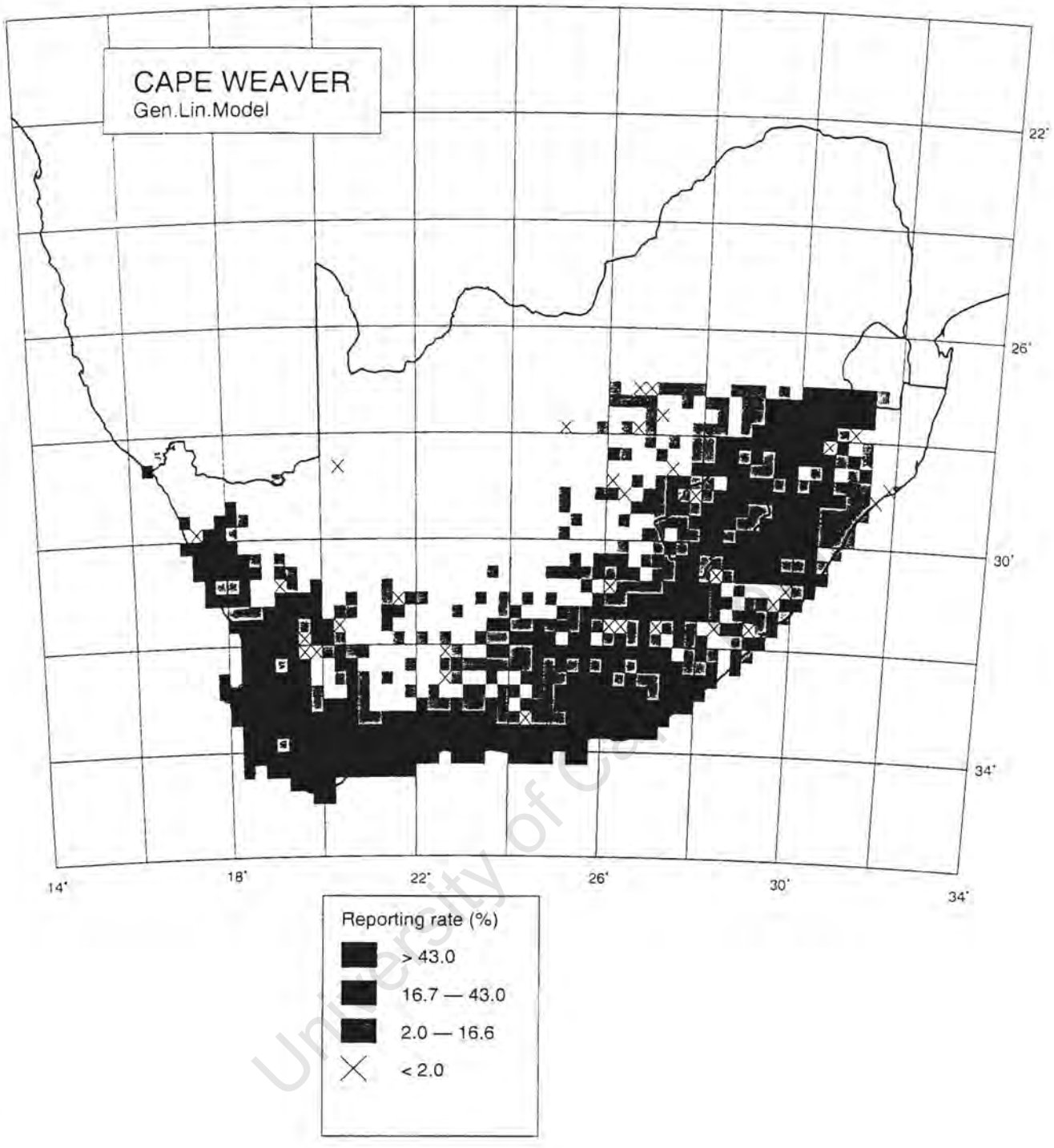


Figure 13. A smoothed distribution map for the Cape Weaver, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.4$.

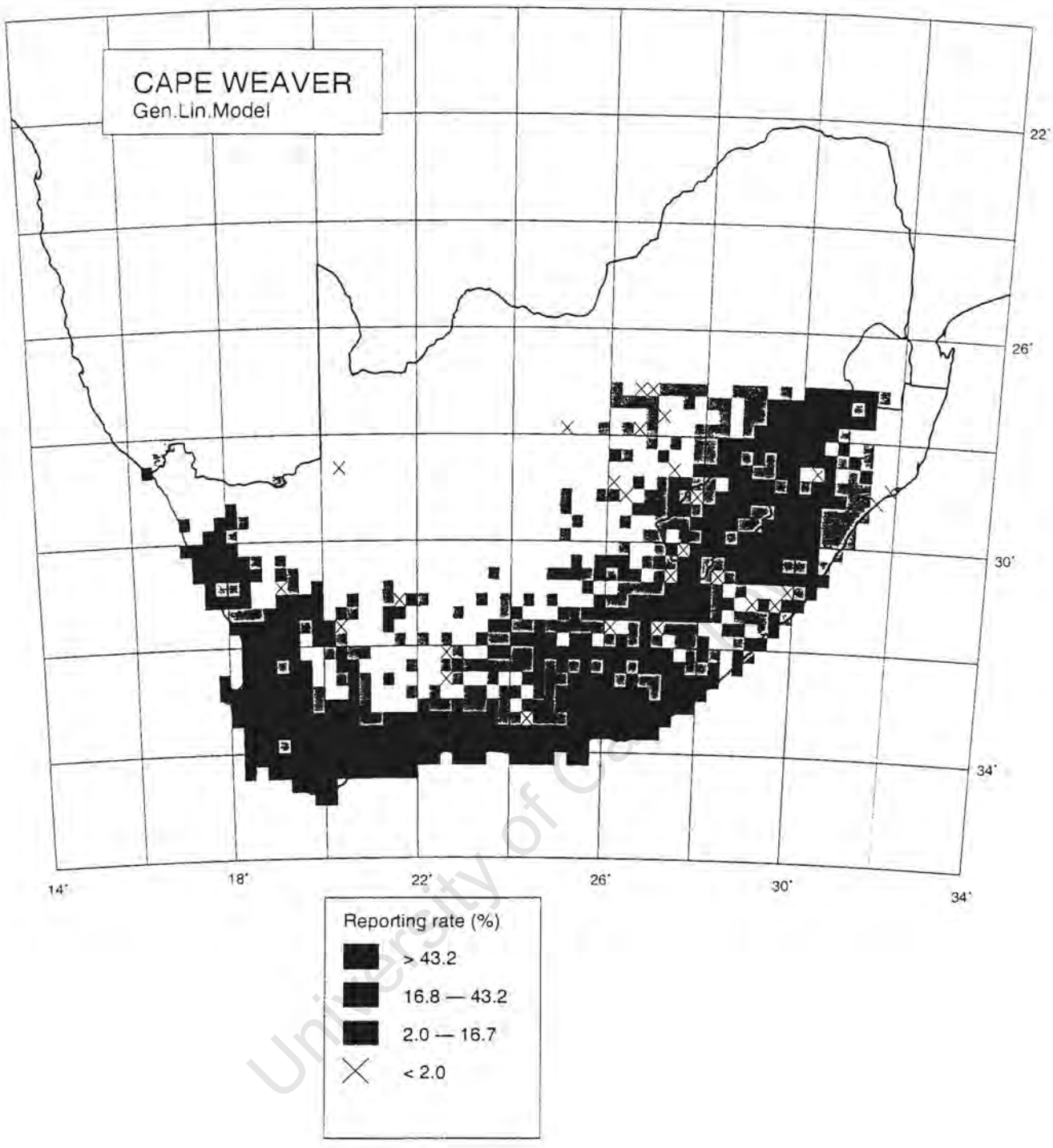


Figure 14. A smoothed distribution map for the Cape Weaver, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.5$.

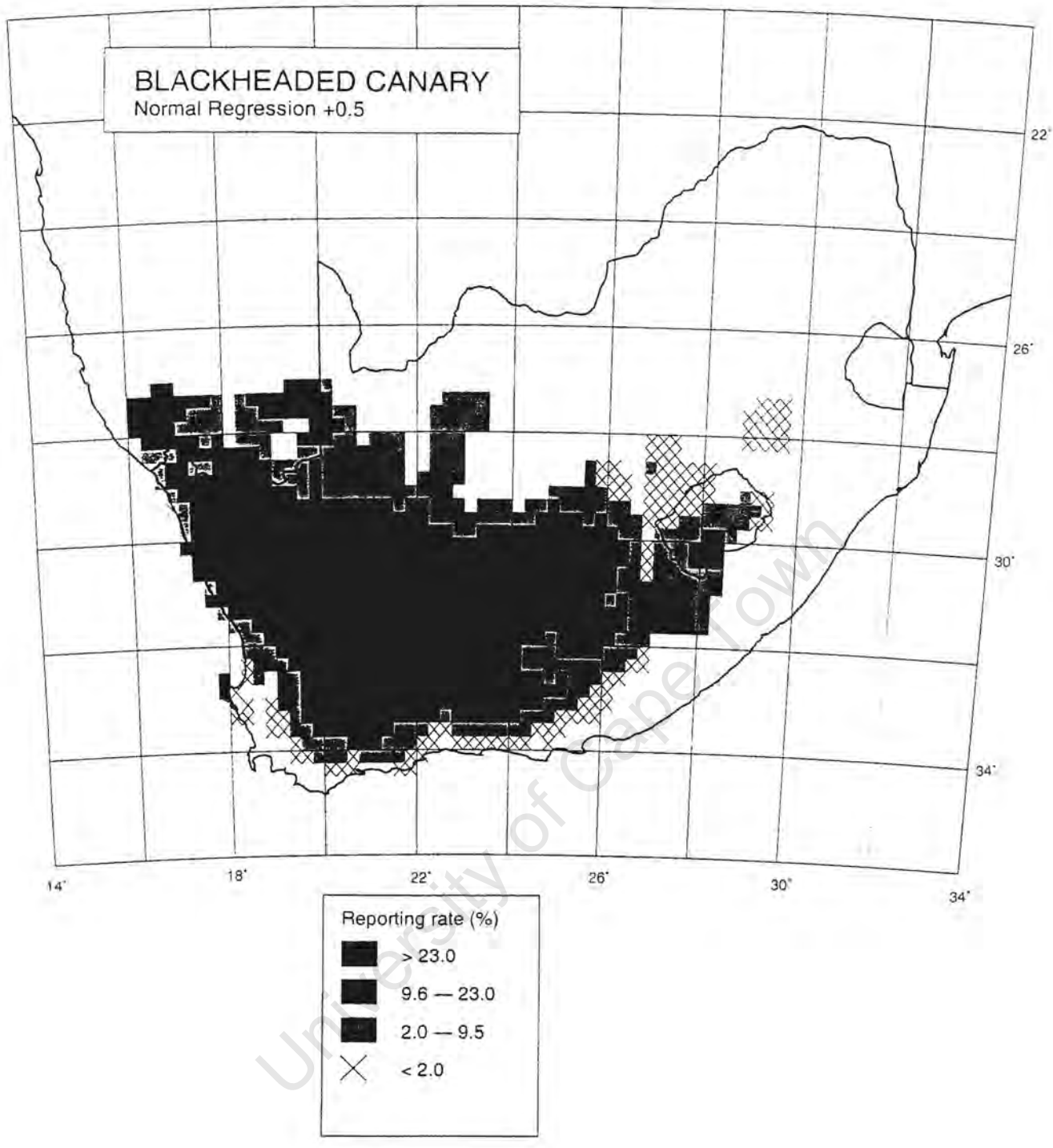


Figure 15. A smoothed distribution map for the Blackheaded Canary, produced by Method NR+0.5.

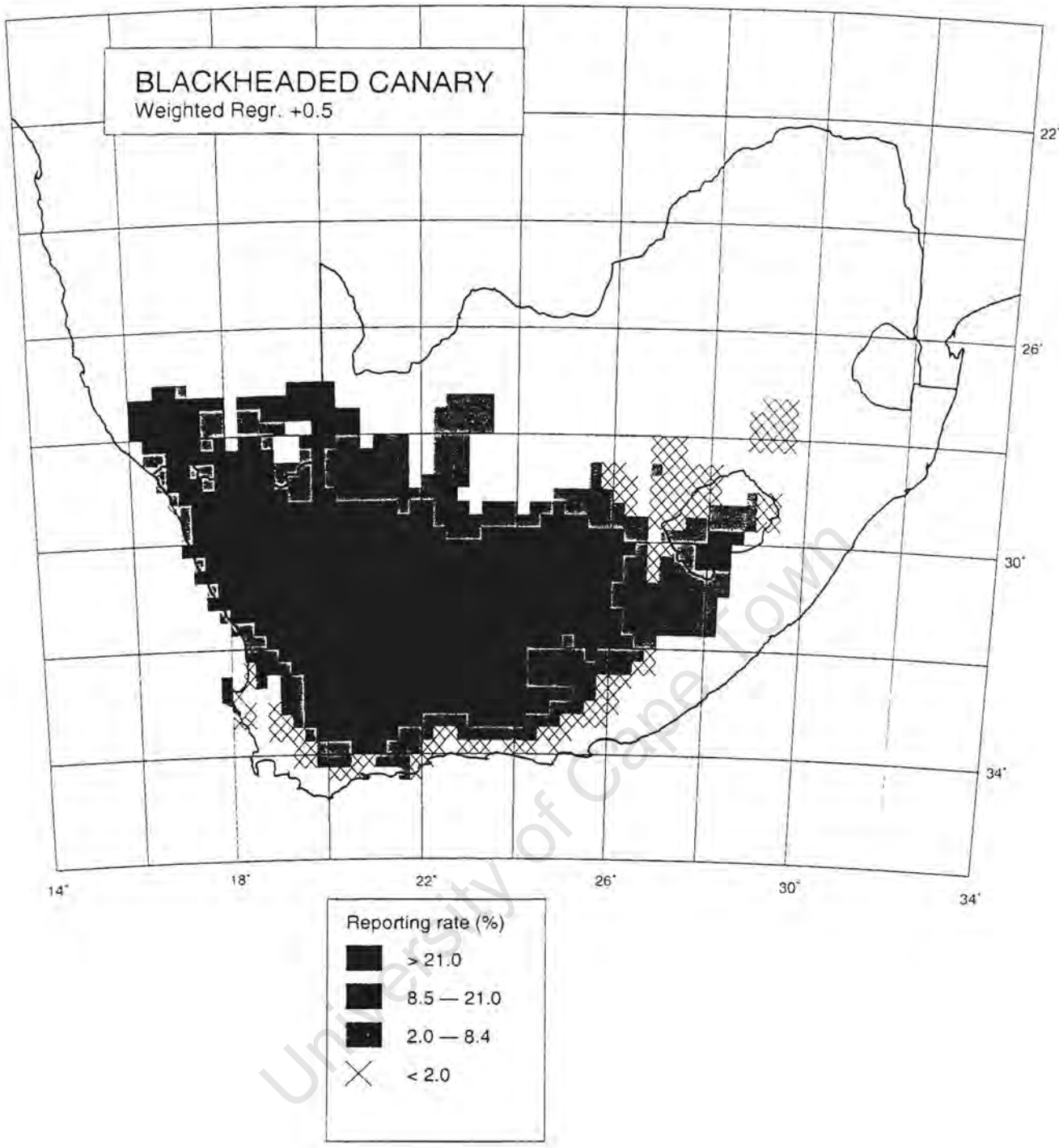


Figure 16. A smoothed distribution map for the Blackheaded Canary, produced by Method WR+0.5.

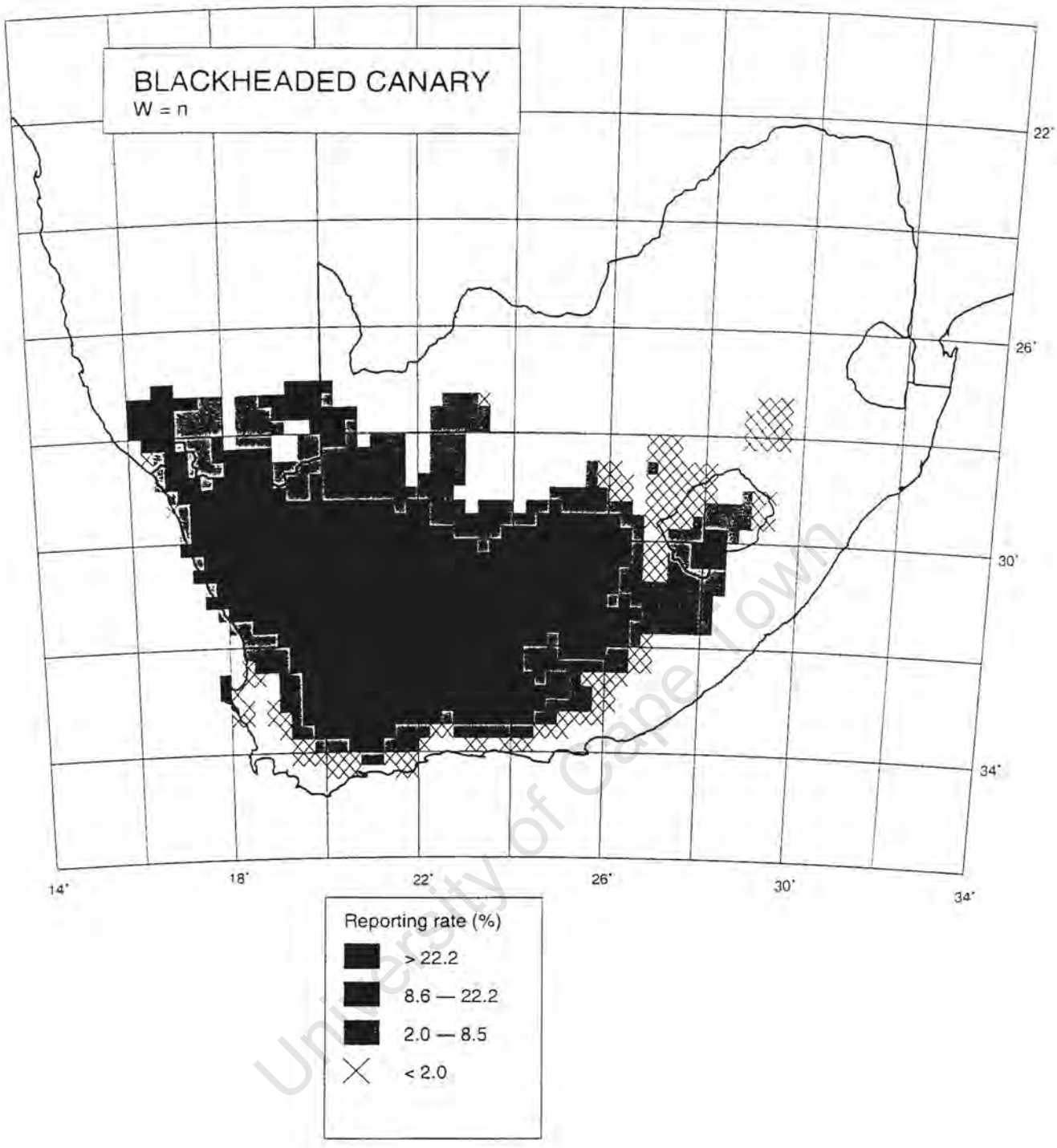


Figure 17. A smoothed distribution map for the Blackheaded Canary, produced by Method W=n.

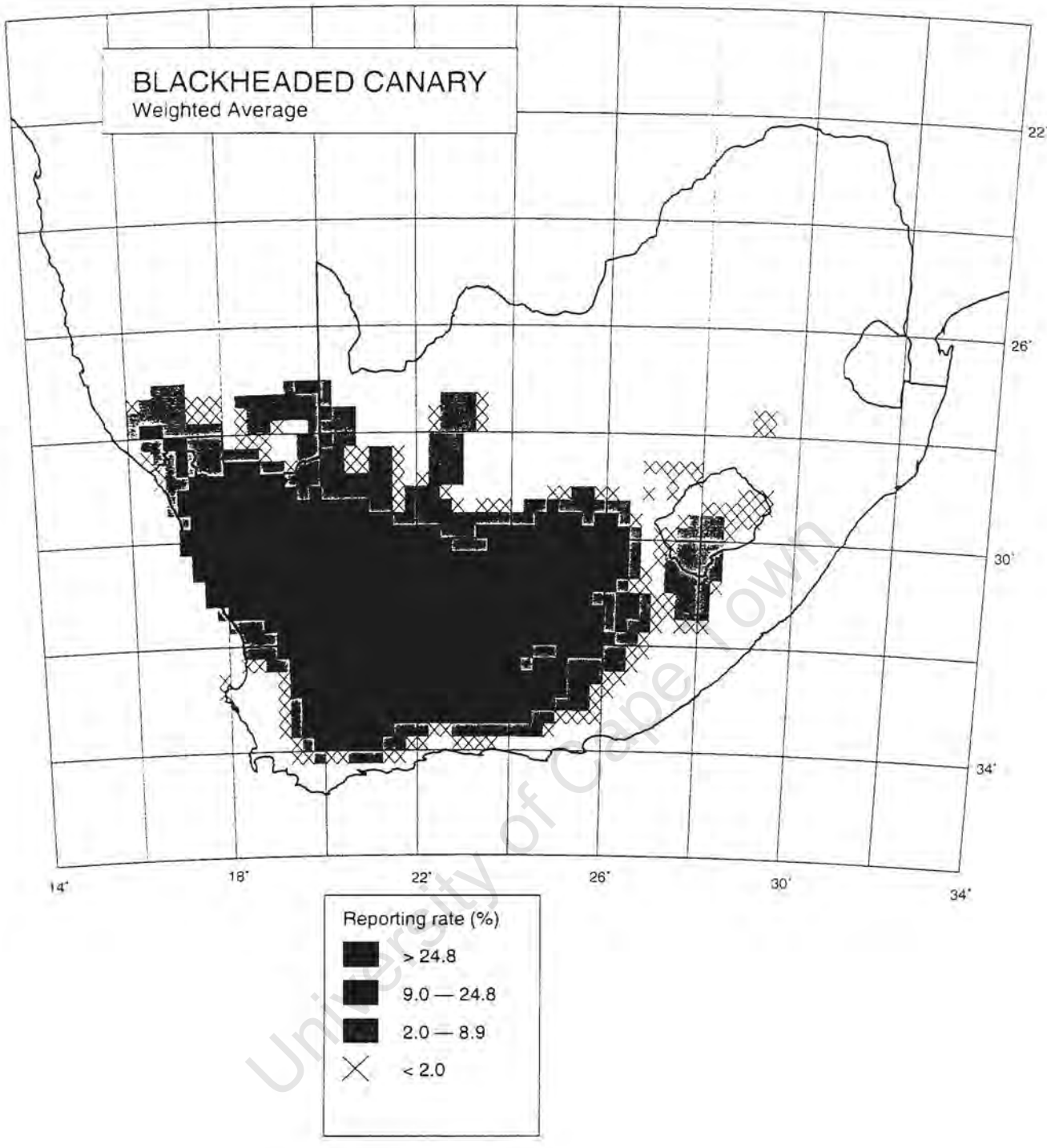


Figure 18. A smoothed distribution map for the Blackheaded Canary, produced by Method WAVG.

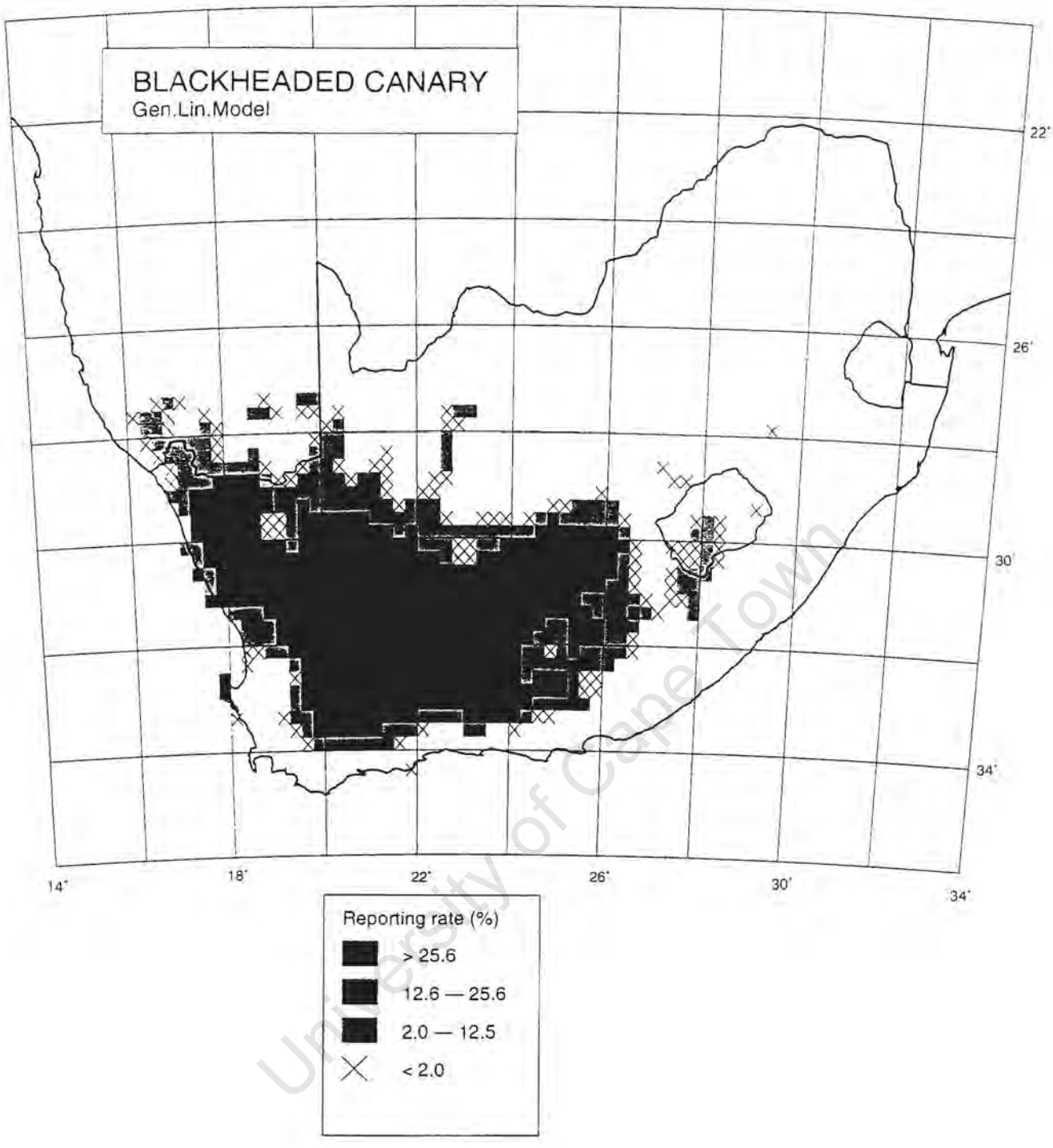


Figure 19. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS, values are pure model-predicted reporting rates.

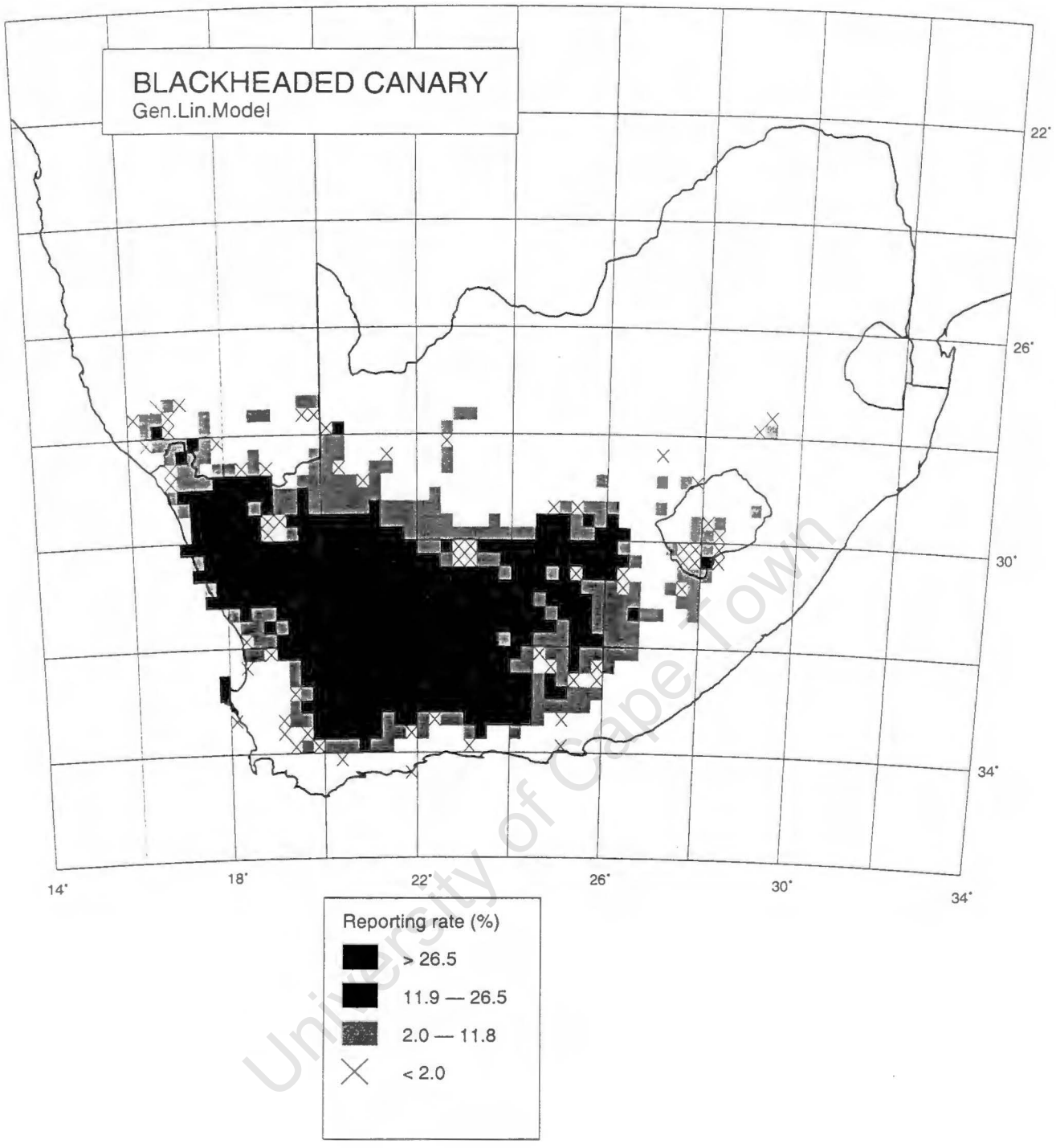


Figure 20. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.05$.

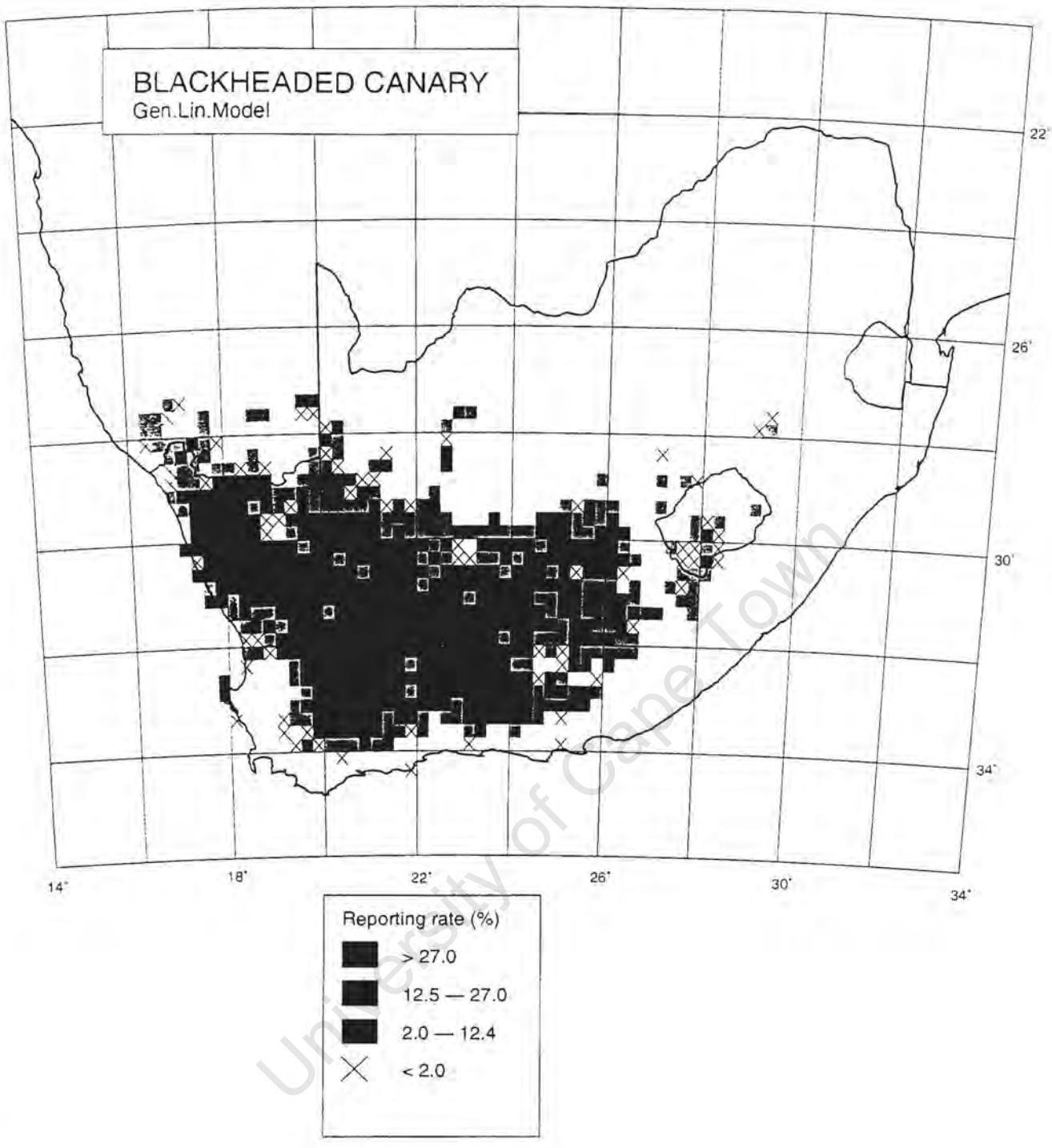


Figure 21. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.1$.

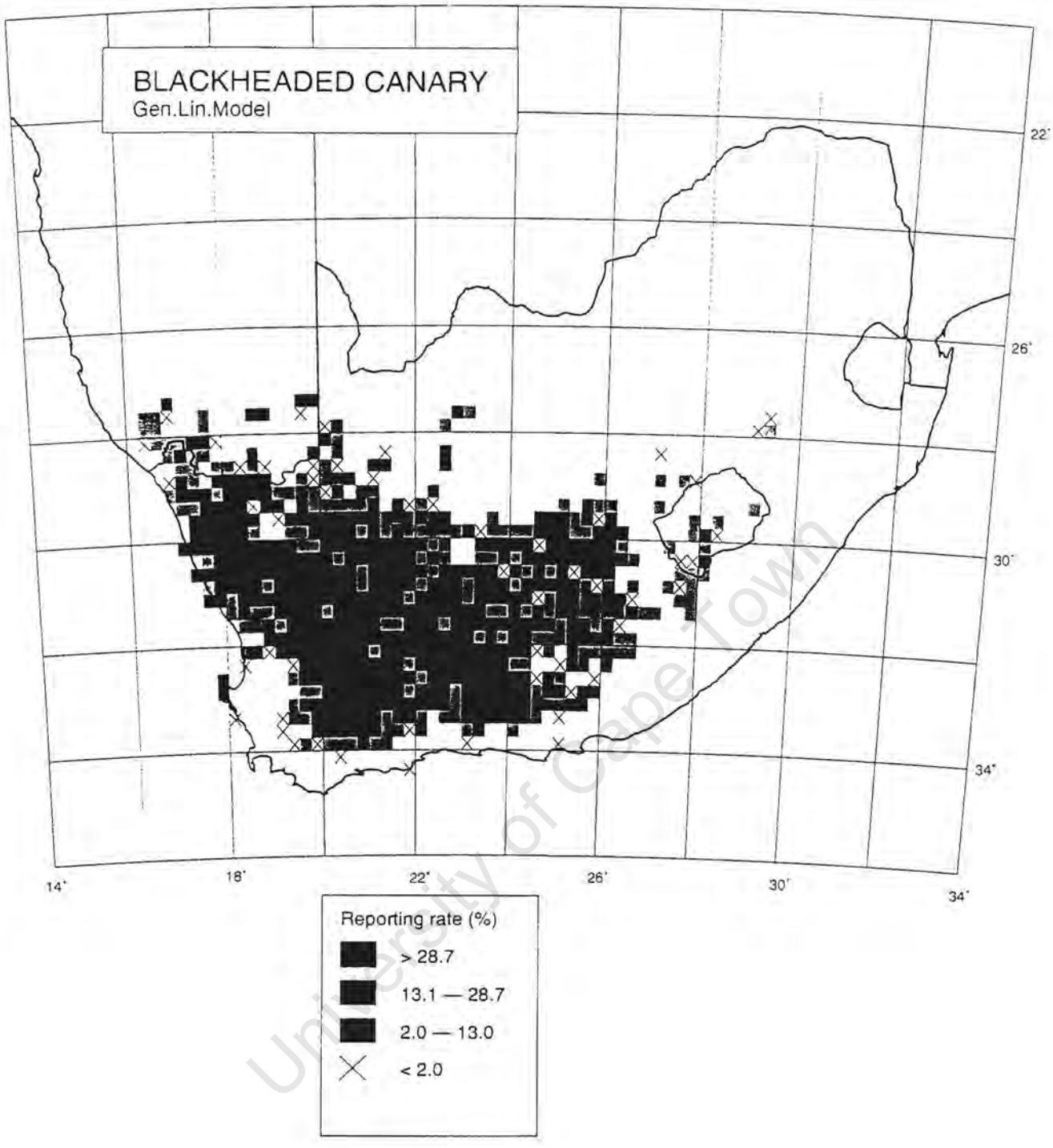


Figure 22. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.2$.

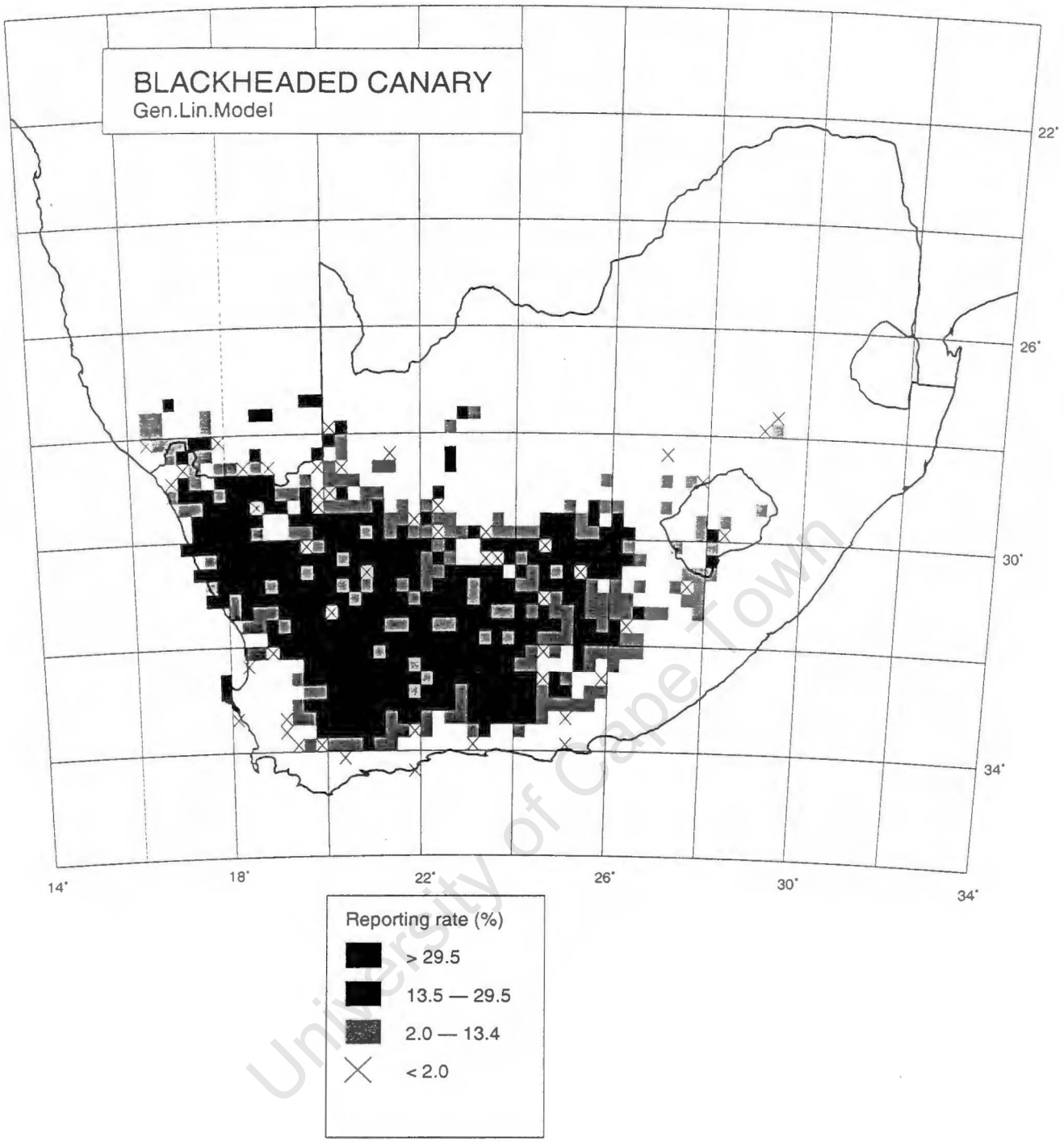


Figure 23. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.3$.

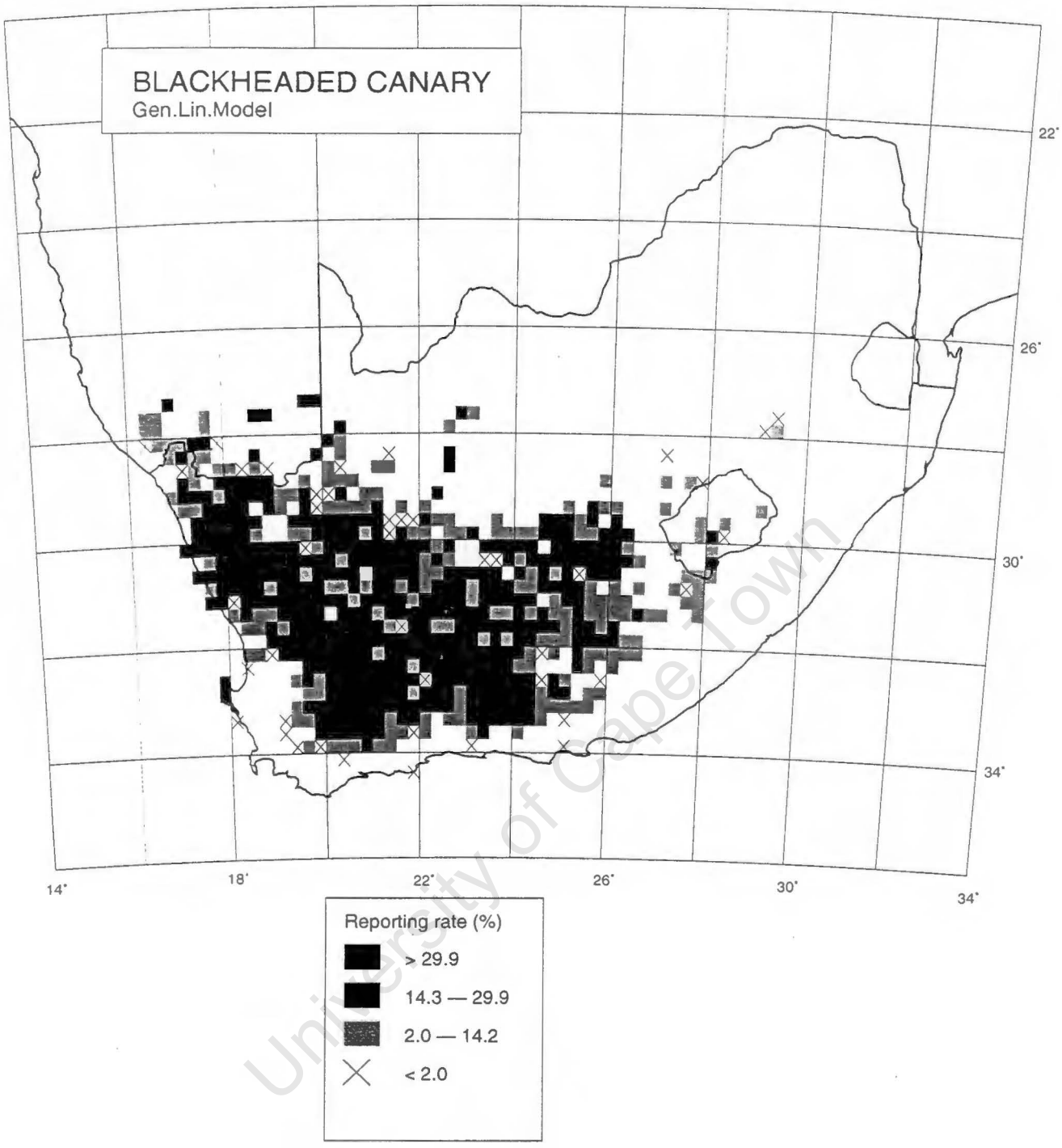


Figure 24. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.4$.

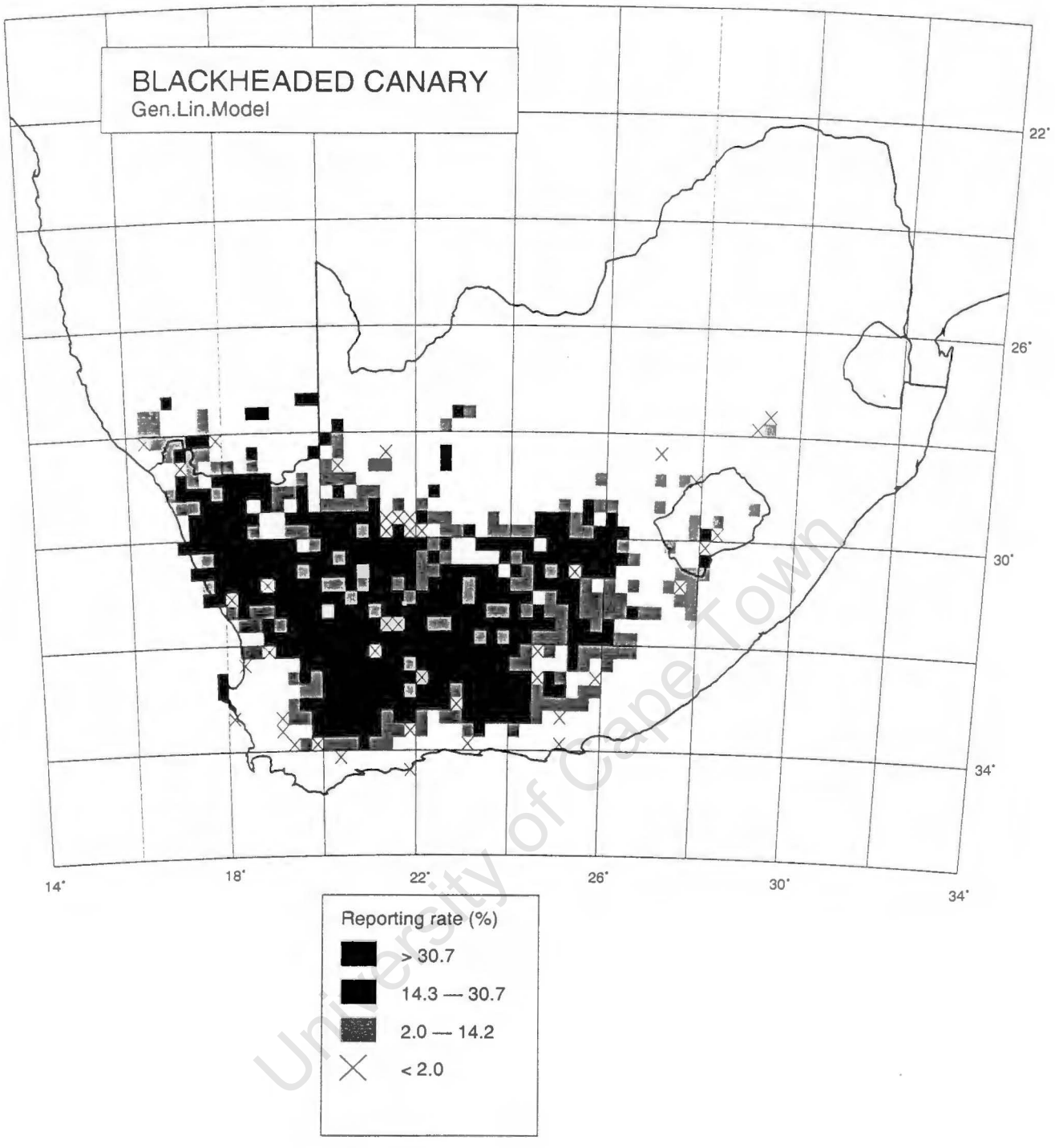


Figure 25. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.5$.

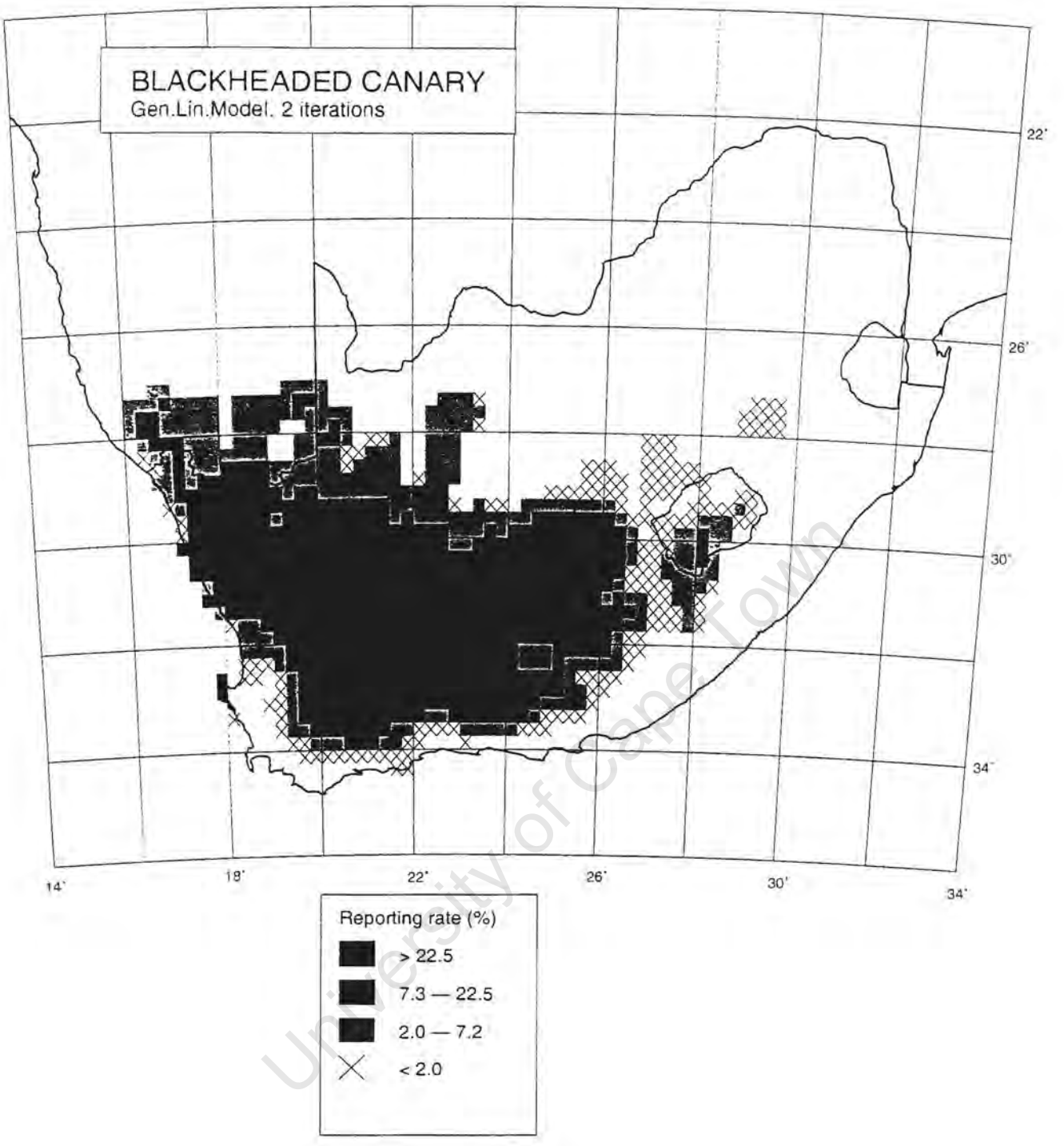


Figure 26. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS, but only using two iterations.

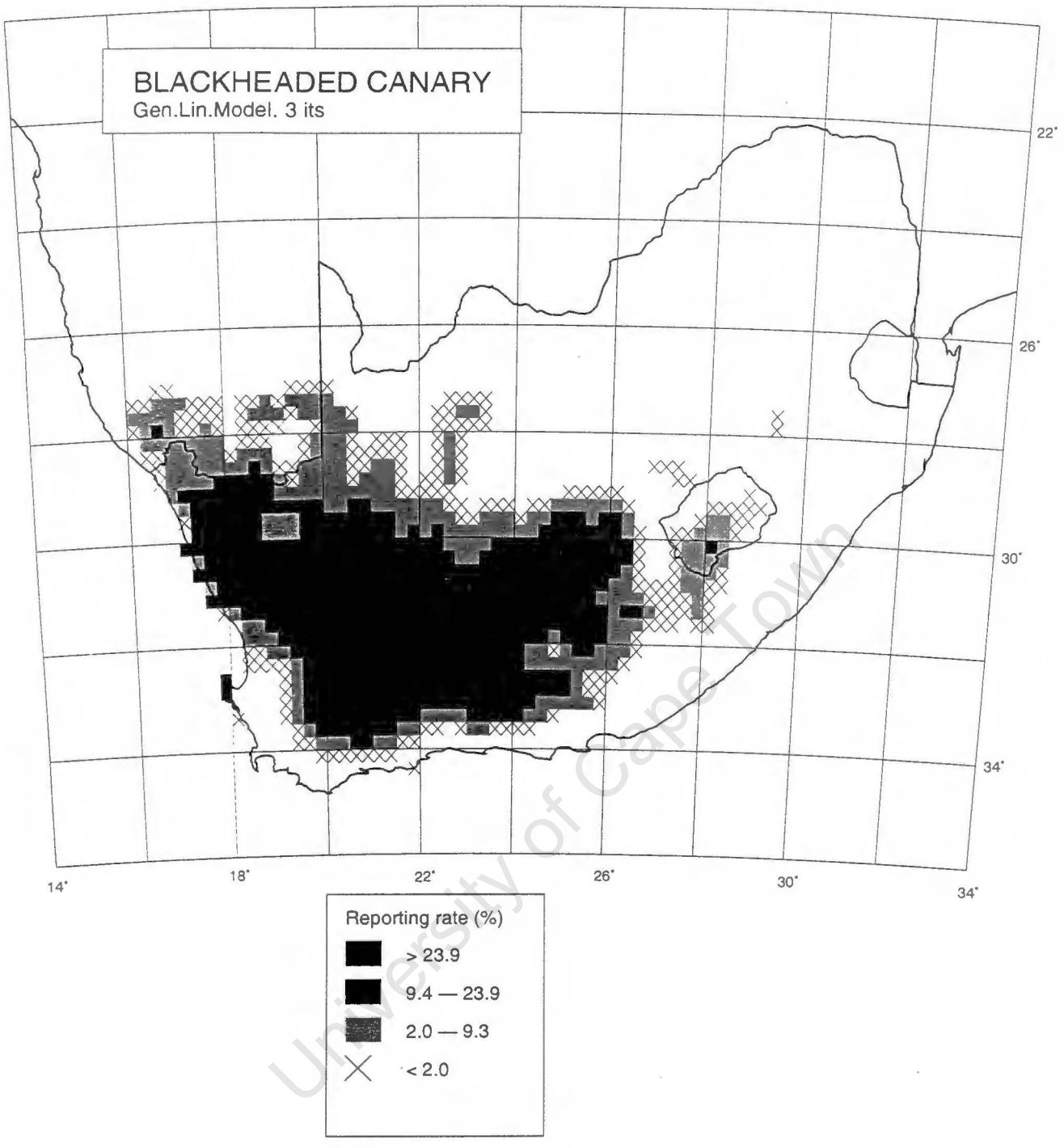


Figure 27. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS, using three iterations.

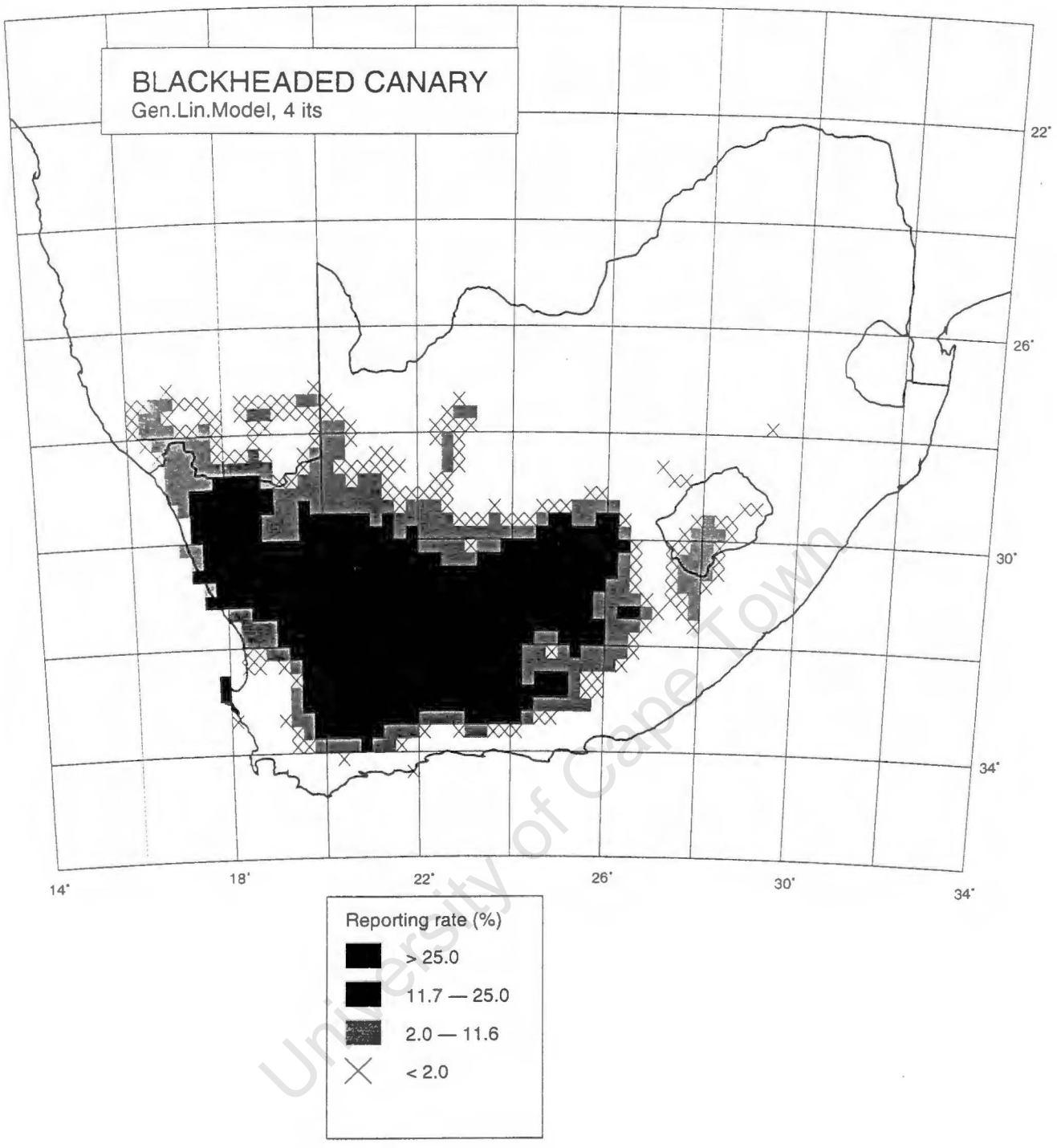


Figure 28. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS, using four iterations.

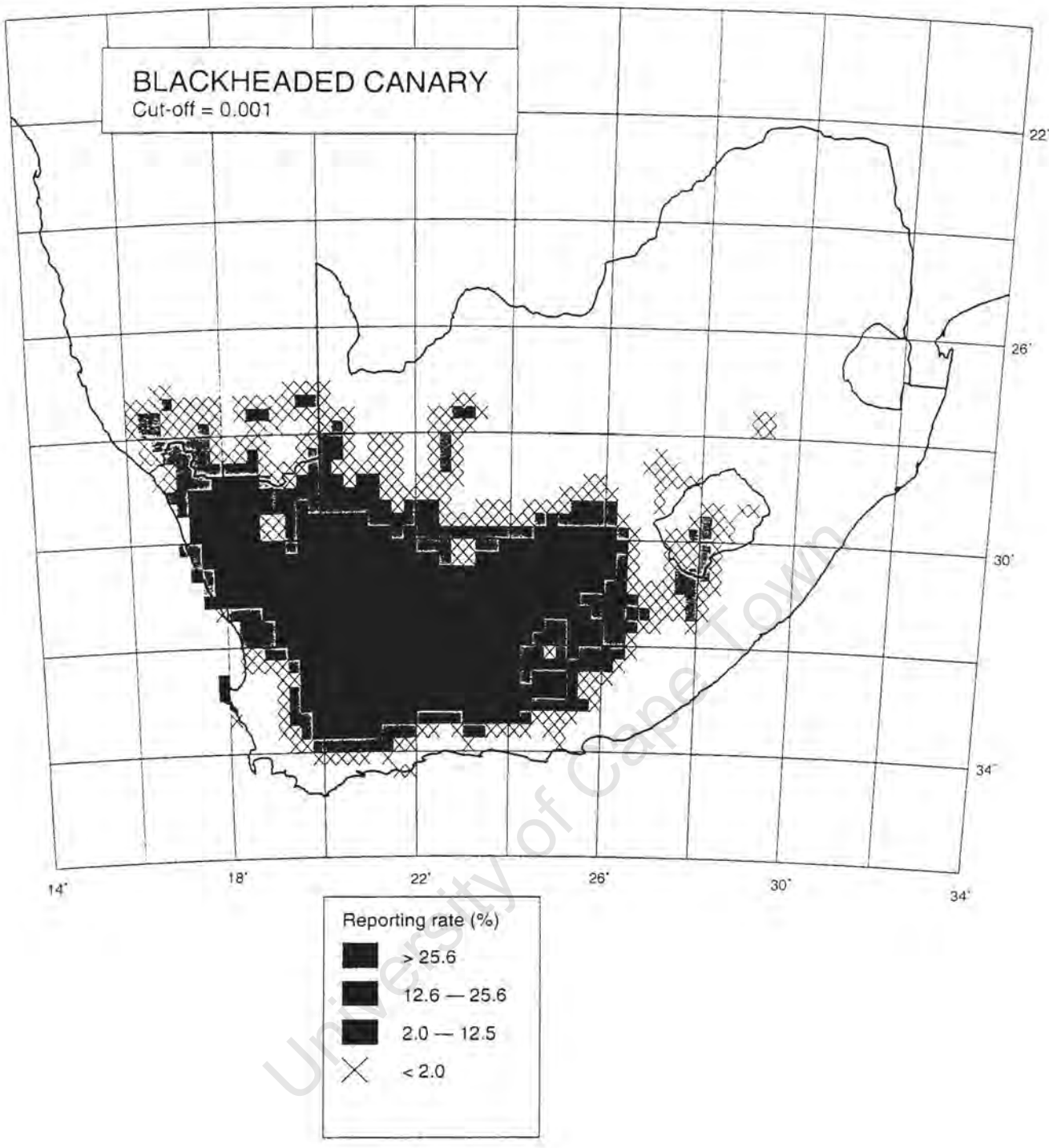


Figure 29. A smoothed distribution map for the Blackheaded Canary, produced by Method IRWLS. The pure model-predicted reporting rates are shown, but a zero cut-off point of 0.001 was used.

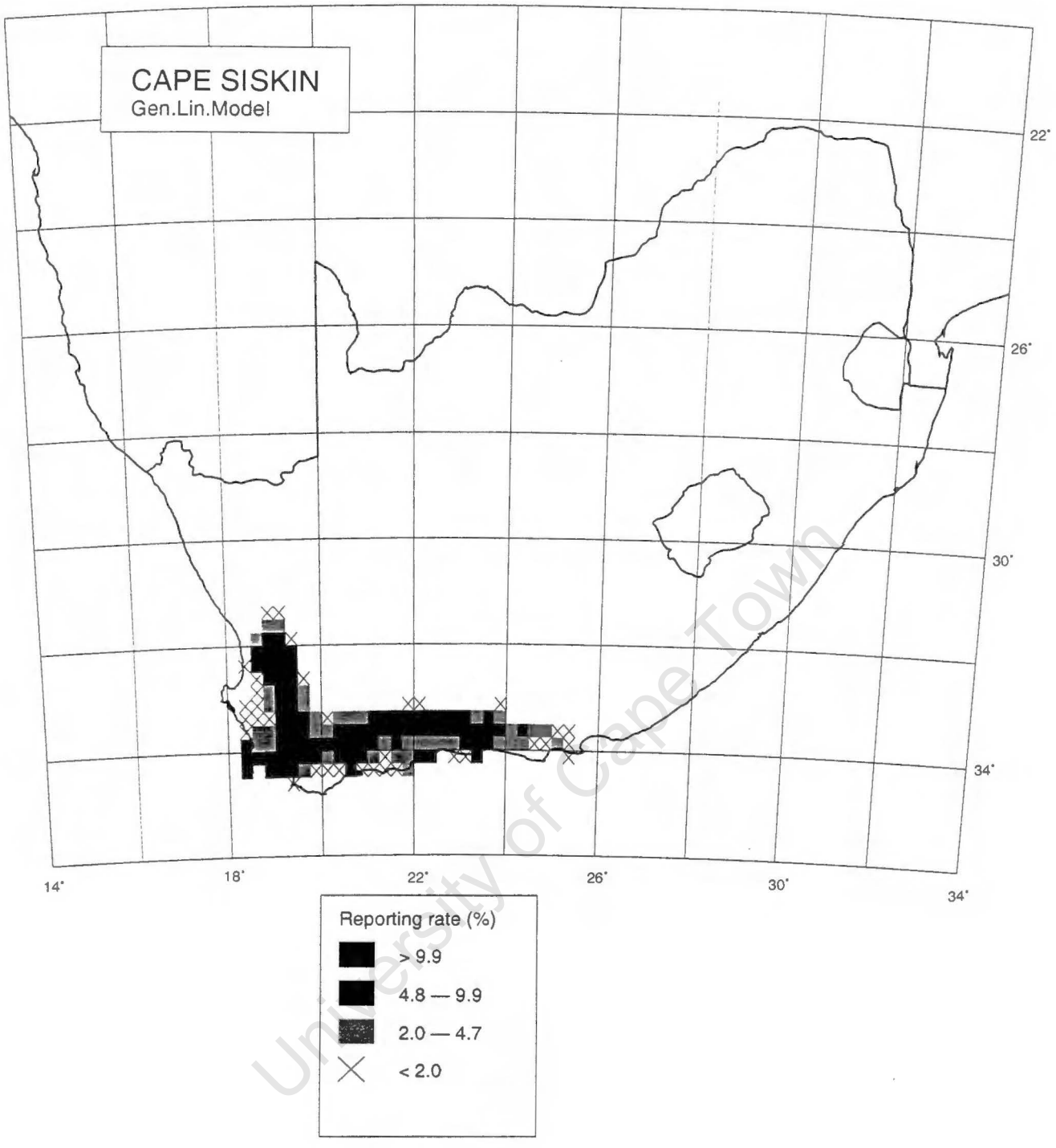


Figure 30. A smoothed distribution map for the Cape Siskin, produced by Method IRWLS, values are pure model-predicted reporting rates.

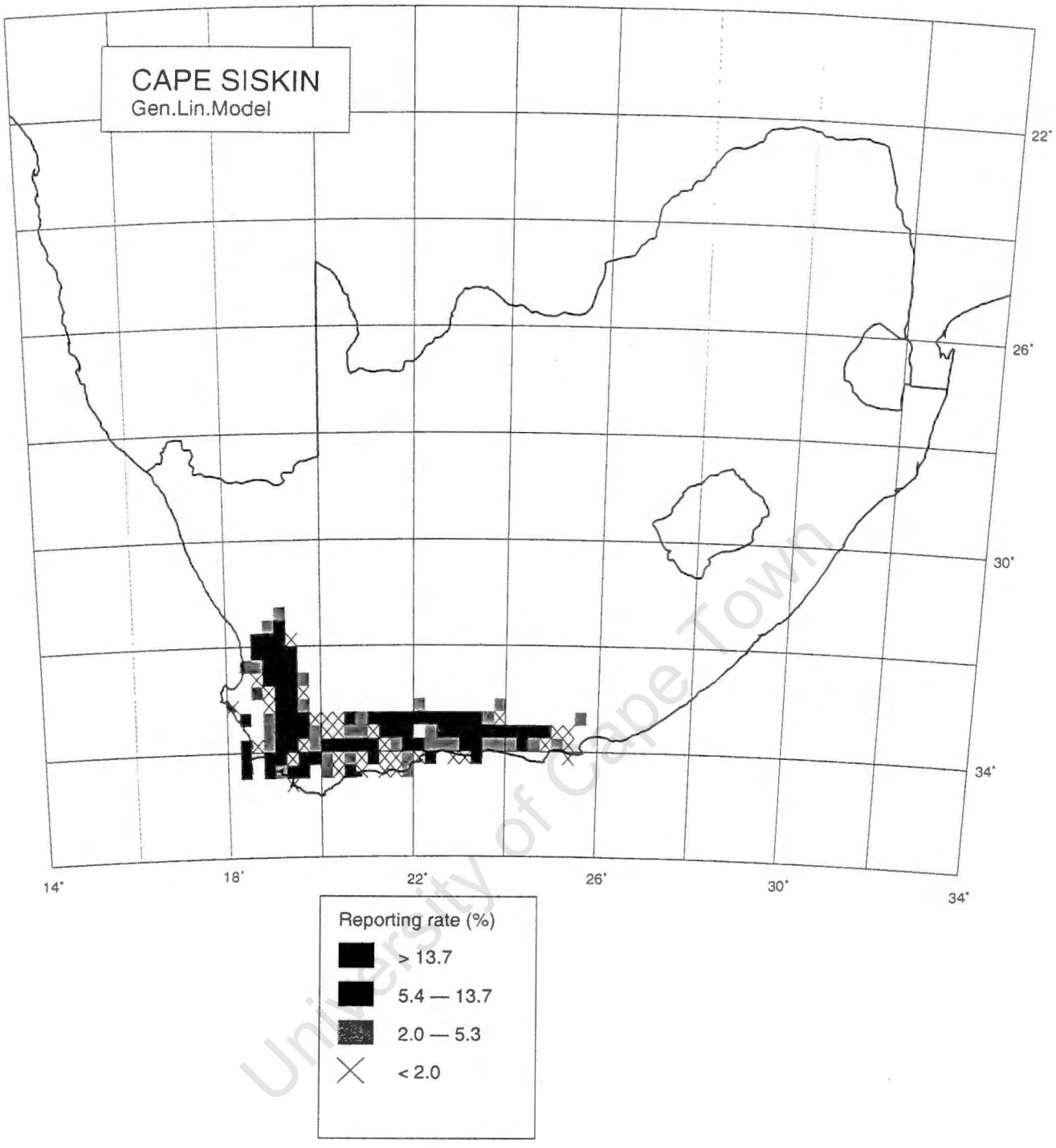


Figure 31. A smoothed distribution map for the Cape Siskin, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.05$.

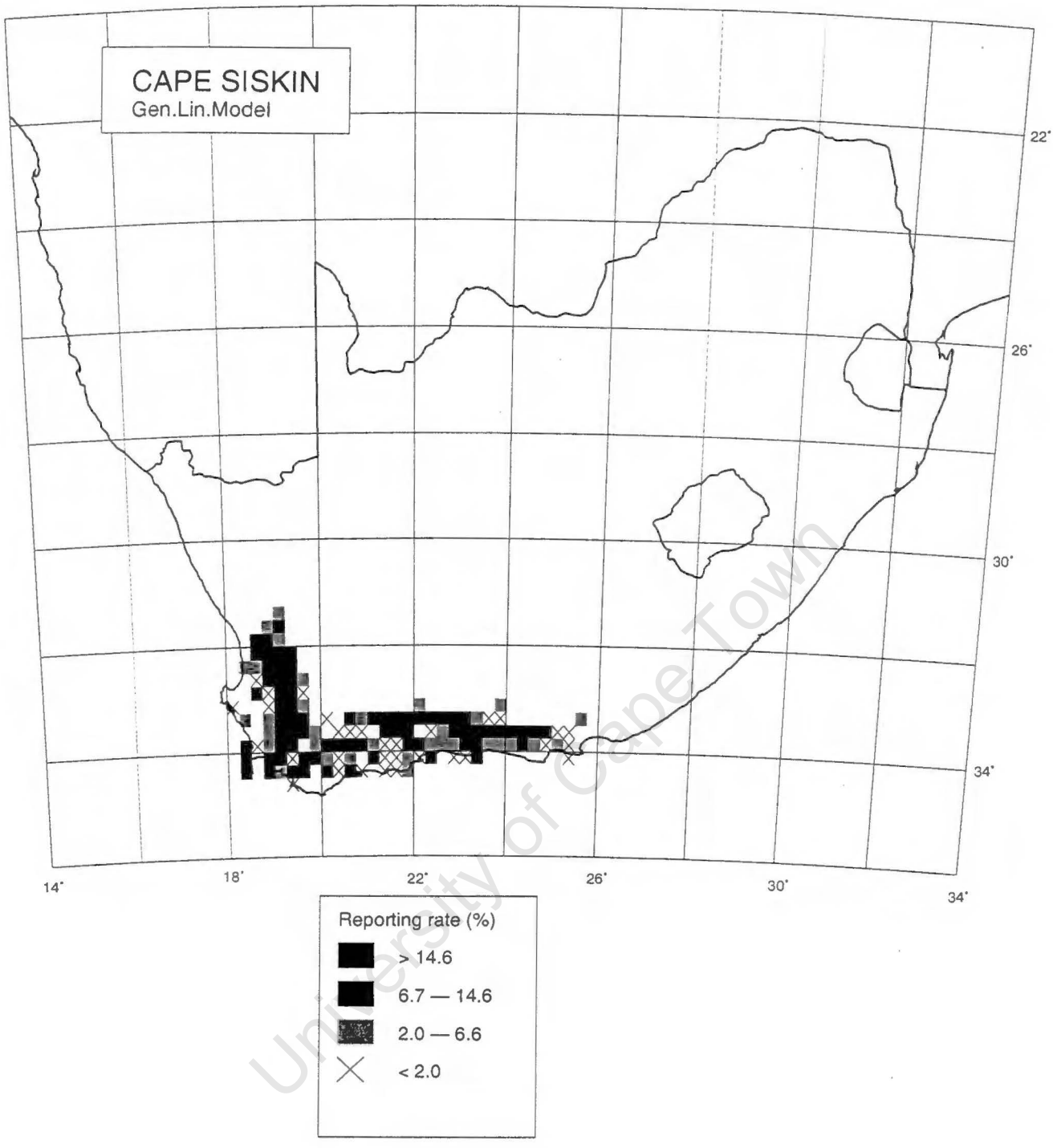


Figure 32. A smoothed distribution map for the Cape Siskin, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.1$.

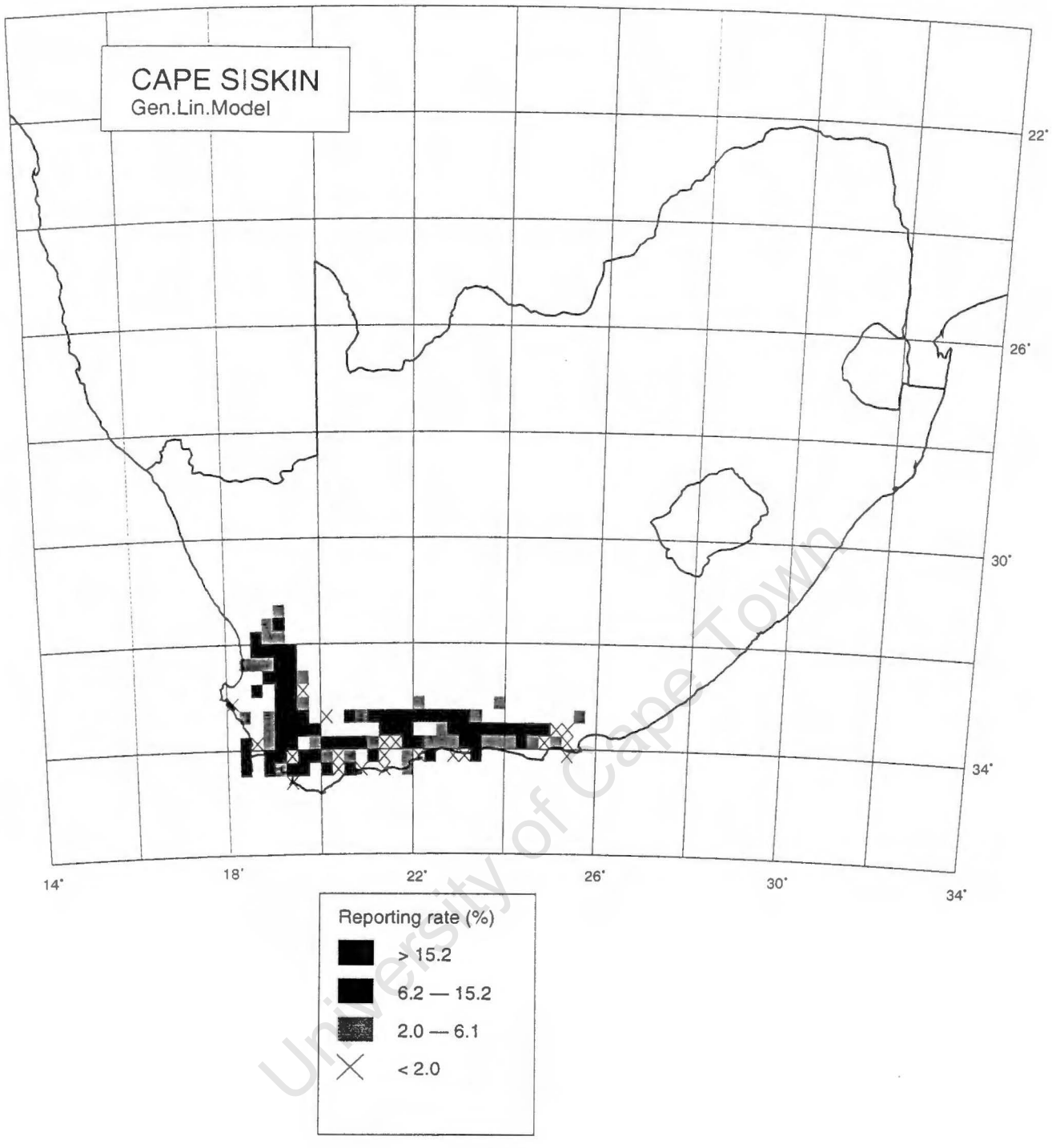


Figure 33. A smoothed distribution map for the Cape Siskin, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.2$.

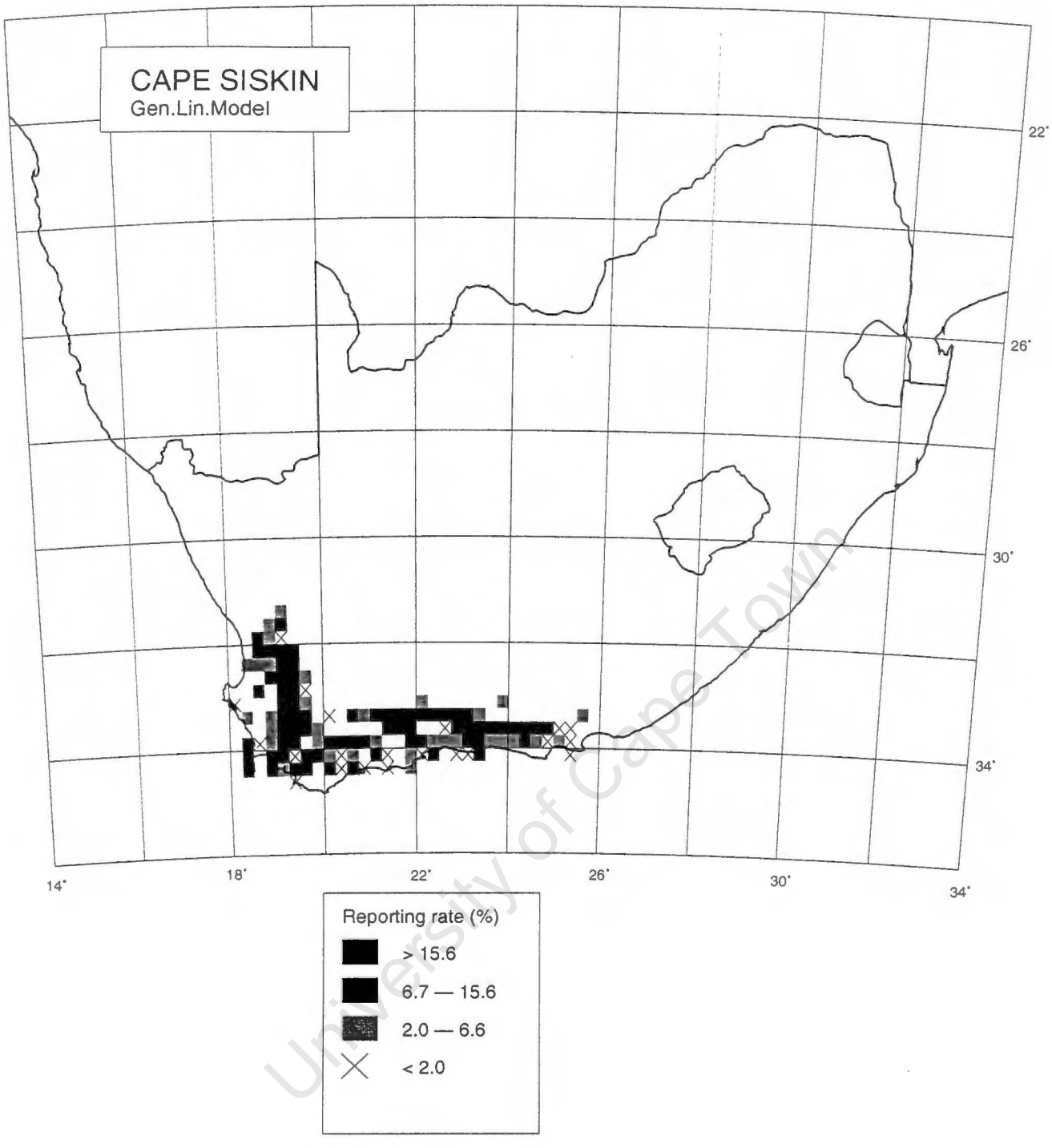


Figure 34. A smoothed distribution map for the Cape Siskin, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.3$.

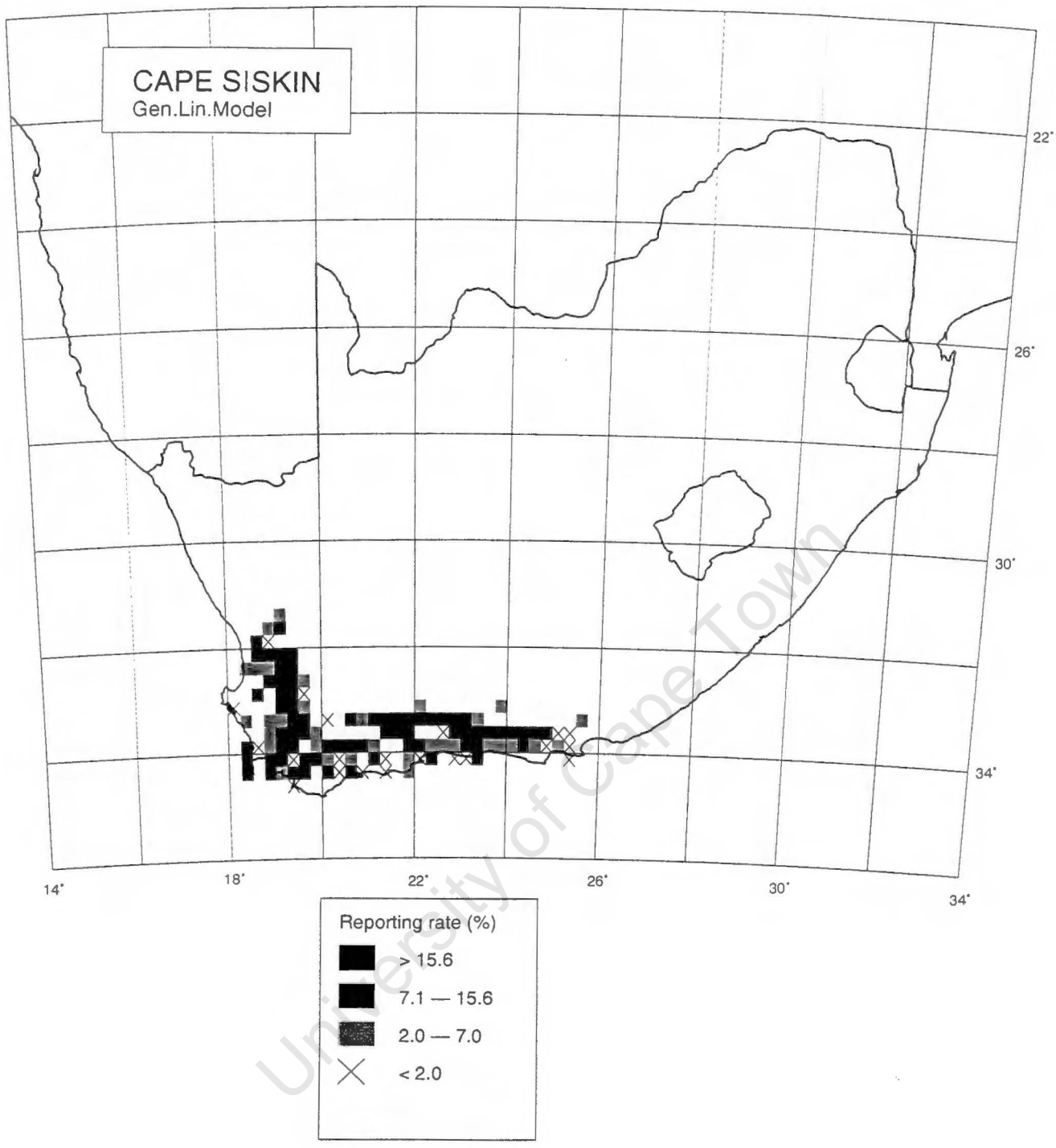


Figure 35. A smoothed distribution map for the Cape Siskin, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.4$.

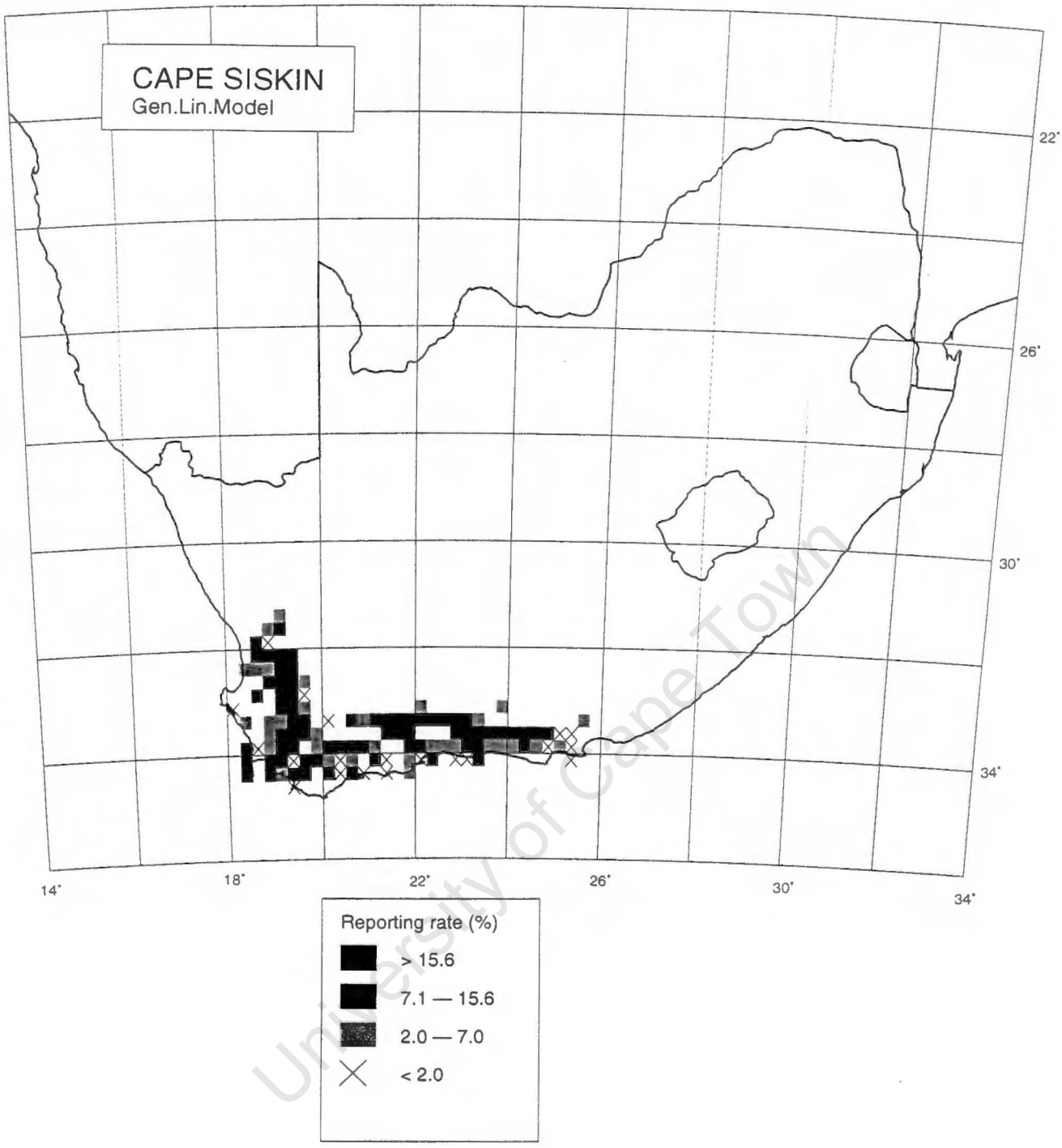


Figure 36. A smoothed distribution map for the Cape Siskin, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.5$.

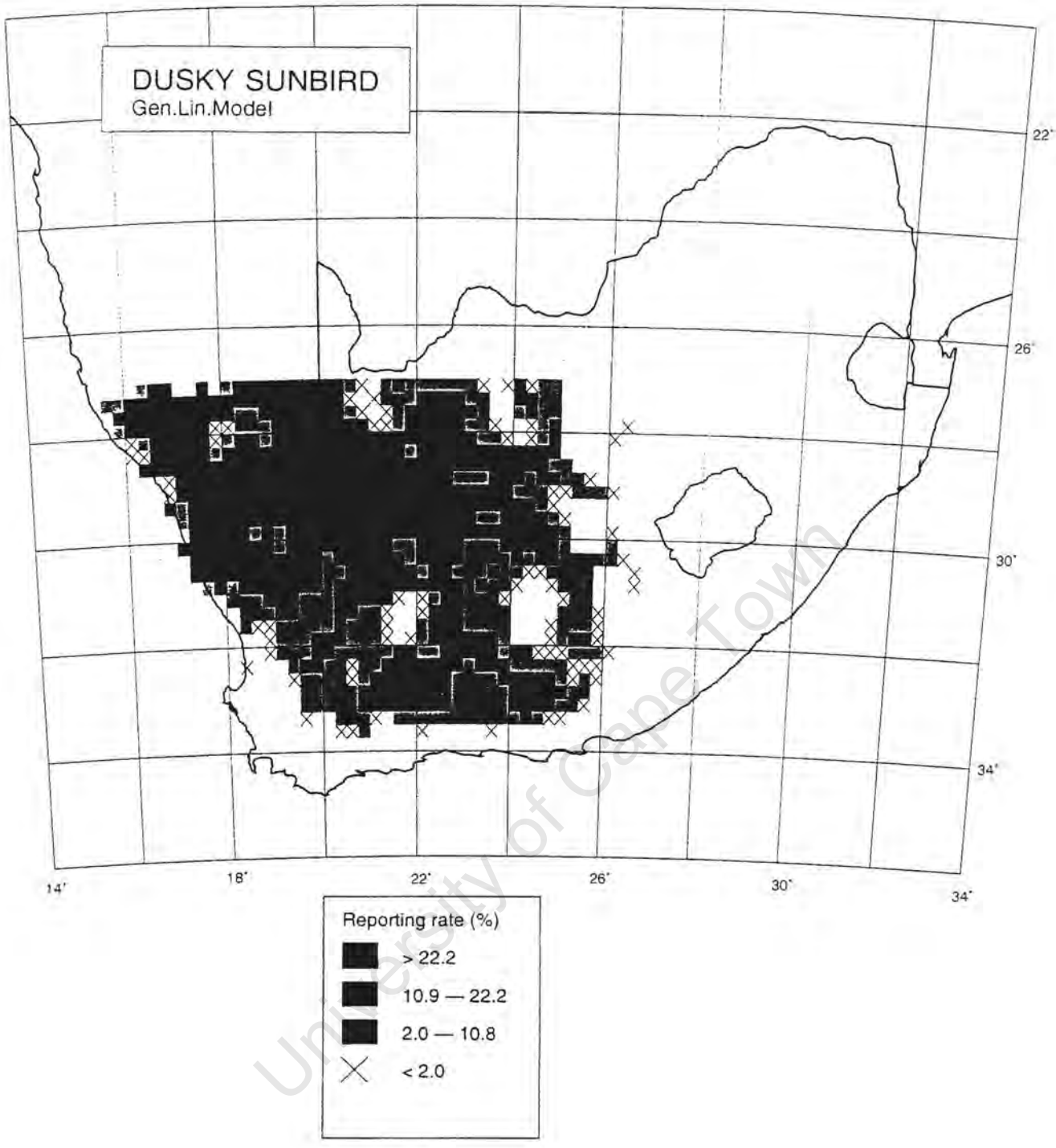


Figure 37. A smoothed distribution map for the Dusky Sunbird, produced by Method IRWLS, values are pure model-predicted reporting rates.

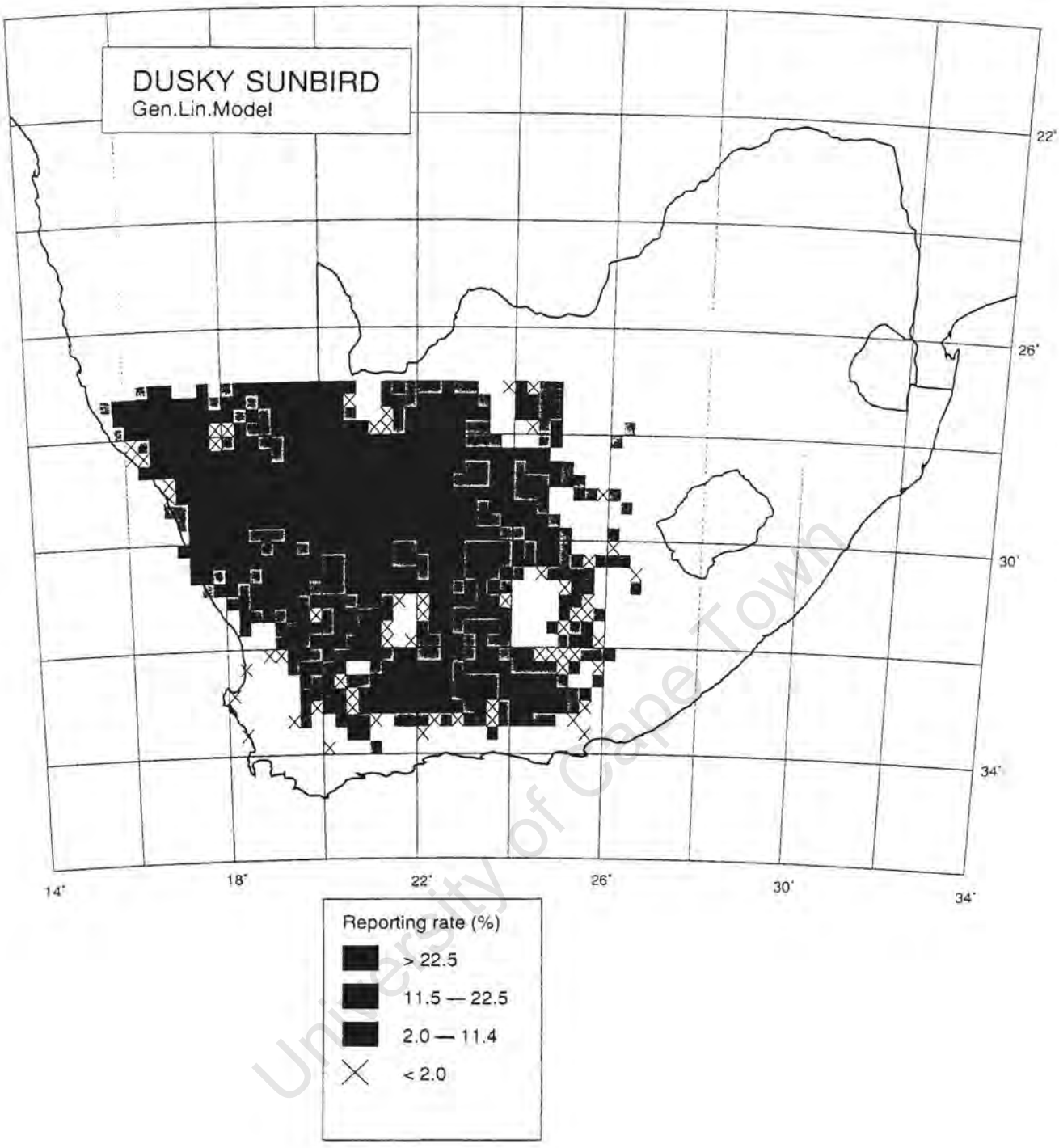


Figure 38. A smoothed distribution map for the Dusky Sunbird, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.05$.

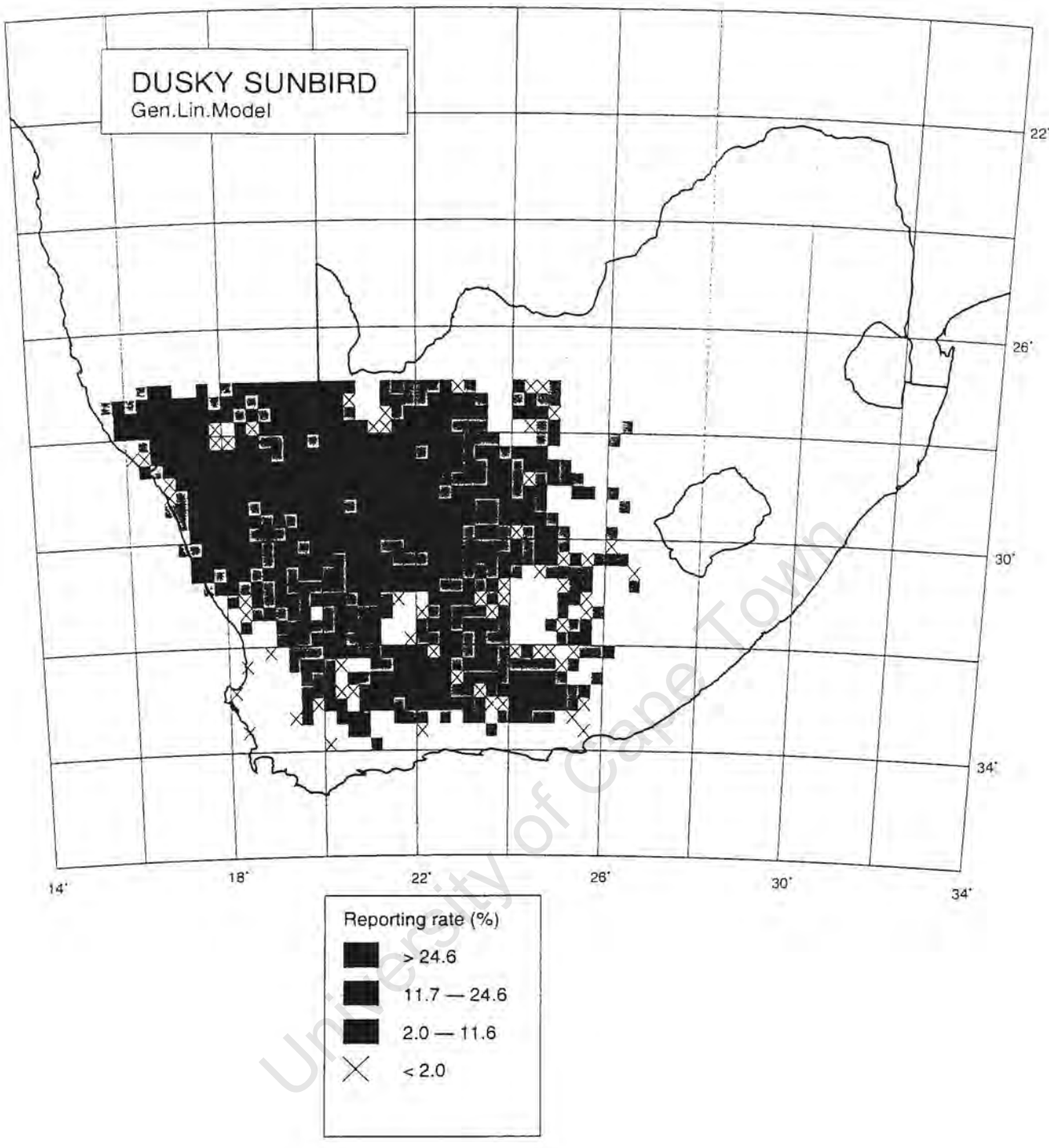


Figure 39. A smoothed distribution map for the Dusky Sunbird, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.1$.

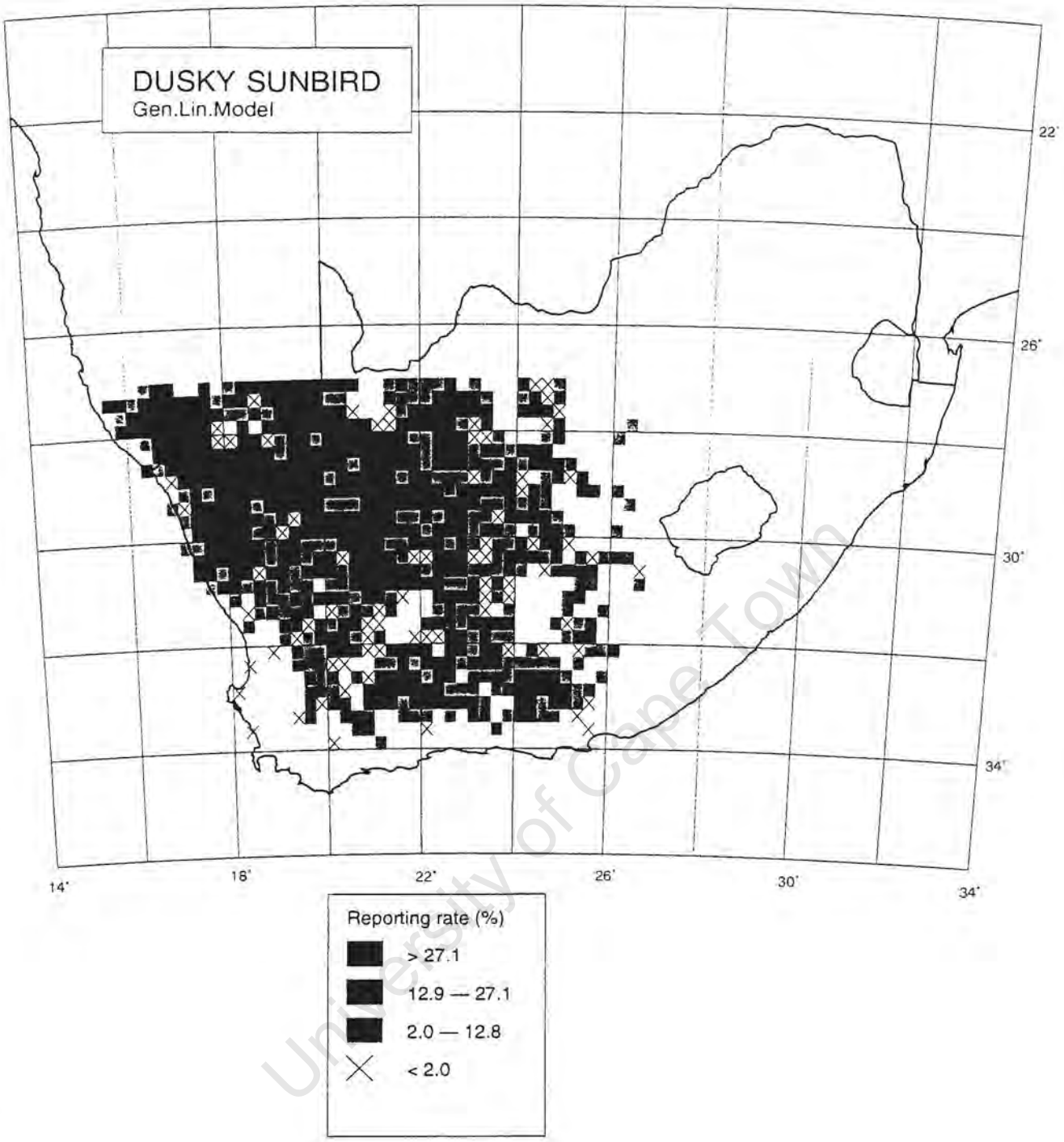


Figure 40. A smoothed distribution map for the Dusky Sunbird, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.2$.

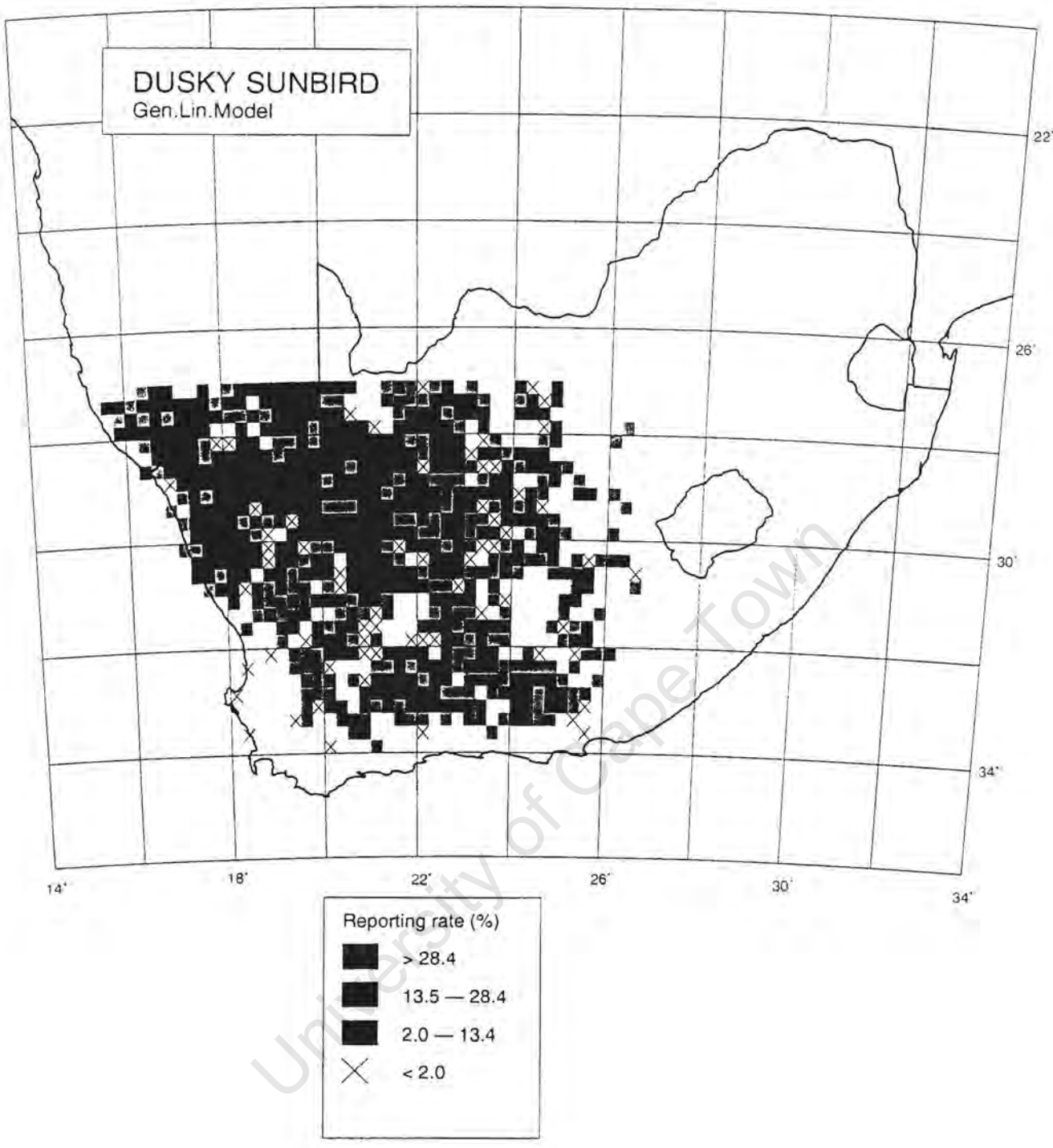


Figure 41. A smoothed distribution map for the Dusky Sunbird, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.3$.

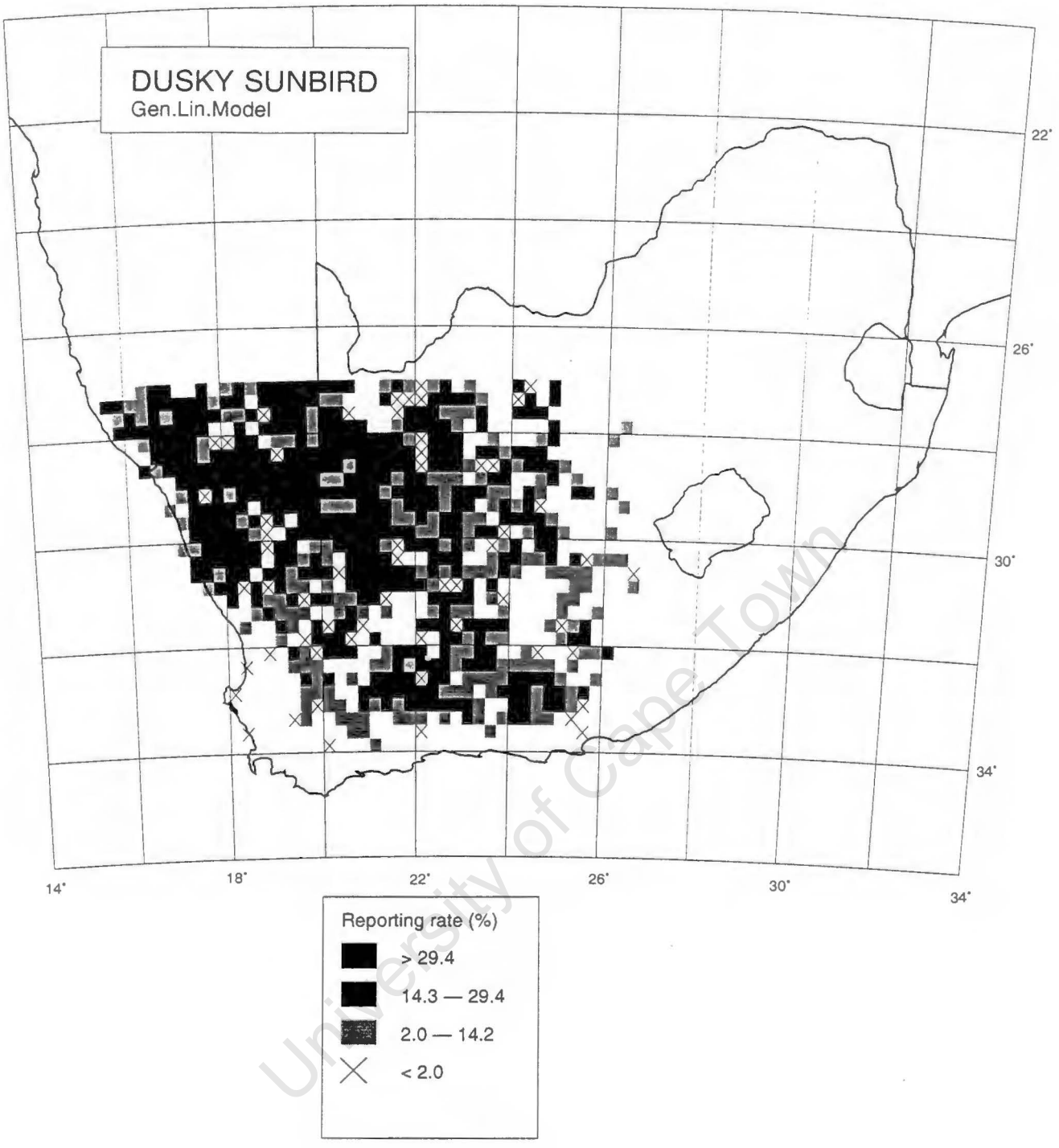


Figure 42. A smoothed distribution map for the Dusky Sunbird, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.4$.

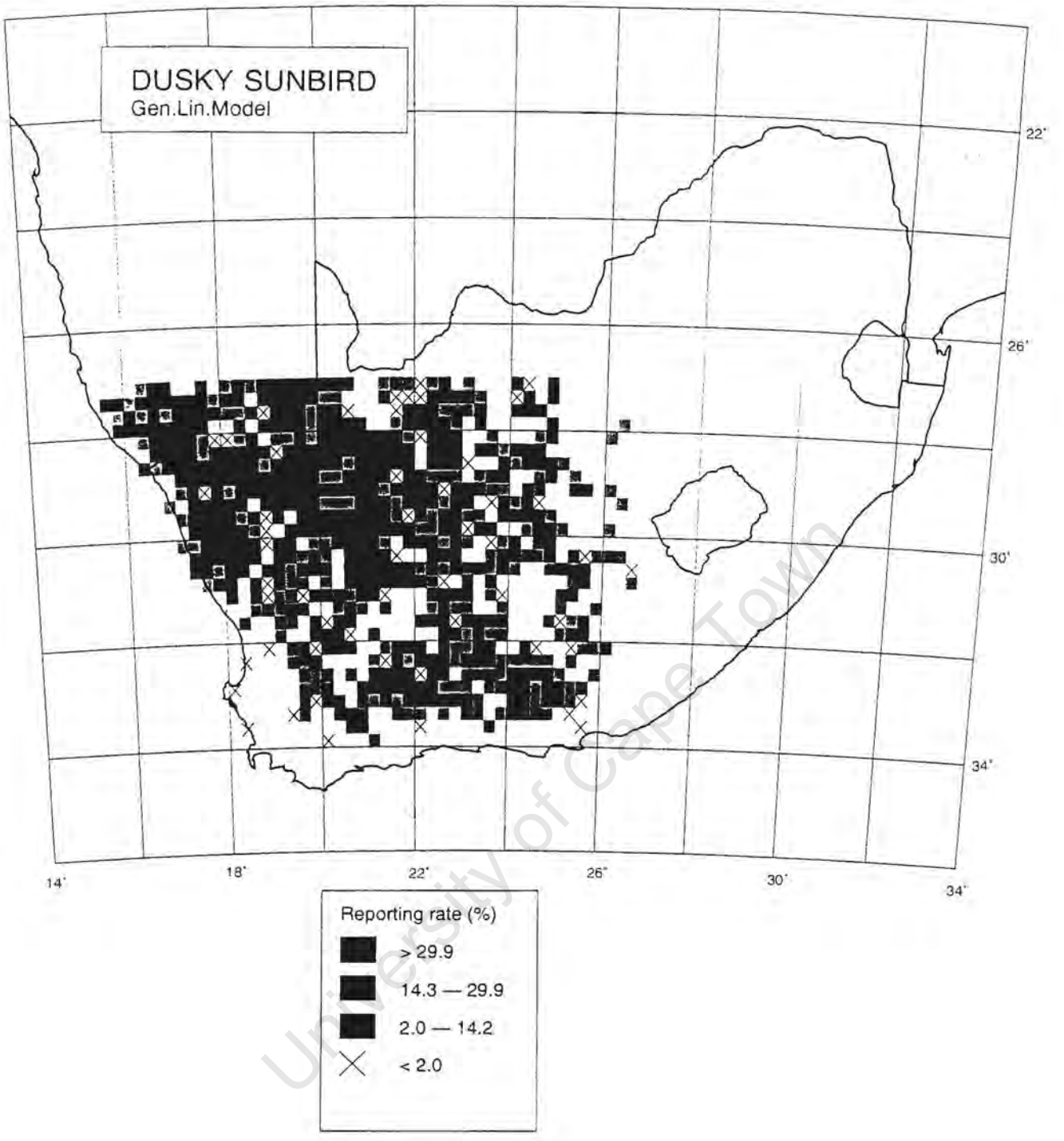


Figure 43. A smoothed distribution map for the Dusky Sunbird, produced by Method IRWLS. A weighted average between model-predicted reporting rates and observed reporting rates was used, $\alpha = 0.5$.

APPENDIX C

This appendix contains various similarity matrices calculated for Chapter 1.

TABLE C1 The 15 species with the most similar distribution to the species indicated on the left of the table are given here. These dissimilarities were calculated with the SQRTW10 coefficient. Only the Roberts numbers are shown for The species. The dissimilarities were calculated for the available species for the area of southern Africa south of 27°S.

188	191	791	189	787	196	815	816	811	810	494	869	205	807	206	793
	4409	5057	5319	5563	6247	6360	6360	6574	6746	7016	7060	7062	7097	7472	7521
189	815	204	780	791	496	779	188	787	871	205	206	191	808	196	793
	1949	3038	3339	3465	3544	4598	5319	5954	5968	6108	6191	6679	6910	7100	7346
190	775	872	813	200	203	507	783	512	814	195	495	500	879	881	878
	3061	3680	4172	4669	4792	4945	5082	5279	5313	5322	5569	5627	5663	5957	6057
191	787	196	791	188	811	810	869	807	790	494	816	793	792	808	205
	2422	3445	4153	4409	4417	4664	4947	5185	5500	5501	5701	6069	6295	6399	6418
192	785	200	792	873	872	198	774	881	813	869	811	494	203	810	775
	4504	4703	4719	4993	5038	5132	5172	5198	5459	5657	5726	5961	6119	6257	6367
193	870	492	495	199	508	506	878	814	494	507	498	203	205	200	497
	4831	5069	5122	5228	6276	6387	6535	6626	6766	7120	7355	7443	7636	7913	8135
194	779	201	497	205	493	193	870	498	788	495	878	814	506	203	200
	7419	8644	8652	8947	8975	9241	9549	9583	9595	9618	9683	9777	9789	9791	9808
195	773	777	783	874	813	190	775	877	880	872	512	879	878	203	507
	2767	3412	4126	4162	5260	5322	5378	6025	6064	6074	6222	6222	6410	6875	6958
196	787	811	810	191	807	494	869	790	791	793	792	808	817	789	199
	2116	3053	3426	3445	3604	3689	3719	4485	4608	4612	4959	5011	5400	5586	5940
198	785	873	792	877	881	192	813	783	872	869	789	810	808	793	811
	2448	2575	3657	3934	4106	5132	5301	5557	5698	5973	6135	6190	6211	6265	6400
199	494	870	508	193	814	203	200	507	495	196	506	192	492	787	498
	3881	3934	5191	5228	5352	5630	5721	5724	5886	5940	6321	6465	6642	6824	7271
200	203	872	813	775	507	814	881	190	192	792	494	785	495	869	783
	2898	3077	3508	3825	4109	4506	4607	4669	4703	4777	4934	5044	5142	5340	5676
201	205	779	194	199	189	508	188	193	870	206	815	494	498	492	497
	8179	8505	8644	8705	8760	8810	8951	9019	9028	9076	9196	9215	9270	9286	9324
203	813	872	814	200	775	507	881	783	495	878	792	190	877	494	785
	2440	2601	2833	2898	2971	3747	4125	4362	4380	4685	4780	4792	4854	4906	4914
204	780	189	815	496	871	791	818	779	808	206	817	793	790	787	789
	2762	3038	3620	4584	4698	5084	5465	5993	6390	6920	7006	7061	7076	7113	7170
205	791	787	815	189	196	191	780	779	494	496	807	810	188	811	790
	5417	5807	6008	6108	6234	6418	6582	6673	6855	6882	6967	6982	7062	7079	7124
206	496	189	815	204	780	791	779	871	188	205	808	191	793	196	787
	5391	6191	6533	6920	7023	7148	7230	7460	7472	7638	8116	8279	8555	8616	8620
492	193	508	199	495	870	506	494	507	814	878	501	498	203	200	512
	5069	5902	6642	6955	6975	7446	7544	7712	8046	8127	8184	8232	8533	8574	8694
493	194	497	779	201	205	193	788	508	498	495	506	878	200	870	507
	8975	9374	9508	9564	9566	9714	9807	9847	9851	9854	9888	9893	9925	9934	9938
494	869	196	787	811	199	810	792	203	200	807	793	870	808	881	191
	3129	3689	3836	3857	3881	3966	4136	4906	4934	5174	5246	5346	5453	5459	5501
495	814	870	506	878	507	203	193	200	498	512	879	500	190	775	199
	3017	3197	3335	3396	3723	4380	5122	5142	5330	5390	5508	5556	5569	5783	5886
496	815	189	204	780	779	791	206	871	205	787	188	808	191	196	793
	3500	3544	4584	4587	4819	5215	5391	6658	6882	7417	7573	7855	7954	8050	8411
497	498	788	506	878	870	511	495	508	205	193	814	779	879	500	194
	5533	6610	7035	7100	7237	7289	7387	7468	8107	8135	8183	8375	8616	8622	8652
498	506	870	788	814	878	495	497	500	879	507	876	203	494	508	512
	4361	4478	5194	5288	5301	5330	5533	5896	5959	6359	6793	6860	6875	7079	7166
499	774	509	201	192	190	500	507	200	792	881	508	872	775	199	813
	9199	9520	9730	9759	9777	9811	9832	9833	9834	9848	9849	9852	9864	9877	9881

500	879	878	512	506	507	814	878	788	775	495	190	498	783	200	502
	2270	2935	2956	3165	3501	3882	3922	4367	5333	5556	5627	5896	5938	6059	8159
501	492	193	508	199	870	495	498	506	494	878	205	786	507	814	497
	8184	8607	9022	9128	9186	9295	9307	9421	9439	9458	9473	9489	9518	9594	9624
502	876	512	879	500	878	783	788	507	775	506	195	190	814	813	495
	4848	5043	5203	6159	6635	6638	6949	7003	7154	7204	7313	7528	7695	8224	8357
504	510	511	876	506	500	788	512	497	879	507	498	878	814	495	508
	6702	6888	9177	9205	9217	9235	9334	9344	9452	9504	9527	9547	9767	9842	9855
505	779	496	780	819	204	815	818	782	206	189	791	871	816	817	790
	9266	9558	9574	9585	9672	9672	9677	9701	9754	9760	9809	9854	9876	9889	9897
506	878	814	507	500	495	879	870	512	876	498	788	203	508	200	199
	2525	2596	2836	3165	3335	3653	3757	3933	4036	4361	4729	5736	6141	6191	6321
507	814	506	878	512	500	495	203	879	200	775	870	813	872	190	876
	2635	2836	2918	3420	3501	3723	3747	4024	4109	4295	4820	4824	4855	4945	4973
508	199	870	492	495	506	193	507	494	814	498	878	497	200	203	205
	5191	5427	5902	6112	6141	6276	6648	6656	6966	7079	7409	7468	7757	7891	7942
509	499	508	205	199	201	492	870	507	193	498	494	506	192	200	196
	9520	9563	9709	9746	9752	9853	9879	9892	9905	9906	9911	9916	9918	9922	9931
510	511	504	788	506	500	876	498	512	497	879	878	507	814	508	870
	6532	6702	8635	8760	8761	8876	8938	9012	9014	9021	9129	9253	9507	9527	9693
511	510	504	497	788	506	498	500	878	876	879	508	507	512	870	814
	6532	6888	7289	7844	8356	8366	8765	8901	8953	9084	9271	9344	9426	9441	9496
512	879	876	500	507	878	506	814	502	775	783	190	495	788	813	195
	2559	2723	2956	3420	3566	3933	4746	5043	5082	5276	5279	5390	6085	6214	6222
514	502	199	507	493	498	501	504	505	511	512	785	786	788	792	190
	9935	9978	9985	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
773	777	874	195	880	783	877	813	775	872	190	203	873	785	881	198
	836	2329	2767	4184	4863	4952	5723	5843	5870	6211	6994	7059	7085	7119	7255
774	192	873	785	792	198	816	811	808	196	807	869	810	881	200	872
	5172	6446	7066	7246	7527	7589	7593	7685	7699	7786	7790	7817	7848	7900	7925
776	813	872	783	203	190	881	200	814	507	879	877	785	512	878	500
	1816	1880	2348	2971	3061	3726	3825	4138	4295	4460	5009	5072	5082	5325	5333
777	773	874	195	880	783	877	872	813	775	190	873	881	785	203	198
	836	1687	3412	3941	5316	5341	6186	6200	6269	6693	6947	7244	7273	7312	7338
779	189	496	815	204	780	205	791	206	194	787	871	188	498	497	201
	4598	4819	5323	5993	6037	6673	6921	7230	7419	7897	8112	8157	8320	8375	8505
780	204	815	791	189	818	496	871	817	790	808	779	787	793	789	807
	2762	2884	3335	3339	4096	4587	5424	5569	5670	5726	6037	6043	6075	6077	6437
782	204	189	780	818	871	815	496	779	808	206	791	793	188	787	789
	7655	8205	8379	8574	8869	8899	9234	9296	9479	9510	9555	9581	9599	9621	9625
783	775	813	872	877	881	195	879	203	785	773	190	792	512	777	873
	2348	2356	3262	3536	3903	4126	4192	4362	4395	4863	5082	5174	5276	5316	5330
785	792	881	198	877	873	813	872	869	783	192	810	811	203	200	775
	1514	1910	2448	2502	2672	3451	3683	3770	4395	4504	4634	4838	4914	5044	5072
786	501	193	199	494	870	492	506	814	497	495	508	507	787	196	203
	9489	9548	9761	9827	9827	9891	9893	9895	9910	9924	9932	9944	9957	9961	9964
787	196	191	811	810	869	791	807	494	793	790	808	792	817	789	816
	2116	2422	2821	2989	3230	3462	3483	3836	4286	4466	4971	5155	5238	5468	5491
788	879	500	876	506	878	498	814	512	870	497	507	495	502	783	775
	3687	4367	4606	4729	5029	5194	5534	6085	6179	6610	6680	6748	6949	7254	7566
789	793	817	808	807	790	810	811	869	792	877	791	787	873	196	494
	1030	1322	1346	1919	2078	2323	3115	3652	4681	5049	5051	5468	5477	5586	6002
790	817	793	807	789	808	810	811	791	869	787	196	191	818	816	780
	1451	1751	1984	2078	2121	2862	3447	3776	4343	4466	4485	5500	5644	5662	5670
791	815	780	787	189	790	191	793	817	810	807	808	196	811	789	188

	2996	3335	3462	3465	3776	4153	4468	4473	4477	4533	4590	4608	4823	5051	5057
792	785	869	881	811	810	877	873	198	807	813	793	872	494	808	789
	1514	1788	2437	2489	2652	2797	3210	3657	3731	3792	3888	4034	4136	4495	4681
793	808	789	807	810	817	790	811	869	792	787	791	196	873	877	494
	695	1030	1197	1500	1577	1751	2101	2693	3888	4296	4468	4612	4785	4949	5246
807	793	810	811	808	789	790	817	869	787	196	792	791	877	494	191
	1197	1325	1483	1827	1919	1984	2008	2542	3483	3604	3731	4533	5025	5174	5185
808	793	789	807	817	790	810	811	869	792	791	873	787	196	877	494
	695	1346	1827	1969	2121	2281	2782	3331	4495	4590	4769	4971	5011	5416	5453
810	811	869	807	793	808	789	792	817	790	787	196	877	494	791	785
	900	1260	1325	1500	2281	2323	2652	2664	2862	2989	3426	3949	3966	4477	4634
811	810	869	807	793	792	808	787	196	789	817	790	494	191	877	791
	900	1294	1483	2101	2489	2782	2821	3053	3115	3177	3447	3857	4417	4539	4823
813	872	775	783	203	881	785	877	200	792	190	507	814	195	198	192
	1382	1816	2356	2440	2866	3451	3462	3508	3792	4172	4824	5197	5260	5301	5459
814	878	506	507	203	495	870	879	500	775	200	512	876	872	813	498
	2526	2596	2635	2833	3017	3139	3546	3882	4138	4506	4746	5014	5149	5197	5288
815	189	780	791	496	204	779	787	205	818	188	871	808	206	191	793
	1949	2884	2996	3500	3620	5323	5813	6008	6133	6360	6400	6421	6533	6730	6812
816	787	807	790	191	811	810	793	196	818	817	188	791	808	869	789
	5491	5515	5662	5701	5900	6035	6046	6059	6168	6236	6360	6605	6642	6860	6943
817	789	790	793	808	807	810	811	869	791	787	196	780	792	818	877
	1322	1451	1577	1969	2008	2664	3177	4142	4473	5238	5400	5569	5615	5682	6136
818	780	204	790	817	791	815	816	789	793	807	808	787	189	810	811
	4096	5465	5644	5682	5725	6133	6168	6587	6752	6840	6966	7304	7498	7556	7867
819	496	189	188	791	505	206	191	204	779	787	815	782	780	196	807
	9422	9432	9447	9516	9585	9596	9634	9679	9694	9743	9746	9753	9775	9782	9808
869	810	811	792	807	793	494	787	808	789	196	881	785	877	817	790
	1260	1294	1788	2542	2693	3129	3230	3331	3652	3719	3769	3770	3948	4142	4343
870	814	495	506	199	878	498	507	193	203	494	508	200	788	500	879
	3139	3197	3757	3934	4132	4478	4820	4831	5072	5346	5427	6075	6179	6680	6771
871	204	780	189	815	496	791	206	818	779	808	188	789	817	793	790
	4698	5424	5968	6400	6658	7100	7460	8039	8112	8271	8273	8527	8564	8648	8650
872	813	775	203	881	200	783	190	785	877	792	507	192	814	873	869
	1382	1880	2601	2812	3077	3262	3680	3683	3862	4034	4855	5038	5149	5277	5663
873	198	785	792	877	881	808	793	192	872	810	811	783	869	789	807
	2575	2672	3210	4022	4455	4769	4785	4993	5277	5300	5328	5330	5364	5477	5528
874	777	773	880	195	783	877	775	872	813	190	881	873	785	203	198
	1687	2329	3333	4162	5956	6486	6701	6767	6822	6873	7497	7560	7788	7909	7928
875	774	190	872	192	775	200	512	813	881	507	878	785	203	873	198
	8118	8397	8964	9057	9067	9077	9095	9321	9390	9502	9522	9567	9621	9692	9740
876	879	512	500	506	788	878	502	507	814	775	783	190	495	498	200
	2243	2723	2935	4036	4606	4788	4848	4973	5014	5734	5930	6318	6434	6793	7470
877	785	881	792	813	783	872	198	869	810	873	811	203	793	773	775
	2502	2737	2797	3462	3536	3862	3934	3948	3949	4022	4539	4854	4949	4952	5009
878	506	814	507	879	495	512	500	870	203	876	788	498	775	783	200
	2525	2526	2918	3173	3396	3566	3922	4132	4685	4788	5029	5301	5325	5490	5913
879	876	500	512	878	814	506	788	507	783	775	502	495	190	498	813
	2243	2270	2559	3173	3546	3653	3687	4024	4192	4460	5203	5508	5663	5959	6110
880	874	777	773	195	783	877	190	775	872	813	881	879	203	873	785
	3333	3941	4184	6064	7746	7868	7998	8020	8143	8263	8660	8826	8854	8898	9019
881	785	792	877	872	813	775	869	783	198	203	873	200	810	192	811
	1910	2437	2737	2812	2866	3726	3769	3903	4106	4125	4455	4607	4979	5198	5288

TABLE C2. SQRTW10 dissimilarities for the group of Canaries.

distributions south of 27°S were considered.

	869	870	871	872	873	874	875	876	877	878	879	880	881
869	0	9070	9303	5663	5364	9940	9835	9603	3948	9443	9255	9977	3769
870	9070	0	9999	7867	9804	9961	9883	7711	9653	4132	6771	9978	8703
871	9303	9999	0	9983	9923	10000	10000	10000	9788	10000	10000	9999	9903
872	5663	7867	9983	0	5277	6767	8964	7796	3862	6705	6684	8143	2812
873	5364	9804	9923	5277	0	7560	9692	9847	4022	9327	9324	8898	4455
874	9940	9961	10000	6767	7560	0	10000	8998	6486	8379	7950	3333	7497
875	9835	9883	10000	8964	9692	10000	0	9778	9919	9522	9941	9999	9390
876	9603	7711	10000	7796	9847	8998	9778	0	9518	4788	2243	9469	8632
877	3948	9653	9788	3862	4022	6486	9919	9518	0	8097	8137	7868	2737
878	9443	4132	10000	6705	9327	8379	9522	4788	8097	0	3173	9091	7965
879	9255	6771	10000	6684	9324	7950	9941	2243	8137	3173	0	8826	7323
880	9977	9978	9999	8143	8898	3333	9999	9469	7868	9091	8826	0	8660
881	3769	8703	9903	2812	4455	7497	9390	8632	2737	7965	7323	8660	0

TABLE C3. SQRTW10 dissimilarities for the group of Weavers.

Only distributions south of 27°S were considered.

	807	808	810	811	813	814	815	816	817	818	819
807	0	1827	1325	1483	7561	9330	6866	5515	2008	6840	9808
808	1827	0	2281	2782	7875	9548	6421	6642	1969	6966	9941
810	1325	2281	0	900	6759	8875	7118	6035	2664	7556	9884
811	1483	2782	900	0	6655	8894	7220	5900	3177	7867	9858
813	7561	7875	6759	6655	0	5197	9856	9230	8485	9921	9978
814	9330	9548	8875	8894	5197	0	9735	9806	9676	9889	9990
815	6866	6421	7118	7220	9856	9735	0	7976	6965	6133	9746
816	5515	6642	6035	5900	9230	9806	7976	0	6236	6168	9979
817	2008	1969	2664	3177	8485	9676	6965	6236	0	5682	9959
818	6840	6966	7556	7867	9921	9889	6133	6168	5682	0	9967
819	9808	9941	9884	9858	9978	9990	9746	9979	9959	9967	0