Predictive modelling of species' potential geographical distributions

A thesis submitted in fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
Of
Rhodes University

By
Mark Peter Robertson

December 2002

Abstract

Models that are used for predicting species' potential distributions are important tools that have found applications in a number of areas of applied ecology. The majority of these models can be classified as *correlative*, as they rely on strong, often indirect, links between species distribution records and environmental predictor variables to make predictions. Correlative models are an alternative to more complex mechanistic models that attempt to simulate the mechanisms considered to underlie the observed correlations with environmental attributes. This study explores the influence of the type and quality of the data used to calibrate correlative models.

In terms of data type, the most popular techniques in use are group discrimination techniques, those that use both presence and absence locality data to make predictions. However, for many organisms absence data are either not available or are considered to be unreliable. As the available range of profile techniques (those using presence only data) appeared to be limited, new profile techniques were investigated and evaluated. A new profile modelling technique based on fuzzy classification (the Fuzzy Envelope Model) was developed and implemented. A second profile technique based on Principal Components Analysis was implemented and evaluated. Based on quantitative model evaluation tests, both of these techniques performed well and show considerable promise.

In terms of data quality, the effects on model performance of false absence records, the number of locality records (sample size) and the proportion of localities representing species presence (prevalence) in samples were investigated for logistic regression distribution models. Sample size and prevalence both had a significant effect on model performance. False absence records had a significant influence on model performance, which was affected by sample size.

A quantitative comparison of the performance of selected profile models and group discrimination modelling techniques suggests that different techniques may be more successful for predicting distributions for particular species or types of organism than others. The results also suggest that several different model design/ sample size combinations are capable of making predictions that will on average not differ significantly in performance for a particular species. A further quantitative comparison among modelling techniques suggests that correlative techniques can perform as well as simple mechanistic techniques for predicting potential distributions.

This thesis is dedicated to my parents who have always supported me in my endeavours

Acknowledgements

In addition to the acknowledgements at the end of each chapter, I would like to express my sincere thanks to several people who have assisted me with various aspects of this study. I am extremely grateful to my supervisors, Martin Villet and Tony Palmer, for their advice and encouragement. I have really enjoyed working with you and have learnt an enormous amount.

I thank Sarah Radloff for statistical advice, Mike Burton for advice with MATLAB software, Lesley Henderson for access to alien plant locality data from SAPIA, and Clyde Mallinson for access to computer hardware. I thank the School of Bioresources Engineering and Environmental Hydrology (University of Natal), the Water Research Commission and the South African Country Study for Climate Change for the use of the climatic predictor variables, and I am grateful to Greg Kiker for assistance in obtaining this data.

I am grateful to Neil Caithness, Craig Peter, Brad Ripley, Dai Herbert and Jan-Robert Baars for the valuable insights into various aspects of predictive biogeography that I have gained through research undertaken in collaboration with them.

I thank Tammy Smith for her love and encouragement, and for helping me to remain focused on this thesis. I appreciate the inspiration and support that my friends have given me during this time. I thank my parents for their support and encouragement throughout my studies.

Funding from the National Research Foundation and the Rhodes University Joint Research Council is gratefully acknowledged.

Table of Contents

I

Introduction

Preface

This chapter introduces predictive distribution modelling including a broad classification of predictive models, the theoretical background on which these predictive models are based, an introduction to the target organisms and the rationale of the thesis.

Potential distribution models

Potential distribution models of species' ranges are important tools that have found applications in a number of areas of applied ecology. Some examples include the management of plants (Panetta and Dodd, 1987; Panetta and Mitchell, 1991; Sindel and Michael, 1992; Higgins *et al.*, 1999), the management of disease vectors (Rogers and Randolph, 1993; Rogers and Williams, 1993; Rogers *et al.*, 1996; Robinson *et al.*, 1997), the study of climate change (Lindenmayer *et al.*, 1991; Beerling *et al.*, 1995; Schulze and Kunz, 1995; Leathwick *et al.*, 1996; Rutherford *et al.*, 1999; Samways *et al.*, 1999; Peterson *et al.*, 2001), biodiversity studies (Austin, 1998; Cumming, 2000 a), to test biogeographical hypotheses (Leathwick, 1998; Peterson *et al.*, 1999), understanding biogeographic patterns and processes (Leathwick and Mitchell, 1992; Leathwick, 1995; Austin *et al.*, 1996; Leathwick *et al.*, 1998; Leathwick and Austin, 2001; Leathwick and Whitehead, 2001), conservation (Osborne and Tigar, 1992; Margules and Austin, 1994; Augustin *et al.*, 1996; Pfab and Witkowski, 1997; Lloyd and Palmer, 1998; Pearce and Lindenmayer, 1998; Leathwick, 2001; Funk and Richardson, 2002) and biological control (Scott, 1992; Palmer *et al.*, 2000; Baars, 2002).

These studies have focused on a number of target organisms including plants (Austin *et al.*, 1990; Skov and Borchsenius, 1997; Franklin, 1998; Guisan *et al.*, 1998, 1999; Leathwick and Whitehead, 2001), birds (Osborne and Tigar, 1992; Buckland *et al.*, 1996; Manel *et al.*, 1999 a & b; Peterson *et al.*, 1999; Lenton *et al.*, 2000), insects (Williams *et al.*, 1994; Samways *et al.*, 1999; Baker *et al.*, 2000; Erasmus *et al.*, 2000), ticks (Cumming, 2000 a & b), mammals (Lindenmayer *et al.*, 1991; Walker, 1990; Walker and Cocks, 1991; Walton *et al.*, 1992; Carpenter *et al.*, 1993; Bauer *et al.*, 1994; Augustin *et al.*, 1996; Skidmore *et al.*, 1996; Jackson and Claridge, 1999; Hirzel, 2001), reptiles (Nix, 1986; Dorrough and Ash, 1999) and land-snails (Kadmon and Heller, 1998). Numerous other examples are given in reviews by Franklin (1995) and Guisan and Zimmermann (2000).

A wide variety of models have been produced to address a large number of biological issues. These models have different designs, require different input data, make different assumptions about these data and differ in the results that they produce and the way in which these results can be realised and applied. To understand this diversity, it is necessary to classify these models using their attributes.

Classification of modelling techniques

Predictive modelling techniques have been described as *static* or *dynamic* (Beerling *et al.*, 1995). Static models provide time-independent equilibrium or quasi-equilibrium predictions while dynamic models predict time-dependent dynamic responses to a changing environment (Beerling *et al.*, 1995). Static models have in turn been divided into two groups, namely *correlative* and *mechanistic* techniques (Beerling *et al.*, 1995). Correlative models are equivalent to Guisan and Zimmermann's third group of models that have been called *empirical*, *statistical* and *phenomenological* models (Guisan and Zimmermann, 2000). Correlative models rely on strong, often indirect links between species distribution records and environmental predictor variables to make predictions (Beerling *et al.*, 1995).

Mechanistic models are equivalent to Guisan and Zimmermann's second group of models which have been called *mechanistic*, *causal* or *process* models (Guisan and Zimmermann, 2000). Mechanistic models attempt to simulate the mechanisms considered to underlie the observed correlations with environmental attributes

(Beerling *et al.*, 1995) by using a detailed knowledge of the target species' life-history attributes and physiological responses to environmental variables (Stephenson, 1998). Such models have also been referred to as *ecophysiological models* (Stephenson, 1998) and *process orientated* models (Carpenter *et al.*, 1993). Stephenson (1998) maintains that the distinction between correlative and ecophysiological (mechanistic) models is often not clear. For example, he observes that for plants, ecophysiological studies depend on empirical correlations to determine quantitative relationships between physiologically important factors and vegetation distribution. Similarly, correlative models have an ecophysiological basis when they employ predictor variables that are suspected to be of broad physiological importance to plants (Stephenson, 1998).

Although the distinction between correlative and mechanistic models may not be clear it remains a useful theoretical framework for describing these models; they should possibly be viewed as two opposite extremes of a continuum rather than two distinct types.

The classification by Caithness (1995) of correlative modelling techniques into group discrimination and profile techniques separates, respectively, those techniques that make use of absence data from those that do not. Correlative models that use both presence and absence locality records to make predictions have been referred to as *group discrimination techniques*, while those that use only presence locality records have been referred to as *profile techniques* (Caithness, 1995). It is useful to separate these techniques on the basis of the type of data that they use, because the data quality considerations associated with these data types differs considerably.

Predictive modelling

In order to predict the distribution of a species using a correlative approach one needs:

1. Locality records, which document the occurrence of the target species, and consist of either presence/absence data or presence-only data.
2. A set of environmental predictor variables that cover the map region over which predictions are to be made.

3. Ecological theory that gives an expectation of how a species responds to its environment and to other species.
4. Modelling techniques that are compatible with ecological theory.
5. Geographical information systems (GIS) to manage the environmental predictor variables and to display the potential distribution maps produced from the models.

The components listed above are discussed in greater detail in the thesis. In this chapter locality records are introduced and discussed briefly. Locality records are discussed further with respect to data quality in Chapter 2. Environmental data are discussed in Chapter 2, while ecological theory and modelling techniques are reviewed in Chapter 3.

Locality records

There are three broad types of data which record species occurrence:

1. Data collected from plots using stratified survey methods where presence/absence data are determined reliably. This is typical of most vegetation surveys.
2. Data collected from surveys of particular areas where specified amounts of sampling effort per area are undertaken. In these surveys, presence/absence data are collected. While presence data are reliable, the reliability of absence data is conditional on the amount of sampling effort. This is typical of most well-designed fauna surveys.
3. Data collected on an opportunistic or *ad hoc* basis where only the presence and not the absence of the species is recorded. This is typical of herbarium and museum collection records.

In general, locality data are thus either obtained from systematic field surveys (presence/absence data) or from existing collections (presence-only data). For field surveys a sampling strategy is usually implemented to maximise efficiency, reduce bias and meet the requirements of the model's objectives (Austin, 1998; Guisan and Zimmermann, 2000). In contrast, data from collections are usually collected on an opportunistic or *ad hoc* basis (Stockwell and Peters, 1999) by a number of different

collectors over a period of time (Peterjohn, 2001; Peterson, 2001). As a cons
data quality usually differs between these sources of locality data (Chapter 2,
source of locality data available for the target species (and its quality) will ha
major influence on the type of modelling technique is used (profile or grou
discrimination).

Aim

The aim of this thesis is to investigate the use of techniques for predicting species distributions using presence-only data (Profile techniques), to investigate various aspects of data quality, and to compare the performance of various techniques for predicting species distributions.

Rationale

There is a considerable amount of data in herbarium and museum collections (Soberón *et al.*, 1996; Funk and Richardson, 2002) that is potentially useful. However, this data has mostly been collected on an *ad hoc* basis rather than by means of systematic field surveys. As a result this is largely presence-only data, which has a number of weaknesses (Margules and Austin, 1994; Zaniewski *et al.* 2002; Chapter 2). The most serious of these weaknesses is that there is usually geographical bias in these datasets (Margules and Austin, 1994; Austin, 1998; Freitag *et al.*, 1998; Lawes and Piper, 1998; Funk and Richardson, 2002; Ferrier, 2002; Zaniewski *et al.* 2002).

Despite the weaknesses associated with presence-only data, there is pressure to use this data, as it is often the only source of data available (Funk and Richardson, 2002). In addition, resources to conduct systematic field surveys to obtain more reliable presence/absence data are often limited. Given this situation, there is a need to investigate the use of presence-only data for predicting species distributions.

Thesis outline

The majority of modelling techniques rely on presence/absence data in order to make species distribution predictions (Franklin, 1995; Guisan and Zimmermann, 2000) and at the time of commencement of the thesis there were very few presence-only modelling techniques available. Consequently, a review of the available correlative techniques was needed, paying particular attention to profile techniques (Chapter 3). Next, there was a need to investigate, design and implement new profile techniques that represented improvements or refinements over existing techniques (Chapters 4 and 5).

When only presence data are available or when absence data are unreliable, an alternative to using profile techniques is to use a group discrimination technique by making use of pseudo-absence data (Ferrier and Watson, 1997). Aside from the problems of bias in presence-only datasets, false absence data are a potential problem, especially when pseudo-absence records are used. False absence data may also be a problem in surveyed presence/ absence data (Chapter 2), as are problems related to prevalence. The effects of false absence data, sample size, prevalence and the use of pseudo-absence data were investigated in Chapter 6.

The next step was to compare the performance of the profile techniques that had been developed and implemented with group discrimination techniques. Chapter 7 compares the performance of selected profile models (including those described in Chapters 4 and 5) and group discrimination predictive modelling techniques. At the time, few quantitative comparisons between profile and group discrimination techniques had been published (Ferrier and Watson, 1997), although such comparisons have subsequently appeared in the literature (Hirzel *et al.*, 2001; Zaniewski *et al.*, 2002). Another important question that had not been addressed was whether correlative techniques could perform as well as simple mechanistic techniques for predicting potential distributions. This question is addressed in Chapter 8. Chapter 9 provides a general discussion of the findings of the thesis. The appendix consists of a paper that describes a multi-criterion system for prioritising problem plants that are most in need of management action and control. This system represents a potentially important and useful means of making policy decisions for managing invasive plants. It is relevant here as at least one of the criteria in the

system may make use of potential distribution predictions. This system was also used in the process of selecting some of the target species (invasive alien plants).

Target organisms

The target organisms selected to address the issues outlined above were motivated by several considerations. A major consideration was the need to select those organisms for which maximum benefit would be derived by making potential distribution predictions for both applied and theoretical purposes. An attempt was made to select organisms that had a high economic or ecological impact, presented interesting theoretical problems, and for which existing data were available or could easily be collected under the constraints of funding and time. The selection was made as broad as possible (selection of plants and insects) so that a wide range of issues could be explored.

Invasive alien plants

Invasive alien plants were selected because they fulfilled a number of the criteria outlined above. In particular, they are extremely problematic world-wide (Drake *et al.*, 1989; Pimm *et al.*, 1995; Vitousek *et al.*, 1997) and in southern Africa (Macdonald *et al.*, 1986) from an ecological (Richardson *et al.*, 1989; Higgins *et al.*, 1997 a & b; Le Maitre *et al.*, 1996) and an economic (Van Wilgen *et al.*, 1996; Higgins *et al.*, 1997; Richardson *et al.*, 1997; Van Wilgen *et al.*, 1997) perspective.

In view of the negative impacts that invasive alien plants have on the environment, there is thus considerable practical value in being able to predict their potential distributions from a management perspective (Higgins *et al.*, 1997). Potential distribution models enable managers to get an idea of where a species is likely to occur, where it may invade in the future and where it is likely to be most successful (and by implication, most problematic). These predictions can be translated into better management practices by allocating fewer management resources to where the plant is likely to be least problematic and most resources to where it is likely to be most problematic.

Building predictive distribution models for alien plants raises a number of interesting and important theoretical considerations. In particular, there are several data quality issues associated with alien species that may influence various aspects of the modelling process.

From a data availability point of view, various sources of distribution data for alien plants are available in southern Africa. In particular, the Southern African Plant Invaders Atlas (SAPIA; Henderson, 1998) is a good source of distribution data. In addition, it was relatively easy to conduct further surveys to collect distribution records, as most of the selected species were conspicuous and did not require specialised sampling equipment.

In addition to the value that these techniques have for the control and management of alien plants these techniques hold enormous potential for improving and streamlining a number of processes in biocontrol using insect biocontrol agents (Baars, 2002).

Three invasive alien plants were selected as target species namely, *Lantana camara* L., *Ricinus communis* L. and *Solanum mauritianum* Scop. These species were selected due to a combination of their priority ranking in a prioritisation system (see Appendix), data availability in existing databases, and because they could be identified easily and were unlikely to be mistaken for other species.

Cicadas

A major consideration for selecting cicadas (Homoptera: Cicadidae) as a target group is that they are insects and as a result were expected to present a different set of problems for modelling than plants. In particular, from a data quality point of view, the size of locality record data sets that are available for insects are in general likely to be much smaller than those available for plants. This may be because there are likely to be fewer amateur collectors actively collecting distribution records (with the exception possibly of butterflies), the collection of specimens often requires specialised equipment (e.g. traps) and identification of specimens (especially to species level) often requires a specialist. In addition, distribution records cannot be

assembled by means of visual surveys as is the case for plants, by means of visual road-side surveys (Henderson, 1998).

A set of eight cicada species were selected as target species namely, *Albanycada albigera* Walker, *Capicada decora* Germar, *Platypleura capensis* L., *P. deusta* Thunberg, *P. divisa* Germar, *P. haglundi* Stål, *P. mijburghi* Villet and *Pycna semiclara* Germar.

These particular species were selected because reasonable numbers of distribution records could be easily obtained from existing collections, there is consensus about the taxonomy of these groups, they occupy a wide range of habitats in southern Africa, a number are endemic to the map region (South Africa, Lesotho and Swaziland), and their distributions are moderately well understood.

Scaevola plumieri

Scaevola plumieri (L) Vahl. (= *Scaevola thunbergii* Eckl. & Zeyh.) (Goodeniaceae) is a coastal dune pioneer plant that was selected mainly because previous studies had been done on its eco-physiology (Peter and Ripley, 2000), enabling a mechanistic model to be built to predict its potential distribution (Peter *et al.*, 2002). It was relatively easy to collect good quality locality records for this species so that correlative models could be built for comparison with the mechanistic model (Chapter 8).

References

Austin, M.P. 1998. An ecological prespective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*. 85: 2-17.

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

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

Austin, M.P., Pausas, J.G., Nicholls, A.O. 1996. Patterns of tree species richness in relation to environment in south eastern New South Wales, Australia. *Australian Journal of Ecology*. 21: 154-164.

Baars, J-R, 2002. Biological control initiatives against *Lantana camara* L. (Verbenaceae) in South Africa: an assessment of the present status of the programme, and an evaluation of *Coelocephalapion camarae* Kissinger (Coleoptera: Brentidae) and *Falconia intermedia* (Distant) (Hemiptera: Miridae), two new candidate natural enemies for release on the weed. Ph.D. Thesis, Rhodes University, Grahamstown.

Baker, R.H.A., Sansford, C.E., Jarvis, C.H., Cannon, R.J.C., MacLeod, A., Walters, K.F.A., 2000. The role of climatic mapping in predicting the potential geographical distribution of non-indigenous pests under current and future climates. *Agriculture Ecosystems and Environment*. 82: 57-71.

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

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

Buckland, S.T., Elston, D.A., Beaney, S.J., 1996. Predicting distributional change, with application to bird distributions in northeast Scotland. *Global Ecology and Biogeography Letters*. 5: 66-84.

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

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

Cumming, G.S., 2000 a. Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*. 27: 425-440.

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

Dorrough, J., Ash, J.E., 1999. Using past and present habitat to predict the current distribution and abundance of a rare cryptic lizard, *Delmar impar* (Pygopodidae). *Australian Journal of Ecology*. 24: 614-624.

Drake, J.A., Mooney, H.A., Di Castri, F., Groves, R.H., Kruger, F.J., Rejmanek, M., Williamson, M., 1989. Biological invasions: a global prespective. SCOPE Publications, Paris.

Erasmus, B.F.N., Kshatriya, M., Mansell, M.W., Chown, S.L., Van Jaarsveld, A.S., 2000. A modelling approach to antlion (Neuroptera: Myreleontidae) distribution patterns. *African Entomology*. 8: 157-168.

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*. 19: 474-499.

Franklin, J., 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*. 9: 733-748.

Funk, V.A., Richardson, K.S. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*. 51: 303-316.

Guisan, A., Theurillat, J-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*. 9: 65-74.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

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

Higgins, S.I., Azorin, E.J., Cowling, R.M., Morris. M.J, 1997 a. A dynamic ecological-economic model as a tool for conflict resolution in an invasive-alien-plant, biological control and native-plant scenario. *Ecological Economics*. 22: 141-154.

Higgins, S.I., Turpie, J.K., Costanza, R., Cowling, R.M., Le Maitre, D.C., Marais, C., Midgley, G.F., 1997 b. An ecological economic simulation model of mountain fynbos ecosystems, dynamics, valuation and management. *Ecological Economics*. 22: 155-169.

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

Hirzel, A., 2001. When GIS come to life. Linking landscape and population ecology for large population management modelling: the case of Ibex (*Capra ibex*) in Switzerland. Ph.D. Thesis, Institute of Ecology, Laboratory for Conservation Biology, University of Lausanne.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*. 145: 111-121.

Jackson, S.M., Claridge, A., 1999. Climatic modelling of the distribution of the mahogany glider (*Petaurus gracilis*), and the squirrel glider (*P. norfolcensis*). *Australian Journal of Zoology*. 47: 47-57.

Kadmon, R., Heller, J., 1998. Modelling faunal responses to climatic gradients with GIS: land snails as a case study. *Journal of Biogeography*. 25: 527-539.

Le Maitre, D.C., Van Wilgen, B.W., Chapman, R.A., McKelly, D.H., 1996. Invasive plants and water resources in the Western Cape Province, South Africa: modelling the consequences of a lack of management. *Journal of Applied Ecology*. 33: 1-12.

Lenton, S.M., Fa, J.E., Perez Del Val, J., 2000. A simple non-parametric GIS model for predicting species distribution: endemic birds in Bioko Island, West Africa. *Biodiversity and Conservation*. 9: 869-885.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F., Tanton, M.T., 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri*

(McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*. 18: 371-383.

Lloyd, P., Palmer, A.R., 1998. Abiotic factors as predictors of distribution in southern African Bulbuls. *The Auk*. 115: 404-411.

Macdonald, I.A.W., Kruger, F.J., Ferrar, A.A., 1986. The ecology and management of biological invasions in southern Africa. Oxford University Press, Cape Town.

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

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

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

Nix, H.A., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*, Australian Government Publishing Service, Canberra, pp. 4-15.

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

Palmer, W.A., Willson, B.W., Pullen, K.R., 2000. Introduction, rearing, and host range of *Aerenicopsis championi* Bates (Coleoptera: Cerambycidae) for the biological control of *Lantana camara* L. in Australia. *Biological Control*. 17: 227-233.

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

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

Pearce, J., Lindenmayer, D., 1998. Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) in southeastern Australia. *Restoration Ecology*. 6: 238-243.

Peter, C.I., Ripley, B.S., 2000. An empirical formula for estimating the water use of *Scaevola plumieri*. *South African Journal of Science*. 96: 1-4.

Peter, C.I., Ripley, B.S., Robertson, M.P. 2002. The distribution of *Scaevola plumieri* along the South African coast is limited by seasonal water balance and temperature. *Journal of Vegetation Science*. (in press).

Peterjohn, B.G., 2001. Some considerations on the use of ecological models to predict species' geographic distributions. *The Condor*. 103: 661-663.

Peterson, A.T., 2001. Predicting species' geographic distributions based on ecological niche modeling. *The Condor*. 103: 599-605.

Peterson, A.T., Sanches-Cordero, V., Soberon, J., Bartley, J., Buddemeier, R.W., Navarro-Siguenza, A.G., 2001. Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling*. 144: 21-30.

Peterson, A.T., Soberon, J., Sanches-Cordero, V., 1999. Conservatism of ecological niches in evolutionary time. *Science*. 285: 1265-1267.

Pfab, M.F., Witkowski, E.T.F., 1997. Use of Geographical Information Systems in the search for additional populations, or sites suitable for re-establishment, of the

endangered Northern Province endemic *Euphorbia clivicola*. *South African Journal of Botany*. 63: 351-355.

Pimm, S.L., Russell, G.J., Gittleman, J.L., Brooks, T.M., 1995. The future of Biodiversity. *Science*. 269: 347-350.

Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A., Solomon, A.M., 1992. A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*. 19: 117-134.

Richardson, D.M., Macdonald, I.A.W., Forsyth, G.G., 1989. Reductions in plant species richness under stands of alien trees and shrubs in the fynbos biome. *South African Journal of Forestry*. 149: 1-8.

Richardson, D.M., Macdonald, I.A.W., Hoffmann, J.H., Henderson, L., 1997. Alien plant invasions. In: Cowling, R.M., Richardson, D.M., Pierce, S.M. (Eds.), *Vegetation of southern Africa*, Cambridge University Press, Cambridge, pp. 535-570.

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

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

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

Rogers, D.J., Williams, B.G., 1993. Tsetse distribution in Africa: seeing the wood and the trees. In: Edwards, P.J., May, R. (Eds.), *Large-scale ecology and conservation biology*, Blackwell Scientific Publications, Oxford, pp. 247-271.

Rutherford, M.C., Powrie, L.W., Schulze, R.E., 1999. Climate change in conservation areas of South Africa and its potential impact on floristic composition: a first assessment. *Diversity and Distributions*. 5: 253-262.

Samways, M.J., Osborn, R., Hastings, H., Hattingh, V., 1999. Global climate change and accuracy of prediction of species' geographical ranges: establishment success of introduced ladybirds (Coccinellidae, *Chilocorus* spp.) worldwide. *Journal of Biogeography*. 26: 795-812.

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

Scott, J.K., 1992. Biology and climatic requirements of *Perapion antiquum* (Coleoptera: Apionidae) in southern Africa: implications for the biological control of *Emex* spp. in Australia. *Bulletin of Entomological Research*. 82: 399-406.

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

Skidmore, A.K., Gauld, A., Walker, P., 1996. Classification of kangaroo habitat distribution using three GIS models. *International Journal of Geographical Information Systems*. 10: 441-454.

Skov, F., Borchsenius, F., 1997. Predicting plant species distribution patterns using simple climatic parameters: a case study of Ecuadorian palms. *Ecography*. 20: 347-355.

Soberón, J., Llorente, J., Benítez, H. 1996. An international view of national biological surveys. *Annals of the Missouri Botanical Garden*. 83: 562-573.

Stephenson, N.L., 1998. Actual evapotranspiration and deficit: biologically meaningful correlates of vegetation distribution across spatial scales. *Journal of Biogeography*. 25: 855-870.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science*. 13: 143-158.

Van Wilgen, B.W., Cowling, R.M., Burgers, C.J., 1996. Valuation of ecosystem services: a case study from South African fynbos ecosystems. *BioScience*. 46: 184-189.

Van Wilgen, B.W., Little, P.R., Chapman, R.A., Görgens, A.H.M., Willems, T., Marais, C., 1997. The sustainable development of water resources: History, financial costs, and benefits of alien plant control programmes. *South African Journal of Science*. 93: 404-411.

Vitousek, P.M., D'Antonio, C.M., Loope, L.L., Rejmanek, M., Westbrooks, R., 1997. Introduced species: A significant component of human-caused global change. *New Zealand Journal of Ecology*. 21: 1-16.

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

Walker, P.A., Cocks, K.D., 1991. HABITAT: A procedure for modelling the disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters*. 1: 108-118.

Walton, D.W., Busby, J.R., Woodside, D.P., 1992. Recorded and predicted distribution of the Golden-tipped Bat *Phoniscus papuensis* (Dobson, 1878) in Australia. *Australian Zoologist*. 28: 1-4.

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

II

Input data and data quality

Preface

The intention of this chapter is to introduce some of the important aspects of data quality. This is approached by first describing the sources of input data available for building potential distribution models. Various data quality issues that are addressed in later chapters of the thesis are introduced and discussed.

Abstract

The quality of data used to build potential distribution models directly influences the prediction success of these models. An understanding of data quality issues is thus essential for building potential distribution models as it allows one to assess the fitness of the data for a particular use. The necessary locality records and predictor variables required to build potential distribution models are described. Aspects of data quality relating to these data are then discussed with particular reference to sources of error. The implications of different errors in data are discussed in relation to predictive modelling. Finally some of the important considerations of model evaluation are discussed. Important data quality issues, which are addressed in later chapters of the thesis, are introduced and explained here.

Introduction

The accepted definition for data quality is couched in terms of "fitness for use" (Chrisman, 1991). The quality of data is determined by their relative accuracy and precision, in the context of a specific application or use. Accuracy refers to the closeness of an observation to a true value or one that is accepted as true and precision refers to the level of measurement and exactness of description. Error encompasses both imprecision and inaccuracy. Chrisman (1991) considers error to be a critical component to be used in judging the fitness of data for a particular use. He further suggests that error should be recognised as a fundamental dimension of the data (Chrisman, 1991), although attempts should be made at all times to minimise sources of error within the data.

It is thus necessary to know the quality of the data that will be used to make a prediction so that its fitness for this intended use can be assessed. In particular, there is usually a need to choose among alternative sources of data for model building and model evaluation, which has to be based on fitness-for-use judgements. Researchers have started to recognise the importance of data quality for predicting geographic distributions (Peterson, 2001). During the evaluation phase, the modeller usually wants to know why the model has not performed as well as expected and how to improve the predictions of the model. The quality of the data used to build the model, while not the only component, has a large impact on the quality of the model's predictions (Peterjohn, 2001; Peterson, 2001). The saying "garbage in, garbage out" is particularly relevant here. Errors in the input will propagate through the model to influence the output (Heuvelink *et al.*, 1989; Heuvelink, 1998) and may result in prediction errors (Fielding and Bell, 1997). An awareness of sources of error can thus be considered essential to systematically minimising error or excluding the use of error-prone data in models.

In their review, Fielding and Bell (1997) described prediction errors as *algorithmic* errors or *biotic* errors. Algorithmic errors arise as a result of "limitations imposed by the classification algorithm and the data-gathering process" and biotic errors arise because "not all of the ecologically-relevant processes have been specified in the model." A number of the data quality issues discussed in this chapter will address the issues implicit in the "data-gathering process" of Fielding and Bell (1997),

while issues relating to "the limitations imposed by the classification algorithm" will be addressed elsewhere (Chapter 3). Fielding and Bell (1997) have reviewed the subject of biotic errors, and some of these issues will be revisited here.

To provide a context for discussing data quality, it is necessary to outline the types of input data that are typically used to build predictive models.

Input data sources

A model that predicts the distribution of a target organism relies on maps of variables that are important for the survival of that organism (called predictor variables). A model contains a "description" of the conditions required by the organism for survival and the predictor variable maps are used to show where these conditions are met. The type of model (correlative or mechanistic) describes the way in which that "description" was obtained. A predictor variable map can be conceptualised as a grid of cells (of equal size) covering a map region, with each grid-cell containing a single value of that variable. These predictor variable maps are typically stored in a GIS as geo-referenced raster grids. The terms *grid* and *surface* are both used to describe these maps (Fotheringham *et al.*, 2000).

These are usually maps of environmental data typically consisting of climatic data (e.g. temperature, rainfall, humidity, frost) and physical data (e.g. elevation, aspect, slope). Climatic variable maps are generally derived from interpolations of point data (Busby, 1991; Hutchinson *et al.*, 1995; Schulze *et al.*, 1997). Predictor variable data may also consist of remotely sensed satellite data (Flather and King, 1992; Rogers and Williams, 1993; Rogers *et al.*, 1996; Packer *et al.*, 1999). Plummer (2000) discusses some of the climate variables that can be derived from remotely sensed satellite data.

In addition to environmental predictor variable maps, correlative models rely on distribution records (locality records) to make predictions. Locality data typically consist of presence/absence data or presence-only data, which are either collected by means of field surveys or are obtained from existing collections (Chapter 1). Data quality considerations differ between these two sources of locality data. The subject of locality data quality examines positional accuracy and precision, and attribute accuracy of locality data. Issues such as sample size and sampling bias are discussed

under the heading of sampling. A section is devoted to the data quality considerations of data obtained from collections (museum and herbarium) that were collected on an *ad hoc* basis.

Pre-analytical data reduction of predictor variable data

Various pre-analytical data reduction techniques have been used to reduce large volumes of predictor variables into fewer dimensions. Data reduction removes redundancy from multicolinear datasets, which serves to considerably reduce computing time. In addition, when predictor variables are highly correlated the model may agree closely with the observations but give poor predictions when extrapolated to unsurveyed sites (Buckland and Elston, 1993). Principal Components Analysis (PCA: James and McCulloch, 1990) has been quite a popular pre-analytical data reduction technique used in distribution modelling (Osborne and Tigar, 1992; Buckland and Elston, 1993; Robinson *et al.*, 1997; Guisan *et al.*, 1998; Robertson *et al.*, 2001). An alternative data reduction technique is that of Fourier Analysis which has been used to reduce the dimensionality of satellite-derived data for predicting tsetse fly distribution (Rogers *et al.*, 1996).

Quality of predictor variable data

Climatic predictor variable maps are generally interpolated in relation to altitude, which is generally represented in the form of a digital elevation model (DEM: Lindenmayer *et al.*, 1991). A number of complex topographic variables can be derived from DEMs (reviewed by Franklin, 1995), the quality of which will be affected by the quality of the DEM used. The quality of these maps is thus influenced by the quality of the digital elevation model used in the interpolation (Dent *et al.*, 1989; Lennon and Turner, 1995). Errors in one surface may lead to errors in another, which is known as error propagation (Heuvelink *et al.*, 1989; Heuvelink and Burrough, 1993). Errors in DEMs are amplified when algorithms are applied to them to derive other topographic variables such as slope and aspect variables (Franklin, 1995). The quality of interpolated climatic surfaces will be dependent on the number and distribution of meteorological stations for which data are available and the

accuracy of the point data recorded at these stations. Busby (1991) points out that frequent errors in surfaces can be expected in areas where steep or complex climatic gradients are poorly sampled by the meteorological network. Similarly, Plummer (2000) suggested that interpolated data inevitably contain artefacts that are a function of the spatial distribution of measurements. Interpolated climatic surfaces may have limitations as they will probably not account for biologically relevant microclimates (Guisan and Zimmermann, 2000). Perhaps additional metadata maps should be included to indicate the network of points used in the interpolation as well as some sort of error surface indicating where errors in the interpolated surface are likely to be greatest. These metadata would greatly assist in assessing the fitness-for-use of particular regions of interpolated surfaces.

In the case of remotely sensed images, a number of image processing procedures usually have to be carried out following image acquisition (Lillesand and Kiefer, 1994). These procedures include image rectification and restoration; image enhancement; and image classification (Lillesand and Kiefer, 1994). One of the major advantages of remotely sensed data is that data values in each cell of the image are obtained by direct measurement, while values of surfaces obtained by interpolation from point measurements represent estimates rather than direct measurements. Remotely sensed data are not limited by a meteorological network and hence do not experience the problems faced by interpolated data associated with steep or complex climatic gradients. The choice of data will be dependent on the organism, the requirements of the model and constraints such as data availability, data resolution and cost.

Resolution

Guisan and Zimmermann (2000) have suggested that the modelling process involves formulating a conceptual model that leads to the choice of an appropriate spatial scale for conducting the study and to the selection of an appropriate set of predictor variables for the model. The selection of scale will largely depend on the scale at which phenomena relating to the distribution of the organism are perceived to operate (operational scale) and the overall goal of the study.

Scale has several meanings. Scale can mean the spatial extent, domain or the map region of the study area (Bian, 1997). It can also mean *resolution*, which is equivalent to *sampling interval* (Bian, 1997) in ecology. Bian (1997) defines spatial resolution as "the size of the smallest distinguishable part of a spatial dataset". In raster-based geographical information systems, resolution refers to the size of the grid-cells of a raster grid. The smaller the size of the grid-cell, the higher the spatial resolution of the grid or surface.

The relationship between resolution and operation scale is such that only those processes that operate at scales larger than the resolution of the grid can be revealed (Bian, 1997). As a result it is necessary to ensure that the resolution of the predictor variables is higher than that of the operational scale of factors affecting the distribution of the target organism if these factors are to be considered in the model. Much attention has been given to issues of scale in biology (Turner, 1989; Wiens, 1989; Levin, 1992). Issues of spatial scale in the context of modelling in a GIS have also been addressed (e.g. Heuvelink 1998; Collingham *et al.*, 2000). Pearce *et al.*, (2001) have suggested that research should be undertaken to determine the most appropriate spatial resolution for modelling individual species distribution at the spatial scale of relevance to the life history of the target species. The study by Collingham *et al.* (2000) is an example of the type of approach that is likely to be able to address these issues.

The choice of scale at which the model is finally implemented is often dependent on, and limited by, the resolution of the available predictor variable and the accuracy and precision of the locality data (e.g. Collingham *et al.*, 2000). The type of environment in which the study is conducted may also influence the resolution of the predictor variables selected. For example in mountainous areas a high spatial resolution is required to obtain reliable results, due to rugged topography and consequent steep environmental gradients (Guisan *et al.*, 1998). The resolution at which potential distribution maps are produced may be determined at least in part by the resolution that is convenient and suitable for the purposes of ground-truthing the predictions or the scale at which planning decisions are made e.g. conservation planning (Pearce *et al.*, 2001).

Although predictor variables recorded at high and low resolution may be used to make predictions at low resolution, only high-resolution predictor variables can be

used to make meaningful predictions at high resolution (Buckland and Elston, 1993). Similarly, the distribution of a species cannot meaningfully be modelled at a higher resolution than the resolution of the original presence/absence distribution data (Buckland and Elston, 1993). As a result, the spatial resolution of a model should always be equal to, or lower than that of the predictor variables or the presence/absence survey data. Buckland and Elston (1993) describe modelling methods for handling variables recorded at different resolutions. In order to build accurate, high-resolution distribution models the input data (both locality data and predictor variables) must be recorded at high precision and have a high level of accuracy. Guisan and Zimmermann (2000) highlighted the need for predictor variable maps with higher resolution and accuracy in order to improve model predictions. Although high-resolution predictor variable maps can be acquired, these data have to be matched with locality data of the same resolution and accuracy in order to make reliable predictions.

Quality of locality data

Various aspects of data quality are discussed which pertain both to data obtained by means of systematic field surveys and to data obtained from collections, such as museum and herbarium collections, where data are collected on an opportunistic or *ad hoc* basis. Since there are a number of problems associated specifically with data obtained from collections a section is devoted to data quality problems associated with collections data and a section on using this data in predictive modelling.

Sampling bias

Stockwell and Peters (1999) define sampling bias as any departure of the data from a random sample of the possible data points. Bias can refer to geographical space or environmental space but this distinction is generally not made.

Correlative distribution models make the assumption that the sample of localities used to build the models is representative of the full range of environments in which the species can occur. If bias (environmental) occurs in the sample then certain environmental conditions will be over- or under-represented in the sample.

The reliability of distribution predictions will be influenced by bias in the data (Margules and Pressey, 2000).

Geographical bias in datasets is fairly easy to detect by looking at maps that plot collection or survey localities, an example is provided by Austin (1998) for Elapid snakes in Australia. Plots of sampled points displayed in two-dimensional environmental space (e.g. Funk and Richardson, 2002) can assist in detecting environmental bias by detecting regions of the environmental space that have not been sampled.

Geographical bias in datasets is likely to result in, or at least indicate bias in environmental space. Funk and Richardson (2002) recently demonstrated that regions in a two dimensional environmental space, which had not been sampled, mapped into distinct regions in geographical space. This illustrates the link between geographical space and environmental space.

Representative sampling of species occurrence in environmental space can be achieved by designing and conducting a systematic field survey (Margules and Austin, 1994; Austin, 1998) using techniques such as the gradsect approach (Gillison and Brewer, 1985) or modifications this approach (Austin and Heyligers, 1989). These approaches have been designed primarily for biodiversity surveys, for a comparison of sampling strategies for predictive modelling of single species, see Hirzel and Guisan (2002). Systematic surveys such as these are expensive and time-consuming to conduct. The alternative is to use data from museum or herbarium collections (collections data), which has usually been collected opportunistically or by means of *ad hoc* surveys. Datasets obtained from these sources have several problems (outlined below), one of these is bias.

Data quality problems with collections data

Data that have been collected on an opportunistic or *ad hoc* basis (collections data) suffers from several weaknesses. Generally only the presence and not the absence of the species is recorded (Margules and Austin, 1994; Austin, 1998; Stockwell and Peters, 1999; Zaniewski *et al.* 2002), however the majority of predictive modelling techniques rely on both presence and absence data. Several authors have reported geographical bias to be a problem in samples of records

obtained from collections (Margules and Austin, 1994; Soberón *et al.*, 1996; Austin, 1998; Freitag *et al.*, 1998; Lawes and Piper, 1998; Funk and Richardson, 2002; Ferrier, 2002; Zaniewski *et al.* 2002). This means that there is usually more sampling effort areas that are easily accessible e.g. along road networks, near cities. Geographical bias occurs because collectors return to known sites, and tend to stay near roads and settlements (Osborne and Tigar, 1992; Rich and Woodruff, 1992; Austin, 1998; Freitag *et al.*, 1998).

Ferrier and Watson (1997) found a bias towards rarer species in presence-only datasets. They suggest that the reason for this is that observers tend to record rare and or more interesting species more often than common species. Funk and Richardson (2002) suggested that collections data were often taxonomically incomplete, meaning that researchers tend to concentrate on taxa that are easy to study at the expense of those that are not. They further suggest that collections data may be temporally biased, based on one survey that was not carried out in the most appropriate season. Margules and Austin (1994) also mention the problem of taxonomic bias in these datasets. Further problems with presence-only data includes, imprecision in recording localities (Austin 1998), the presence of other species and of environmental variables is inconsistently recorded (Austin 1998), the plot size is unknown, there is uncertainty as to the precision of species identification, effects of habitat disturbance, species competition and species dispersal rates (Zaniewski *et al.*, 2002).

Using collections data in predictive modelling

Various data quality problems have been outlined with regard to collections data, which makes it potentially more problematic for use in predictive modelling.
The first problem (outlined above) is that these data are mostly presence-only data. Several modelling techniques have been designed specifically to make predictions using presence–only data. One of the earliest techniques that was developed specifically for this purpose is BIOCLIM (Nix, 1986; Busby, 1991), subsequently several other, more sophisticated, techniques have been developed (reviewed in Chapter 3). One of the arguments against using presence-only data is that bias is likely to be a problem in these datasets, and that the extent of this bias is usually

unknown. However, presence-only data are often the only source of data available as systematic field surveys are costly and time-consuming to conduct. In addition, there are vast sources of presence-only data in collections (Soberón *et al.*, 1996) that are potentially valuable. Funk and Richardson (2002) suggest that collections data should be used in conservation planning, despite its limitations. Integrated spatial analysis systems that make use of collections data (presence-only) for predicting distributions are making these data and modelling techniques accessible to a large number of users through the World Wide Web (Kaiser, 1999; Stockwell and Peters, 1999). This means that presence-only data are easily available to a large number of users (who are not necessarily modellers) for the purposes of predictive modelling, despite its limitations.

While bias in presence-only datasets is likely to reduce the accuracy and value of predictions that make use of it, the extent of these problems is currently unknown, and requires further investigation. The results of predictive modelling studies that are based on presence-only data are unlikely to perform as well as those based on presence-absence data, as better biological survey data will produce better models (Ferrier and Watson, 1997). The success of presence-only models will be related to the degree of bias in samples and the sensitivity of these modelling techniques to this bias. There is an urgent need for the development of methods for detecting the degree of bias in samples, so that these problems can be assessed. The approaches of Funk and Richardson (2002) show some promise, although require further development.
I suggest that presence-only data has the potential to be very useful, although it may have several problems which are likely to reduce the usefulness of predictions made from it.

Absence data

Good quality absence data are generally considered to be more difficult to obtain than good quality presence data. When a survey is conducted to collect absence data for a given species then each of the grid-cells in which the species is suspected to be absent have to be fully surveyed to ensure that these cells represent true absences as opposed to undetected presence. Absence is thus conditional on the sampling effort made at a site (Austin, 1998).

False negatives may be recorded when a target organism has not yet realised the full extent of its potential distribution. For example, introduced species may take time to realise the full extent of their potential distributions (Groves, 1992; Wilson *et al.*, 1992; Hirzel *et al.*, 2001), making it very difficult to collect good quality absence records for these species. In general, presence records are considered to be more reliable than absence records (Fielding and Bell, 1997). Although suggestions have been made that false absence records influence model performance, this needs to be confirmed quantitatively. The aim of Chapter 6 is to address this issue and to assess the impact of prevalence and sample size on the performance of logistic regression models. Another important question is to establish whether profile techniques can perform as well as group discrimination techniques. Chapter 7 attempts to address this question by quantitatively comparing the performance of some profile and group discrimination techniques (see also Hirzel *et al.*, 2001).

Positional accuracy

Positional accuracy refers to the accuracy of an observation or feature on the earth's surface. In the case of locality data, a locality record typically consists of the geographic position of a point (recorded as a co-ordinate) at which an observation (e.g. present, absent, abundant) relating to a target organism was made. In this context positional accuracy refers to the closeness of the point's recorded position to its actual position on the earth's surface.

For any locality, values of a number of predictor variables can be obtained that correspond with the position of that locality in the grid. The position of the locality determines to which grid-cell the locality is linked and as a result the values that are returned for each of the variables of interest. The positional accuracy of locality records will thus influence the values that are associated with those records in the model. Positional accuracy of locality records is particularly important in the case of climatic predictor variables where climatic gradients are steep (Nix, 1986). For example, in the Andes mountains of Equador, Skov and Borchsenius (1997) claim that a 2 km horizontal error can cause a vertical displacement of up to 1000 m, which would seriously affect the estimates of climatic attributes of a given locality. With the advent of Geographic Positioning Systems (GPS) positional accuracy of locality

records is likely to increase in the future, although historically inaccurate records will remain a problem.

Positional precision

Positional precision refers to the number of significant digits to which the geographic position of an observation is recorded or measured. The geographic position of a locality that has been measured to the nearest second has been recorded at a higher precision than one that was measured to the nearest quarter degree square. While the position of a given point on the earth may have been recorded at high precision (e.g. to the nearest second) it may not necessarily be accurate.

It is possible to scale presence locality records from a higher to a lower precision, but this is not valid for absence localities. For example, an organism that is recorded as being present in a 1-minute grid square will also be present in a quarter-degree or one degree grid square in which that 1-minute square is nested. In contrast, if the organism is recorded absent in the smallest nested grid square (e.g. 1 minute) it is not necessarily absent in either of the two larger grid squares. Absence data are therefore specific to the precision at which they were recorded.

Presence locality data can thus be scaled up from a higher precision to lower precision whereas absence locality data cannot be scaled in this way. Neither presence nor absence locality records should be scaled from low precision to higher precision. As a result, the distribution of a target organism should only be predicted at the same or a lower precision as the precision at which its locality records were originally recorded, even if predictor variable data are available at a higher resolution. The precision of locality records will influence the final resolution at which distribution maps can be produced. Nix (1986) maintains that locality records collected during surveys should always be recorded at maximum precision so that these records do not limit the resolution of possible future distribution maps. This may be true for presence data although, as discussed above, absence data should not be scaled from a higher to a lower precision. A similar problem may exist for certain presence data, although the problem is likely to be less frequent than for absence data. Lawes and Piper (1998) illustrated this problem, which they suggest should be called the "oasis effect". The argument presented is as follows: if a grid-cell that contains a

desert oasis with given species in it, is made larger, it becomes increasingly likely that incongruous associations will arise.

Attribute accuracy

Attribute accuracy refers to the accuracy of attribute data at a given location (Chrisman, 1991). In the context of potential distribution models, attribute data for a given locality usually consist of whether the target organism was recorded present or absent at that locality. Two types of error are generally associated with this sort of attribute data: false positives and false negatives (Fielding and Bell, 1997), which are conceptually equivalent to Type I and Type II errors respectively. While these terms are generally used to describe prediction errors (Fielding and Bell, 1997), they are used here in the context of data quality to describe errors in locality data.

A false positive (FP) error is made when an organism which is actually absent is recorded as being present, and a false negative (FN) error occurs when a organism which is present is recorded as being absent at a particular locality. There are a number of potential sources of false negative and false positive errors in any dataset. The number of errors in a given dataset is probably linked to the source of the data.

Nix (1986) points out that errors arise through misidentification and or incorrect labelling of specimens. This is largely applicable to locality data obtained from museum or herbarium specimen collections. Misidentification errors are likely to be more frequent in datasets which have been assembled by a number of amateur observers on an *ad hoc* basis, than by one expert observer surveying to gather data for a specific model. Both types of error can arise through misidentification or incorrect labelling. FN errors occur when individuals of the target species are mistaken for members of a different species, with the result that the target species is recorded as being absent at that locality when it is actually present. FP errors occur when individuals of a different species are mistaken for member of the target species. Errors also arise when a species complex is mistaken for a single species, when there is taxonomic uncertainty within a group (Nix 1986) or when taxonomic revisions have taken place since the original data were collected.

High quality locality data for a rare or cryptic species are likely to be limited because sampling requires considerably more effort and the prevalence of FN errors is

likely to greater than it would be for abundant or conspicuous species (Peterjohn, 2001). Certain features of the biology of organisms may make them more conspicuous, including: breeding plumage in birds (Lawes and Piper, 1998); calling of birds and insects; and brightly coloured fruits or flowers of plants. In addition, the likelihood of committing FN errors for conspicuous species is also lower than for cryptic species (Dorrough and Ash, 1999) because the chances of a conspicuous species being present but undetected is much lower than for a cryptic species. Features of the biology of organisms, such as hibernation or migration, may increase the chances of FN errors in locality datasets.

Locality records taken from specimens are likely to be more reliable than those obtained simply by observation, since the chances of misidentification are reduced when a specimen is available for (re) examination. Peterjohn (2001) mentions the problem of changes in identification skills of observers in relation to data collected for the North American Breeding Bird Survey. Specimen collection for sessile species (e.g. plants) is generally easier than for highly mobile species (e.g. small mammals, birds or insects) since specialised equipment (e.g. traps or nets) is generally not required.

Inter-annual species range expansion or contraction may occur due to such factors as resource fluctuations or disturbance mediated by certain events e.g. El Niño climate shifts (Hayward, 1997), making FN errors more likely. The area of occupancy and abundance of species is reported to vary with time (Gaston *et al.*, 2000), which will affect the accuracy of locality attribute data, especially if this has been collected over a long period of time (Lawes and Piper, 1998; Peterjohn, 2001). The effects of climate change may cause similar problems in the future and may already be having an influence.

Sample size

Sample size refers to the number of locality records in the sample. A sufficient number of records must be collected so that the sample is representative of the full range of environments in which the species can occur. It is usually unclear how large a sample should be in order to be representative. This depends in part on how much variation has to be captured and how many predictor variables are used. Various

studies have investigated the effect of sample size on model performance (Manel *et al.*, 1999 a & b; Cumming, 2000; Pearce and Ferrier, 2000; Hirzel and Guisan, 2002), although most of these have concentrated on a fairly narrow range of sample sizes. In Chapter 6 the effect of a large range of sample sizes on model performance is investigated.

Conclusion

The interpretation of the model outputs and the application of model predictions should be done with care and should consider the quality of the input data used to calibrate the model. Consideration should also be given to the quality of the data used to evaluate the model as this will influence the apparent error rate, which may influence the conclusions drawn about the performance of the model. From this review, it should be clear how data quality (fitness-for-use) affects the whole modelling process.

References

Austin, M.P. 1998. An ecological prespective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*. 85: 2-17.

Austin, M.P., Heyligers, P.C. 1989. Vegetation survey design for conservation: gradsect sampling of forests in north-eastern New South Wales. *Biological Conservation*. 50: 13-32.

Bian, L., 1997. Multiscale nature of spatial data in scaling up environmental models. In: Quattrrochi, D.A., Goodchild, M.F. (Eds.), *Scale in Remote Sensing and GIS*, Lewis Publishers, New York, pp. 13-26.

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

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

Chrisman, N.R., 1991. The error component in spatial data. In: Maguire, D.J., Goodchild, M.F., Rhind, D.W. (Eds.), *Geographical Information Systems: principles and applications*, John Wiley and sons, New York, pp. 165-174.

Collingham, Y.C., Wadsworth, R.A., Huntley, B., Hulme, P.E., 2000. Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology*. 37: 13-27.

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

Dent, M.C., Lynch, S.D., Schulze, R.E., 1989. Mapping mean annual and other rainfall statistics in southern Africa. Department of Agricultural Engineering, Pietermaritzburg, pp. 250.

Dorrough, J., Ash, J.E., 1999. Using past and present habitat to predict the current distribution and abundance of a rare cryptic lizard, *Delmar impar* (Pygopodidae). *Australian Journal of Ecology*. 24: 614-624.

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*. 51: 331-363.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra, p. 193.

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

Flather, C.H., King, R.M., 1992. Evaluating performance of regional wildlife habitat models: implications to resource planning. *Journal of Environmental Management*. 34: 31-46.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2000. Quantitative geography: perspectives on spatial data analysis. Sage Publications, London, pp. 270.

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*. 19: 474-499.

Freitag, S., Hobson, C., Biggs, H.C., Van Jaarsveld, A.S., 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*. 1: 119-127.

Funk, V.A., Richardson, K.S. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*. 51: 303-316.

Gaston, K.J., Blackburn, T.M., Greenwood, J.J.D., Gregory, R.D., Quinn, R.M., Lawton, J.H., 2000. Abundance-occupancy relationships. *Journal of Applied Ecology*. 37: 39-59.

Gillison, A.N., Brewer, K.R.W. 1985. The use of gradient directed transects or gradsects in natural resource survey. *Journal of Environmental Management*. 20: 103-127.

Groves, R.H., 1992. Weed ecology, biology and spread. Proceedings of the First International Weed Control Congress, Melbourne, pp. 83-88.

Guisan, A., Theurillat, J-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*. 9: 65-74.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

Hayward, T.L., 1997. Pacific ocean climate change: atmospheric forcing, ocean circulation and ecosystem response. *Trends in Ecology and Evolution*. 12: 150-154.

Heuvelink, G.B.M., 1998. Uncertainty analysis in environmental modelling under a change of spatial scale. *Nutrient Cycling in Agroecosystems*. 50: 255-264.

Heuvelink, G.B.M., Burrough, P.A., 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*. 7: 231-246.

Heuvelink, G.B.M., Burrough, P.A., Stein, A., 1989. Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*. 3: 303-322.

Hirzel, A., Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*. 157: 331-341.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*. 145: 111-121.

Hutchinson, M.F., Nix, H.A., McMahon, J.P., Ord, K.D., 1995. A topographic and climatic database. Centre for Resource and Environmental Studies, The Australian National University, Canberra.

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

Kaiser, J., 1999. Searching museums from your desktop. *Science*. 284: 888.

Lawes, M.J., Piper, S.E., 1998. There is less to binary maps than meets the eye: the use of species distribution data in the southern African sub-region. *South African Journal of Science*. 94: 207-210.

Leathwick, J.R. 1995. Climatic relationships of some New Zealand forest tree species *Journal of Vegetation Science*. 6: 237-248.

Leathwick, J.R. 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*. 9:719-732.

Leathwick, J.R., Austin, M.P. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*. 82: 2560-2573.

Leathwick, J.R., Whitehead, D. 2001. Soil and atmospheric water deficits and the distribution of New Zealand's indigenous tree species. *Functional Ecology*. 15: 233-242.

Leathwick, J.R., Whitehead, D. McLeod, M. 1996. Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach. *Environmental Software*. 11:81-90.

Leathwick, J.R., Burns, B.R., Clarkson, B.D. 1998. Environmental correlates of tree alpha-diversity in New Zealand primary forests. *Ecography*. 21: 235-346.

Leathwick, J.R., Mitchell, N.D. 1992. Forest pattern, climate and vulcanism in central North Island, New Zealand. *Journal of Vegetation Science*. 3: 603-616.

Lennon, J.J., Turner, J.R.G., 1995. Predicting the spatial distribution of climate: temperature in Great Britain. *Journal of Animal Ecology*. 64: 370-392.

Levin, S.A., 1992. The problem of pattern and scale in ecology. *Ecology*. 73: 1943-1967.

Lillesand, T.M., Kiefer, R.W., 1994. Remote sensing and image interpretation. John Wiley & Sons, New York, pp. 750.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F., Tanton, M.T., 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*. 18: 371-383.

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

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

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

Margules, C.R., Pressey, R.L. 2000. Systematic conservation planning. *Nature*. 405:243-253.

Nix, H.A., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*, Australian Government Publishing Service, Canberra, pp. 4-15.

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

Packer, M.J., Canney, S.M., McWilliam, N.C., Abdallah, R., 1999. Ecological mapping of a semi-arid savanna. In: Coe, M.J., McWilliam, N.C., Stone, G.N., Packer, M.J. (Eds.). Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna, Royal Geographical Society (with The Institute of British Geographers), London, pp. 43-68.

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

Pearce, J., Ferrier, S., 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*. 128: 127-147.

Peterjohn, B.G., 2001. Some considerations on the use of ecological models to predict species' geographic distributions. *The Condor*. 103: 661-663.

Peterson, A.T., 2001. Predicting species' geographic distributions based on ecological niche modeling. *The Condor*. 103: 599-605.

Plummer, S.E., 2000. Perspectives on combining ecological process models and remotely sensed data. *Ecological Modelling*. 129: 169-186.

Rich, T.C.G., Woodruff, E.R., 1992. Recording bias in botanical surveys. *Watsonia*. 19: 73-95.

Robertson, M.P., Caithness, N., Villet, M.H., 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*. 7: 15-27.

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

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

Rogers, D.J., Williams, B.G., 1993. Tsetse distribution in Africa: seeing the wood and the trees. In: Edwards, P.J., May, R. (Eds.), *Large-scale ecology and conservation biology*, Blackwell Scientific Publications, Oxford, pp. 247-271.

Schulze, R.E., Maharaj, M., Lynch, S.D., Howe, B.J., Melvil-Thomson, B., 1997. South African Atlas of agrohydrology and climatology. Water Research Commission, Pretoria.

Skov, F., Borchsenius, F., 1997. Predicting plant species distribution patterns using simple climatic parameters: a case study of Ecuadorian palms. *Ecography*. 20: 347-355.

Soberón, J., Llorente, J., Benítez, H. 1996. An international view of national biological surveys. *Annals of the Missouri Botanical Garden*. 83: 562-573.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science*. 13: 143-158.

Turner, M.G., 1989. Landscape ecology: the effect of pattern on process. *Annual Review of Ecology and Systematics*. 20: 171-197.

Wiens, J.A., 1989. Spatial scaling in ecology. *Functional Ecology*. 3: 385-397.

Wilson, J.B., Rapson, G.L., Sykes, M.T., Watkins, A.J., Williams, P.A., 1992. Distributions and climatic correlations of some exotic species along roadsides in South Island, New Zealand. *Journal of Biogeography*. 19: 183-193.

Zaniewski, A.E., Lehmann, A., Overton, J. McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*. 157: 261-280.

III

A review of correlative modelling techniques for predicting species' potential distributions

Preface

This chapter is intended to provide a context for the profile techniques that are described and implemented in chapters 4 and 5, and to complement the review of predictive modelling techniques by Guisan and Zimmermann (2000). The focus of this chapter is thus on profile techniques, but it also describes certain group discrimination techniques not reviewed by Guisan and Zimmermann (2000), and gives further examples of those techniques that they did not review. Another important aspect of the modelling process, model evaluation, is also discussed.

Abstract

The use of models to predict species' potential geographical distributions has increased dramatically. The majority of these techniques are based on correlations between species locality (distribution) records and environmental predictor variables (correlative techniques). Group discrimination correlative techniques, those that use both presence and absence locality data to make predictions, appear to have been more popular than profile correlative techniques, those that use only presence locality data. As a result, previous reviews have tended to concentrate on group discrimination techniques. Several new profile techniques have recently been described and implemented, indicating the need for a review of these. The emphasis of this review is thus on profile techniques with some coverage being given to group discrimination techniques. The profile techniques reviewed include three envelope techniques, one similarity metric technique four PCA-based techniques and Contingency Table Analysis. Further examples of the application of selected group discrimination techniques reviewed elsewhere (including Discriminant Function

Analysis, Classification and Regression Trees, and Artificial Neural Networks), are given. Genetic Algorithms and logistic regression are also described. The selection of an appropriate modelling technique is dependent on a number of factors. Comparative studies assessing model performance for a range of organisms and under various data quality conditions are extremely valuable for selecting appropriate modelling techniques for particular situations.

Introduction

Models that predict species' potential distributions have become increasingly popular over the past few years for addressing a number of biogeographical questions (Chapter 1). Franklin (1995), and more recently, Guisan and Zimmermann (2000), reviewed the techniques available for making these predictions. These reviews have largely concentrated on group-discrimination correlative techniques (Chapter 1), which use both presence and absence locality data to make predictions, rather than on profile correlative techniques, those that use only presence locality data to make predictions. This is probably in part because group discrimination techniques have been used fairly widely (e.g. Walker, 1990; Osborne and Tigar, 1992; Rogers and Randolph, 1993; Rogers and Williams, 1993; Lees, 1994; Michaelsen *et al.*, 1994; Williams *et al.*, 1994; Rogers *et al.*, 1996; Higgins *et al.*, 1999; Cumming, 2000 a & b) and profile techniques e.g. BIOCLIM (Nix, 1986; Busby, 1991) and DOMAIN (Carpenter *et al.*, 1993) appear to have been less popular. A possible explanation may be that the multivariate statistical nature of many group discrimination techniques (e.g. Logistic Regression and Discriminant Analysis) has been considered superior to the simpler, often non-statistical profile techniques (e.g. BIOCLIM), in terms of performance (Ferrier and Watson, 1997). Concerns about the quality and availability of absence data (discussed in Chapter 2) for certain organisms has recently prompted the development of new profile techniques (e.g. Erasmus *et al.*, 2000; Hirzel, 2001; Robertson *et al.*, 2001) that have a multivariate statistical basis.

As a result of these developments, this review focuses largely on correlative profile modelling techniques. Although some consideration is also given to group-discrimination techniques, a number of these techniques were recently reviewed elsewhere (Guisan and Zimmermann, 2000). Further examples of the application of

the reviewed techniques and descriptions of techniques that have not previously been reviewed, are provided here.

In the same way that the quality of the data used to build a predictive model has an impact on the quality of its predictions (Chapter 2), so does the choice of modelling technique (algorithm). Just as data quality can be judged on fitness-for-use criteria, the same approach can be applied to the choice of modelling technique. The only way in which these judgements can be made is by understanding how a particular technique makes predictions and the resultant strengths and weaknesses of the approach.

This review is concerned only with models that are used to predict the potential distribution of a single taxon, rather than models such as those used for predictive vegetation mapping, that predict vegetation composition across a landscape (Franklin, 1995). The classification of techniques used in this review differs slightly from that of Guisan and Zimmermann (2000).

Ecological Niche theory

Most predictive models make theoretical assumptions, many of which are not stated explicitly by the authors. In view of this, it is necessary to examine the theoretical framework on which most predictive models are implicitly built. The importance of this ecological theory has also been acknowledged elsewhere (Franklin, 1995; Austin, 1999; Guisan and Zimmermann, 2000; Austin, 2002).

The spatial distribution of a species is likely to be limited by the distribution of the resources and conditions needed for its existence (Woodward, 1987; Woodward and Williams, 1987). In this century this idea has been expressed in the concept of the ecological niche, which evolved from Grinnel's qualitative description of the place a species occupies in its environment to Hutchinson's quantitative formulation which emphasises the requirements of the organism itself (Schoener, 1990).

Hutchinson's model is composed of an abstract set of axes, each of which represents a resource or condition of importance to the organism. On each axis there will be a range of values within which the species can survive. These ranges essentially describe the physiological tolerances of the species. Plotting three of these axes together yields an abstract three-dimensional space within which the organism

can survive. This space can be generalised mathematically to include as many axes as necessary to completely characterise the species' needs, resulting in an *n-dimensional hyperspace* that is termed the fundamental niche (Schoener, 1990). A refinement of Hutchinson's niche concept; the utilisation distribution niche concept, defines the niche for a particular species population in terms of fractional resource use (Schoener, 1990). These measures of fractional resource use (essentially frequency histograms of resource use) are arranged along one or more dimensions called niche axes (Schoener, 1990). A variety of niche axes, classified by food, space and time, can be incorporated into this scheme. Animal ecologists use *Resource Selection Functions* (RSF) to characterise the selection of resources by animals and these can be used for prediction (Boyce and McDonald, 1999; Boyce *et al.* 2002).

The niche models that are built are never exhaustive, and therefore only approximate Hutchinson's ideal (Schoener, 1990).

Few organisms occupy the whole of their fundamental niche because they may be excluded from parts of it by competition or predation (Begon *et al.*, 1990). The reduced hypervolume in which the organism can survive is termed its realised niche (Schoener, 1990; Begon *et al.*, 1990).

Guisan and Zimmermann (2000) highlighted the importance of distinguishing between models that predict the fundamental niche from those that predict the realised niche of the target organism. Correlative models use actual distribution records to make predictions and these must therefore be drawn from the realised niche of that organism (Malanson *et al.*, 1992; Wilson *et al.*, 1992; Bongers *et al.*, 1999). Thus, although biotic interactions are not explicitly accounted for (Robertson *et al.*, 2001) their influence will be accounted for by sampling the realised niche and the result is a prediction of the realised niche (Malanson *et al.*, 1992; Huntley *et al.* 1995; Franklin, 1995; Guisan and Zimmermann, 2000). In contrast, mechanistic models that are based only on physiological constraints (Peter *et al.*, 2002) and do not explicitly account for biotic interactions tend to predict the fundamental niche of the target organism (Guisan and Zimmermann, 2000). These can be refined to model the realised niche by adding simple rules to account for biotic interactions (see Prentice *et al.*, 1992).

Shape of realised niche responses

A common assumption of niche theory is that of a bell-shaped (Gaussian) species response to a resource gradient (Austin, 1999). However, there is little evidence for symmetric bell-shaped responses (Austin, 1999) and bell-shaped response curves are certainly not universal (Austin *et al.*, 1984; Austin, 1987). Several studies have found skewed responses (Austin *et al.*, 1984; Austin, 1987; Austin *et al.*, 1990; Leathwick and Mitchell, 1992; Austin *et al.*, 1994; Bio *et al.*, 1998; Bongers *et al.*, 1999; Ejrnæs, 2000). However, several modelling techniques rely on the assumption that species responses to environmental variables are bell-shaped (reviewed in this chapter), one reason for this is because gaussian curves are easy to handle statistically (Bio *et al.*, 1998).

Profile techniques

Envelope techniques

The simplest of the profile techniques and some of the earliest models (Chicoine *et al.*, 1985) were range-based models that are referred to here as *envelope* techniques. Envelope techniques are conceptually simple and easy to apply, and are generally based on very few, if any statistical assumptions.

Simple envelope models

Simple envelope models are constructed as follows. A training dataset is produced for the target organism by deriving values from each of a set of predictor variable maps (corresponding with each of the axes in Hutchinson's niche model) using a set of locality records. These locality records represent sites where the target species have been found to be present. For each predictor variable map, the minimum and maximum values are calculated from the training set, which are assumed to represent the realised niche limits of the target organism. For a given map, those grid-cells whose values fall within the upper and lower extremes (the minimum and maximum values respectively) of the target species are taken to represent areas where

the species can survive. The predictor variable maps are reclassified into new maps indicating regions of predicted presence (coded 1) and absence (coded 0) of the target organism. These new maps are superimposed using the Boolean AND function (Burrough, 1989), which is consistent with set theory where individual sets can be intersected to produce a multivariate set which is defined by a joint membership function, the value of which is defined by the minimum value of the individual membership functions of each of the sets (Heuvelink and Burrough, 1993). That part of the map where all of the regions of predicted presence (grid-cells with a value of 1) overlap represents the potential distribution of the target species. In areas where not all conditions are satisfactory, the target organism is assumed to be absent (Chicoine *et al.*, 1985). The potential distribution maps produced using this type of envelope technique are binary, and regions of predicted presence are usually represented by a value of one and regions of predicted absence by a value of zero. The predictor variables are assumed to be equally important in determining the distribution of the target organism, as they are not differentially weighted in the model. In one of the earliest studies that employed this technique, the analysis was performed manually by superimposing a series of paper maps (Chicoine *et al.*, 1985). Later applications of the technique include Pfab and Witkowski (1997) and Skov and Borchsenius (1997). Recently, a simple envelope technique was implemented in a GIS application for the desktop GIS package ArcView (Skov, 2000). An advantage of simple envelope models is that they are relatively simple to implement in a GIS and predictions can be made very quickly, using relatively few locality records and very little processing power. Simple envelope models may be useful as exploratory techniques, particularly at the beginning of a study, prior to building more sophisticated statistical models.

BIOCLIM / ANUCLIM

A predictive modelling package known as BIOCLIM (recently renamed ANUCLIM) uses a refinement of the simple envelope approach described above for predicting potential distributions of target species (Nix, 1986; Busby, 1991). BIOCLIM defines two simple envelopes. The first envelope is constructed using the maximum and minimum values in the training set and then the second envelope,

which is contained within the first envelope, is constructed using selected thresholds e.g. the 5th and 95th percentile of each of the predictor variables.

These envelopes are used to define the "core" and "marginal" environments for the target species. Nix (1986) defined core environments (the second envelope) as those values falling between the 5th and 95th percentile of each of the predictor variables, although they can be estimated using other values, e.g. Lindenmayer *et al.* (1991) used the 10th and 90th percentiles to define the core range. Nix (1986) points out that these thresholds are arbitrary. Marginal environments are those that fall outside of the core range but within the upper and lower limits (minimum and maximum values) of the training set. The marginal range is thus identical to a simple envelope model calculated using minimum and maximum values of the training set.

The output of BIOCLIM consists of a distribution map indicating regions of predicted presence (in terms of core and marginal environments) and regions of predicted absence. Each of these regions is defined by classifying each of the localities (grid-cells) in the map region into one of three Boolean or crisp sets (core, marginal or absent) based on the data in the training set.

BIOCLIM has been used extensively to predict the potential distributions of various target organisms. It has been used to predict the potential distribution of the golden-tipped bat (Walton *et al.*, 1992); various weed species (Panetta and Mitchell, 1991 a & b; Sindel and Michael, 1992), Leadbeater's possum (Lindenmayer *et al.*, 1991), kangaroos (Skidmore *et al.*, 1996), gliders (Jackson and Claridge, 1999) and various snakes (Nix, 1986). BIOCLIM has also been used to assist in the reintroduction of an endangered bird (the helmeted honeyeater: Pearce and Lindenmayer, 1998), and in a climate change study (McKenzie and Busby, 1992).

An advantage of BIOCLIM is that the algorithm can be implemented directly in a GIS and predictions can be made fairly quickly using relatively few locality records. It also has the advantage over the simple envelope model that it predicts a core range for the target organism.

Fuzzy envelope models

Fuzzy envelope models (Chapter 4) represent a refinement to the BIOCLIM envelope approach. Fuzzy envelope models (FEMs) incorporate the notion that within a particular survival range, some conditions are more favourable than others. They further assume that environmental suitability varies on a continuous scale. FEMs are based on fuzzy classification, which has its basis in fuzzy set theory (Zadeh, 1965).

Fuzzy set theory differs from classical mathematical set theory in several ways. In classical mathematical set theory, an object either belongs to a particular set or not. These sets are termed crisp sets (Lark and Bolam, 1997) or Boolean sets (Burrough, 1989) because they are characterised by a clearly defined value or criterion. This type of classification assumes that all change between classes takes place at the class boundary and that very little significant change occurs within classes (Burrough, 1989), although this is often not the case with continuous data.

A fuzzy set is described by a fuzzy membership function, with values ranging from 0 to 1, corresponding with non-membership through to complete membership (Eastman, 1999). Fuzzy sets thus have continuous membership functions and are thus suited to situations where clearly defined class membership values are absent (Zadeh, 1965; Altman, 1994). The term "continuous classification" is used by some authors instead of "fuzzy classification" (Heuvelink and Burrough, 1993). Although fuzzy membership functions may appear to be similar to probability functions, these two concepts are quite different (Zadeh, 1965): fuzzy membership functions define possibility rather than probability (Zadeh, 1987).

A Fuzzy Envelope Model is typically developed as follows. A set of locality records (representing the presence of the target organism) is used to derive a set of values from each of the predictor variable maps to produce a training data set. Each predictor variable map is reclassified using a fuzzy membership function. The data in the training set define the shape of the membership functions for each predictor variable. These reclassified maps (fuzzy sets) are then superimposed using fuzzy algebra (Heuvelink and Burrough, 1993) to produce a map indicating the potential distribution of the target organism (a multivariate fuzzy set). The potential distribution map contains a continuum of possibility values indicating conditions of

varying suitability for the target organism. Localities with high possibility values are interpreted as representing conditions that are more favourable for the organism than those with low possibility values.

The use of fuzzy classification as a possible technique for potential distribution modelling has been investigated (Fairbanks and McKelly, 1994), although it appears not to have seen much use in this field of biology. Fuzzy classification has been used in soil science applications (Burrough, 1989; Lark and Bolam, 1997) and in remote sensing image classification techniques (Eastman, 1999).

FEMs deliver credible results and they represent refinements to the approach used in the BIOCLIM modelling package. These refinements are related to the way in which FEMs deal with uncertainty, the way in which this uncertainty is represented in the resultant potential distribution maps, and the way that these maps can be interpreted and applied (Chapter 4).

Criticisms of envelope techniques

The predictor variables in all three of the envelope models outlined above are unweighted (or implicitly equally weighted) and thus contribute equally to the predicted distribution. Thus, the implicit assumption is that all the predictor variables are equally important in predicting (or determining) the distribution of the target species, which is likely to be unrealistic and may lead to inaccuracies.

For all envelope models, the extent of the predicted range is highly dependent on the sample being representative of the population of the target organism and the influence of outliers on the predictions is likely to be quite large (Pearce and Lindenmayer, 1998), although the FEM will be less affected by outliers. If the full range of the target organism is not represented in the sample of locality records then sites that are quite similar to the majority of the sampled sites may be excluded from the total range (Carpenter *et al.*, 1993).

Envelope techniques do not consider the multivariate structure of the data and each predictor variable is treated independently. One of the possible consequences of this can be illustrated by an example. An organism may be able to survive in areas where conditions are hot and in areas that are dry but not in areas where hot and dry conditions occur simultaneously because they dehydrate too rapidly.

One of the criticisms levelled at BIOCLIM is that there is no biological justification for the use of either of the above percentile ranges to define the core range. A discussion of the limitations of BIOCLIM is provided by Carpenter *et al.* (1993). BIOCLIM and the simple envelope model both produce categorical outputs in their potential distribution maps, which may have important implications for interpretation of these maps and for the type of accuracy assessment measures that can be used to evaluate these models (Chapter 2).

Similarity metric techniques

Carpenter *et al.* (1993) described a technique that makes use of a point-to-point similarity metric (the Gower metric) to assign a classification value to an unsurveyed site (grid-cell) based on its proximity in environmental space to the most similar site of recorded presence for an organism. This technique is implemented in a modelling package known as DOMAIN (Carpenter *et al.*, 1993). Similar approaches have been used elsewhere (Skov, 2000). These techniques are referred to as *similarity metric techniques*.

The Gower metric is one of several Manhattan distance measures. These measures are distinguished by the type of standardisation procedure used; range standardisation being used in the case of the Gower metric (Booth *et al.*, 1987). The effect of range standardisation is to equalise the contribution from each predictor variable. Booth *et al.* (1987) claim that this is preferable to variance standardisation because it is considered to be less susceptible to bias arising from dense clusters of sample points. The range standardisation of the Gower metric has the opposite effect to that of variance standardisation, namely that it increases the influence of outliers.

In DOMAIN, the Gower metric is used to calculate the distance between two points (dAB) in p-dimensional Euclidean space. Each predictor variable corresponds with one of the dimensions of this space. Two similarity measures are then defined, based on the distance between the two points (dAB). The first is the *complementarity similarity measure* (R), which is calculated by subtracting dAB from one. The second, a *maximum similarity measure* (S), is used to determine to which of the surveyed points an unsurveyed point is most similar i.e. which point is closest in p-dimensional space. Each unsurveyed point (grid-cell) in the map region is assigned a

maximum similarity value to produce a map of continuously varying similarity values. The values in the map are not probabilities but are described as degrees of classification confidence. Skov (2000) suggests that the map of continuous values can be converted into a binary potential distribution map by selecting appropriate thresholds to define presence or absence of the target organism.

Carpenter *et al.* (1993) do not state how similar unsurveyed sites should be to surveyed sites, in order for the target species to survive at these unsurveyed sites. It is also not known how suitable the conditions at each of the surveyed sites are for the organism. As a result, the surveyed sites cannot be ranked in order of suitability before similarity values are assigned to unsurveyed sites, which is likely to result in overestimation of the target organism's range. This represents a major deficiency in the technique. Unsurveyed sites are assigned high similarity values based on their proximity in predictor variable space to surveyed sites. If a surveyed site represents the limits of the organism's range then unsurveyed sites that are similar but beyond the limits of the organism's range will be incorrectly classified as being suitable. As a result, this approach is likely to be particularly sensitive to sampling intensity, sampling bias and outliers. Centroid-based techniques (Caithness, 1995; Jones and Gladkov, 1999; Erasmus *et al.*, 2000; Hirzel *et al.*, 2001 a; Robertson *et al.*, 2001) overcome this problem because surveyed-site suitability is explicit in these models. DOMAIN has recently been used to predict plant and animal distributions in Guyana to assist in conservation planning (Funk and Richardson, 2002).

Carpenter *et al.* (1993) suggest that the selection of presence/absence thresholds should be based on expert knowledge. This is likely to differ considerably among experts and will probably result in widely differing binary maps. Similarly, potential distribution maps are likely to differ widely as predictions are iteratively improved with the addition of new distribution records to a dataset for a target species. In the case of envelope models, extreme values in the training set are taken to represent the physiological limits of the target organism in the map region. This is likely to yield a more realistic and possibly a more conservative potential distribution map, especially in the absence of expert knowledge. Similarity metric techniques are likely to be computationally more intensive and time-consuming than envelope techniques and are likely to offer few advantages over these techniques. The algorithm used by

DOMAIN has recently been implemented in a GIS application for use in the desktop GIS package Arcview (Skov, 2000).

Contingency Table Analysis

The Contingency Table Analysis (CTA) technique can be described as follows. Two predictor variables (usually elevation and rainfall) are divided into classes using a suitable class interval (Palmer, 1991; Palmer and Van Staden, 1992). A scatterplot of values associated with species presence of the two predictor variables is used to define the class intervals. The categorical predictor variables are cross-tabulated to produce a contingency table that is used to test the null-hypothesis of independence. Where the null hypothesis of independence is rejected, then those cells in the contingency table with the highest frequencies provide the conditions that are considered to be most suitable for the community or target species (Palmer, 1991). This technique was originally developed to predict the potential distribution of plant communities based on environmental variables (Palmer, 1991; Palmer and Van Staden, 1992). It has also been used to predict the distribution of individual species, using elevation and rainfall (Gibson, 1995). Lenton *et al.* (2000) also use a contingency table approach as a basis for making predictions using a simple indexed overlay model.

This method deals with non-linear responses of species to predictor variables and is capable of dealing with continuous or categorical data. This is a simple technique that is fairly easy to understand and implement, although the limitation is that predictions have been based on only two predictor variables. It may be possible to perform such an analysis with more than two predictor variables, using loglinear analysis, which is an extension of chi-square analysis of two-way contingency tables for which there are more than two variables (James and McCulloch, 1990). Fienberg (1989) provides a good introduction to loglinear analysis.

PCA-based techniques

Principal components analysis (PCA) is a multivariate dimension-reduction technique that produces a set of abstract variables (called principal components) that

are weighted linear combinations of the original variables (James and McCulloch, 1990).

PCA has frequently been applied in vegetation science to species compositional data obtained from quadrats, for the purposes of ordination (Randerson, 1993; Jongman *et al.*, 1995). It has also been used in genetics and morphometrics (James and McCulloch, 1990), mainly for the purposes of making quantitative comparisons and for dimension reduction. PCA has been quite popular for data reduction of predictor variables in distribution modelling (Osborne and Tigar, 1992; Buckland and Elston, 1993; Robinson *et al.*, 1997; Guisan *et al.*, 1998; Robertson *et al.*, 2001; Chapter 2).

The use of PCA with ecological data has been criticised (Austin, 1999), as various problems have been experienced in analysing ecological data using linear multivariate statistical techniques such as PCA (Noy-Meir and Austin, 1970; Swan, 1970). PCA has traditionally been used as an ordination technique (Randerson, 1993; Jongman *et al.*, 1995) in plant ecology and thus its suitability has been investigated in this context (Noy-Meir and Austin, 1970). Swan (1970) found that bell-shaped response curves and zero values for species composition data distorted ordination trends when using a linear ordination technique with simulated data. Noy-Meir and Austin (1970) found similar distortions when using PCA as the ordination technique with the simulated data of Swan (1970). These problems with PCA occur when it is applied to species compositional data (obtained from quadrats) in relation to environmental gradients.

The application of PCA described below differs from its traditional use in that it is used here as a technique for prediction for a single species only. The problems described above for PCA thus do not seem to apply to the application of PCA that is described here.

Robertson *et al.*, (2001) described and applied a technique, originally developed by Caithness (1995), which uses PCA for prediction of species distributions.

First, the technique of Robertson *et al.* (2001) is described and then compared with a similar approach by Erasmus *et al.* (2000). This is followed by descriptions of related methods that have been implemented in modelling packages known as FloraMap (Jones and Gladkov, 1999) and Biomapper (Hirzel *et al.*, 2001 a).

The technique described by Robertson *et al.* (2001), can be summarised as follows. A PCA is performed on correlation matrix derived from the values of the predictor variables associated with the presence locality data (training data set) to construct a mathematical hyperspace in which each orthogonal dimension is defined by an orthogonal principal component axis. The origin of this hyperspace is taken to characterise the centre of the niche of the organism in terms of the predictor variables. The distance from any point to the origin gives a measure of the "centrality" of the point in this niche hyperspace.

The principal components of a PCA are constructed so that most of the variance in the original variables is accounted for in the first few components. Using too many components results in overfitting of the model which usually results in loss of generality. At this point a stopping rule is used to determine the optimum number of principal components that should be included in the model so that overfitting is avoided (Caithness, 1995).

Using the retained eigenvectors of the PCA and the predictor variables associated with unsurveyed sites, one can map these sites into the niche hyperspace and calculate the distance from each unsampled site to the origin of the hyperspace. The squared distance between a point and the origin of the n-dimensional hyperspace is thus calculated by taking the sum of squares of the component scores (using Pythagoras' theorem). This distance can be used to calculate a probability of environmental suitability for each locality (grid-cell) as follows. Based on the assumption that the realised niche requirements of an organism are generally considered to follow Gaussian curves (Austin and Smith, 1989; Austin, 1999), a normal distribution is appropriate. As the distance of a point from the origin of the hyperspace is calculated from the sum of its squared component scores, and as the sum of squares of n standard normal random variates is distributed as chi-square with n degrees of freedom (Sokal and Rohlf, 1987), a chi-square distribution can be used to describe the overall distribution of observations in the hyperspace relative to the origin. This assumes that the further a point is from the origin of the hyperspace, the less suitable it is for the target species. The probability associated with each chi-square value can thus be determined by referring to a chi-square distribution (a chi-square distribution is equivalent to a squared normal distribution). These values can be mapped back to the cells of the original real-world map, to produce a probability

map. These probability values can be interpreted as an indication of the suitability of a given grid-cell for the target organism (Robertson *et al.*, 2001).

In order to calculate the centrality of a point in the hyperspace this technique relies on a chi-square distribution, which is equivalent to a squared normal distribution. While niche theory has suggested that species responses to environmental variables can be described using a symmetric bell-shaped curve (Austin and Smith, 1989; Austin, 1999), several studies have found skewed species responses to environmental variables (Austin *et al.*, 1984; Austin, 1987; Austin *et al.*, 1990; Leathwick and Mitchell, 1992; Austin *et al.*, 1994; Bio *et al.*, 1998; Ejrnæs, 2000) suggesting that the symmetric bell-shaped curve may be inadequate. The success of this technique and those that make similar simplifying assumptions is dependent on how robust it is to violations of these assumptions.

The implementation of this technique is outlined below in a series of steps. In the first step, the values of the training set are standardised by subtracting the mean and dividing by the standard deviation for each variable. Next, a PCA is performed on the standardised training set. In the third step, the observations of the prediction data set are standardised by the means and standard deviations calculated from the training data set in the first step of the analysis. The effect of standardising the prediction set (using means and standard deviations of the training set) is to centre it on the origin of the hyperspace, which allows the origin to be viewed as the niche optimum for the target organism. The eigenvectors resulting from the PCA performed on the standardised training set are then multiplied by the standardised prediction set to yield the component scores. Conceptually, this step projects the prediction set into the hyperspace defined by the training set. In the fifth step, the variances of each component axis are standardised by dividing the component scores of each component by their respective eigenvalues to produce a matrix of standardised component scores. This step is necessary because the variance on each PCA axis is different, and spherical probability contours that are concentric about the origin of the hyperspace, can only be assumed if the variance on each component axis is first standardised. In step six, the probability associated with each observation is calculated by summing the squares of the standardised component scores and substituting this value into the chi-square probability distribution function. In the final step, the probability values

for each grid cell are mapped back to their associated original geographical coordinates of each observation.

The approach described by Erasmus *et al.* (2000) differs from that described above in several ways. Firstly, Erasmus *et al.* (2000) do not explicitly define a separate training and a prediction set. The training set *sensu* Robertson *et al.* (2001) is equivalent to "the values of climate variables for each KR (known record) grid cell". Erasmus *et al.* (2000) start by subtracting the means of the values associated with the KR grid cells (training set). This is similar to the standardisation applied by Robertson *et al.* (2001), although the predictor variables are not divided by their standard deviations - with the result that the effect of different measurement units of the predictor variables are not removed. This step centers the values of the training set around the origin of the hyperspace but it does not center the values of the prediction set around the origin of this hyperspace. Unlike the approach used by Robertson *et al.* (2001) the prediction set is not standardised using the means and standard deviations of the training set. As a result, the prediction set is not projected into the hyperspace defined by the training set. While the authors have standardised the component scores, they do not report summing the squares of standardised component scores. Erasmus *et al.* (2000) imply that the standardised component scores themselves (which represent distances from the origin) follow a χ^2 distribution. They have not used a stopping rule and as a result, overfitting of the model is likely to occur. Probability values are interpreted as probability of occurrence, while these should rather be taken to indicate probability of environmental suitability. Similarly, Aspinal and Veitch, (1993) have suggested that probability values can be interpreted as an index of habitat suitability or quality.

FloraMap

FloraMap is a PCA-based software package for predicting distributions of plants and animals (Jones and Gladkov, 1999). For a brief review of the package see Arnold (2000). In the manual describing the package, various theoretical aspects of the modelling process are described (Jones and Gladkov, 1999). The first section deals with the climate predictor variables and how these can be standardised to account for

differences in seasonality. This allows comparisons to be made between northern and southern hemisphere localities.

The next section gives an overview of the predictive modelling technique. The term "calibration set" used by Jones and Gladkov (1999) is equivalent to the "training set" used here. Implementing the technique involves doing a PCA on the variance-covariance matrix of the calibration set (training set) in which the data are first standardised, although the standardisation procedure differs from most standard applications. In the example used, the PCA is performed on a set of 36 climatic variables. This set consists of three groups of 12 predictor variables (rainfall, temperature and diurnal temperature range). The climatic variable values in each group are standardised by the common variance (pooled variance) for that group e.g. rainfall values are standardised by the common variance for rainfall. The values from each variable are thus standardised by subtracting the mean of that variable and dividing by its group variance. Once the PCA has been performed, the authors state that a subset of the components resulting from this PCA can be selected, although they do not describe a procedure for making this selection.

Next, a system of equations is used to derive a formula for calculating the probability that a point occurs in an infinitely thin spherical shell centered on the origin of an n-dimensional hyperspace. The number of dimensions defining the hyperspace is determined by the number of principal components selected by the user.

How the prediction set is fitted into the hyperspace defined by the calibration set (training set) so that probabilities can be calculated for the whole map region, is not addressed. One has to assume that the probability calculated for a point is used as a measure of centrality of that point in the hyperspace. The FloraMap technique is similar to those PCA-based techniques described by Caithness (1995, Robertson *et al.*, 2001) and Erasmus *et al.* (2000), although there is a marked difference in the method of calculating the probabilities.

Various hierarchical cluster analysis algorithms have been incorporated into FloraMap as a means of dealing with multiple populations. The use of phenetic clustering in biology has been questioned, particularly for studies of variation among populations, as its use is considered to be inappropriate (De Queiroz and Good, 1997).

One of the potential weaknesses of the PCA-based techniques outlined above are that they are unable to cope with skewed responses of target organisms to

environmental variables. This is also a weakness of the of *Ecological Niche Factor Analysis* technique (Hirzel, 2001), which is described below.

Ecological Niche Factor Analysis

A technique related to PCA-based techniques is that of *Ecological Niche Factor Analysis* (ENFA: Hirzel, 2001) which has been implemented in the BIOMAPPER package (Hirzel *et al.*, 2001a). ENFA calculates uncorrelated factors (similar to components of a PCA) that are used to explain the distribution of a target species. The first factor is called the *marginality factor* and the remaining factors (which are all orthogonal) are called *specialisation factors*. Some of the terminology differs from that used in this review, for example, the output of the model is described as a *habitat suitability map*; the predictor variables are referred to as *ecogeographical variables*, the *global distribution* is equivalent to the prediction set and the *species distribution* is equivalent to the training set (Hirzel, 2001).

The technique can be outlined briefly as follows, but readers are referred to Hirzel (2001) for a full description. Conceptually, a hyperspace can be defined using the predictor variables as axes. The values of the training (species) and prediction (global) sets can be plotted into this hyperspace as points. Two hyper-ellipsoids can be visualised, a large hyper-ellipsoid defined by the prediction set and a smaller hyper-ellipsoid (contained within the first) which is defined by the training set. In ENFA, the first axis of a factor analysis (the marginality factor) can be visualised as a line going through the centroids of the two hyper-ellipsoids. Once the marginality factor has been defined, the training set hyperspace is transformed into a sphere, which is necessary to calculate the specialisation factors. The first specialisation factor accounts for maximum variance of the global distribution while being orthogonal to the marginality factor. Subsequent specialisation factors are extracted so that they are orthogonal and account for less and less of the variance as is the case with PCA. As most of the variation is accounted for in the first few factors, only a subset of factors is selected (using either a broken-stick distribution or a threshold value for cumulative variance) from which to calculate the habitat suitability map. The habitat suitability value of each grid-cell in the map region is calculated from the combination of the scores associated with that grid-cell on each of the factors. The

details of the approach used to calculate habitat suitability values are not explicit in the description of the technique (Hirzel, 2001). In order to account for differences in the ecological importance equal weight is attributed to the marginality and specialisation, but the weighting of the individual specialisation factors is done according to their eigenvalues. Again, the details of this approach have not been given (Hirzel, 2001). Habitat suitability values are calculated for each grid-cell in the map region to produce a habitat suitability map with values constrained between 0 and 1. Hirzel *et al.* (2001 b) suggest that ENFA is robust to various data quality and quantity scenarios, based on a study undertaken using hypothetical data. However, the weaknesses of this technique have not been documented in the literature yet.

Group discrimination techniques

Group-discrimination techniques have been used more frequently than profile techniques for predicting distributions. Guisan and Zimmermann (2000) reviewed a number of group discrimination techniques used for predicting distributions. Further examples of the application of these techniques are given here. In particular, consideration is given to Discriminant Function Analysis, Maximum Likelihood Classification, Classification and Regression Trees, and Artificial Neural Networks, Genetic Algorithms and an alternative implementation of Logistic Regression.

Discriminant Function Analysis

James and McCulloch (1990) provide a good introduction to Discriminant Function Analysis (DFA) and describe some of its applications. Guisan and Zimmermann (2000) described the use of DFA as a predictive technique. Additional examples of its application for predicting species distributions include Flather and King (1992); Rogers and Williams (1993); Williams *et al.* (1994); Rogers *et al.* (1996); Robinson *et al.* (1997); Manel *et al.*, (1999 a & b) and Cumming (2000 b). DFA can be used to distinguish amongst more than two categories, for example in predicting the distributions of different races of a species (Lloyd and Palmer, 1998; Steele *et al.* 1998). DFA assumes multivariate normally-distributed predictor variables and similar covariance structures around the group means of each of the

classification groups (presence and absence categories). This assumption is likely to be violated by distributional data since species presumably select, or are selected by, a rather well-defined and non-random subset of environmental conditions (Rogers and Williams, 1993). Predictor variables are not always normally distributed (Flather and King, 1992) and this is likely to have an effect on the efficiency of the method (James and McCulloch, 1990).

Maximum likelihood classification

Lillesand and Kiefer (1994) describe Maximum Likelihood Classification (MLC) and its application in digital image processing. Guisan and Zimmermann (2000) mentioned this technique in their review but could not find examples of its application for predicting plant or animal distributions. This technique has been used successfully to predict tsetse fly distributions (Robinson *et al.*, 1997). MLC has been proposed as an alternative to DFA, with the advantage that it is not constrained by the assumption of common covariances in the presence and absence categories within multivariate space and is therefore considered to have greater predictive power (Robinson *et al.*, 1997). In a quantitative comparison, Robinson *et al.* (1997) found that MLC yielded better predictions than DFA. The weaknesses of this technique have not been documented.

Logistic Regression

Logistic Regression (LR: Hosmer and Lemeshow, 1989) is a form of Generalised Linear Model (GLM: McCullagh and Nelder, 1989) in which a binomial error distribution and a logistic link function are used (Guisan and Zimmermann, 2000). GLM have been used frequently in biology (e.g. Austin *et al.* 1984; Nicholls 1989; Austin *et al.*, 1990; Leathwick and Mitchell, 1992; Osborne and Tigar, 1992; Austin *et al.* 1994; Austin and Meyers 1996; Ferrier and Watson, 1997; Guisan *et al.*, 1998, 1999; Higgins *et al.*, 1999; Manel *et al.*, 1999 a & b; Collingham *et al.*, 2000; Cumming, 2000 a & b; Pearce and Ferrier, 2000a; Hirzel *et al.*, 2001a). GLM are a more flexible family of regression models than classical least square regression models that are restricted to cases where the response variable is normally distributed

and the variance does not change as a function of the mean (Guisan and Zimmerman, 2000). The combination of predictors (the linear predictor) is related to the mean of the response variable by means of a link function. The link function allows transformation to linearity and for the predictions to be constrained within the range of values of the response variable. GLM allow the use of Gaussian, Poisson, Binomial or Gamma distributions to be used, for which the appropriate link function is required.

If the relationship of the response variable to a predictor variable is not linear then quadratic, cubic or higher order terms can be included in the linear predictor (e.g. Austin *et al.*, 1984; Guisan *et al.*, 1999; Higgins *et al.*, 1999; Hirzel *et al.*, 2001a). Interaction terms can also be included (Austin *et al.*, 1996; Guisan *et al.*, 1999). GLM are flexible in that they allow both continuous or categorical predictors to be used (Austin *et al.*, 1984). Nicholls (1989) provides a good example of the application of GLM to predicting species distributions.

A number of studies investigating the shape of species responses to environmental predictors found skewed responses (Austin *et al.*, 1984; Austin, 1987; Austin *et al.*, 1990; Leathwick and Mitchell, 1992; Austin *et al.*, 1994; Bio *et al.*, 1998; Ejrnæs, 2000). Austin and Meyers (1996) suggested that methods based on symmetric Gaussian curves were likely to give biased predictions. This highlighted the need to use techniques that were flexible enough to handle more complex species responses than those available in GLM.

Generalised Additive Models (GAM; Hastie and Tibshirani, 1990) are a non-parametric extension of GLM in which it is not necessary to specify the functional form of the relationship (linear, quadratic, cubic) and hence the response curve is more data- than model-driven (Bio *et al.*, 1998). GAMs allow both linear and complex response shapes, as well as combinations of both in a single model. They allow all the functions of the GLM family to be used and they include a variety of smooth functions, which can fit any shape of response curve (Bio *et al.*, 1998). GAM have also been used extensively (Yee and Mitchell, 1991; Austin and Meyers 1996; Leathwick, 1995; 1998; Leathwick *et al.* 1996; Ferrier and Watson, 1997; Austin, 1998; Bio *et al.*, 1998; Franklin, 1998; Ejrnæs, 2000; Pearce and Ferrier, 2000a; Leathwick and Austin, 2001; Leathwick, 2001; Leathwick and Whitehead, 2001). Yee and Mitchell (1991) were the first to use GAM for predicting species distributions.

Since then GAM has been very popular. For a recent review of GLM and GAM in the context of species distribution modelling see (Guisan *et al.*, 2002).

Austin and Meyers (1996) demonstrated some of the advantages of the flexible nature of non-parametric GAM functions over GLM, using eucalyptus species. In a recent study, Bio *et al.* (1998) compared GLM models containing linear and quadratic functions with non-parametric data-driven GAM models and found that the majority of models fitted contained at least one environmental variable that was better fitted by a non-parametric than a linear or quadratic function.

Pearce and Ferrier (2000a) found that GAM models performed slightly but significantly better overall than GLM models applied to a number of species of birds, reptiles and plants. Ferrier and Watson (1997) found that GAM models performed significantly better than GLM models in a comparison among several modelling techniques using presence/absence data. However, when presence-only data was used they found that there was no significant difference in performance between the GAM and GLM models. This was attributed to the quality of the data, as the nonparametric curve fitting procedures used in GAM were more likely to fit spurious response functions to the poor quality (biased) data in presence-only datasets than to the high quality presence/absence data. Pearce and Ferrier (2000a) found that the characteristics of the data (data quality) had the strongest effect on predictive performance of models in a comparison among modelling techniques.

When presence/absence data from systematic field surveys is not available then the alternative is to use profile techniques or to make use of pseudo-absence data in order to use GLM and GAM (Ferrier and Watson, 1997; Cumming, 2000 a & b; Zaniewski *et al.*, 2002). Pseudo-absence data are best described as absence data that have not been obtained by means of a survey designed specifically to establish the absence of the target organism at a number of sites. Pseudo-absence records can be defined randomly (Ferrier and Watson, 1997; Zaniewski *et al.*, 2002), by using presence data collected for other species (Zaniewski *et al.*, 2002), or by using all the un-surveyed grid-cells in the map region as absence records (Cumming 2000 a & b).

One of the disadvantages of using pseudo-absence data is that false absence records are likely to be included in the dataset. The influence of false absence records on the performance of LR models has been investigated elsewhere (Chapter, 6). Another potentially serious problem with using presence-only data, which comes from

museum or herbarium collections, is that there is often bias in these datasets (Chapter 2).

GLM and GAM that are used to make species predictions usually rely on stepwise procedures to decide which predictor variables should be included in the model. The stepwise procedure used in the model may influence the efficiency and performance of the model (Pearce and Ferrier, 2000a), and James and McCulloch (1990) suggest that stepwise procedures should be avoided altogether.

Classification and Regression Trees

Classification and Regression Tree (CART: Breiman *et al.*, 1984) techniques have been used successfully to predict the distributions of kangaroos (Walker, 1990; Skidmore *et al.*, 1996) and Tsetse flies (Williams *et al.*, 1994). Walker and Cocks (1991) describe a modelling procedure, called HABITAT, that models potential distribution using classification and regression trees. Vayssieres *et al.* (2000) compared the performance of CART models with LR models for predicting the distributions of three oak species. Despite the fact that the LR models were optimised to account for non-linearity and factor interactions, CART models performed significantly better than these models. Franklin (1998) used CART, generalised linear models (GLM) and generalised additive models (GAM) for predicting species distributions and found that while CART models were sometimes difficult to interpret, they yielded the lowest prediction errors.

One of the most important advantages of tree-based methods is that they can deal with interactions between variables and represent these in a relatively simple form (Michaelsen *et al.*, 1994). They are also capable of dealing with both continuous and categorical predictor variables (Michaelsen *et al.*, 1994). They require a larger learning sample to produce the correct output than Artificial Neural Networks (Lees, 1994). Michaelsen *et al.*, (1994) claim that datasets need to be in the region of 300-400 observations in order to fully realise their strengths, although they claim that success can be achieved with datasets as small as 100 observations. Skidmore *et al.* (1996) claim that models based on CART can be extremely time consuming to develop, with the quality of the result being dependent on the skill of the analyst and on the quality of the data comprising the sample.

Artificial Neural Networks

Barth (1991) gives a good overview Artificial Neural Networks (ANN). A number of the ANN used to make distribution predictions are backpropagation networks. A potentially useful review of backpropagation networks is provided by Paola and Schowengerdt (1995), in the context of remote sensing. Schultz and Wieland (1997) discuss the use of ANN in agro-ecological modelling.

Besides those studies reviewed by Guisan and Zimmermann (2000), additional studies can be listed in which ANN were used to predict distributions. Mastrorillo *et al.* (1997) compared the performance of ANN models with DFA models for three species of freshwater fish and found that ANN performed consistently better than DFA. Manel *et al.*, (1999 b) compared of the performance of techniques based on Discriminant Analysis (DA), Logistic Regression (LR) and ANN for predicting the distributions of Himalayan river birds and found that their performance differed only marginally. This expands upon an earlier study (Manel *et al.*, 1999 a). One of the problems with the comparison by Manel *et al.* (1999 a) is that only linear terms were used in the logistic regression. As a result only sigmoidal species responses to environmental variables could be modelled. This may have reduced the performance of the LR models in the comparison, especially if non-monotonic species responses occurred. Higher order terms such as quadratic, cubic or higher order terms are usually included in the linear predictor to account for more complex relationships (e.g. Austin *et al.*, 1984; Guisan *et al.*, 1999; Higgins *et al.*, 1999; Hirzel *et al.*, 2001).

Williams *et al.*, (1994) used several modelling techniques, including ANN, to predict the distribution of tsetse flies in Zimbabwe. They found that although the predictions made using ANN were very good, this technique took several orders of magnitude longer to converge than other multivariate techniques (DFA), and the output was difficult to interpret biologically. Özesmi and Özesmi (1999) used ANN models to predict habitat selection of marsh-breeding bird species, which performed better than equivalent models built using LR.

Genetic Algorithms

Genetic Algorithms (GA) are based on evolutionary theory and find solutions to problems by an iterative process (consisting of several generations) in which the best solution to the problem are selected by a process similar to natural selection in natural populations. The basic concepts of GA were developed by Holland (1975). Lees (1994) and Franklin (1995) provide a good introduction to GA, in the context of predictive distribution modelling. Stockwell and Peters (1999) present a modelling system, called GARP (Genetic Algorithm for Rule-set Production), that make use of GA to make potential distribution predictions. The modelling system has a feature that enables generation and testing of a range of model types, including categorical, range-type and logistic models (Stockwell and Peters, 1999). Recently, Peterson *et al.* (2001) used GA (implemented in GARP) to investigate the effects of climate change on bird distributions. Other applications of GARP include Peterson *et al.* (1999), Peterson and Cohoon (1999) and Peterson (2001). GA are considered to produce reliable results under a wide range of operating conditions and are designed to handle poorly structured domains (Stockwell and Peters, 1999).

Model evaluation

One of the most important aspects of the modelling process is evaluating the prediction success of the model (termed model evaluation). In general, prediction success is determined by measuring the success with which a model is able to correctly predict the species as being present or absent from independent localities not used to calibrate the model. This involves estimating the apparent error rate for the prediction (Guisan and Zimmermann, 2000), which is calculated using an appropriate error or accuracy measure (Fielding and Bell, 1997). Evaluation assists in determining the suitability of a model for a specific purpose, allows quantitative comparisons among alternative models and assists in identifying aspects of a model that most need improvement (Pearce and Ferrier, 2000 b).

Guisan and Zimmermann (2000) outlined two approaches for model evaluation. In the first approach a single dataset is used to calibrate and evaluate the model (e.g. Franklin, 1998; Manel *et al.*, 1999 a & b; Cumming, 2000 a & b). Guisan and

Zimmermann (2000) reviewed techniques available for this type of evaluation e.g. cross validation, jackknifing and bootstrapping. Verbyla and Litvaits (1989) described various resampling strategies including: resubstitution, cross-validation, ten-fold validation, jackknife and bootstrap resampling procedures.

The second approach is to use two independent datasets, one for model calibration (training) and the other for model evaluation (testing). The use of an independent dataset of localities with which to evaluate the model is considered to provide the best assessment of prediction success (Power, 1993; Chatfield, 1995; Fielding and Bell, 1997; Guisan and Zimmermann, 2000; Pearce and Ferrier, 2000 a & b). As it is not always possible to obtain two completely independent datasets, it is often necessary to partition the original dataset into a training and testing set (Fielding and Bell, 1997). However, Chatfield (1995) maintains that partitioning the dataset is not the same as collecting new data and has questioned this approach. Although data partitioning may not be optimal, it is superior to using the same dataset for calibration and evaluation (resubstitution). Instead of partitioning the data into only one training and one evaluation set, a superior approach may be the 10-fold cross validation described by Verbyla and Litvaits (1989). This procedure can be outlined as follows. The dataset is partitioned into ten equal-sized sub-samples. The model is calibrated using the data from nine of the sub-samples and the tenth sub-sample is used for evaluation. The process is repeated ten times so that each of the ten sub-samples can be used for evaluation.

Fielding and Bell (1997) reviewed a number of error or accuracy measures that are appropriate for evaluating the prediction success of distribution models. The choice of accuracy measure will be determined by the type of data available for evaluation, the type of output produced from the model and the overall goal of the study (Fielding and Bell, 1997). Most of these measures are based on a confusion matrix in which the observed (actual) and predicted presence/absence patterns are cross-tabulated (Fielding and Bell, 1997). Fielding and Bell (1997) and Guisan and Zimmermann (2000) reviewed various threshold-dependent and threshold-independent accuracy measures that can be calculated from a confusion matrix. Pearce and Ferrier (2000 b) described approaches appropriate for evaluating the performance of logistic regression models, which are also applicable to other models that generate probabilistic predictions. Threshold-independent measures (e.g.

Receiver Operator Characteristic curves) are considered to be more robust and more objective than threshold-dependent measures (e.g. Kappa statistics) since they do not rely on a single threshold to distinguish between predicted presence and predicted absence (Fielding and Bell, 1997). In this thesis, wherever possible, model performance has been assessed using the area under the curve (AUC) of Receiver Operator Characteristic (ROC) curves, as this is a threshold-independent measure (Fielding and Bell, 1997).

If only presence data are available for evaluation then only parameters *a* and *c* in the confusion matrix (Table 1) can be calculated. This means that only the False Negative Rate [FNR = c/(a+c)] or Sensitivity [Sn = a/(a+c)] can be calculated (Fielding and Bell, 1997). As a result, only the false positive errors can be assessed. The best accuracy measures are those that use all of the data available in the confusion matrix (parameters a, b, c and d) since these take both FP and FN errors into account, although these measures require both presence and absence testing data.

Models that only produce predictions of presence and absence (e.g. simple envelope techniques) or other categorical outputs (e.g. core and marginal range of BIOCLIM) instead of continuous outputs, cannot make use of threshold independent measures. This is because threshold-independent measures assess the performance of the model at all possible thresholds and thus require models that produce continuous output.

Problems arise when one is comparing the performance of models that have categorical outputs with those that have continuous outputs. In such cases one is confined to threshold-dependent measures (Chapter 7). One of the problems with these measures is that a threshold has to be defined that separates predicted values that one considers to represent species presence from those that one considers to represent species absence (Fielding and Bell, 1997), which is essentially arbitrary. Often the threshold of 0.5 is used (Manel *et al.*, 1999 a & b; Cowley *et al.*, 2000) although other arbitrary thresholds have also been used (Chapter 5). The choice of threshold may be determined by the consequences of various kinds of correct or incorrect decisions (Fielding and Bell, 1997; Pearce and Ferrier, 2000 b). If one is comparing the performance of two models using a threshold-dependent technique, there is no guarantee that the same arbitrary threshold will have equivalent meaning in both models; it is more likely that it will be different (Chapters 7 and 8). Instead, an

optimal threshold should be used to compare the performance of the models. This is the threshold at which the best agreement between predicted and observed values is achieved. This approach is described by Guisan and Zimmermann (2000), and examples of its application include Franklin (1998), Guisan *et al.* (1998), and chapters 7 and 8. In addition to using optimal thresholds to evaluate performance, these thresholds can also be used to convert potential distribution maps from a continuous (probabilistic) scale into presence-absence maps.

Discussion

In this review a number of techniques have been described that are available for producing models to predict species potential geographical distributions. Each technique has a number of strengths and weaknesses, which the researcher should take into account when selecting an appropriate technique. The problem is that the strengths and weaknesses of each technique are not always made explicit by those who develop or implement them. This is probably partly because the capabilities of a new technique are not always immediately known. An understanding of these strengths and weaknesses generally develops over a period of time, usually by means of quantitative comparisons with other techniques.

Comparative studies documenting the performance of two or more modelling techniques using the same dataset will assist in selecting an appropriate modelling technique (e.g. Rogers *et al.*, 1996; Ferrier and Watson, 1997; Robinson *et al.*, 1997; Franklin, 1998; Manel *et al.*, 1999 a & b; Özesmi and Özesmi, 1999; Cumming, 2000 b; Vayssieres *et al.*, 2000; Hirzel *et al.*, 2001 b; Zaniewski *et al.*, 2002; Chapter 7). Similarly, the influence of various data quality issues on the performance of modelling techniques (Manel *et al.*, 1999 b; Peterson and Cohoon, 1999; Cumming, 2000 b; Pearce and Ferrier, 2000 b; Hirzel *et al.*, 2001 b; Chapter 6) should also be investigated so that recommendations for appropriate use can be made. In order for these studies to be useful, comparisons of model performance should be done using quantitative evaluation measures preferably using independent testing data. An important component in developing an understanding of the strengths and weaknesses of a technique is to determine how robust a technique is to violations of its assumptions (e.g. Chapter 6). Hypothetical data are often useful for testing particular

data quality problems or data assumptions (Swan, 1970; Cumming, 2000 b; Hirzel *et al.*, 2001; Hirzel and Guisan, 2002).

The approach selected for making predictions will depend on several criteria such as the purpose for which the model is required; the spatial resolution and level of accuracy required; the biology of the target organism; the nature and quality of the data available; the expertise of the person producing the models, and logistical resource constraints. The choice between using a profile or a group discrimination technique will largely depend on the availability and quality of suitable absence data. Absence data are also considered to be less reliable than presence data (there are more chances of committing FN errors than FP errors), and is specific to the scale at which it was collected (Chapter 2).

Several modelling techniques may meet the decision criteria listed above, in which case the researcher will probably have to choose among various alternatives. One approach is to develop two or more models using different techniques and then to select one of these based on model performance (measured by assessing prediction success). An alternative approach is to select an appropriate technique based on previous studies documenting the application of techniques to particular problems in the literature. These types of studies are considered to be particularly important for choosing among alternative techniques (Guisan and Zimmermann, 2000).

Table 1. A confusion matrix (Fielding and Bell, 1997). Where + indicates presence and – indicates absence. The parameters a, b, c and d represent counts rather than percentages.

		Observed	
		+	-
Predicted	+	a	b
	-	c	d

References

Altman, D., 1994. Fuzzy set theoretic approaches for handling imprecision in spatial analysis. *International Journal of Geographical Information Systems*. 8: 271-289.

Arnold, H.R., 2000. FloraMap - predicting distributions. *Diversity and Distributions*. 6: 271-272.

Aspinall, R., Veitch, N., 1993. Habitat mapping from satellite imagery and wildlife survey data using a Bayesian modeling procedure in a GIS. *Photogrammetric Engineering and Remote Sensing*. 59: 537-543.

Austin, M.P. 1987. Models for the analysis of species' response to environmental gradients. *Vegetatio*. 69: 35-45.

Austin, M.P. 1998. An ecological prespective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*. 85: 2-17.

Austin, M.P. 1999. A silent clash of paradigms: some inconsistencies in community ecology. *Oikos*. 86: 170-178.

Austin, M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*. 157: 101-118.

Austin, M.P., Cunningham, R.B., Fleming, P.M., 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio*. 55: 11-27.

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

Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a beta-function. *Journal of Vegetation Science*. 5: 215-228.

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

Austin, M.P., Pausas, J.G., Nicholls, A.O. 1996. Patterns of tree species richness in relation to environment in south eastern New South Wales, Australia. *Australian Journal of Ecology*. 21: 154-164.

Austin, M.P., Smith, T.M., 1989. A new model for the continuum concept. *Vegetatio*. 83: 35-47.

Barth, G., 1991. What is all the fuss about neural networks? In: Maurer, H. (Ed.), *Proceedings of new results and new trends in computer science*, Springer-Verlag, Berlin, pp. 1-14.

Begon, M., Harper, J.L., Townsend, C.R., 1990. Ecology - Individuals, Populations and Communities, second Edition. Blackwell Scientific Publications, Oxford, pp. 945.

Bio, A.M.F., Alkemade, R., Barendregt, A. 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. *Journal of Vegetation Science*. 9: 5-16.

Bongers, F., Poorter, L., Van Rompaey, R.S.A.R., Parren, M.P.E., 1999. Distribution of twelve moist forest canopy tree species in Liberia and Cote d'Ivoire: response curves to a climatic gradient. *Journal of Vegetation Science*. 10: 371-382.

Booth, T.H., Nix, H.A., Hutchinson, M.F., Busby, J.R., 1987. Grid matching: A new method for homoclime analysis. *Agricultural and Forest Meteorology*. 39: 241-255.

Boyce, M.S., McDonald, L.L. 1999. Relating populations to habitats using resource selection functions. *Trends in Ecology and Evolution*. 14: 268-272.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schiegelow, F.K.A. 2002. Evaluating resource selection functions. *Ecological Modelling*. 157: 281-300.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. 1984. Classification and regression trees. The Wadsworth statistics/probability series. Chapman & Hall, New York.

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

Burrough, P.A., 1989. Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science*. 40: 477-492.

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

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

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

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A*. 158: 419-446.

Chicoine, T.K., Fay, P.K., Nielsen, G.A., 1985. Predicting weed migration from soil and climate maps. *Weed Science*. 34: 57-61.

Collingham, Y.C., Wadsworth, R.A., Huntley, B., Hulme, P.E., 2000. Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology*. 37: 13-27.

Cowley, M.J.R., Wilson, R.J., Leon-Cortes, J.L., Gutierrez, D., Bulman, C.R., Thomas, C.D., 2000. Habitat-based statistical models for predicting the spatial distribution of butterflies and day-flying moths in a fragmented landscape. *Journal of Applied Ecology*. 37: 60-72.

Cumming, G.S., 2000 a. Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*. 27: 425-440.

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

De Queiroz, K., Good, D.A., 1997. Phenetic clustering in biology: a critique. The Quarterly Review of Biology. 72: 3-30.

Eastman, J.R., 1999. Guide to GIS and image processing, Volume 2, Clark Labs, Worcester, pp. 170.

Ejrnæs, R. 2000. Can we trust gradients extracted by detrended correspondence analysis? *Journal of Vegetation Science*. 11: 565-572.

Erasmus, B.F.N., Kshatriya, M., Mansell, M.W., Chown, S.L., Van Jaarsveld, A.S., 2000. A modelling approach to antlion (Neuroptera: Myreleontidae) distribution patterns. *African Entomology*. 8: 157-168.

Fairbanks, D.H.K., McKelly, D., 1994. Investigating fuzzy set classification methods for use in GIS decision support modelling to determine land suitability. FOR-I 515, Division of Forest Science and Technology, CSIR.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra, p. 193.

Fienberg, S.E. 1989. The analysis of cross-classified categorical data. MIT press, Cambridge. p. 198.

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

Flather, C.H., King, R.M., 1992. Evaluating performance of regional wildlife habitat models: implications to resource planning. *Journal of Environmental Management*. 34: 31-46.

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*. 19: 474-499.

Franklin, J., 1998. Prediction the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*. 9: 733-748.

Funk, V.A., Richardson, K.S. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*. 51: 303-316.

Gibson, D., 1995. Modelling the distribution of *Portulacaria afra* in the Eastern and Western Cape Provinces, South Africa, in relation to environmental variables and the normalised difference vegetation index. Honours Thesis, Rhodes University, Grahamstown.

Guisan, A., Edwards, T.C., Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*. 157: 89-100.

Guisan, A., Theurillat, J.-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*. 9: 65-74.

Guisan, A., Weiss, S.B., Weiss, A.D., 1999. GLM versus CCA spatial modelling of plant species distribution. *Plant Ecology*. 143: 107-122.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

Hastie, T.J., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.

Heuvelink, G.B.M., Burrough, P.A., 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*. 7: 231-246.

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

Hirzel, A., 2001. When GIS come to life. Linking landscape and population ecology for large population management modelling: the case of Ibex (*Capra ibex*) in Switzerland. Ph.D. Thesis, Institute of Ecology, Laboratory for Conservation Biology, University of Lausanne.

Hirzel, A.H., Hausser, J., Perrin, N., 2001a. Biomapper 1.0 Lausanne, Laboratory for Conservation Biology, URL: http://www.unil.ch/biomapper

Hirzel, A.H., Helfer,V., Metral, F., 2001 b. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*. 145: 111-121.

Hirzel, A., Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*. 157: 331-341.

Holland, J.H., 1975. Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor.

Hosmer, D.W., Lemeshow, S., 1989. Applied logistic regression. John Wiley & Sons, New York, pp. 307.

Huntley, B., Berry, P.M., Cramer, W., Macdonald, A.P., 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. *Journal of Biogeography*. 22: 967-1001.

Jackson, S.M., Claridge, A., 1999. Climatic modelling of the distribution of the mahogany glider (*Petaurus gracilis*), and the squirrel glider (*P. norfolcensis*). *Australian Journal of Zoology*. 47: 47-57.

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

Jones, P.G., Gladkov, A., 1999. FloraMap - a computer tool for predicting the distribution of plants and other organisms in the wild. International Center for Tropical Agriculture, Cali, Columbia, pp. 99.

Jongman, R.H.G., Ter Braak, C.J.F., Van Tongeren, O.F.R. 1995. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge.

Lark, R.M., Bolam, H.C., 1997. Uncertainty in prediction and interpretation of spatially variable data on soils. *Geoderma*. 77: 263-282.

Leathwick, J.R. 1995. Climatic relationships of some New Zealand forest tree species *Journal of Vegetation Science*. 6: 237-248.

Leathwick, J.R. 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*. 9:719-732.

Leathwick, J.R. 2001. New Zealand's potential forest pattern as predicted from current species-environment relationships. *New Zealand Journal of Botany*. 39:447-464.

Leathwick, J.R., Austin, M.P. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*. 82: 2560-2573.

Leathwick, J.R., Whitehead, D. 2001. Soil and atmospheric water deficits and the distribution of New Zealand's indigenous tree species. *Functional Ecology*. 15: 233-242.

Leathwick, J.R., Whitehead, D. McLeod, M. 1996. Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach. *Environmental Software*. 11:81-90.

Leathwick, J.R., Mitchell, N.D. 1992. Forest pattern, climate and vulcanism in central North Island, New Zealand. *Journal of Vegetation Science*. 3: 603-616.

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

Lenton, S.M., Fa, J.E., Perez Del Val, J., 2000. A simple non-parametric GIS model for predicting species distribution: endemic birds in Bioko Island, West Africa. *Biodiversity and Conservation*. 9: 869-885.

Lillesand, T.M., Kiefer, R.W., 1994. Remote sensing and image interpretation. John Wiley & Sons, New York, pp. 750.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F., Tanton, M.T., 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*. 18: 371-383.

Lloyd, P., Palmer, A.R., 1998. Abiotic factors as predictors of distribution in southern African Bulbuls. *The Auk*. 115: 404-411.

Malanson, G.P., Westman, W.E., Yan, Y.-L., 1992. Realized versus fudamental niche functions in a model of chaparral response to climate change. *Ecological Modelling*. 64: 261-277.

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

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

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

Mastrorillo, S., Lek, S., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology*. 38: 237-246.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models, 2nd Edition. Chapman & Hall, London, pp. 511.

McKenzie, G.M., Busby, J.R., 1992. A quantitative estimate of Holocene climate using a bioclimatic profile of *Nothofagus cunninghamii* (Hook.) Oerst. *Journal of Biogeography*. 19: 531-540.

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

Nicholls, A.O. 1989. How to make biological surveys go further with generalised linear models. *Biological Conservation*. 50: 51-75.

Nix, H.A., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*, Australian Government Publishing Service, Canberra, pp. 4-15.

Noy-Meir, I., Austin, M.P. 1970. Principal Components ordination and simulated vegetational data. *Ecology*. 51: 551-552.

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

Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*. 116: 15-31.

Palmer, A., 1991. The potential vegetation of the upper Orange River, South Africa: concentration analysis and its application to rangeland assessment. *Coenoses*. 6: 131-138.

Palmer, A.R., Van Staden, J.M., 1992. Predicting the distribution of plant communities using annual rainfall and elevation: an example from southern Africa. *Journal of Vegetation Science*. 3: 261-266.

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

Panetta, F.D., Mitchell, N.D., 1991 b. Homoclime analysis and the prediction of weediness. *Weed Research*. 31: 273-284.

Paola, J.D., Schowengerdt, R.A., 1995. A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. *International Journal of Remote Sensing*. 16: 3033-3058.

Pearce, J., Ferrier, S., 2000 a. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*. 128: 127-147.

Pearce, J., Ferrier, S., 2000 b. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*. 133: 225-245.

Pearce, J., Lindenmayer, D., 1998. Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) in southeastern Australia. *Restoration Ecology*. 6: 238-243.

Peter, C.I., Ripley, B.S., Robertson, M.P. 2002. The distribution of *Scaevola plumieri* along the South African coast is limited by seasonal water balance and temperature. *Journal of Vegetation Science*. (in press).

Peterson, A.T., 2001. Predicting species' geographic distributions based on ecological niche modeling. *The Condor*. 103: 599-605.

Peterson, A.T., Cohoon, K.P., 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling*. 117: 159-164.

Peterson, A.T., Sanches-Cordero, V., Soberon, J., Bartley, J., Buddemeier, R.W., Navarro-Siguenza, A.G., 2001. Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling*. 144: 21-30.

Peterson, A.T., Soberon, J., Sanches-Cordero, V., 1999. Conservatism of ecological niches in evolutionary time. *Science*. 285: 1265-1267.

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

Power, M., 1993. The predictive validation of ecological and environmental models. *Ecological Modelling*. 68: 33-50.

Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A., Solomon, A.M., 1992. A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*. 19: 117-134.

Randerson, P.F. 1993. Ordination. In: Fry, J.C. (ed.), *Biological data analysis: a practical approach*. Oxford University Press, New York. Pp 173-217.

Robertson, M.P., Caithness, N., Villet, M.H., 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*. 7: 15-27.

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

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

Rogers, D.J., Williams, B.G., 1993. Tsetse distribution in Africa: seeing the wood and the trees. In: Edwards, P.J., May, R. (Eds.), *Large-scale ecology and conservation biology*, Blackwell Scientific Publications, Oxford, pp. 247-271.

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

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

Schultz, A., Wieland, R., 1997. The use of neural networks in agroecological modelling. *Computers and Electronics in Agriculture*. 18: 73-90.

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

Skidmore, A.K., Gauld, A., Walker, P., 1996. Classification of kangaroo habitat distribution using three GIS models. *International Journal of Geographical Information Systems*. 10: 441-454.

Skov, F., 2000. Potential plant distribution mapping based on climatic similarity. *Taxon*. 49: 503-515.

Skov, F., Borchsenius, F., 1997. Predicting plant species distribution patterns using simple climatic parameters: a case study of Ecuadorian palms. *Ecography*. 20: 347-355.

Sokal, R.R., Rholf, F.J., 1987. Introduction to biostatistics, 2nd Edition. W.H. Freeman and Co., New York, pp. 363.

Steele, G.R., Villet, M.H., Radloff, S.E., Hepburn, H.R., 1998. Clinal morphometric variation in wild honey bees (Hymenoptera: Apidae) in South Africa. *Diversity and Distributions*. 4: 17-25.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science*. 13: 143-158.

Swan, J.M.A. 1970. An examination of some ordination problems by use of simulated vegetation data. *Ecology*. 51: 89-102.

Vayssieres, M.P., Plant, R.E., Allen-Dias, B.H., 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*. 11: 679-694.

Verbyla, D.L., Litvaitis, J.A., 1989. Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management*. 13: 783-787.

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

Walker, P.A., Cocks, K.D., 1991. HABITAT: A procedure for modelling the disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters*. 1: 108-118.

Walton, D.W., Busby, J.R., Woodside, D.P., 1992. Recorded and predicted distribution of the Golden-tipped Bat *Phoniscus papuensis* (Dobson, 1878) in Australia. *Australian Zoologist*. 28: 1-4.

Williams, B.G., Rogers, D.J., Staton, G., Ripley, B., Booth, T., 1994. Statistical modelling of georeferenced data: mapping tsetse distributions in Zimbabwe using climate and vegetation data. In: Perry, B.D., Hansen, J.W. (Eds.),

Modelling vector-borne and other parasitic diseases, ILRAD, Nairobi, pp. 267-280.

Wilson, J.B., Rapson, G.L., Sykes, M.T., Watkins, A.J., Williams, P.A., 1992. Distributions and climatic correlations of some exotic species along roadsides in South Island, New Zealand. *Journal of Biogeography*. 19: 183-193.

Woodward, F.I., 1987. Climate and plant distribution. Cambridge University Press, Cambridge.

Woodward, F.K., Williams, B.G., 1987. Climate and plant distribution at global and local scales. *Vegetatio*. 69: 189-197.

Yee, T.W., Mitchell, N.D. 1991. Generalized additive models in plant ecology. *Journal of Vegetation Science*. 2: 587-602.

Zadeh, L.A., 1965. Fuzzy sets. *Information and Control*. 8: 338-353.

Zadeh, L.A., 1987. Fuzzy sets as a basis for a theory of possibility. In: Yager, R.R., Ovchinnikov, S., Tong, R.M., Nguyen, H.T. (Eds.), *Fuzzy sets and applications: Selected papers by L.A. Zadeh*, John Wiley and Sons, New York, pp. 193-218.

Zaniewski, A.E., Lehmann, A., Overton, J. McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*. 157: 261-280.

IV

A correlative modelling technique for predicting potential distributions of organisms from presence records using fuzzy classification

Preface

This chapter describes the first of two profile techniques that are implemented and evaluated in the thesis. This chapter is being prepared for submission to *Diversity and Distributions* (Robertson, M.P., Villet, M.H., Palmer, A.R. A correlative modelling technique for predicting potential distributions of organisms from presence records using fuzzy classification).

Abstract

A new predictive modelling technique is introduced, called the fuzzy envelope model (FEM), to predict potential distributions of organisms using presence-only locality records and a set of environmental predictor variables. The technique uses fuzzy logic to classify a set of predictor variable maps based on the values associated with presence records and combines the results to produce a potential distribution map for a target species. This technique represents several refinements of the envelope approach used in the BIOCLIM modelling package. These refinements are related to the way in which FEMs deal with uncertainty, the way in which this uncertainty is represented in the resultant potential distribution maps, and the way that these maps can interpreted and applied. The FEM technique was applied to predicting the potential distribution of three alien invasive plant species (*Lantana camara* L., *Ricinus communis* L. and *Solanum mauritianum* Scop.) and three cicada species (*Capicada decora* Germar, *Platypleura deusta* Thun. and *P. capensis* L.) in South Africa, Lesotho and Swaziland. These models were quantitatively compared with

models produced by means of the algorithm used in the BIOCLIM modelling package, which is referred to as a Crisp Envelope Model (the CEM design). The average performance of models of the FEM design was consistently higher than those of the CEM design. There were significant differences in model performance among species but there was no significant interaction between model design and species. The average maximum kappa value ranged from 0.698 to 0.900 for FEM design and from 0.565 to 0.887 for the CEM design, which can be described as "good" to "excellent" using published ranges of agreement for the kappa statistic.

Introduction

Biogeographical distribution models have been applied to a number of biological problems, and numerous examples can be found in recent reviews (Franklin, 1995; Guisan and Zimmermann, 2000). The majority of these models can be classified as *correlative*, as they rely on strong, often indirect links between species distribution records and environmental predictor variables to make predictions (Beerling *et al.*, 1995). Correlative models are an alternative to more complex mechanistic models that attempt to simulate the mechanisms considered to underlie the observed correlations with environmental attributes (Beerling *et al.*, 1995).

Correlative distribution models can be divided into two groups based on the input data used to build them. Models that use both presence and absence locality records have been termed *group discrimination techniques* and those which use only presence locality records have been termed *profile techniques* (Caithness, 1995).

Presence/absence data are typically obtained by means of systematic field surveys (Margules and Austin, 1994; Austin, 1998) that are usually expensive and time-consuming to conduct (Austin, 1998). The result is that presence-only data (obtained from museum or herbarium collections) is often the only data available for modelling (Chapter 2).

Profile modelling techniques have a conceptual base in Hutchinson's niche model (Schoener, 1990). Hutchinson's niche model consists of an abstract space defined by a set of axes, each of which represents a resource or condition of importance to the organism (Schoener, 1990). On each there will be a range of values

within which the organism can survive. This space can be generalised mathematically to include as many axes as necessary to completely characterise the species' needs, resulting in an n-dimensional hyperspace that is termed the fundamental niche (Schoener, 1990). Few organisms occupy the whole of their fundamental niche because they may be excluded from parts of it by competition or predation (Begon *et al.*, 1990). The reduced hypervolume in which the organism can survive is termed its realised niche (Schoener, 1990; Begon *et al.*, 1990).

The simplest of the profile techniques and some of the earliest models (Chicoine *et al.*, 1985) were range-based models that are referred to as *envelope* models. These models are constructed as follows. A set of locality records is used to derive values from each of a set of predictor variable maps (corresponding with each of the axes in Hutchinson's model) to produce a training data set for a target organism. On each map the contours representing the upper and lower values between which the organism can survive are plotted using the maximum and minimum values in the training set. It is assumed that the minimum and maximum values (obtained from the training set) of each variable represent the physiological limits of the target organism. The predictor variable maps are thus reclassified into new maps indicating regions of predicted presence (coded 1) and absence (coded 0) of the target organism. These maps are then superimposed using the Boolean AND function (Burrough, 1989). If the organism can survive somewhere in the map area, there will be a region where all of the "survival ranges" overlap. In areas where not all conditions are satisfactory, the organism should be absent (Chicoine *et al.*, 1985). The output consists of a binary potential distribution map indicating regions of predicted presence and absence. This approach can be referred to as a Simple Envelope Model (SEM).

A minor modification of the SEM has been implemented in a generic profile modelling package known as BIOCLIM (renamed ANUCLIM) uses an envelope approach for predicting potential distributions of species (Nix, 1986; Busby, 1991). BIOCLIM has been used extensively to predict the potential distributions of various target organisms. It has been used to predict the potential distribution of various weed species (Panetta and Mitchell, 1991a, b; Sindel and Michael, 1992), various snakes (Nix, 1986), kangaroos (Skidmore *et al.*, 1996), gliders (Jackson and Claridge, 1999), the Helmeted Honeyeater (Pearce and Lindenmayer, 1998), the Golden-tipped bat

(Walton *et al.*, 1992), Leadbeater's possum (Lindenmayer *et al.*, 1991). BIOCLIM has also been used in a climate change study (McKenzie and Busby, 1992).

In BIOCLIM, the distribution of a target organism is predicted by characterising the organism's tolerances in relation to a number of climatic parameters to produce a "climate profile" for that organism (Busby, 1991). The parameters (predictor variables) are considered to provide a broad characterisation of annual variations in temperature and levels of moisture availability (Nix, 1986). BIOCLIM predicts "core" and "marginal" environments for the organism under consideration, based on selected threshold values. Nix (1986) defined core environments as those values falling between the 5^{th} and 95^{th} percentile of each of the predictor variables, and marginal environments as those which fall outside of the core range but within the upper and lower limits of the variables, for the species. Core and marginal environments can be estimated using other reasonable values, e.g. Lindenmayer *et al.* (1991) used the 10^{th} and 90^{th} percentiles to define the core range. Nix (1986) points out that these thresholds are arbitrary. The output of BIOCLIM consists of a distribution map indicating regions of predicted presence (in terms of core and marginal habitat) and regions of predicted absence. Each of these regions is defined by classifying each of the localities in the map region into one of three Boolean or crisp sets (core, marginal or absent) based on the data in the training set. In this study, techniques that use this approach are referred to as Crisp Envelope Models (CEM).

Fuzzy Envelope Model

A refinement to the CEM technique is introduced, called the Fuzzy Envelope Model (FEM) that incorporates the notion that within a particular survival range, some conditions are more favourable than others and that the differences are continuous. This refinement uses a technique known as fuzzy classification, which is based on fuzzy set theory (Zadeh, 1965).

The use of fuzzy classification for potential distribution modelling has been investigated (Fairbanks and McKelly, 1994), although it appears not to have seen much use in this field of biology. Fuzzy classification has been used in soil science applications (Burrough, 1989; Lark and Bolam, 1997) and in remote sensing image classification techniques (Eastman, 1999).

Fuzzy classification is an approach quite closely related to expert systems that has its own algebra which is an extension of Boolean algebra (Burrough, 1989). In classical mathematical set theory, an object either belongs to a particular set or not. These sets (classes) are typically characterised by a clearly defined value or criterion and are termed crisp sets (Lark and Bolam, 1997) or Boolean sets (Burrough, 1989). This type of classification assumes that all change between classes takes place at the class boundary and that very little significant change occurs within classes (Burrough, 1989), although this is often not the case with continuous data. In cases where a clearly defined criterion or value of class membership does not exist, fuzzy set theory can be used (Zadeh, 1965; Altman, 1994). A fuzzy set has a continuum of grades of membership, allowing for situations where clearly defined class membership values are absent. A fuzzy set is described by a fuzzy membership function, with values ranging from 0 to 1, corresponding with non-membership through to complete membership (Eastman, 1999). Fuzzy sets thus have continuous membership functions and for this reason the term "continuous classification" is used by some authors instead of "fuzzy classification" (Heuvelink and Burrough, 1993). Although fuzzy membership functions may appear to be similar to probability functions, these two concepts are quite different (Zadeh, 1965): fuzzy membership functions define possibility rather than probability (Zadeh, 1987).

In practice, a set of locality records is used to derive a set of values from each of the predictor variable maps to produce a training data set. Each predictor variable map is reclassified using a fuzzy membership function. The shape of the membership function can be defined by the data in the training set. These reclassified maps (fuzzy sets) are then superimposed using fuzzy algebra (Heuvelink and Burrough, 1993) to produce a final map indicating the potential distribution of the target organism (a multivariate fuzzy set). The final potential distribution map contains a continuum of possibility values indicating conditions of varying suitability for the target organism. Localities with high possibility values are interpreted as representing more favourable conditions for the organism than those with low possibility values.

The FEM technique is explored by predicting the distribution of three alien invasive weed species and three cicada species in South Africa, Lesotho and Swaziland. In addition, an established and popular predictive modelling technique,

the CEM, is used to quantitatively compare the performance of the new FEM technique.

Methods and materials

The data

The map region for this study included South Africa, Lesotho and Swaziland (Fig. 1). Localities representing the presence of three alien invasive plants, *Lantana camara* L., *Ricinus communis* L. and *Solanum mauritianum* Scop. and three cicada species *Capicada decora* Germar, *Platypleura deusta* Thun. and *Platypleura capensis* L. were partitioned so that they could be used for model training and model evaluation respectively (Table 1). Data were partitioned randomly in a ratio of 3:1 training to testing data and this was repeated five times to ensure that different combinations of records were available for model building and evaluation. Localities representing the absence of these species were used for model evaluation but not for training. There were thus five sets of data for training (each set selected randomly from the available record set), consisting only of presence records. There were also five sets of data for evaluation, which consisted of presence and absence records.

Digital maps of environmental variables (climatic variables and altitude) developed by Schulze *et al.* (1997) were selected as predictor variables (Table 2). Each of the climatic predictor variables was interpolated from point data obtained from a network of weather recording stations distributed throughout South Africa, to produce continuous digital maps at a 1-minute spatial resolution (Schulze *et al.*, 1997). The values of the 10 predictor variables (Table 2) associated with the training localities are referred to as the training data set.

Implementation

The Crisp Envelope Model

The CEM algorithm was implemented in MATLAB and the distribution maps were visualised in IDRISI32. Distribution maps containing regions indicating the core and marginal ranges for each species were produced. The marginal range was determined by reclassifying each predictor variable map using the maximum and minimum values of the training data for each predictor variable. The reclassified maps were superimposed using the intersection (AND) function in Boolean logic (Heuvelink and Burrough, 1993) to produce a map indicating the marginal range of the species. Similarly, a second map of the core range was produced by reclassifying each predictor variable map using the 10^{th} and 90^{th} percentiles of the training data as boundaries of the core area, as defined by Lindenmayer *et al.* (1991). The core and marginal range maps were superimposed to produce a single map indicating the core and marginal ranges for each species. I used the same approach described by Nix (1986) for BIOCLIM, with the exception that the predictor variables were different.

The Fuzzy Envelope Model

The design of the Fuzzy Envelope Model was based on some of the concepts implemented in the fuzzy classification module of IDRISI32 (Eastman, 1999). The FEM program was written and implemented entirely in MATLAB.

The FEM algorithm classifies the grid cells in each predictor variable map using an appropriate sigmoidal fuzzy membership function. A sigmoidal membership function was considered to be most appropriate for describing the response of an organism to a set of environmental variables. This is because a species might be expected to exhibit tolerance over part of a climatic gradient, decreasing tolerance once a threshold has been reached, and then intolerance over the remainder (Osborne and Tigar, 1992).

The sigmoidal membership function can have symmetric, monotonically increasing or monotonically decreasing forms (Fig. 2; The equations defining these functions are listed in Appendix 1). The selection of the appropriate form (symmetric,

monotonically increasing or monotonically decreasing) of membership function is done by examining a frequency histogram of the training data for that predictor variable.

The shape of the membership function is governed by four control points (Fig. 2) which are ordered from low to high on the measurement scale of the predictor variable axis. For the symmetric membership function, point "a" marks the location where the membership function begins to rise above zero, point "b" indicates where it reaches a value of one, point "c" indicates the location where the membership grade begins to drop below one, and point "d" marks the region where it again approaches zero (Fig. 2a). In the case of the monotonically increasing function, point "a" indicates the location where the membership function rises above zero and all values on the measurement scale above point "b" take a value of one. In the case of the monotonically decreasing function, all values below point "c" on the measurement scale take on a value of one, while point "c" indicates the location where the membership grade begins to drop below one.

For the symmetric membership function, points "a" and "d" are assigned the minimum and maximum values respectively and points "b" and "c" are both assigned the median value from the training data set. This approximates the normal distribution that is often assumed to underly resource utilisation axes (Austin and Smith, 1989), but allows for skewed distributions. When the monotonically increasing function is selected then all values greater than or equal to the median are assigned a value of one. When the monotonically decreasing function is selected then all values less than or equal to the median are assigned a value of one.

Finally, all of the fuzzily classified predictor variable maps are superimposed using a minimum overlay function to produce the final distribution map. This allows the individual fuzzy sets to be intersected to produce a multivariate set (Heuvelink and Burrough, 1993; Eastman, 1999) representing the potential distribution of the organism. Both Boolean and fuzzy multivariate sets are defined by a joint membership function (JMF), which describes the combined effect of the individual membership functions (Heuvelink and Burrough, 1993). The value of the JMF is given by the minimum value of the individual membership functions (Heuvelink and Burrough, 1993) and thus a minimum overlay function is appropriate (Eastman, 1999). The minimum overlay function provides a means of combining all of the

fuzzily classified predictor variable maps into a single map that indicates where the conditions in all of these maps are suitable for the target species. Figure 3 provides a summary of the implementation of the FEM technique and describes the process used to partition records into sets for training and evaluation.

Model Evaluation

Most quantitative model performance tests are based on a confusion matrix in which the observed (actual) and predicted presence/absence patterns are cross-tabulated (Fielding and Bell, 1997). Various threshold-dependent and threshold-independent accuracy measures can be calculated from the confusion matrix (reviewed by Fielding and Bell, 1997). Threshold-independent measures (e.g. ROC curves) are considered to be more robust and more objective than threshold-dependent measures (e.g. Kappa statistics) since they do not rely on a single threshold to distinguish between predicted presence and predicted absence (Fielding and Bell, 1997). However, threshold-independent accuracy measures could not be calculated for the CEM, since it does not produce a map of continuous values. As a result, the kappa statistic (a threshold-dependent measure) was calculated using those localities reserved for model evaluation. In the case of the CEM design, the value of kappa was calculated at the threshold defined by the marginal range of the model. For FEM, kappa values were calculated at all thresholds and the maximum value of kappa (κmax.) was selected as a measure of performance for the model. The threshold at which performance was highest was thus selected for each model, which can be regarded as an optimum threshold for evaluating predictions on an independent data set (Guisan and Zimmermann, 2000). The equation for calculating the kappa statistic (Fielding and Bell, 1997) is:

$$\kappa = [(a+d)-(((a+c)(a+b)+(b+d)(c+d))/N]/[N-(((a+c)(a+b)+(b+d)(c+d))/N)]$$

where:

a= the number of cases predicted present when actually present

b= the number of cases predicted present when actually absent

c= the number of cases predicted absent when actually present

d= the number of cases predicted absent when actually absent

N= a + b + c + d

Results

Results of a 2-way ANOVA suggest that there was a significant difference in performance between models produced using the CEM and FEM designs (Table 3). The average performance of the FEM design was consistently higher than the CEM design. There were significant differences in model performance among species but no significant interaction between model design and species (Table 3).

Potential distribution maps produced using the FEM and CEM techniques for all the species appear to correspond fairly well with localities used to build these models (Fig. 4 and 5). This can be confirmed by the results of the quantitative tests of model performance. The average maximum kappa value ranged from 0.698 to 0.900 for FEM design and from 0.565 to 0.887 for the CEM design, which can be described as "good" to "excellent" using the ranges of agreement for the kappa statistic proposed by Monserud and Leemans (1992).

There is an obvious difference in the appearance of the distribution map produced by the FEM design compared with the CEM design. Distribution maps produced from the FEM design have a continuous grade of values while those produced from the CEM have only two categories, the core and marginal range (Fig. 4 and 5).

There was very good visual agreement between maps generated from the FEM and CEM designs for *R. communis* and *S. mauritianum* (Fig. 4). The maximum kappa value (κmax) was the same for the CEM and FEM designs for *S. mauritianum*, but the FEM design had a higher κmax value than the CEM model for *R. communis*. There was less visual agreement between the FEM and CEM designs for *L. camara* and *P. deusta*, and a considerable difference for *C. decora* (Fig. 5). The FEM predicted the presence of *L. camara* along the western border of the Northern Province, while the CEM predicted it to be absent in this region. The κmax value was only slightly higher for the FEM design (0.648) than the CEM design (0.646). The CEM design appeared to make a much more conservative prediction than the FEM design in the case of *C. decora*. The κmax value was higher for the FEM design (0.814) than the CEM design

(0.621), indicating better performance. For *P. deusta*, the FEM design predicted the species as being present in the southern and western regions of Lesotho, the northern regions of the Eastern Cape and in parts of the south-eastern Free State, whereas it was predicted as being absent in these regions by the CEM design. Again, the κmax value was higher for the FEM design (0.824) than the CEM design (0.438), indicating much better performance. In general, the core ranges of the CEM maps corresponded fairly well with areas of high possibility in the FEM maps.

There was fairly good visual agreement between the distribution maps produced for *L. camara*, *R. communis* and *S. mauritianum* using the FEM technique (Fig. 4) and models produced using an independent PCA-based modelling technique (Robertson *et al.*, 2001).

Discussion

Performance of FEMs

Models of the FEM design performed on average significantly better than those produced using the CEM design. For the selected distribution maps examined (Fig. 4 and 5) there was either no difference in performance between the FEM and CEM designs or the FEM design performed better than the CEM design for models produced for the same species (see individual κmax values). These results suggest that the FEM design is capable of performing as well or better than the CEM design. In addition, the average performance of models of both the FEM and CEM designs was "good" to "excellent" using the ranges of agreement for the kappa statistic proposed by Monserud and Leemans (1992).

Design features of FEMs

One of the most important features of the design of FEMs is that various fuzzy membership functions are used to construct a model for a target species. Different forms of the membership function (symmetrical, monotonically increasing or monotonically decreasing) may be appropriate for expressing the response of the

target species to a particular predictor variable. For example, a species that is intolerant of frost would survive at localities where no frost occurs or where there are very few days of frost per year, but it would not survive at localities with moderate to large numbers of frost days per year. A monotonically decreasing membership function would be appropriate to express this response. For this function, the possibility value is highest when the number of frost days are low (values less than the median) and decreases as the number of frost days increases (values greater than the median) to a minimum possibility value corresponding with the maximum value in the training set (d).

A monotonically increasing membership function would be appropriate for expressing the response of a species that is associated with high altitudes. In this example, survivorship of the species increases with increasing altitude with no decrease in survivorship at very high altitudes. For a monotonically increasing membership function, possibility values are lowest when the altitude is lowest (corresponding with the minimum value of the training set, a) increasing to a maximum at higher altitudes (corresponding with the median value of the training set, b).

A similar example of the response of a species to increasing altitude can be used to illustrate the use of a symmetric membership function. In this example, survivorship of the species increases with increasing altitude up to some maximum threshold, beyond which survivorship would again decline at very high altitudes. A symmetric membership function would be appropriate to express this response as possibility values decrease for values larger than the median (c) to a minimum possibility value when the maximum training set value is reached (d).

The higher average performance of the FEM models over the CEM models is largely due to the greater flexibility of function selection of the FEM models and that it is possible to select an optimal threshold for the FEM but not for the CEM. When all of the functions selected in the FEM are symmetric then the ranges of an FEM and CEM model built on the same data will coincide e.g. Fig. 4 e & f. It is possible for FEMs to predict wider species ranges than CEMs. This occurs when a monotonically increasing or monotonically decreasing membership function is selected in the FEM for the predictor variable that is most limiting to the target species e.g. Fig. 5 c & d. The ability to vary the threshold allows the predicted range of a FEM model to be

reduced or expanded. The higher the suitability value at which the threshold is selected, the greater the reduction in the predicted range. A reduction in the predicted range results in a more conservative prediction.

Fuzzy sets vs. crisp sets

Fuzzy sets (used in FEM) rather than crisp sets (used in CEM) may be more appropriate for building envelope distribution models for a number of reasons. A locality with attributes that place it close to the threshold between two crisp classes (e.g. the core-marginal threshold) could be assigned to either class depending on the values of the thresholds in the training set. Heuvelink and Burrough (1993) suggest that it is not sensible to use crisp sets to classify continuous variables (the predictor variables) because attributes with very similar values may be assigned to different classes which have very different meaning e.g. core and marginal.

It is often difficult or inappropriate to define attributes in terms of exact thresholds and when this is possible there is often uncertainty as to the exact values of these thresholds (Heuvelink and Burrough, 1993). In such cases, fuzzy methods are appropriate because they are designed to handle this inexactness and uncertainty in a definable way (Burrough, 1989). In bioclimatic modelling this uncertainty arises because distribution models tend to predict the "average" distribution of a species because climatic variables used as predictors are usually calculated using long-term means (Dent *et al.*, 1989; Schulze *et al.*, 1997). The locality records used to predict the potential distribution of an organism are usually collected over a period of time (usually several years or seasons). Inter-annual species range expansion or contraction may occur due to such factors as resource fluctuations or disturbances mediated by certain events e.g. El Niño climate shifts (Hayward, 1997). This is likely to alter the values of the thresholds defining the distribution in the training set, depending on the temporal period during which the locality records were collected. Uncertainty also arises due to measurement and interpolation errors (Heuvelink and Burrough, 1993). There is thus some uncertainty as to the exact value of these thresholds and hence the spatial extent of the core and marginal ranges of the target organism. I suggest that fuzzy classes more realistically represent this "average" potential distribution and the uncertainty associated with the data than crisply defined

techniques do. In addition, error propagation through models is reduced when continuous classes (fuzzy classes) are used rather than crisply defined classes (Heuvelink and Burrough, 1993).

Advantages of a continuous output

Through the use of fuzzy classification (fuzzy sets) a continuous output can be produced in the resultant distribution map. A continuous output allows one to calculate the kappa value for model evaluation at an optimal, rather than an arbitrary, threshold for that model. This is likely to be particularly important when comparing among model designs, species or sample sizes as it appears unlikely that a single arbitrary threshold has the same meaning among different model designs, species or sample sizes. It has been suggested that distribution maps produced from different modelling techniques (model designs) have different meanings (Zaniewski *et al.*, 2002; Chapter 8).

Once the optimal threshold has been calculated then the continuous distribution map can be objectively reclassified into a categorical distribution map containing only two classes (e.g. present and absent), which is often useful for further analysis and interpretation (Guisan and Zimmermann, 2000).

A continuous representation of the predicted distribution can be used in the final distribution map produced by the model to indicate to the user that there is a level of uncertainty in the prediction. The interpretation of fuzzy potential distribution maps by managers is likely to be different from the interpretation of binary potential distribution maps or maps with crisply defined core and marginal ranges. Fuzzy maps more realistically display the uncertainty associated with the input data used to generate them.

I suggest that the continuous output of the FEM design provides more scope for interpreting the predicted distribution of the target organism in terms of its biology than the output of the CEM design. Interpretation of the predicted distribution can be done in the following way. If one has data on the relative performance of the target species at a set of localities in the map region, such as density or fecundity, then the values in the model associated with those localities can be extracted. The relationship between the predicted values (extracted from the model) and the species performance

measure can then be used to reclassify the continuous distribution map to produce a new map based on a feature of the biology of the target species. For example this approach was used to reclassify continuous maps of potential distribution predictions made for a number of biocontrol agents (insects) released in South Africa for the control of an invasive weed (*Lantana camara*; Baars, 2002). Data on the level of damage caused to the weed and the abundance of these agents at a number of localities were used as a measure of species performance for reclassifying the original continuous distribution maps into categorical maps to facilitate further analysis of the data. A further advantage of the FEM technique is that the individual fuzzily classified predictor variable maps that are used to produce the final distribution map can be examined and interpreted.

Criticisms of CEMs and FEMs

Various criticisms, have been levelled at CEMs like BIOCLIM (Carpenter *et al.*, 1993). One of these criticisms is that BIOCLIM does not account for interactions among predictor variables and each predictor variable axis is treated independently (Carpenter *et al.*, 1993). Techniques that are based on multivariate statistics such as PCA (Robertson, *et al.* 2001, Erasmus *et al.*, 2000), discriminant analysis (Rogers and Williams, 1993; Rogers *et al.*, 1996; Robinson *et al.*, 1997) and logistic regression (Austin *et al.*, 1984; Osborne and Tigar, 1992; Cumming, 2000 a & b) take the multivariate structure of the data into consideration and may thus produce better results than the CEM. This criticism could also be levelled at the FEM technique because the predictor variables are also treated independently. A second criticism of both the CEM and FEM designs is the implicit assumption that all the predictor variables are equally important in predicting (or determining) the distribution of the target species. This is likely to be unrealistic and may lead to inaccuracies in predictions. Despite these criticisms, the models generated using the FEM and CEM techniques appear to have performed well.

Although criticisms can be levelled at the FEM technique, it appears to be useful and to produce reasonable results but it should be compared quantitatively with other competing techniques. Such quantitative comparisons would enable one to determine the absolute and relative performance of competing techniques under a

range of conditions. These comparisons help to establish when a given technique may be most useful and or reliable.

A major advantage of the FEM design is that it requires presence only locality data and does not rely on absence data, as required by many other multivariate techniques. The FEM design may be particularly suited to predicting distributions of species for which absence data are either not available or are unreliable, such as alien species. FEMs deliver credible results and they represent refinements to the CEM approach used in the BIOCLIM modelling package. These refinements are related to the way in which FEMs deal with uncertainty, the way in which this uncertainty is represented in the resultant potential distribution maps, and the way that these maps can be interpreted and applied.

Acknowledgements

I thank the School of Bioresources Engineering and Environmental Hydrology (University of Natal), the Water Research Commission and the South African Country Study for Climate Change for the use of the climatic predictor variables; Lesley Henderson at the Southern African Plant Invaders Atlas for locality data; Craig Peter for collecting locality data, The National Botanical Institute for the use of data from the National Herbarium, Pretoria (PRE) Computerised Information System (PRECIS). This work was funded by the National Department of Agriculture, Directorate of Agricultural Land Resource Management (previously the Directorate of Resource Conservation) and by the National Research Foundation.

Table 1. The number of presence (Pres.) and absence (Abs.) localities used for model training (Train.) and model evaluation (Eval.) of CEM and FEM models. For both the CEM and FEM techniques only localities representing the presence of the target species are used to train the models.

	Train.	Eval.	
Species	Pres.	Pres.	Abs.
Lantana camara	322	64	46
Ricinus communis	237	47	30
Solanum mauritianum	324	65	33
Capcicada decora	27	5	6
Platypleura deusta	78	16	16
Platypleura capensis	23	5	7

Table 2. Predictor variables selected for building the distribution models.

No.	Predictor variable
1	Monthly potential evaporation - January
2	Monthly potential evaporation - July
3	Monthly maximum temperature - January
4	Monthly minimum temperature - July
5	Monthly rainfall – January
6	Monthly rainfall – April
7	Monthly rainfall – July
8	Monthly rainfall – October
9	Number of days with frost
10	Digital elevation model

Table 3. Results of a 2-way ANOVA. Model performance is the mean of five maximum kappa values (based on 5 replicates) for each species.

	df	SS	MS	F	p-level
Model design	1	0.058	0.058	5.017	0.03
Species	5	0.411	0.082	7.136	0
Model*species	5	0.056	0.011	0.976	0.442
Residuals	48	0.553	0.012		

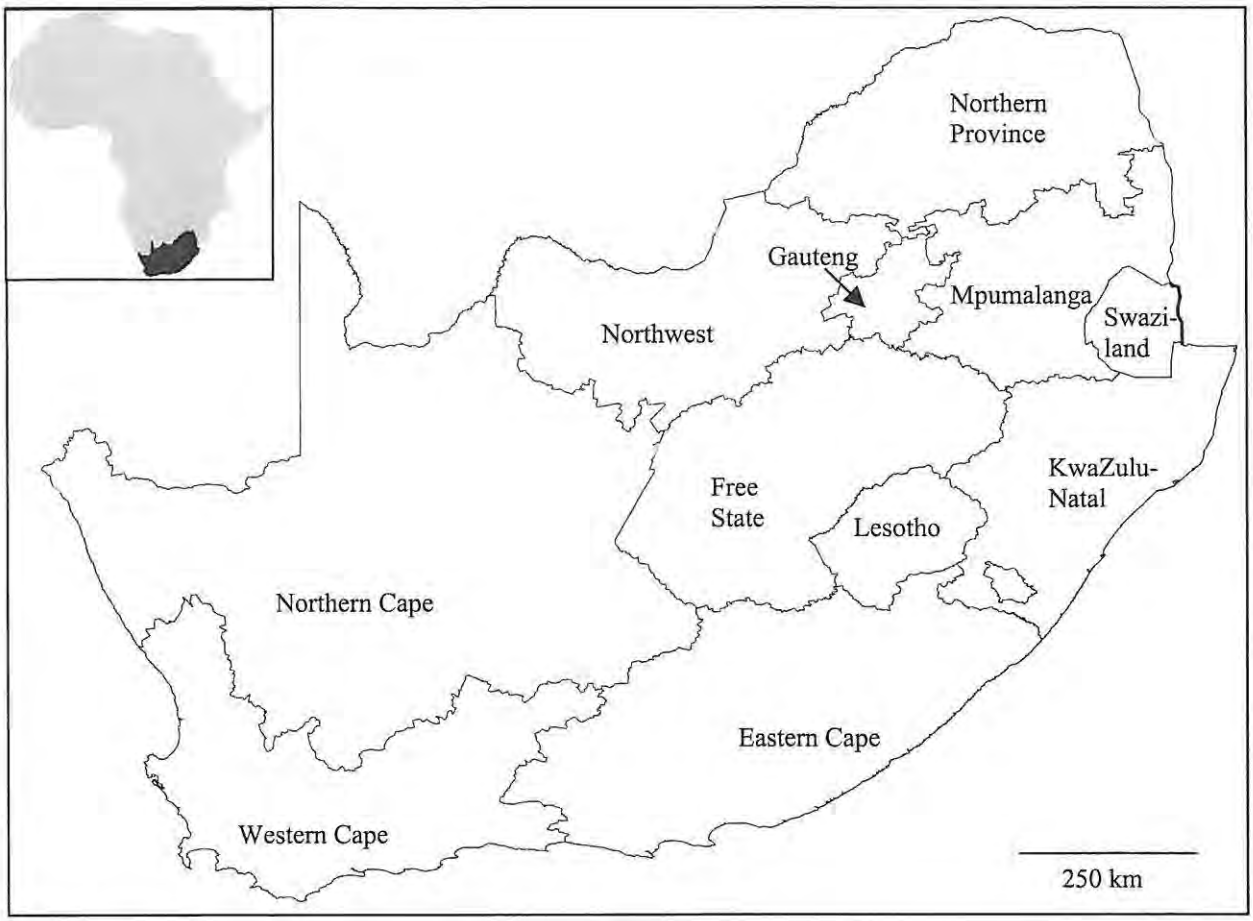


Figure 1. The map region, including South Africa (with province boundaries shown), Lesotho and Swaziland. In the inset, black indicates southern Africa relative to Africa.

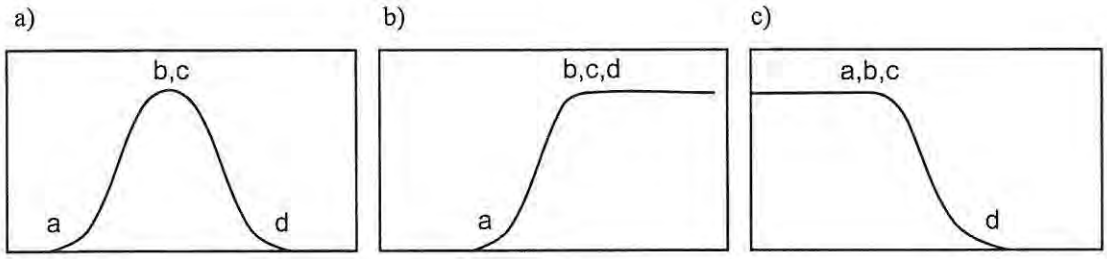


Figure 2. Sigmoidal fuzzy membership functions: (a) a symmetric membership function; (b) a monotonically increasing membership function and (c) a monotonically decreasing membership function (from Eastman, 1999). Control points are indicated by the letters a to d.

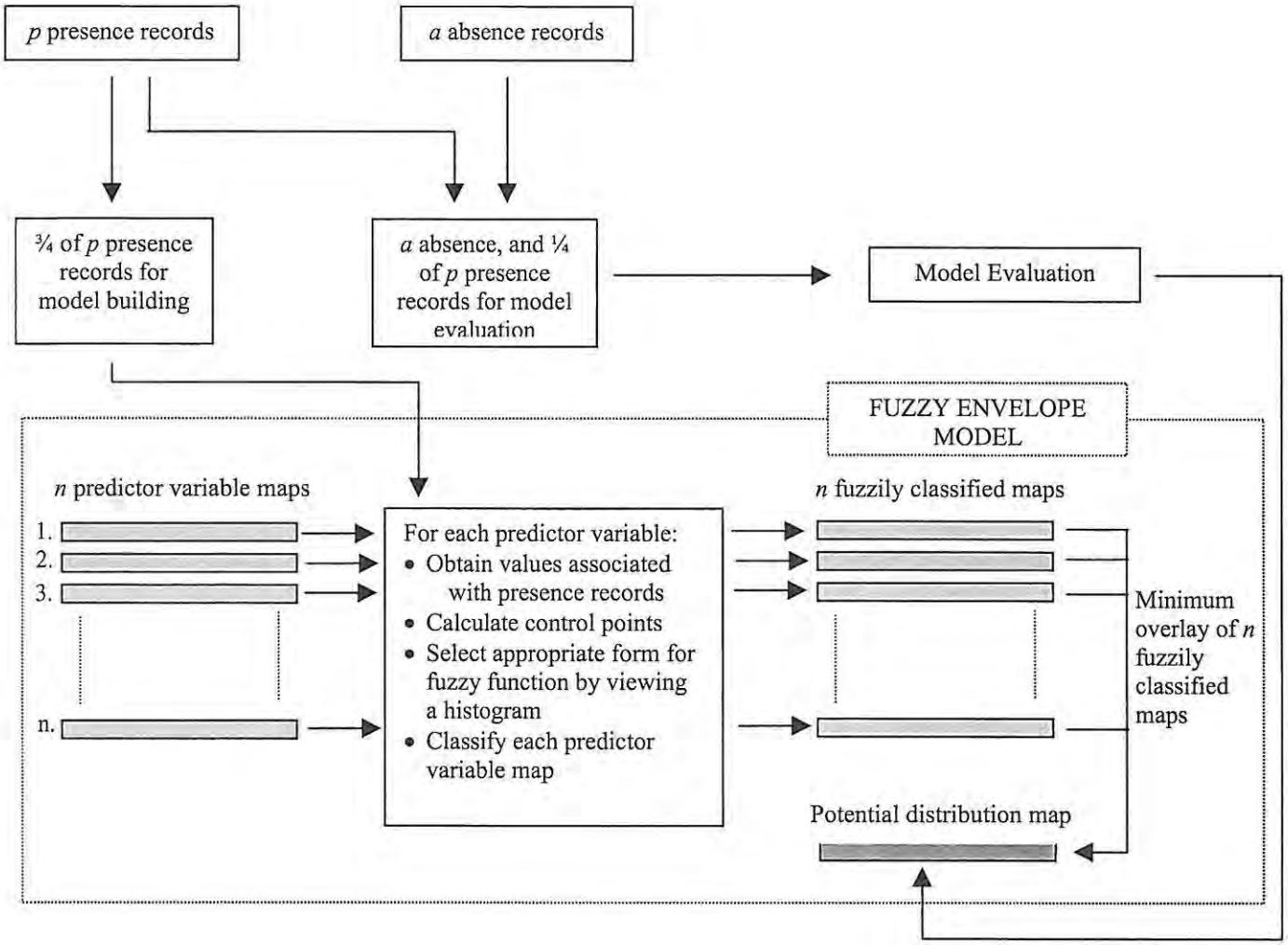


Fig. 3. Implementation of the Fuzzy Envelope Model (FEM) indicating how the original set of locality records were partitioned into a training and a testing (evaluation) set. The components of the FEM appear in the box. The procedure shown in this figure describes how a single potential distribution prediction is made. This procedure was repeated 5 times for each species so that different combinations of presence records could be used for model training and evaluation.

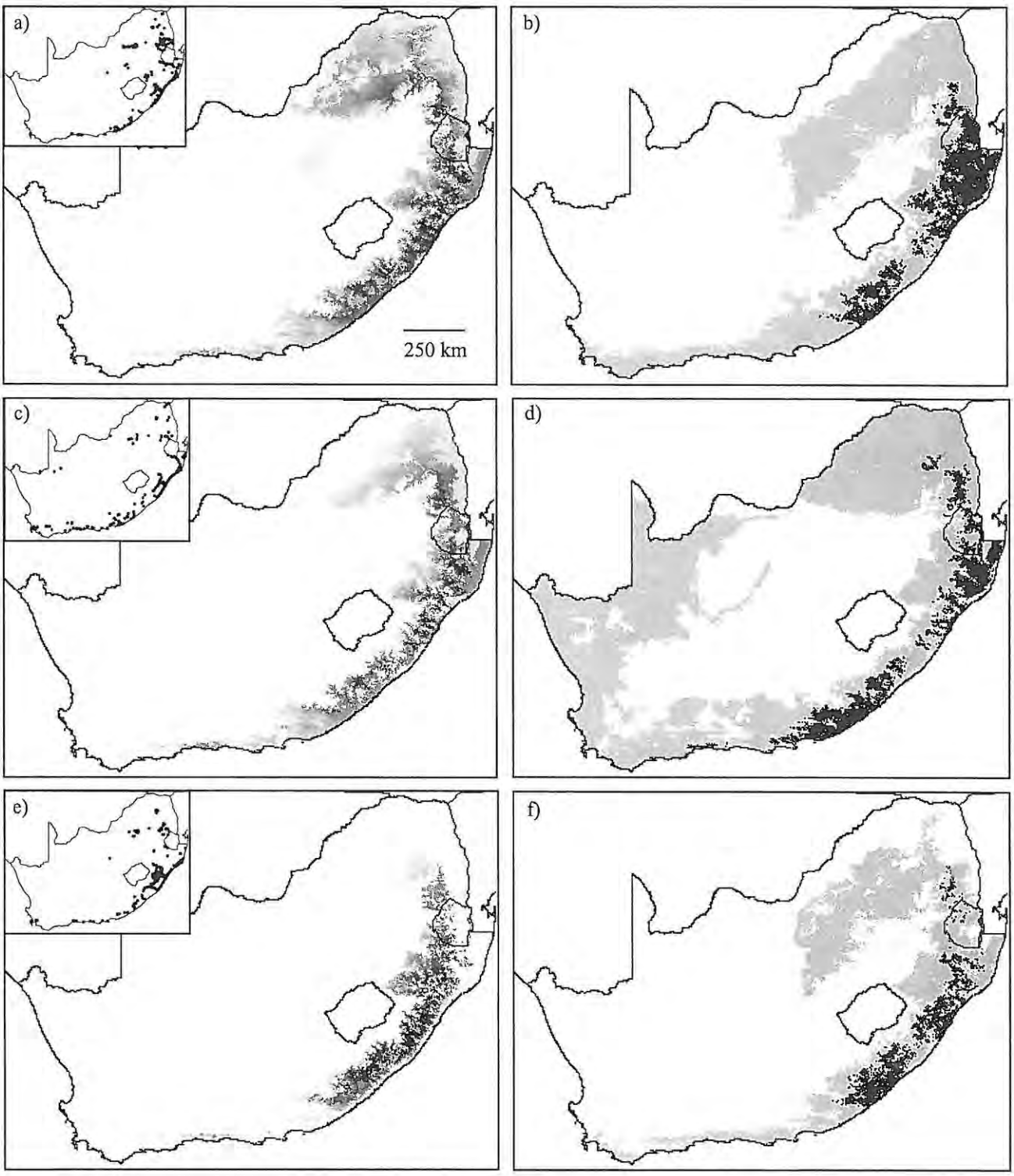


Figure 3. Potential distribution maps for three alien plant species generated using FEM and CEM predictive modelling techniques; a. *L. camara* FEM (κmax = 0.648, n = 322); b. *L. camara* CEM (κmax = 0.646); c. *R. communis* FEM (κmax = 0.864, n = 237); d. *R. communis* CEM (κmax = 0.582); e. *S. mauritianum* FEM (κmax = 0.739, n = 324); f. *S. mauritianum* CEM (κmax = 0.739). All localities representing the presence of the species, which were available for model building and model evaluation, appear in the insets. The number of presence localities used to build each model (n) and the maximum kappa value (κmax) calculated for the model are listed.

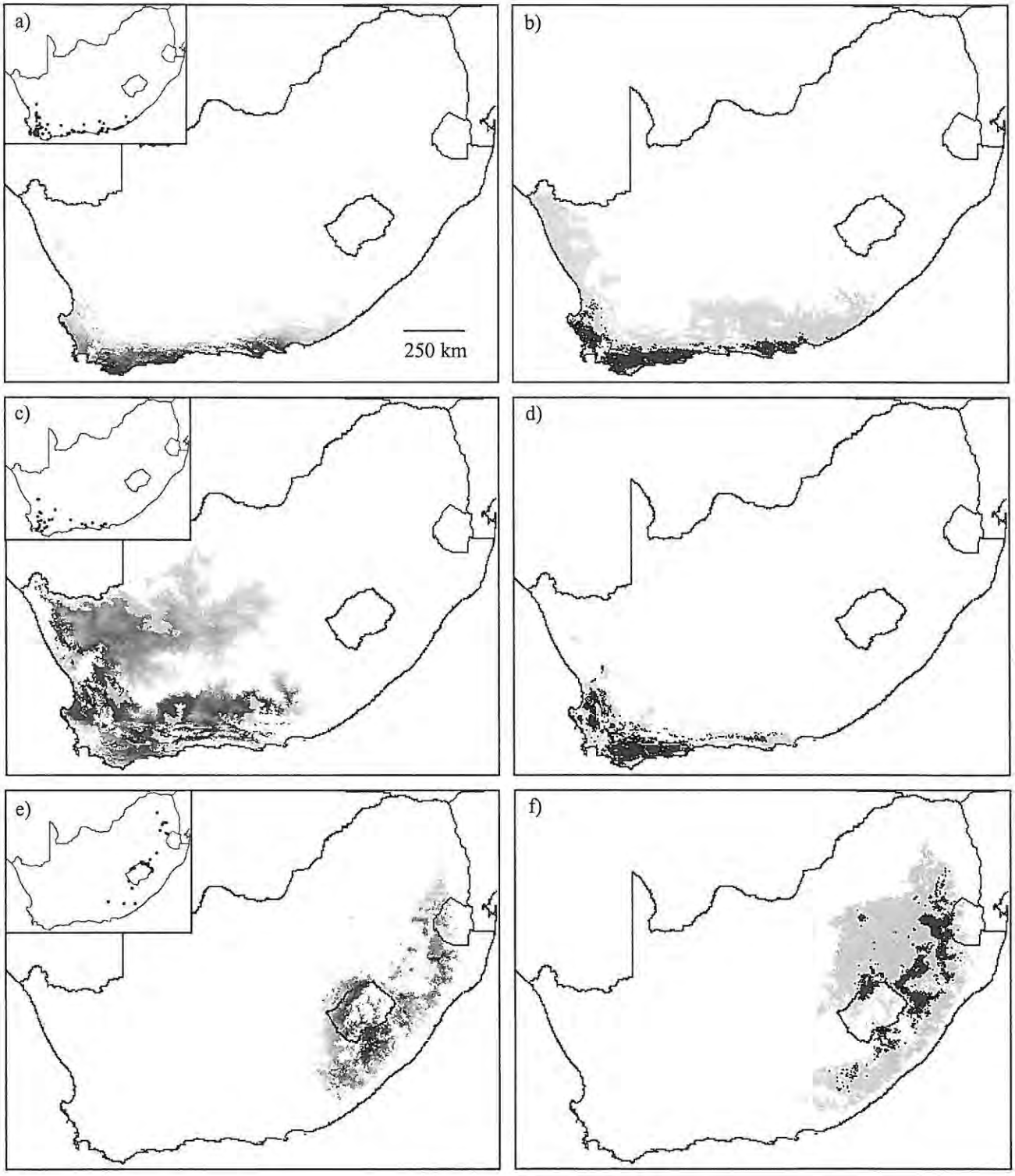


Figure 4. Potential distribution maps for three cicada species generated using FEM and CEM predictive modelling techniques; a. *P. capensis* FEM (κmax = 0.938, n = 23); b. *P. capensis* CEM (κmax = 0.875); c. *C. decora* FEM (κmax = 0.814, n = 27); d. *C. decora* CEM (κmax = 0.621); e. *P. deusta* FEM (κmax = 0.824, n = 78); f. *P. deusta* CEM (κmax = 0.438). All localities representing the presence of the species, which were available for model building and model evaluation, appear in the insets. The number of presence localities used to build each model (n) and the maximum kappa value (κmax) calculated for the model are listed.

References

Altman, D., 1994. Fuzzy set theoretic approaches for handling imprecision in spatial analysis. *International Journal of Geographical Information Systems*. 8: 271-289.

Austin, M.P. 1998. An ecological prespective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*. 85: 2-17.

Austin, M.P., Cunningham, R.B., Fleming, P.M., 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio*. 55: 11-27.

Austin, M.P., Smith, T.M., 1989. A new model for the continuum concept. *Vegetatio*. 83: 35-47.

Baars, J-R, 2002. Biological control initiatives against *Lantana camara* L. (Verbenaceae) in South Africa: an assessment of the present status of the programme, and an evaluation of *Coelocephalapion camarae* Kissinger (Coleoptera: Brentidae) and *Falconia intermedia* (Distant) (Hemiptera: Miridae), two new candidate natural enemies for release on the weed. Ph.D. Thesis, Rhodes University, Grahamstown.

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

Begon, M., Harper, J.L., Townsend, C.R., 1990. Ecology - Individuals, Populations and Communities, second Edition. Blackwell Scientific Publications, Oxford, pp. 945.

Burrough, P.A., 1989. Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science*. 40: 477-492.

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

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

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

Chicoine, T.K., Fay, P.K., Nielsen, G.A., 1985. Predicting weed migration from soil and climate maps. *Weed Science*. 34: 57-61.

Cumming, G.S., 2000 a. Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*. 27: 425-440.

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

Dent, M.C., Lynch, S.D., Schulze, R.E., 1989. Mapping mean annual and other rainfall statistics in southern Africa. 1st Edition. Department of Agricultural Engineering, Pietermaritzburg, pp. 250.

Eastman, J.R., 1999. Guide to GIS and image processing, Volume 2. Clark Labs, Worcester, pp. 170.

Erasmus, B.F.N., Kshatriya, M., Mansell, M.W., Chown, S.L., Van Jaarsveld, A.S., 2000. A modelling approach to antlion (Neuroptera: Myreleontidae) distribution patterns. *African Entomology*. 8: 157-168.

Fairbanks, D.H.K., McKelly, D., 1994. Investigating fuzzy set classification methods for use in GIS decision support modelling to determine land suitability. FOR-I 515, Division of Forest Science and Technology, CSIR.

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

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*. 19: 474-499.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

Hayward, T.L., 1997. Pacific ocean climate change: atmospheric forcing, ocean circulation and ecosystem response. *Trends in Ecology and Evolution*. 12: 150-154.

Heuvelink, G.B.M., Burrough, P.A., 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*. 7: 231-246.

Jackson, S.M., Claridge, A., 1999. Climatic modelling of the distribution of the mahogany glider (*Petaurus gracilis*), and the squirrel glider (*P. norfolcensis*). *Australian Journal of Zoology*. 47: 47-57.

Lark, R.M., Bolam, H.C., 1997. Uncertainty in prediction and interpretation of spatially variable data on soils. *Geoderma*. 77: 263-282.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F., Tanton, M.T., 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*. 18: 371-383.

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

McKenzie, G.M., Busby, J.R., 1992. A quantitative estimate of Holocene climate using a bioclimatic profile of *Nothofagus cunninghamii* (Hook.) Oerst. *Journal of Biogeography*. 19: 531-540.

Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*. 62: 275-293.

Nix, H.A., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*, Australian Government Publishing Service, Canberra, pp. 4-15.

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

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

Panetta, F.D., Mitchell, N.D., 1991 b. Homoclime analysis and the prediction of weediness. *Weed Research*. 31: 273-284.

Pearce, J., Lindenmayer, D., 1998. Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) in southeastern Australia. *Restoration Ecology*. 6: 238-243.

Robertson, M.P., Caithness, N., Villet, M.H., 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*. 7: 15-27.

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

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

Rogers, D.J., Williams, B.G., 1993. Tsetse distribution in Africa: seeing the wood and the trees. In: Edwards, P.J., May, R. (Eds.), *Large-scale ecology and conservation biology*, Blackwell Scientific Publications, Oxford, pp. 247-271.

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

Schulze, R.E., Maharaj, M., Lynch, S.D., Howe, B.J., Melvil-Thomson, B., 1997. South African Atlas of agrohydrology and climatology, 1st Edition. Water Research Commission, Pretoria.

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

Skidmore, A.K., Gauld, A., Walker, P., 1996. Classification of kangaroo habitat distribution using three GIS models. *International Journal of Geographical Information Systems*. 10: 441-454.

Walton, D.W., Busby, J.R., Woodside, D.P., 1992. Recorded and predicted distribution of the Golden-tipped Bat *Phoniscus papuensis* (Dobson, 1878) in Australia. *Australian Zoologist*. 28: 1-4.

Zadeh, L.A., 1965. Fuzzy sets. *Information and Control*. 8: 338-353.

Zadeh, L.A., 1987. Fuzzy sets as a basis for a theory of possibility. In: Yager, R.R., Ovchinnikov, S., Tong, R.M., Nguyen, H.T. (Eds.), *Fuzzy sets and applications: Selected papers by L.A. Zadeh*, John Wiley and Sons, New York, pp. 193-218.

Zaniewski, A.E., Lehmann, A., Overton, J. McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*. 157: 261-280.

Appendix

The following equations describe the monotonically increasing and monotonically decreasing sigmoidal membership functions, from which the symmetric membership function is comprised. These equations are identical to those used in the fuzzy classification module in IDRISI32 (Eastman, 1999).

The monotonically increasing function:

$$\mu = \cos \alpha^2 \qquad \alpha = (1 - (x - \text{point a})/(\text{point b} - \text{point a})) * pi/2$$

When x > point b, $\mu = 1$.

μ = the fuzzy possibility value of a given grid-cell in a particular predictor variable map
x = the value of the predictor variable in a given grid-cell
point a and point b refer to control points a and b (Fig. 2), the minimum and median values, respectively of the training set.

The monotonically decreasing function:

$$\mu = \cos \alpha^2 \qquad \alpha = (x - \text{point c})/(\text{point d} - \text{point c}) * pi/2$$

When x < point c, $\mu = 1$.

μ = the fuzzy possibility value of a given grid-cell in a particular predictor variable map
x = the value of the predictor variable in a given grid-cell
point c and point d refer to control points c and d (Fig. 2), the median and maximum values, respectively of the training set.

V

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

Preface

This chapter describes the second of two profile techniques that are implemented and evaluated in the thesis. This chapter was published in *Diversity and Distributions* in January 2001 (Robertson, M.P.; Caithness, N., Villet, M.H. 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*. 7: 15-27).

Abstract

A correlative modelling technique that uses locality records (associated with species presence) and a set of predictor variables to produce a statistically justifiable probability response surface for a target species is presented. The probability response surface indicates the suitability of each grid cell in a map for the target species in terms of the suite of predictor variables. The technique constructs a hyperspace for the target species using principal component axes derived from a principal components analysis performed on a correlation matrix derived from a training data set. The training data set comprises the values of the predictor variables associated with the localities where the species has been recorded as present. The origin of this hyperspace is taken to characterise the centre of the niche of the organism. All the localities (grid-cells) in the map region are then fitted into this hyperspace using the values of the predictor variables at these localities (the prediction data set). The Euclidean distance from any locality to the origin of the hyperspace gives a measure of the "centrality" of that locality in the hyperspace.

These distances are used to derive probability values for each grid cell in the map region. The modelling technique was applied to bioclimatic data to predict bioclimatic suitability for three alien invasive plant species (*Lantana camara* L., *Ricinus communis* L. and *Solanum mauritianum* Scop.) in South Africa, Lesotho and Swaziland. The models were tested against independent test records by calculating Area Under Curve (AUC) values of Receiver Operator Characteristic (ROC) curves and Kappa statistics. There was good agreement between the models and the independent test records. The pre-processing of climatic variable data to reduce the deleterious effects of multicollinearity, and the use of stopping rules to prevent overfitting of the models are important aspects of the modelling process.

Introduction

In response to the needs of environmental managers, a wide variety of recent biogeographical distribution models have been applied to a selection of biological problems. They have been used to predict the potential distribution of problem organisms such as weeds (Panetta and Dodd, 1987; Panetta and Mitchell, 1991; Sindel and Michael, 1992; Beerling *et al.*, 1995), and disease vectors, including Tsetse flies (Rogers and Williams, 1993; Rogers *et al.*, 1996; Robinson *et al.*, 1997) and ticks (Rogers and Randolph, 1993; Cumming, 2000). They have been used to assess the potential impacts of climate change on species distributions (Rogers and Randolph, 1993; Lindenmayer *et al.*, 1991; Beerling *et al.*, 1995; Schulze and Kunz, 1995), and to determine where new populations of a threatened species could be established (Pfab and Witkowski, 1997) or where extinct populations may have occurred (Bauer *et al.*, 1994). These models have found applications in conservation (Lindenmayer *et al.*, 1991; Osborne and Tigar, 1992; Austin *et al.*, 1996; Lloyd and Palmer, 1998) and ecoclimatic site matching of forestry species (Richardson and McMahon, 1992).

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

discrimination techniques, which use both presence and absence locality records, or *profile techniques*, which use only presence locality records (Caithness, 1995).

Examples of group-discrimination techniques include those models based on discriminant analysis (Rogers and Randolph, 1993; Rogers and Williams, 1993; Rogers *et al.*, 1996), logistic regression (Osborne and Tigar, 1992; Cumming, 2000; Higgins *et al.*, 1999) and decision-tree-based methods (Walker, 1990; Lees, 1994; Michaelsen *et al.*, 1994; Williams *et al.*, 1994). Examples of profile techniques include the approaches used in the modelling packages known as BIOCLIM (Nix, 1986; Busby, 1991) and DOMAIN (Carpenter *et al.*, 1993).

A profile technique for predicting suitability based on principal components analysis is described. The technique is then used in combination with climatic predictor variables to illustrate the prediction of bioclimatic suitability for three alien plant species in South Africa, Lesotho and Swaziland.

The PCA technique

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

PCA has traditionally been used as a mathematical technique for dimension reduction (James and McCulloch, 1990). It has seen considerable use in vegetation science, for the purposes of ordination (Randerson, 1993; Jongman *et al.*, 1995), where it is applied to species compositional data obtained from quadrats. It has also been used in genetics and morphometrics (James and McCulloch, 1990), mainly for the purposes of making quantitative comparisons and for dimension reduction. The application of PCA described here differs from its traditional use in that it is used here for prediction.

A map of the area for which one wants to predict an organism's distribution is subdivided into regular grid cells. This allows the map to be represented as a matrix of values. The values of the predictor variables associated with grid-cells in which the target organism has been recorded as present are referred to as the "training data set".

The values of the predictor variables associated with all the grid-cells within the map region comprise the "prediction data set" (i.e. the training data set plus the values associated with the remaining unsampled grid-cells).

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

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

This distance can be used to calculate a probability of bioclimatic suitability for each locality (grid-cell) as follows. Based on the assumption that the fundamental niche of an organism is generally considered to follow a broad Gaussian curve (Austin and Meyers, 1996), a normal distribution would be most appropriate for this purpose. As the distance of a point from the origin of the hyperspace is calculated from the sum of its squared component scores, and as the sum of squares of *n* standard normal random variates is distributed as chi-square with *n* degrees of freedom (Sokal and Rohlf, 1987), a chi-square distribution can be used instead of a normal distribution. This assumes that the further a point is from the origin of the hyperspace, the less suitable it is for the target species. The probability associated with each chi-square value can thus be determined by referring to a chi-square distribution (a chi-square

distribution is equivalent to a squared normal distribution). These values can be mapped back to the cells of the original real-world map. An output of the model is therefore a map of grid-cells, with each grid-cell containing a probability value, and these probability values can be interpreted as an indication of the suitability of that grid-cell for the target organism.

Methods

The target species

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

The data

Data sources

Digital predictor variable maps (climatic variables and altitude) of South Africa, Lesotho and Swaziland developed by Schulze *et al.* (1997) were selected for the purpose of illustrating this method of predictive modelling. Each of the climatic predictor variables was interpolated from point data obtained from a network of weather recording stations distributed throughout South Africa, to produce continuous digital maps at a resolution of 60 pixels per degree (Schulze *et al.*, 1997). Localities representing species presence were obtained from the Southern African Plant Invader's Atlas (Henderson, 1998) and the National Herbarium's Computerised Information System (PRECIS). Additional records of presence or absence were collected, using a GPS, on road transects selected to sample major climatic gradients

represented in the map region. If a target species occurred continuously along any part of a transect then its position was recorded approximately every 2 to 4 km to represent that species' presence. Absence records were only recorded if they were at least more than 10 km from any presence localities. The absence records were used only for model assessment and not for model building. The presence data used to predict the distribution of an organism obviously represent records collected from that organism's realised niche.

Climatic variable pre-processing

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

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

Locality data

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

Implementation

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

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

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

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

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

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

deviations of these estimates were used in these models (Fig. 1), following Caithness (1995).

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

Model assessment

In order to have confidence in a predictive model or in the approach used to build it, the model's predictions should be assessed by some objective means. This is usually done quantitatively using a set of independent testing locality records and an accuracy assessment measure. Fielding and Bell (1997) reviewed a number of model assessment measures for quantitatively assessing a model's prediction success. One of the most robust measures described by them is derived from a Receiver Operator Characteristic (ROC) Plot.

ROC Plots

If those testing localities where a target species has been recorded as present are termed "positives" and those localities where it has been recorded as absent are termed "negatives", then sensitivity is defined as the probability that the model produces a positive result in a positive locality and specificity is the probability that the model produces a negative result in a negative locality (Table 2). A ROC plot is

obtained by plotting all sensitivity values on the *y*-axis against their equivalent (1-specificity) values for all available decision thresholds on the *x*-axis (Fielding and Bell, 1997). The area under the ROC function (AUC) provides a single measure of overall accuracy that is not dependent on a particular decision threshold (Fielding and Bell, 1997). The value of the AUC ranges between 0.5 and 1, where 0.5 indicates randomness and 1 indicates a perfect fit.

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

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

Results

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

Regions of high bioclimatic suitability for *Lantana camara* include the coastal regions of the Eastern Cape, parts of KwaZulu-Natal, Mpumalanga, Gauteng, Northern Province and Swaziland (Figs. 2 and 4). The Free State Province, Lesotho,

North-West, Northern Cape and Western Cape provinces demonstrate low bioclimatic suitability. The regions of high suitability correspond approximately with the Savanna (excluding the Kalahari Thornveld) and Forest biomes (Low and Rebelo, 1996). Those areas of lower suitability appear to be associated with the Grassland biome (Low and Rebelo, 1996). *Lantana camara* is reported to invade forests, plantation margins, savanna and watercourses (Henderson, 1995) which would explain the correspondence between the model's predictions and the Savanna and Forest biomes. An AUC value of 0.991 was calculated for this species (using 78 presence and 172 absence localities) which indicates a good fit between the distribution predicted by the model and the independent test localities. An AUC value of 0.991 indicates that in 991 out of 1000 cases, random selection of a point from the group of known occurrences will be associated with a probability that is greater than that of a random selection from the negative group (Fielding and Bell, 1997). A Kappa value of 0.909 was calculated, which can be considered to indicate 'excellent' agreement between the model and the test data (Monserud and Leemans, 1992).

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

Regions of high bioclimatic suitability for *Solanum mauritianum* include the higher altitude regions of Eastern Cape, KwaZulu-Natal, Mpumalanga and Swaziland

(Fig. 6). The coastal regions appear to be less suitable for this species than the higher altitude regions although the high-altitude regions of Lesotho and the Free State are unsuitable. The highest suitability areas appear to be associated with the Forest biome (Low and Rebelo, 1996). *Solanum mauritianum* is reported to be associated with forest margins, plantations and wooded valleys (Henderson, 1995) which may explain the correspondence between areas predicted as high suitability for this species and the Forest biome. In addition, this species is considered to be the principal weed of South Africa's timber plantations (Bromilow, 1995) which are situated within the areas of high predicted suitability. An AUC value of 0.950 was calculated for this species (using 97 presence and 149 absence localities) which indicates a good fit between the model and the independent test localities. This model (AUC = 0.950) performed slightly better than the *R. communis* model (AUC = 0.948) but not as well as the *L. camara* model (AUC = 0.991). A Kappa value of 0.726 was calculated, which can be considered to indicate 'very good' agreement between the model and the test data (Monserud and Leemans, 1992).

Discussion

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

Climatic variable pre-processing

In addition to data reduction, pre-processing of the original variables is intended to remove or considerably reduces multicollinearity in the predictor variables eventually used to build the models. When one or more linear relationships exist among the original variables they are said to be linearly dependent or multicollinear (Bernstein *et al.*, 1988). Multicollinearity produces highly unstable results, especially in factor analysis and multiple regression, with the result that slight differences in sampling error or rounding may lead to substantially different results (Bernstein *et al.*, 1988). While this may not be considered to be a serious problem when PCA is used

for data reduction, it becomes particularly important when it is used as a predictive tool and when one intends to analyse the resulting principal components further, as has been done here.

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

The PCA model

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

Sources of presence-only data typically include records from museum and herbarium collections where data have been collected on an opportunistic or *ad hoc* basis. Geographical bias has been reported to be a problem in samples of records obtained from collections (Margules and Austin, 1994; Soberón *et al.*, 1996; Austin,

1998; Freitag *et al.*, 1998; Lawes and Piper, 1998; Funk and Richardson, 2002; Ferrier, 2002; Zaniewski *et al.* 2002). The bias in these datasets is likely to influence the quality of predictions made using these data.

The fundamental niche of an organism was defined by Hutchinson (cited in Schoener, 1990) as a *n-dimensional hypervolume* defined by *n* environmental dimensions within which the organism can survive and reproduce. The organism may be excluded from parts of its fundamental niche due to competition or other biotic interactions. The reduced hypervolume in which the organism can survive is its realised niche. The organism's occurrence along each axis of the fundamental niche is generally considered to follow a broad Gaussian curve (Austin and Meyers 1996). In contrast, occurrence in the realised niche has been shown to exhibit various skewed shapes (Austin *et al.*, 1984; Austin, 1987; Austin *et al.*, 1990; Leathwick and Mitchell, 1992; Austin *et al.*, 1994) which is often attributed to competition. In a correlative model such as the one presented here, the locality records used to build the model are drawn from the realised niche of the organism and as a result are likely to demonstrate skewed responses which will differ among predictor variables as well as among species (Austin *et al.*, 1990). In the modelling technique described here, a normal distribution[1] is used to describe the shape of the response on each component axis as a compromise among several possible responses. However, the responses of the species modelled here are not normal (Fig. 3). Despite this, the models have performed well against independent test records and also correspond with known habitat associations indicating that the departures of the data from the modelling assumptions may not be serious. The data certainly seem to occupy a central cluster in the hyperspace, and do thin out away from the origin (Fig. 3). The success of the technique is dependent on how robust it is to violations of its assumptions.

Model assessment

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

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

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

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

---

[1] The chi-squared distribution (which is equivalent to a squared normal distribution) can be used instead of the normal distribution because the component scores have to be squared in order to

In this study, alien plants were used to demonstrate the application of the modelling technique. One of the necessary assumptions in correlative modelling is that the target species is in equilibrium with the environment (Guisan and Zimmermann, 2000), which is not always true (Leathwick, 1998). It is thus necessary to assume that the alien plants have invaded all suitable environments but not necessarily all locations. However, it is possible that certain combinations of environmental conditions that are suitable for these species have not been invaded yet. If this is the case then the models would tend to under-predict the distributions of these species. Environments that are suitable but that have not yet been invaded will thus be predicted as being unsuitable. This is likely to be a problem when modelling alien plant distributions and may require predictions to be revised periodically, as new distribution records become available.

calculate the distance of a point from the origin of the hyperspace.

Acknowledgments

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

Table 1. Predictor variables selected for building the distribution models.

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

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

		Observed	
		+	-
Predicted	+	a	b
	-	c	d

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

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

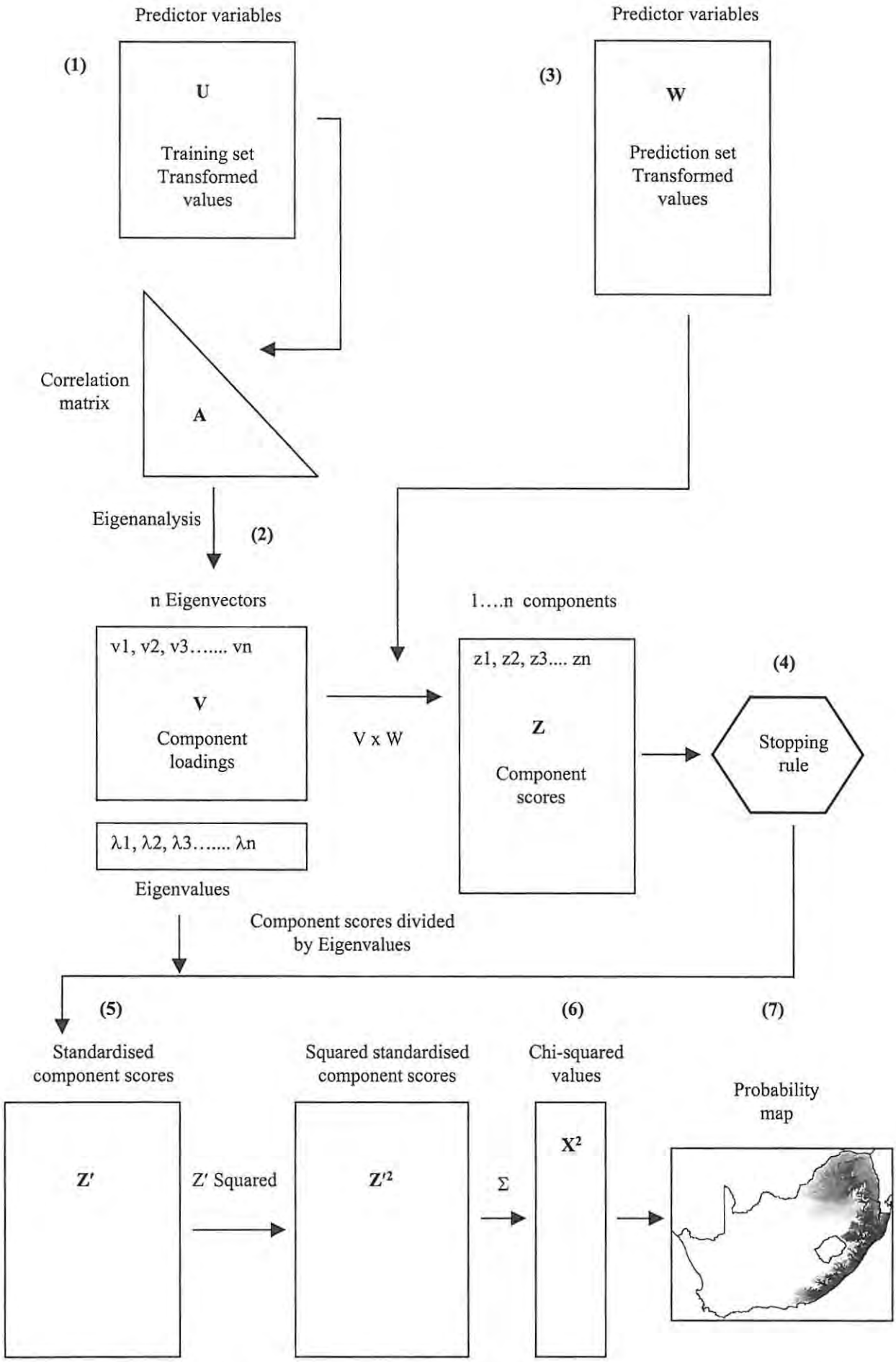


Figure 1. Implementation of the PCA modelling technique. The numbers in round brackets correspond with the steps described in the methods section.

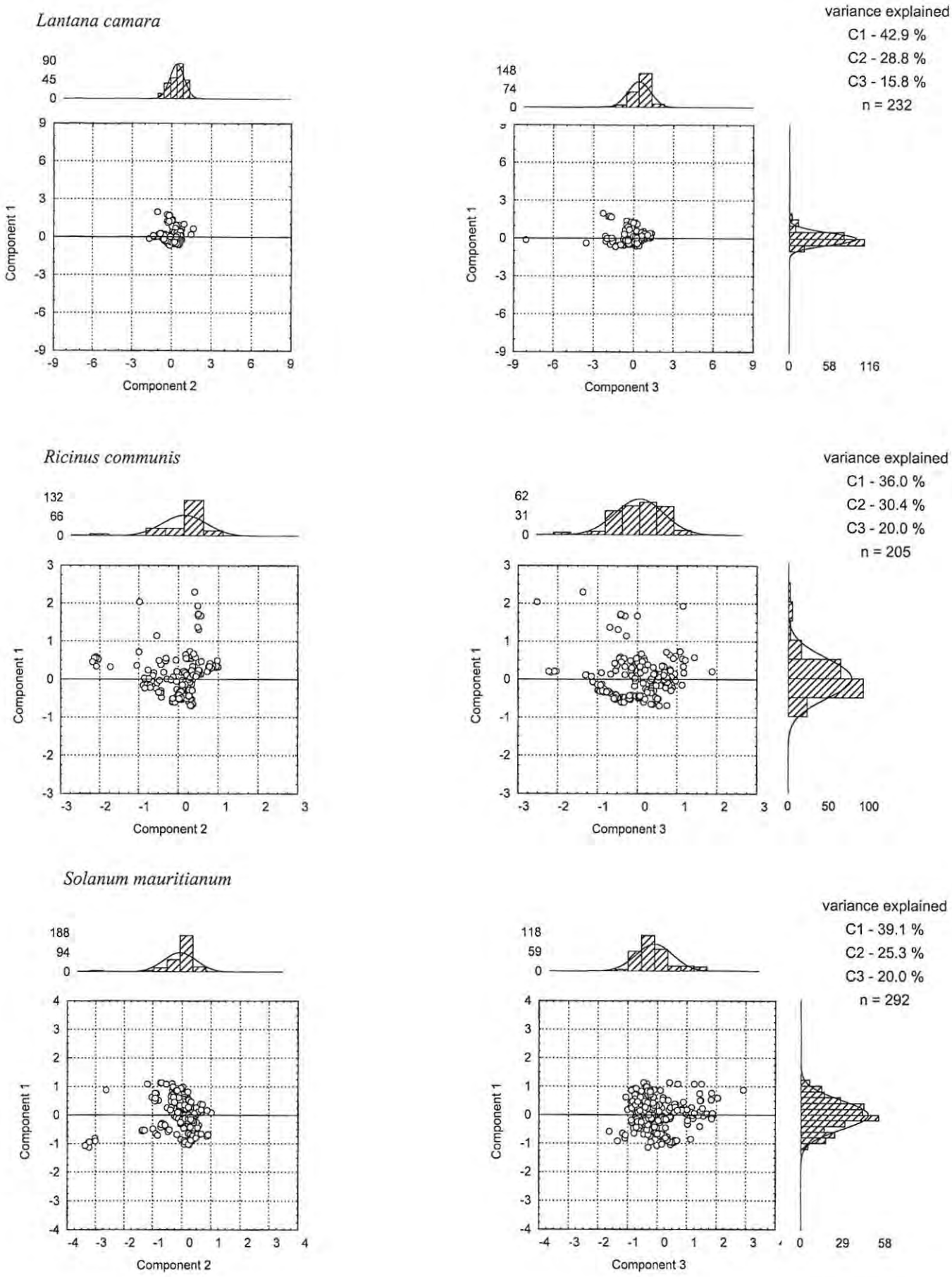


Figure 3. Plots of component scores and histograms (with normal distribution curves), calculated from the training sets for each species. Plots are for components 1 vs. 2 and components 1 vs. 3 for *L. camara* (a & b); *R. communis* (b & c) and *S. mauritianum* (d & e). Only the first three components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

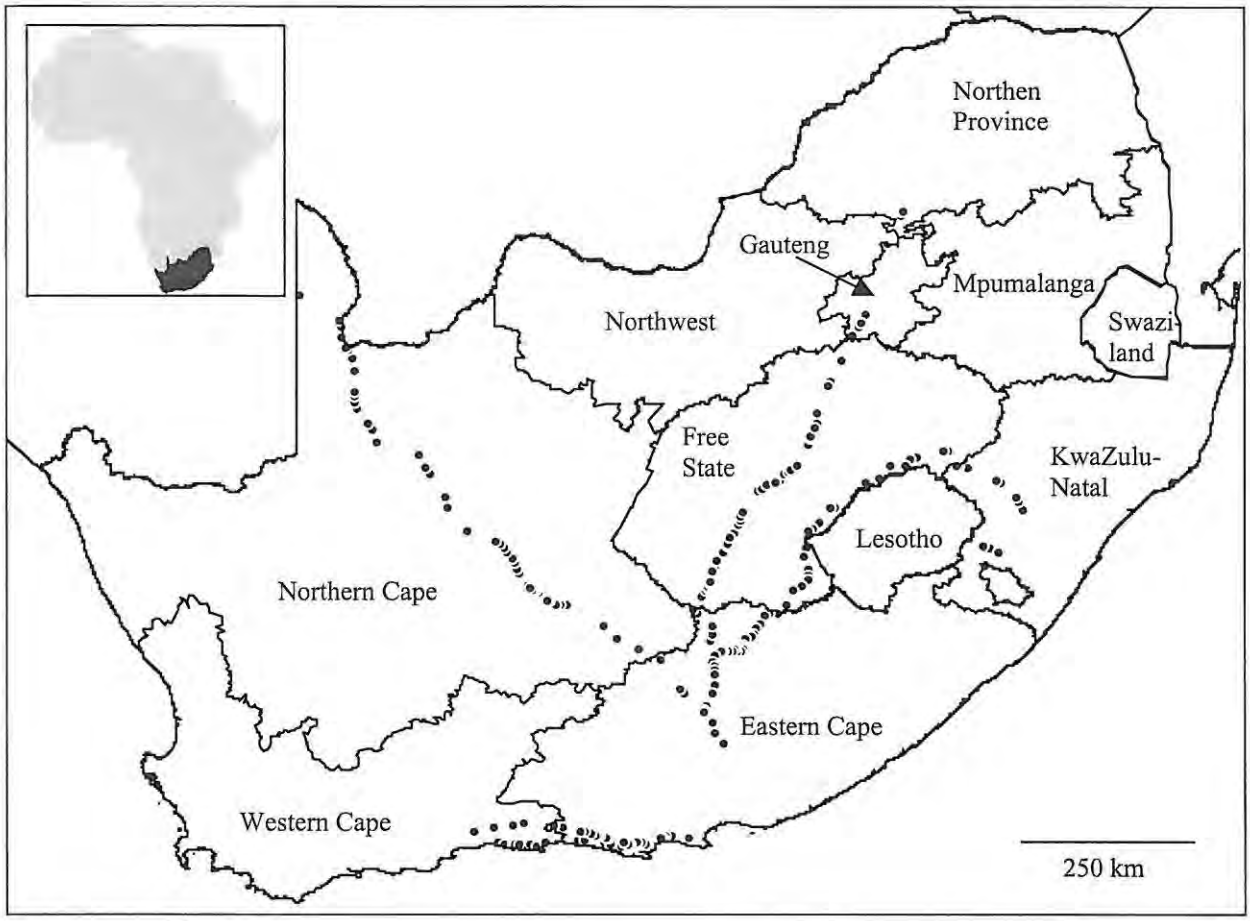


Figure 2. Map of South Africa (indicating the provinces), Lesotho and Swaziland. Black symbols indicate localities from which the absence test data were drawn for model testing. These localities indicate absence for all three species. In the inset, black indicates southern Africa relative to Africa.

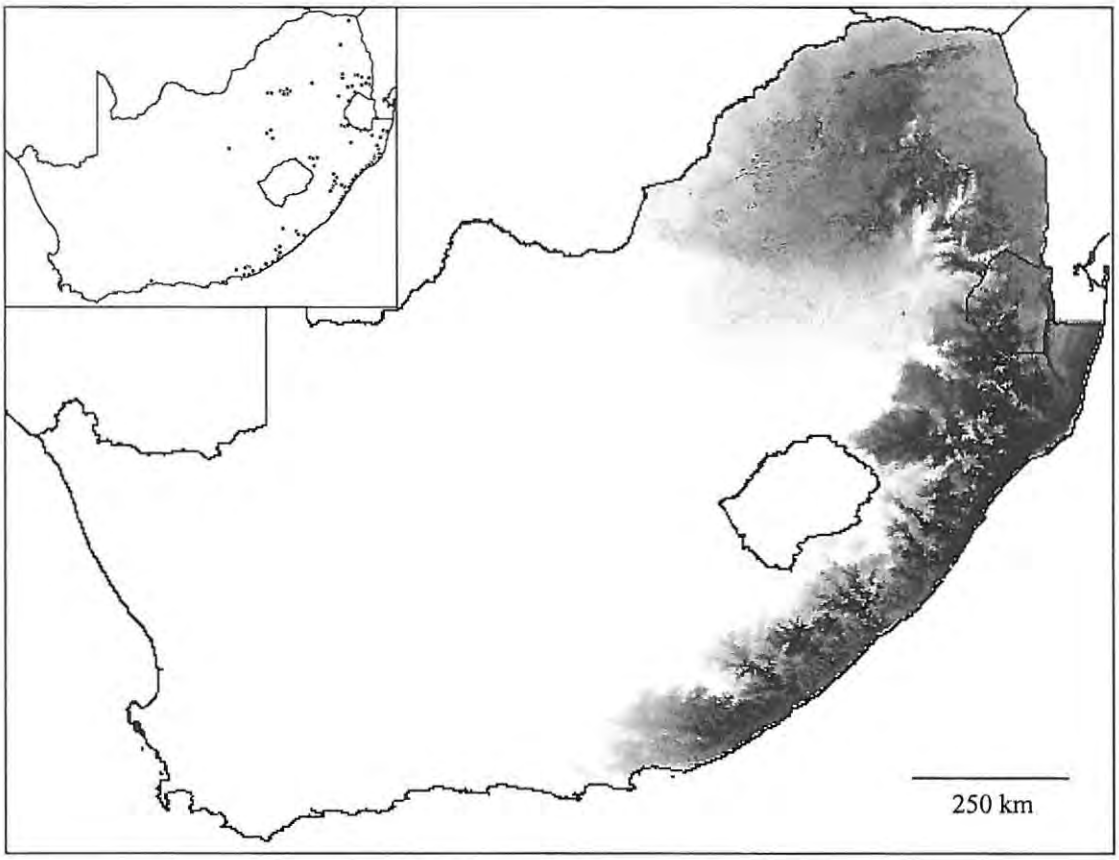


Figure 4. Bioclimatic suitability map for *Lantana camara* in South Africa, Lesotho and Swaziland produced from 232 localities (see inset) where the species was recorded present (condition number = 10). Darker shades indicate higher probabilities.

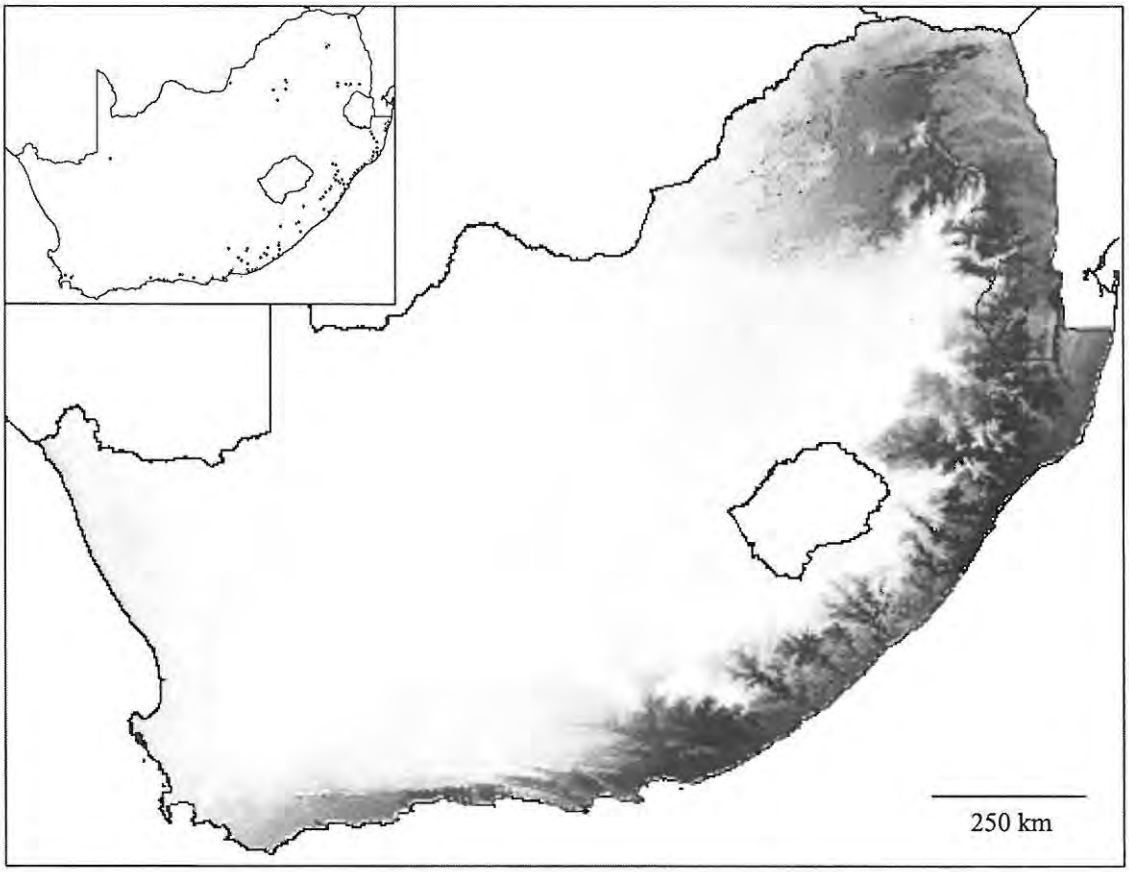


Figure 5. Bioclimatic suitability map for *Ricinus communis* in South Africa, Lesotho and Swaziland produced from 205 localities (see inset) where the species was recorded present (condition number = 5). Darker shades indicate higher probabilities.

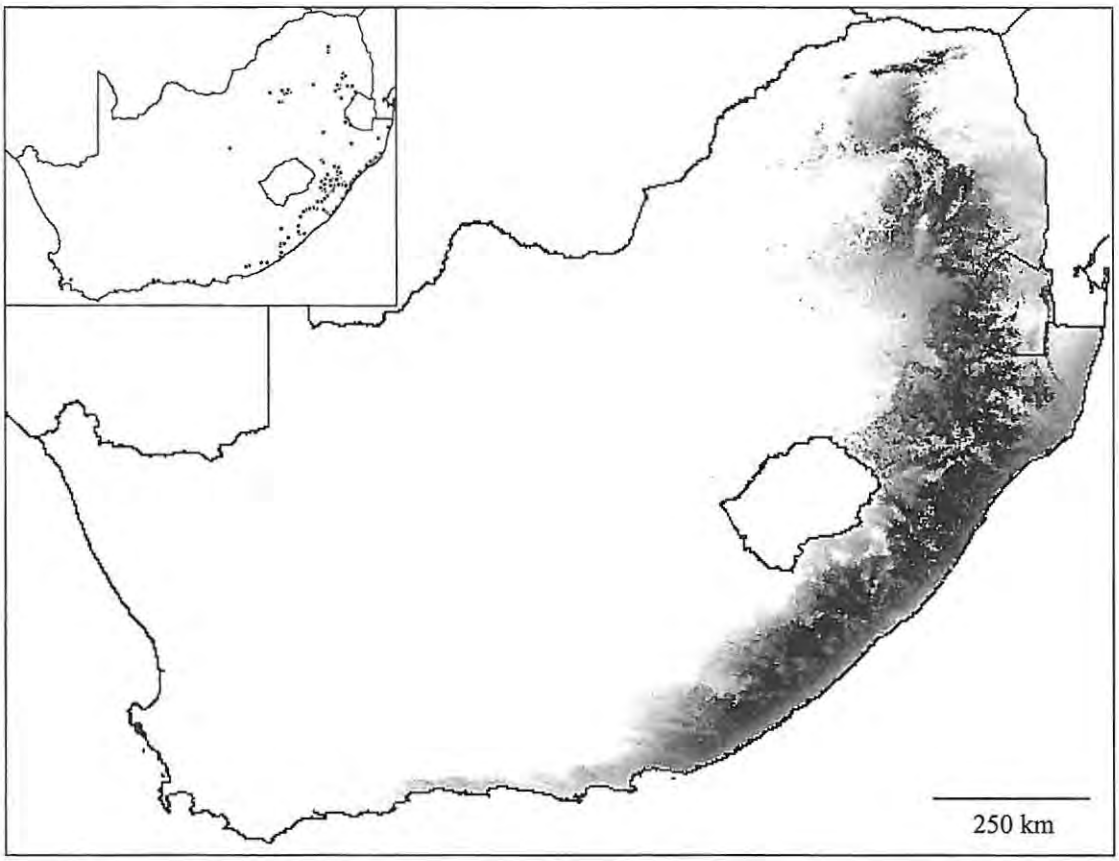


Figure 6. Bioclimatic suitability map for *Solanum mauritianum* in South Africa, Lesotho and Swaziland produced from 292 localities (see inset) where the species was recorded present (condition number = 11). Darker shades indicate higher probabilities.

References

Austin, M.P. 1987. Models for the analysis of species' response to environmental gradients. *Vegetatio*. 69: 35-45.

Austin, M.P. 1998. An ecological prespective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*. 85: 2-17.

Austin, M.P., Cunningham, R.B., Fleming, P.M., 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio*. 55: 11-27.

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

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

Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a beta-function. *Journal of Vegetation Science*. 5: 215-228.

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

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

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

Bernstein, I.H., Garbin, C.P., Teng, G.K., 1988. Applied multivariate analysis. 1st Edition. Springer-Verlag, New York.

Bromilow, C., 1995. Problem plants of South Africa, 1st Edition. Briza Publications, Arcadia, pp. 315.

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

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

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

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

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

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*. 51: 331-363.

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

Freitag, S., Hobson, C., Biggs, H.C., Van Jaarsveld, A.S., 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*. 1: 119-127.

Funk, V.A., Richardson, K.S. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*. 51: 303-316.

Henderson, L., 1995. Plant invaders of southern Africa. Plant Protection Research Institute Handbook no.5, 1st Edition. Agricultural Research Council, Pretoria, pp. 177.

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

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

Huberty, C.J., 1994. Applied discriminant analysis, 1st Edition. Wiley Interscience, New York, pp. 466.

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

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

Johnston, J., 1984. Econometric methods, 3rd Edition. McGraw-Hill International Book Company, Auckland, pp. 568.

Jongman, R.H.G., Ter Braak, C.J.F., Van Tongeren, O.F.R. 1995. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge.

Lawes, M.J., Piper, S.E., 1998. There is less to binary maps than meets the eye: the use of species distribution data in the southern African sub-region. *South African Journal of Science*. 94: 207-210.

Leathwick, J.R. 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*. 9:719-732.

Leathwick, J.R., Mitchell, N.D. 1992. Forest pattern, climate and vulcanism in central North Island, New Zealand. *Journal of Vegetation Science*. 3: 603-616.

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

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F., Tanton, M.T., 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*. 18: 371-383.

Lloyd, P., Palmer, A.R., 1998. Abiotic factors as predictors of distribution in southern African Bulbuls. *The Auk*. 115: 404-411.

Low, A.B., Rebelo, A.G., 1996. Vegetation of South Africa, Lesotho and Swaziland, 1st Edition. Department of Environmental Affairs and Tourism, Pretoria, pp. 85.

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

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

Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*. 62: 275-293.

Nix, H.A., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*, Australian Government Publishing Service, Canberra, pp. 4-15.

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

Packer, M.J., Canney, S.M., McWilliam, N.C., Abdallah, R., 1999. Ecological mapping of a semi-arid savanna. In: Coe, M.J., McWilliam, N.C., Stone, G.N., Packer, M.J. (Eds.), *Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna*, Royal Geographical Society (with The Institute of British Geographers), London, pp. 43-68.

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

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

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

Randerson, P.F. 1993. Ordination. In: Fry, J.C. (ed.), *Biological data analysis: a practical approach*. Oxford University Press, New York. Pp 173-217.

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

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

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

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

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

Rogers, D.J., Williams, B.G., 1993. Tsetse distribution in Africa: seeing the wood and the trees. In: Edwards, P.J., May, R. (Eds.), *Large-scale ecology and conservation biology*, Blackwell Scientific Publications, Oxford, pp. 247-271.

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

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

Schulze, R.E., Maharaj, M., Lynch, S.D., Howe, B.J., Melvil-Thomson, B., 1997. South African Atlas of agrohydrology and climatology, 1st Edition. Water Research Commission, Pretoria.

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

Soberón, J., Llorente, J., Benítez, H. 1996. An international view of national biological surveys. *Annals of the Missouri Botanical Garden*. 83: 562-573.

Sokal, R.R., Rholf, F.J., 1987. Introduction to biostatistics, 2nd Edition. W.H. Freeman and Co., New York, pp. 363.

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

Williams, B.G., Rogers, D.J., Staton, G., Ripley, B., Booth, T., 1994. Statistical modelling of georeferenced data: mapping tsetse distributions in Zimbabwe using climate and vegetation data. In: Perry, B.D., Hansen, J.W. (Eds.), *Modelling vector-borne and other parasitic diseases*, ILRAD, Nairobi, pp. 267-28

Zaniewski, A.E., Lehmann, A., Overton, J. McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*. 157: 261-280.

VI

The effects of false absence records, sample size and prevalence on the performance of single-species potential distribution models

Preface

This chapter quantitatively assesses the influence of several important aspects of data quality (outlined in Chapter Two) on the performance of a group discrimination technique in the form of logistic regression. In particular the influence of false absence records, sample size and prevalence on model performance is evaluated. A revised version of this chapter will be submitted to *Ecological Modelling*.

Abstract

Models based on relationships between distribution records and environmental predictor variables that are used to make potential distribution predictions for a target species are useful tools in biology. However, the influences of certain aspects of data quality on model performance are not known. This study investigated the effects on model performance of false absence records, the number of locality records (sample size) and the proportion of localities representing species presence (prevalence) in samples of records used to build logistic regression distribution models using an hypothetical distribution. Sample size and false absence record effects were tested using three model designs. In the first design, the sample of locality records consisted of equal numbers of surveyed presence and surveyed absence records (the surveyed-absence design, SA). The second design was similar to the first except that the locality record sample contained a 30% false absence fraction (the false-absence design, FA). In the third design, a sample of surveyed presence records and an equal number of "pseudo-absence" records drawn from the remaining un-surveyed grid-

cells in the map region were used to represent "presence" and "absence" respectively (pseudo-absence design, PSA).

False absence records and sample size had a significant effect on model performance. The use of pseudo-absence data appears to be viable in certain cases, but this depends on the extent of the range of the target species and the method used to select pseudo-absence records. This study suggests that the PSA design could be used to produce models that would perform on average no worse than those using the SA design on condition that presence samples are random and unbiased, and that the proportion of false absence records is kept to a minimum (<30%). However, this study did not make use of real presence-only data, where sampling is likely to be biased. Prevalence was found to significantly affect model performance. Samples with very low (10%) or very high (90%) prevalence produced models that were significantly lower in performance than those built using samples with less extreme prevalence (30%-80%). Prevalence did not appear to have a negative effect on model performance when smaller samples of records were used (160-320 records) but were more serious when large samples were used (2560-5120 records).

Introduction

Various biogeographical distribution models have been used to make species' potential distribution predictions using distribution records and associated environmental predictor variables (Franklin, 1995; Guisan and Zimmermann, 2000; Chapter 1). A number of these models make use of distribution records that indicate localities where a target organism has been found to be present (presence records) and localities where it has been found to be absent (absence records). These models have been referred to as *group discrimination* models (Caithness, 1995) while those that make use of only presence records have been referred to as *profile* models (Caithness, 1995).

Group discrimination techniques have been more popular than profile techniques. Examples of group-discrimination techniques include those models based on Discriminant Analysis (Rogers and Randolph, 1993; Rogers and Williams, 1993; Rogers *et al.*, 1996), Generalised Linear Models (Nicholls, 1989; Austin *et al.* 1984; Austin *et al.* 1990; Osborne and Tigar, 1992; Austin *et al.*, 1994; Guisan *et al.* 1998; Higgins *et al.*, 1999; Manel *et al.*, 1999 a & b; Cumming, 2000 a & b), Generalised Additive Models (Yee and Mitchell, 1991; Austin and Meyers 1996; Leathwick *et al.* 1996; Bio *et al.*, 1998; Leathwick, 1998; Pearce and Ferrier, 2000 a; Leathwick and Whitehead, 2001) and decision-tree-based methods (Walker, 1990; Lees, 1994; Michaelsen *et al.*, 1994; Williams *et al.*, 1994).

These models generally rely on presence and absence data collected in a systematic way by means of a field survey that uses a specific sampling strategy (Austin, 1998). These field surveys tend to expensive, labour intensive and time-consuming. As a result, alternative sources of data have been used in predictive modelling. These sources of data include museum and herbarium collection records where usually only the presence and not absence of organisms is recorded (Margules and Austin, 1994; Stockwell and Peters, 1999; Hirzel *et al.*, 2001; Peterson, 2001).

When no absence data are available then one can use modelling techniques that make use of presence-only data. Examples of these techniques include models developed by Palmer and Van Staden (1992), Erasmus *et al.* (2000), Robertson *et al.* (2001), Hirzel *et al.* (2001) and the approaches used in the modelling packages known as BIOCLIM (Nix, 1986; Busby, 1991) and DOMAIN (Carpenter *et al.*, 1993).

An alternative to using profile techniques is to make use pseudo-absence data in order to use group discrimination techniques (Ferrier and Watson, 1997; Cumming, 2000 a & b; Zaniewski *et al.*, 2002). Pseudo-absence data are best described as absence data that have not been obtained by means of a survey designed specifically to establish the absence of the target organism at a number of sites. Pseudo-absence records can be defined randomly (Ferrier and Watson, 1997; Zaniewski *et al.*, 2002), by using presence data collected for other species (Zaniewski *et al.*, 2002), or by using all the un-surveyed grid-cells in the map region as absence records (Cumming 2000 a & b).

One of the disadvantages of using pseudo-absence data is that false absence records are likely to be included in the dataset. Fielding and Bell (1997) refer to two types of prediction error, namely false positives (FP) and false negatives (FN). These are associated with presence and absence predictions respectively (see Chapter 2), and are equivalent to Type I and Type II statistical errors.

Although these terms are used to describe prediction errors, they can be extended to describe errors associated with locality records used to build predictive models. I refer to locality records that incorrectly indicate target organism presence instead of absence as *false presence* records and those records that incorrectly indicate absence instead of presence as *false absence* records. Although both types of error are problematic, it appears that an observer is less likely to commit a FP than a FN error while sampling. Taxonomic errors result in FN and FP errors, probably with roughly equal frequency. However, FP errors are less likely than FN because the observer can be more certain about recording the presence of an organism than recording its absence at a given site. This is because an organism is recorded as being "present" when it is detected and recorded as being "absent" when it is not detected. The organism may not be detected because it is genuinely not at that site or because the observer did not search thoroughly enough for it. In addition, the organism may be recorded as being absent because insufficient time has elapsed for it to colonise that area (Hirzel *et al.*, 2001) and not because it is unable to occur at that site e.g. alien organisms. Similarly, seasonal migration and local extinction may also result in FN errors.

Even if a systematic sampling strategy is used, false absences may still be recorded. Although it has been suggested that false absence records have an influence

on model performance, no quantitative assessments have appeared in the literature except for that of Hirzel *et al.* (2001).

When pseudo-absence records are used, one of the strategies of reducing the number of potential false absence records in samples is to reduce the number of pseudo-absence records relative to the number of presence records. As a result the ratio of presence to pseudo-absence records would not be equal and could influence model performance (Fielding and Bell, 1997; Manel *et al.*,1999 b; Pearce and Ferrier, 2000 a). The term *prevalence* refers to the ratio of presence to absence records (group size) in the sample (Fielding and Bell, 1997; Manel *et al.*, 1999 b), although this has also been referred to as *rarity* (Pearce and Ferrier, 2000 b).

The effects of false absence records and problems with unequal group sizes (prevalence) are likely to vary with the size of the sample of presence and absence records (sample size).

The aim of this study were to:

1. Investigate the effect of false absence records on model performance at different sample sizes
2. Compare the performance of pseudo-absence (PSA) models with surveyed-absence models (SA) and models built using samples with a high proportion of false absence records (FA models).
3. Investigate the effect of prevalence on model performance.

Methods

Distribution data

In order to produce a biologically feasible distribution map, a hypothetical distribution was based on a prediction made for a real organism. A potential distribution prediction was made for an alien invasive weed, *Lantana camara,* over a map region including South Africa, Lesotho and Swaziland. A simple envelope approach was used to generate a presence-absence range map for *L. camara* using the predictor variables listed in Table 1, and presence locality records used elsewhere

(Robertson *et al.*, 2001; Chapter 5). The choice of target organism was largely based on the fact that sufficient data were easily available from a previous study (Robertson *et al.*, 2001; Chapter 5). The envelope approach is identical to the "marginal range" predictions of the BIOCLIM modelling package (Nix, 1986; Busby, 1991). This presence-absence map was taken to represent the "known" distribution of a hypothetical target organism. Grid-cells in the map representing presence were coded 1 and those representing absence were coded 0. The entire map region consisted of 422503 grid-cells with the presence region of the hypothetical distribution comprising just over 37% of this region (158817 grid-cells).

Model design and sample size investigation

Potential distribution predictions were made for the hypothetical target organism using samples of different sizes for the three different designs of logistic regression model. A summary of these designs is given in Table 2.

The first design is referred to as the "surveyed absence" (SA) design. In this design, samples of 40, 80, 160, 320, 640, 1280, 2560 and 5120 grid-cells were selected from the presence region of the hypothetical distribution to represent "presence" and the same numbers of grid-cells were selected from the absence region of the hypothetical distribution to represent "surveyed absence". These sample sizes are referred to in this paper as presence-sample sizes and are also given as a proportion of the total number of grid-cells in the map region and as a proportion of the presence region of the hypothetical distribution (Table 3). In the case of the SA design, the same number of records was used for both the presence and absence groups (prevalence = 50%).

In the second design, samples of 40, 80, 160, 320, 640, 1280, 2560 and 5120 grid-cells were selected from the presence region of the hypothetical distribution to represent "presence". The same procedure was followed to select grid-cells to represent "absence" except that for each absence sample, 30% of the grid-cells were drawn from grid-cells in the hypothetical distribution representing "presence" instead of "absence". This was done in order to simulate cases where false absence records are incorporated into samples used to fit distribution models. This design is referred to as the false-absence (FA) sampling design.

In the third design, samples of 40, 80, 160, 320, 640, 1280, 2560 and 5120 grid-cells were selected from the presence region of the hypothetical distribution to represent "presence". Next, samples of 40, 80, 160, 320, 640, 1280, 2560 and 5120 grid-cells were selected from all the remaining grid-cells that had not already been selected for the presence samples were taken to represent "pseudo-absence". This is referred to as the pseudo-absence (PSA) design.

For each design, the selection of records was repeated seven times so that seven replicate predictions were performed for all presence-sample size categories. In total 168 models were produced for the sample size investigation (3 designs x 8 sample size categories x 7 replicates = 168). The independent variables for the models consisted of the predictor variables and the squares of selected variables (Table 4). Scatter plots were examined to determine which predictor variables should be squared. Models were fitted using the logistic regression procedure described in the section on logistic regression (see below).

Prevalence investigation

The influence of the proportion of presence records (prevalence) in a sample of surveyed localities (comprising presence and absence records) on model performance was investigated using the SA design. Predictions were made using total sample sizes (presence and absence) of 160, 320, 640, 1280, 2560, 5120, 10240 and 81920 records. In each case the proportion of presence records (prevalence) comprised 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% of the total sample size. Five replicates were performed for each prevalence category for each sample size. For the prevalence investigation, a total of 270 models were produced (9 prevalence categories x 6 sample size categories x 5 replicates = 270). The independent variables for the models consisted of the predictor variables listed in Table 4 as well as the squares of selected variables. Scatter plots were examined to determine which predictor variables should be squared. Models were fitted using the logistic regression procedure described in the section below.

Logistic regression

Logistic Regression is a form of Generalised Linear Model (GLM: McCullagh and Nelder, 1989) in which a binomial error distribution and a logistic link function are used (Guisan and Zimmermann, 2000).

Model fitting was performed using GLMFIT, a generalised linear modelling function within MATLAB. A binomial distribution was used for the error function and a logit link function was used to calculate the values of the regression coefficients (β_i) of the linear predictor (Eq. 1).

A backwards elimination procedure was used to remove variables that did not significantly improve model fit. A predictor variable was removed if it did not result in a significant reduction in the deviance calculated from the full model (a model containing all the predictor variables). Significance tests were based on the assumption that the change in deviance followed a chi-square distribution. A threshold of 0.20 was used for exclusion of variables as several authors have suggested that a threshold of 0.05 may be too stringent (Pearce and Ferrier, 2000 a).

Probability values for each grid cell in the map region were calculated by substituting the values of the predictor variables associated with that cell into the linear predictor (Eq. 1) and then transforming these values using the inverse logistic transform (Eq. 2).

$$\beta_1 \text{variable}_1 + \beta_2 \text{variable}_2 + \ldots + \beta_8 \text{variable}_8 = \eta \qquad (1)$$

$$P_{(y=1)} = \exp(\eta)/(1+\exp(\eta)) \qquad (2)$$

Model evaluation

Most quantitative model performance tests are based on a confusion matrix in which the observed (actual) and predicted presence/absence patterns are cross-tabulated (Fielding and Bell, 1997). Various threshold-dependent and threshold-independent accuracy measures can be calculated from the confusion matrix (reviewed by Fielding and Bell, 1997). Threshold-independent measures (e.g. ROC

curves) are considered to be more robust and more objective than threshold-dependent measures (e.g. Kappa statistics) since they do not rely on a single threshold to distinguish between predicted presence and predicted absence (Fielding and Bell, 1997).

Model performance was evaluated using the area under the curve (AUC) of a receiver operator characteristic (ROC) curve. This threshold-independent technique is emerging as a popular measure of model performance (Packer *et al.*, 1999; Cumming, 2000 a & b; Pearce and Ferrier, 2000 a & b; Robertson *et al.*, 2001). The hypothetical distribution was used to define localities representing "actual presence" and "actual absence" for the target organism, against which the "predicted presence" and "predicted absence" of each of the models could be compared.

A major advantage of using a hypothetical distribution is that it can be made to represent the "known" distribution of that organism (Hirzel *et al.*, 2001; Hirzel and Guisan, 2002). As a result, for each grid-cell in the entire map region either the "actual presence" or "actual absence" for the target organism is known. This is particularly useful for model evaluation since predictions generated using the model can be compared with an independent set of records representing the "actual" distribution over the entire map region. This results in a rigorous, objective test of model performance.

AUC values obtained from ROC plots were used as a single measure of overall model performance rather than deconstructing prediction success by examining specificity and sensitivity separately at a single threshold (cf. Manel *et al.*, 1999 b). The choice of threshold will influence the ratio of sensitivity to specificity values calculated (Fielding and Bell, 1997). As a result, any relationships between specificity or sensitivity and model performance at one threshold are unlikely to hold at others. The choice of threshold to use is often arbitrary and may be influenced by the aims of the study or by the target organism (Fielding and Bell, 1997). In addition it has been suggested that logistic regression is sensitive to threshold effects due to species prevalence (Fielding and Bell, 1997).

Due to serious violations of the assumptions of ANOVA (unequal variances among groups and departure from normality) that could not be corrected by transformations, Kruskal-Wallis tests (based on rank-sums) were used to assess the effects of sample size and model design on model performance (measured using AUC

values). Non-parametric Tukey-type multiple comparisons were made among model designs for each sample size category using the Nemenyi test (Zar, 1996). For the prevalence investigation, Kruskal-Wallis tests were used to assess the effects of prevalence and sample size on model performance and non-parametric multiple comparisons (Zar, 1996) were made among prevalence categories for each of the sample size categories.

Results

Model design and sample size

The ranges of agreement proposed by Pearce and Ferrier (2000 b) for interpreting AUC values are useful. They have suggested that AUC values between 0.5 and 0.7 indicate *poor* discrimination capacity, values between 0.7 and 0.9 indicate *reasonable* discrimination and values higher than 0.9 indicate *very good* discrimination of models.

Average model performance was very good (AUC>0.91) for all presence-sample size categories and sampling designs tested (Fig. 1). Model performance increased with increasing presence-sample size for all three model designs. The rate of increase in model performance was greater for models built using smaller numbers of presence records (between 20 and 160) than when larger numbers (>160) of presence records were used (Fig 1).

There was generally fairly good visual agreement between the hypothetical distribution map (Fig. 2a) and individual maps derived from models built using various combinations of model design and presence-sample size (Fig. 2c-k). Note, however, that the fit between the hypothetical distribution and the predicted distribution maps is still not exact (Fig. 2). There was no agreement (AUC=0.500) between the random distribution (Fig. 2b) and the hypothetical distribution map (Fig. 2a). The SA design showed better visual agreement with the hypothetical distribution than the FA or PSA designs at all three sample sizes (Fig. 2). This was confirmed by the lower AUC values for each of the FA and PSA maps compared with the AUC values for the SA maps. For each design, the maps generated using models developed from only 40 records (Fig 2c, f & i) did not agree with the hypothetical distribution as

well as maps generated using models developed from 320 (Fig. 2 d, g & j) or 5120 (Fig 2e, h & k) presence records. This was confirmed by the AUC values (Fig. 1). For each design for models built using 320 presence records the maps agreed more closely with the maps from models built using 5120 presence records than they did with maps from models using only 40 records. Similar trends were observed for the AUC values associated with these maps (Fig. 1). For the SA design maps, a greater proportion of grid-cells had high probabilities (0.75-1.00) than the maps of the FA or PSA designs, across all three sample sizes.

Kruskal-Wallis tests indicated that both presence-sample size ($H=60.40$, d.f.=7, $N=168$) and model design ($H=67.76$, d.f.=2, $N=168$) had significant effects on model performance. Non-parametric multiple comparisons using the Nemenyi test (Zar, 1996) revealed that models built from all three model designs demonstrated significant increases in performance with increasing presence-sample size (Table 5). There was no significant increase in model performance for models built with 80 or more presence records for the PSA and SA designs (Table 3). There was no significant increase in model performance for models built with 320 or more presence records in the case of the FA design.

The influence of model design on model performance for each presence-sample size category was investigated using Nemenyi tests (Table 6). There was no significant difference in model performance among the model designs when only 80 presence records were used (Table 6). Models built using the SA design performed significantly better than those performed using the FA design for all presence-sample size categories from 160 to 5120 records (Table 6). There was no significant difference in model performance between the PSA and FA designs for all presence-sample size categories (Table 6).

Prevalence

Mean model performance was highest for the 50% prevalence category and lowest for the 10% and 90% prevalence categories (Fig. 3). Average model performance increased from just above 0.963 at 10% prevalence to a maximum of just above 0.970 at 50%, followed by a decrease to a minimum value just below 0.962 at 90%. Performance was slightly higher at prevalence categories below 50% than at

prevalence categories above 50% (Fig. 3). Kruskal-Wallis tests indicated that prevalence (H=64.756, d.f.=8, N=270) had a significant effect on model performance. Note that the term "sample size" here refers to the total number of records (presence and absence) used to build the models.

Non-parametric multiple comparisons (Zar, 1996) indicated that model performance for the 10% and 90% prevalence categories was significantly lower than all other categories with the exception of the 20% category (Table 7). Model performance did not differ significantly among the 30%, 40%, 50%, 60%, 70% or 80% prevalence categories (Table 7).

Kruskal-Wallis tests were performed on each sample size group to determine at which sample sizes prevalence had a significant effect on model performance (Table 8). These tests indicated that significant differences occurred only at sample sizes of 640, 2560 and 5120 records but not at sample sizes of 160, 320 and 1280 records.

For those sample size groups where prevalence had a significant effect on model performance (640, 2560 and 5120), non-parametric multiple comparisons (Zar, 1996) were made to determine which prevalence categories differed significantly from one another (Table 9). For the 640-record category, only models from the 10% prevalence category differed significantly from models from the remaining prevalence categories (Table 9). For the 2560-record category models from the 90% and 10% prevalence categories differed significantly from models of the 50% and 60% categories. While for the 5120-record category models from the 90% prevalence category demonstrated significantly lower model performance than models from the 50%, 60% and 70% prevalence categories. Models from the 10% category differed significantly from the models of the 60% prevalence category.

Discussion

Model performance

Mean AUC values were greater than 0.91 indicating *very good* performance for all model designs and presence-sample size categories, [using the Pearce and Ferrier (2000 b) scale]. This suggests that on average, logistic regression is reasonably robust

to the problems associated with small sample sizes and false absence records. However, individual predicted distribution maps (Fig. 2c-k) still show departure from the hypothetical distribution map (Fig. 2a).

The AUC measurement scale has a much smaller range than more commonly used measurement scales such as percentages or probabilities. AUC values can range between 0.5 for no agreement (a random distribution, Fig. 2b) and 1.0 for perfect agreement between a predicted distribution and an observed distribution (Zwieg and Campbell, 1993). Therefore, care should be taken when interpreting these values. In addition, the difference between mean AUC values for those models that performed worst (just greater than 0.91; Fig. 1) and those that performed the best (just less than 0.97), was very small (less than 0.06), but significant (Table 5), suggesting that significant differences in model performance can occur within a very narrow range of AUC values.

Model design

The three model designs were used to illustrate three possible situations in which predictive models could be applied, and to test the influence of false absence records on performance of models using these three designs. Models of the SA design represent situations in which good quality (i.e. contains no false presence or false absence records) surveyed presence and surveyed absence records are available. Models of the PSA design represent situations when no surveyed absence data are available for a target species and un-surveyed grid-cells are used as putative absence records. Models of the FA design represent situations in which a surveyed presence and surveyed absence records are available but the sample of presence records contains a fairly large proportion (30%) of false absence records which have inadvertently been included due to survey errors (see Chapter 2).

Models of the PSA and FA designs contained various proportions of false absence records and these were compared with models of the SA design, which contained no false absence records. The FA samples contained a 30% proportion of false absence records while the PSA design contained an unknown proportion of false absence records which is dependent on the extent of the range of the species. The presence region of the hypothetical distribution used here comprised just over 37% of

the entire map region. In cases where the range of the species is more restricted the proportion of the map region occupied would be less and the proportion of false absence records in a random sample of un-surveyed grid-cells would be lower.

Model design and sample size had a significant effect on model performance. The model design results suggest that false absence records do have an effect on model performance. Comparisons among designs at various sample sizes indicated that no significant differences occurred among designs at low sample sizes (40 presence records). At larger sample sizes (80 or more records) the SA design performed significantly better than the FA design but did not perform significantly better than the PSA design. This contrasts with the findings of Ferrier and Watson (1997) that GLM and GAM models built using presence and absence records performed significantly better than GLM and GAM models built using presence and randomly generated pseudo-absence records. These differences between the two studies may be as a result of differences in the numbers of false absences included in the pseudo-absence records and also possibly because a non-parametric test was used in the current study, which is likely to be less sensitive than a parametric test.

Hirzel *et al.* (2001) compared the performance of GLM (built with linear and quadratic terms) and ENFA models (reviewed in chapter 3) using different sample sizes (300 and 1200 records) and under different data quality scenarios using hypothetical distributions. They found that when the quality of the absence data was more reliable (fewer false absences), the GLMs performed better.

The finding that models of the SA design performed on average significantly better than those of the FA design but not the PSA design suggests that the proportion of false absence records included in samples of the PSA design was probably lower on average than that of the FA design (<30%). Assuming that the Kruskal-Wallis test is sufficiently sensitive to detect differences, these results suggest that, on average, models produced using the PSA design will not differ significantly from models produced using the SA design provided that the proportion of false absence records in the samples used to build the PSA models is below 30%. This implies that in situations when one has unreliable absence data (i.e. with many false absences) then equivalent or better results may be achieved on average by using the PSA design. Conversely, if one had no absence data then one could make predictions using the PSA design and these models would perform no worse on average than if one had

unreliable absence data (with a false absence fraction of 30%), provided that the presence data were a representative random sample.

However, this is unlikely to be true for real presence-only data. Sources of presence-only data usually include museum and herbarium collections where these data are usually collected on an *ad hoc* basis and thus have a number of problems (Ferrier and Watson, 1997; Funk and Richardson, 2002; Zaniewski *et al.*, 2002). One of the most serious of these problems is that samples obtained from these sources usually contain geographical bias, and the extent of this bias is always unknown. Several authors have reported geographical bias to be a problem in samples of records obtained from collections (Margules and Austin, 1994; Soberón *et al.*, 1996; Austin, 1998; Freitag *et al.*, 1998; Lawes and Piper, 1998; Funk and Richardson, 2002; Ferrier, 2002; Zaniewski *et al.* 2002). This means that easily accessible areas tend to be sampled more often e.g. along road networks, near cities. The result is that data obtained from these sources are unlikely to be random. In addition, a recent study by Hirzel and Guisan (2002) suggests that random sampling may not result in models with the highest predictive performance.

Real pseudo-absence datasets are likely to be quite small, approximately 40 or fewer presence records (cf. Ferrier and Watson, 1997). This means that the pseudo-absence design is likely to be applied in situations where the number of presence records is small. The results of this study found that the variability of the PSA design is greatest at small sample sizes, considerably greater than the SA design (Fig. 1). Greater variability in performance suggests that the risk of producing models with poor performance is greater (Hirzel and Guisan, 2002). This may reduce the usefulness of this design.

Based on their comparison of GAM presence/absence, GAM pseudo-absence and ENFA models, Zaniewski *et al.* (2002) suggested that the use of pseudo-absence records may be viable when only presence data are available. However, their study was based on data derived from planned surveys and thus they were unable to assess some of the important data quality aspects associated with true presence-only data e.g. sampling bias.

The method used to define pseudo-absence records may influence the reliability of the records. In the current study pseudo-absence records were drawn at random from the un-surveyed grid-cells in the map region. A similar approach was taken by

Ferrier and Watson (1997). In contrast, Zaniewski *et al.* (2002) used presence records for several non-target species (drawn from a presence/absence dataset) to define pseudo-absence records for the target species. This approach is likely to produce pseudo-absence records that contain fewer false absences because if the target species was present in the plot containing the non-target species then it is also likely to have been recorded as being present. Data collected in this way can be used as indirect evidence that the target species is absent at a particular locality. However, this method of defining pseudo-absence records relies on multi-species surveys, which are not always available.

Sample size

Sample size had a significant effect on model performance for all three model designs. Model performance increased rapidly with increasing sample size for smaller sample sizes followed by very small increases beyond 320 presence records.

Similarly, Hirzel *et al.* (2001) found only very small differences in performance between GLM models (containing linear and quadratic terms) built using 300 records and those built using 1200 records.

There was a greater variability in performance for models built using smaller samples than larger samples (Fig. 1). Hirzel and Guisan (2002) reported a similar finding for GLM (containing linear and quadratic terms) using sample sizes of approximately 100, 200, 400 and 800 records. Pearce and Ferrier (2000 a), using total sample sizes (presence and absence) of 50, 250 and 500 records, also found that sample size significantly affected model performance.

In the current study comparisons were made using presence samples ranging from 40 to 5120 presence records, which for the SA design would equate to a range in total sample size (presence and absence) from 80 to 10240 records. Although significant differences in model performance were found across this range of sample sizes, models from a range of different sample sizes did not differ significantly in performance. These results suggest that, on average, models of equivalent performance can be obtained using samples with a range of different sizes. For example, for the SA design, the average performance of models built using only 80

records (presence + absence) did not differ significantly from models built using 1280 records (Table 5).

Pearce and Ferrier (2000 a) suggested that total sample sizes (presence and absence) of 50 records were too small to build reliable models, particularly with rare species (low prevalence). The smallest sample size tested in this study contained 80 records (40 presence + 40 absence for the SA design), which appeared to produce models of adequate performance.

The minimum number of records required to make a reliable prediction is probably a function of several factors. One of these factors is likely to be the number of grid-cells available in the map region, which can be both a function of grid-cell size and the geographical extent of the map region. For a map region of a given extent, fewer large grid-cells (e.g. quarter degree squares) than small grid-cells (e.g. minute squares) are required to cover the map region. Similarly, for a grid comprising cells of a given size, more grid-cells are required to cover a region of a large extent (e.g. Africa) than a region of smaller extent (e.g. southern Africa). Other factors that may influence the minimum sample size, requiring further investigation, include the extent of the geographical range occupied by the target species and the dimensionality of the environmental hyperspace constructed from the predictor variables. From this study, a sample of 0.025% of occurrence (Table 2) may be adequate to produce an acceptable model (SA design).

Prevalence

The proportion of presence records in samples of locality records (prevalence) had a significant effect on model performance. For data pooled across all sample sizes, models built with samples that had very low (10%) or very high (90%) prevalence performed significantly worse than those built with samples where the prevalence was less extreme (30-80%). This confirms that unequal group sizes can influence model performance as suggested by Fielding and Bell (1997) and as reported in other studies (Manel *et al.*, 1999 b; Pearce and Ferrier, 2000 a).

These prevalence results suggest that one could reduce the uncertainty in pseudo-absence predictions by building models with fewer pseudo-absence records

relative to presence records, provided that prevalence was not very high (90%) or very low (10%).

Prevalence did not have a significant effect on model performance for models built using total sample sizes of 160, 320 or 1280 records (Table 8). This suggests that prevalence effects may only be felt when large total sample sizes are used to build models.

The results of this prevalence investigation suggest that samples of very low or very high prevalence should be avoided as they may produce unreliable models, especially when the total sample size is large. This situation may arise when the un-surveyed grid-cells in the map region are taken to represent pseudo-absence (Cumming, 2000 a & b), especially if the number of surveyed grid-cells (presence) is very small relative to the total number of grid-cells in the map region.

The hypothetical distribution approach

The use of a hypothetical distribution has been used previously to evaluate model performance under specific data quality regimes (Cumming, 2000 b; Hirzel *et al.*, 2001; Hirzel and Guisan, 2002). Data derived from hypothetical distributions have a number of advantages over real data (Hirzel *et al.*, 2001; Hirzel and Guisan, 2002). The type of hypothetical distribution approach taken will depend on the aim of the study, e.g. compare Cumming (2000 b) with Hirzel *et al.* (2001).

The hypothetical distribution approach used in this study was aimed at testing the effects of sample size, false absence records and prevalence on model performance while controlling for as many the other factors that may influence model performance as possible. This approach has a number of advantages. Firstly, the hypothetical distribution map was used to define localities representing "true presence" and "true absence" for a hypothetical target organism. As a result, no false presence or false absence errors (see Chapter 2) could have been accidentally incorporated into the training sample, except where this was intentionally done to test a specific effect e.g. the FA design. Some of the factors contributing to these errors in real data are mentioned by Hirzel *et al.* (2001). Secondly, the localities used to build the models were drawn at random from the map region thus eliminating sampling bias and ensuring that various combinations of environmental conditions were equitably

sampled. In many ways this approach can be considered to represent an optimal sampling strategy (Guisan and Zimmermann, 2000). As a result, it is likely that fewer records are required to produce credible predictions than if sampling was not random. Real samples are often biased (Margules and Austin, 1994; Soberón *et al.*, 1996; Austin, 1998; Freitag *et al.*, 1998; Lawes and Piper, 1998; Funk and Richardson, 2002; Ferrier, 2002; Zaniewski *et al.* 2002). Thirdly, the models could be rigorously evaluated by using the entire hypothetical distribution in the model performance tests as a set of independent localities.

The hypothetical distribution approach used here enabled the influence of sample size, model design and prevalence on model performance to be tested with a minimum of possible confounding factors. However, care should be taken when applying these findings to real organisms. Firstly, the conclusions drawn in this study are based on the results of statistical tests performed on several replicates for particular treatments and thus tend to describe the "average" response rather than the range. However, when predictions are made for real organisms, these predictions are generally not replicated using different sets of locality records.

Secondly, the minimum number of presence records required to make reliable predictions is likely to be slightly higher for real organisms since sampling in the map region is unlikely to be completely random, thus possibly introducing some redundancy. Other factors such as the number of grid-cells in the map region, the geographical extent of the target species' range and the dimensionality of the environmental hyperspace may also influence the minimum number of records required to make reliable predictions, although these factors require further investigation.

Although the results of this study are important, they are based on only one type of hypothetical distribution, and thus their generality requires testing. The effects of sample size, false absence records and prevalence on model performance are likely to vary with different types of distribution, in particular their effects may be more profound for those species that have restricted ranges e.g. local endemics. In addition, model performance is likely to be more seriously affected in situations where false absence records occur in samples of locality records with high prevalence. This points to the need for further studies to be conducted using a variety of different types of hypothetical distributions including those with restricted ranges. The results of

hypothetical distribution studies would also be more convincing if they were complemented by suitable examples using real data, where possible (Chapter 7).

Conclusion

This study found that false absence records and sample size had a significant effect on model performance. However, logistic regression appears to be robust to a certain proportion of false absence records. The use of pseudo-absence data appears to be viable in certain cases, but this depends on the extent of the range of the target species and the method used to select pseudo-absence records. This study suggests that the PSA design could be used to produce models that would perform on average no worse than those using the SA design on condition that presence samples are random and unbiased, and that the proportion of false absence records is kept to a minimum (<30%). However, this study did not make use of real presence-only data, where sampling is almost certain to be biased.

Prevalence was found to significantly effect model performance. Samples with very low (10%) or very high (90%) prevalence produced models that were significantly lower in performance than those built using samples with less extreme prevalence (30%-80%). Prevalence did not appear to have a negative effect on model performance when smaller samples of records were used (160-320 records) but were more serious when large samples were used (2560-5120 records). Although hypothetical distributions may be useful for investigating various aspects of data quality on model performance, where possible, conclusions drawn from these studies should be supported by studies using real organisms.

Acknowledgements

I thank Sarah Radloff for statistical advice and Mike Burton for advice with MATLAB software. The School of Bioresources Engineering and Environmental Hydrology (University of Natal), the Water Research Commission and the South African Country Study for Climate Change are thanked for the use of the climatic predictor variables. Funding from the National Research Foundation and the Rhodes University Joint Research Council is gratefully acknowledged.

Table 1. Predictor variables selected for building the hypothetical distribution model.

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

Table 2. Descriptions of the three model designs of logistic regression (SA – surveyed absence, FA – false absence and PSA – pseudo-absence). The number of presence records (p) can be 40, 80, 160, 320, 640, 1280 or 5120. Category indicates to which category the records are assigned for fitting the logistic regression model.

Design	Design description	Formulae	Category
SA	p presence records drawn from presence region		presence
	a absence records drawn from absence region	a = p	absence
FA	p presence records drawn from presence region		presence
	fa false absence records drawn from presence region	fa = p x 0.33	absence
	ar absence records drawn from absence region	ar = p - fa	absence
PSA	p presence records drawn from presence region		presence
	psa pseudo-absence records drawn from anywhere in map region	psa = p	absence

Table 3. The number of presence records (n) used in each of the sampling designs, expressed as a percentage of the total number of grid-cells (% total grid-cells) in the map region and as a percentage of the number of grid-cells comprising the presence region of the hypothetical distribution (% pres. grid-cells).

n	% total grid-cells	% pres. grid-cells
40	0.009	0.025
80	0.019	0.05
160	0.038	0.101
320	0.076	0.201
640	0.151	0.403
1280	0.303	0.806
2560	0.606	1.612
5120	1.212	3.224

Table 4. Predictor variables selected for building the distribution models.

No.	Predictor variable	
1	Evap	Component axis 1 of a PCA on 12 monthly potential evaporation surfaces
2	Maxt	Component axis 1 of a PCA on 12 monthly maximum temperature surfaces
3	Mint	Component axis 1 of a PCA on 12 monthly minimum temperature surfaces
4	Rain1	Component axis 1 of a PCA on 12 monthly rainfall surfaces
5	Rain2	Component axis 2 of a PCA on 12 monthly rainfall surfaces
6	$Evap^2$	(Component axis 1 of a PCA on 12 monthly potential evaporation surfaces)^2
7	$Rain1^2$	(Component axis 1 of a PCA on 12 monthly rainfall surfaces)^2
8	$Rain2^2$	(Component axis 2 of a PCA on 12 monthly rainfall surfaces)^2

Table 5. Nonparametric multiple comparisons among sample size categories for each of the three designs (SA, FA, PSA), using the Nemenyi test (α=0.05). The multiple comparisons were calculated separately for each of the three designs. Sample size (n) refers to the number of presence records used to build each of the models. Those models that appear in the same group (e.g. G1 or G2) do not differ significantly in performance from one another. Group membership is indicated by **.

SA			FA			PSA		
n	G1	G2	n	G1	G2	n	G1	G2
40	**		40	**		40	**	
80	**		80	**	**	80	**	**
160	**		160	**	**	160	**	**
320	**	**	320	**	**	320		**
640	**	**	640	**	**	640		**
1280		**	1280	**	**	1280		**
2560		**	2560	**	**	2560		**
5120		**	5120		**	5120		**

Table 6. Nonparametric multiple comparisons among model designs performed separately at each sample size category (n), using the Nemenyi test (α=0.05). Sample size (n) refers to the number of presence records used to build each of the models. Those models that appear in the same group (e.g. G1 or G2) do not differ significantly in performance from one another. Group membership is indicated with by **. The mean and standard deviation (S.D.) refer to the AUC values derived from ROC curves.

n	Design	Mean	S.D.	G1	G2	n	Design	Mean	S.D.	G1	G2
40	FA	0.927	0.026	**		640	FA	0.948	0.006	**	
	PSA	0.913	0.039	**			PSA	0.956	0.003	**	**
	SA	0.938	0.012	**			SA	0.966	0.001		**
80	FA	0.934	0.013	**		1280	FA	0.95	0.003	**	
	PSA	0.946	0.007	**	**		PSA	0.957	0.002	**	**
	SA	0.953	0.009		**		SA	0.967	0		**
160	FA	0.942	0.009	**		2560	FA	0.951	0.002	**	
	PSA	0.952	0.005	**	**		PSA	0.957	0.002	**	**
	SA	0.962	0.004		**		SA	0.968	0		**
320	FA	0.946	0.009	**		5120	FA	0.952	0.002	**	
	PSA	0.956	0.004	**	**		PSA	0.957	0.001	**	**
	SA	0.965	0.002		**		SA	0.968	0		**

Table 7. Nonparametric multiple comparisons among prevalence categories, using the Nemenyi test (α=0.05). Those models that appear in the same group (e.g. G1 or G2) do not differ significantly in performance from one another. Group membership is indicated by **. The order in which the prevalence categories appear in the table is based on the rank-sums of AUC values calculated for these models, with the lowest performing models appearing at the top of the table.

Prevalence (%)	G1	G2	G3
90	**		
10	**		
20	**	**	
30		**	**
80		**	**
40		**	**
70		**	**
50			**
60			**

Table 8. Kruskal Wallis tests performed on models built with a range of prevalence categories (10%-90%) at a range of total sample sizes (n= presence + absence). Those models for which prevalence had a significant effect model performance (α=0.05) are indicated by "*".

n	H	N	df	Signif.
160	13.74	45	8	ns
320	17.02	45	8	ns
640	24.19	45	8	*
1280	15.15	45	8	ns
2560	38.55	45	8	*
5120	41.24	45	8	*

Table 9. Nonparametric multiple comparisons among prevalence categories for models built using total sample sizes (n) of 640, 2560 and 5120 records, using the Nemenyi test (α=0.05). Those models that appear in the same group (e.g. G1 or G2) do not differ significantly in performance from one another. Group membership is indicated by **. Only those models for which prevalence had a significant effect on model performance (Table 8) were tested here.

Total n	Prevalence	G1	G2	Total n	Prevalence	G1	G2	G3	G4	Total n	Prevalence	G1	G2	G3
640	10	**		2560	90	**				5120	90	**		
	90	**	**		10	**	**				10	**	**	
	20	**	**		20	**	**	**			20	**	**	**
	30	**	**		80	**	**	**	**		80	**	**	**
	40	**	**		30	**	**	**	**		30	**	**	**
	80	**	**		40		**	**	**		40	**	**	**
	50	**	**		70		**	**	**		50		**	**
	70	**	**		50			**	**		70		**	**
	60		**		60				**		60			**

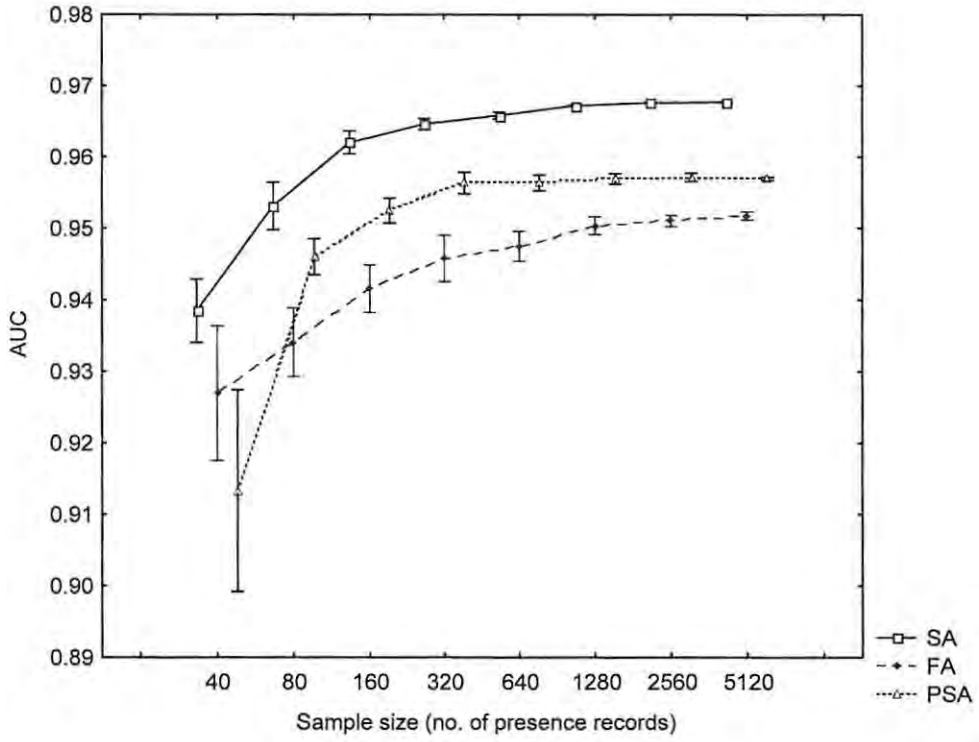


Figure 1. Mean AUC values vs. sample size for three sample designs: Pseudo-Absence (PSA) Surveyed Absence (SA) and False Absence (FA).

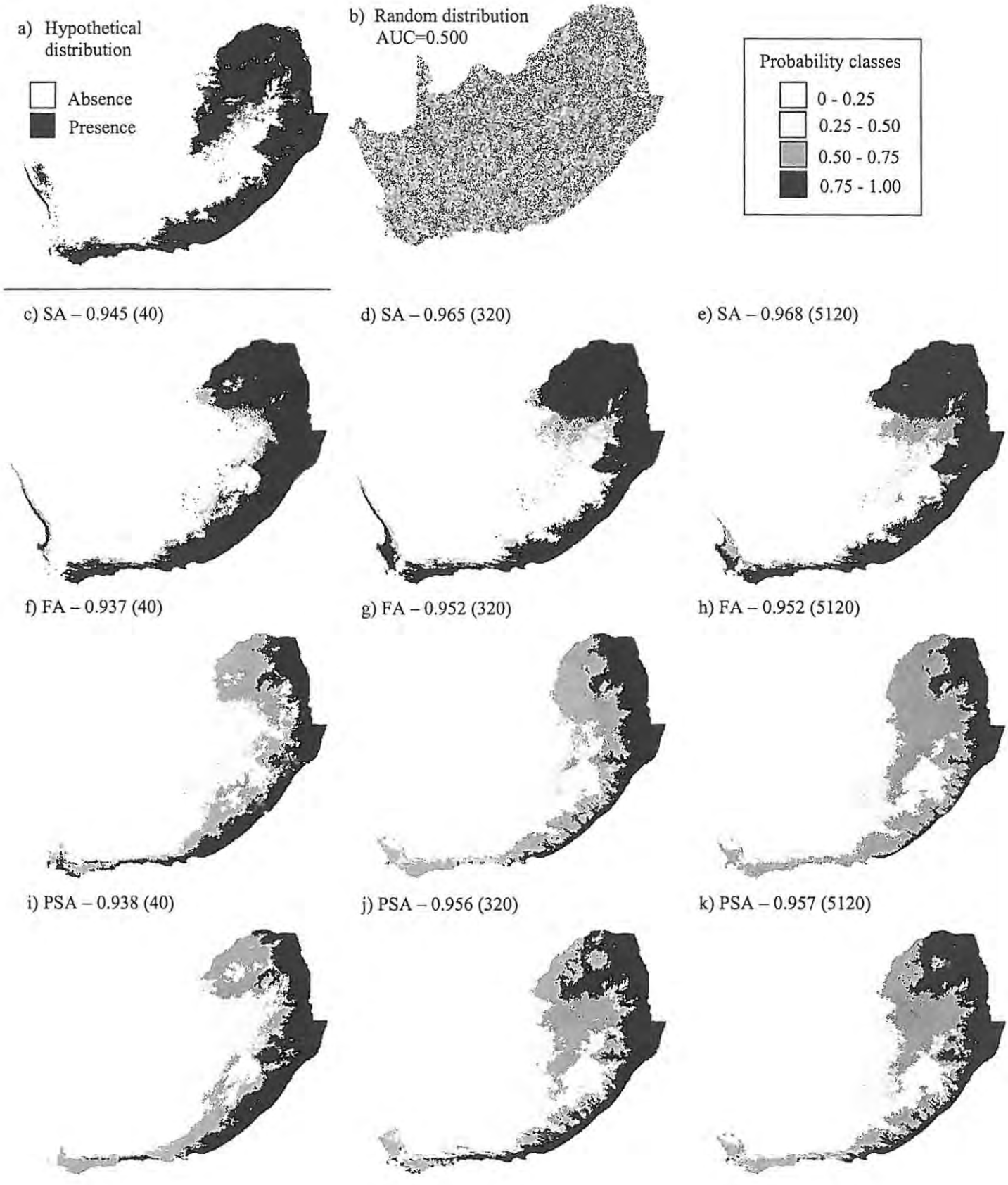


Figure 2. Potential distribution maps generated from logistic regression models in South Africa, Lesotho and Swaziland using three sample designs. The hypothetical distribution (a) has only two classes, namely presence and absence. The distribution maps (b-k) have been reclassified into 4 probability classes for the purposes of display. A random distribution (b), which has an AUC value of 0.500, has been included for the purposes of comparison. AUC values and sample sizes, appearing in round brackets, are given for each sample design (c-k). The maps presented in this figure are derived from individual models and the AUC values represent the performance of that model, whereas AUC values in Fig. 1 represent mean AUC values calculated from seven models fitted for each sample size and model design.

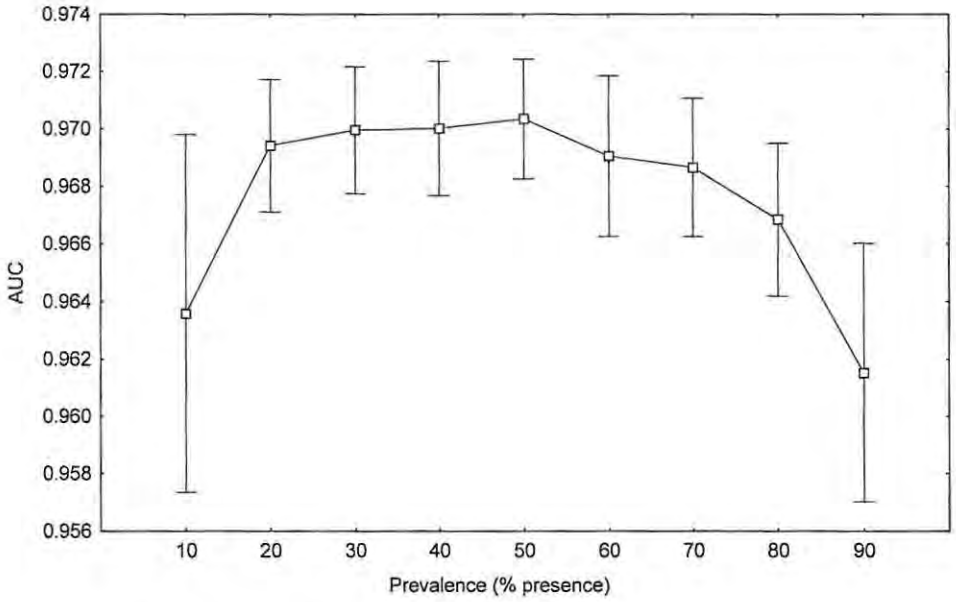


Figure 3. Mean AUC values with standard error bars at various prevalence categories.

References

Austin, M.P. 1998. An ecological prespective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*. 85: 2-17.

Austin, M.P., Cunningham, R.B., Fleming, P.M., 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio*. 55: 11-27.

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

Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a beta-function. *Journal of Vegetation Science*. 5: 215-228.

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

Bio, A.M.F., Alkemade, R., Barendregt, A. 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. *Journal of Vegetation Science*. 9: 5-16.

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

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

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

Cumming, G.S., 2000 a. Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*. 27: 425-440.

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

Erasmus, B.F.N., Kshatriya, M., Mansell, M.W., Chown, S.L., Van Jaarsveld, A.S., 2000. A modelling approach to antlion (Neuroptera: Myreleontidae) distribution patterns. *African Entomology*. 8: 157-168.

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*. 51: 331-363.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra, p. 193.

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

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*. 19: 474-499.

Freitag, S., Hobson, C., Biggs, H.C., Van Jaarsveld, A.S., 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*. 1: 119-127.

Funk, V.A., Richardson, K.S. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*. 51: 303-316.

Guisan, A., Theurillat, J-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*. 9: 65-74.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

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

Hirzel, A., Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*. 157: 331-341.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*. 145: 111-121.

Lawes, M.J., Piper, S.E., 1998. There is less to binary maps than meets the eye: the use of species distribution data in the southern African sub-region. *South African Journal of Science*. 94: 207-210.

Leathwick, J.R. 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*. 9:719-732.

Leathwick, J.R., Whitehead, D. 2001. Soil and atmospheric water deficits and the distribution of New Zealand's indigenous tree species. *Functional Ecology*. 15: 233-242.

Leathwick, J.R., Whitehead, D. McLeod, M. 1996. Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach. *Environmental Software*. 11:81-90.

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

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

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

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

McCullagh, P., Nelder, J.A. 1989. Generalized Linear Models. Chapman and Hall, London. p. 511.

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

Nicholls, A.O. 1989. How to make biological surveys go further with generalised linear models. *Biological Conservation*. 50: 51-75.

Nix, H.A., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*, Australian Government Publishing Service, Canberra, pp. 4-15.

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

Packer, M.J., Canney, S.M., McWilliam, N.C., Abdallah, R., 1999. Ecological mapping of a semi-arid savanna. In: Coe, M.J., McWilliam, N.C., Stone, G.N., Packer, M.J. (Eds.), *Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna*, Royal Geographical Society (with The Institute of British Geographers), London, pp. 43-68.

Palmer, A.R., Van Staden, J.M. 1992. Predicting the distribution of plant communities using annual rainfall and elevation: an example from southern Africa. *Journal of Vegetation Science*. 3: 261-266.

Peterson, A.T., 2001. Predicting species' geographic distributions based on ecological niche modeling. *The Condor*. 103: 599-605.

Pearce, J., Ferrier, S., 2000 a. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*. 128: 127-147.

Pearce, J., Ferrier, S., 2000 b. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*. 133: 225-245.

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

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

Rogers, D.J., Williams, B.G., 1993. Tsetse distribution in Africa: seeing the wood and the trees. In: Edwards, P.J., May, R. (Eds.), *Large-scale ecology and conservation biology*, Blackwell Scientific Publications, Oxford, pp. 247-271.

Rich, T.C.G., Woodruff, E.R., 1992. Recording bias in botanical surveys. *Watsonia*. 19: 73-95.

Robertson, M.P., Caithness, N., Villet, M.H., 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*. 7: 15-27.

Soberón, J., Llorente, J., Benítez, H. 1996. An international view of national biological surveys. *Annals of the Missouri Botanical Garden*. 83: 562-573.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science*. 13: 143-158.

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

Williams, B.G., Rogers, D.J., Staton, G., Ripley, B., Booth, T., 1994. Statistical modelling of georeferenced data: mapping tsetse distributions in Zimbabwe using climate and vegetation data. In: Perry, B.D., Hansen, J.W. (Eds.), *Modelling vector-borne and other parasitic diseases*, ILRAD, Nairobi, pp. 267-28

Yee, T.W., Mitchell, N.D. 1991. Generalized additive models in plant ecology. *Journal of Vegetation Science*. 2: 587-602.

Zaniewski, A.E., Lehmann, A., Overton, J. McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*. 157: 261-280.

Zar, J.H., 1996. Biostatistical analysis, 3rd Edition. Prentice-Hall, London, pp. 662.
Zweig, M.H., Campbell, G., 1993. Receiver-Operating Characteristic (ROC) Plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*. 39: 561-577.

VII

A quantitative comparison of the performance of selected profile and group discrimination predictive modelling techniques

Preface

This chapter quantitatively compares the performance of three profile techniques and two group-discrimination techniques using data from hypothetical and real organisms. Two of the profile techniques, the FEM technique and the PCA-based technique are those described and implemented in Chapters 4 and 5 respectively. The third profile technique is the CEM algorithm used in the BIOCLIM modelling package. The group discrimination techniques include the two model designs of logistic regression described in Chapter 6.

Abstract

This study quantitatively compares the performance of correlative modelling techniques that make use of presence-only locality records to predict species' potential distributions (profile techniques) with techniques that use both presence and absence locality records (group discrimination techniques). The performance of three profile techniques and two group discrimination techniques was evaluated, using four hypothetical distributions and eight real target organisms (cicadas, Homoptera: Cicadidae). The profile techniques included the algorithm implemented in the BIOCLIM modelling package, which is described as a Crisp Envelope Model (CEM), an envelope technique based on fuzzy classification (FEM) and a PCA-based technique (PCA). The group discrimination models included two logistic regression techniques, the Logistic Regression Surveyed Absence (LRSA) and the Logistic Regression Pseudo-absence (LRPSA). Random samples of locality records of varying size (40, 80, 160, 320, 640) were drawn from each of the hypothetical distributions

and used to make potential distribution predictions for each of these hypothetical organisms using all five model designs.

For the hypothetical distributions, the CEM and FEM model designs showed unexpectedly high performance relative to the other model designs. The way in which the hypothetical distributions were generated appeared to confer an unfair advantage on these two designs, thus reducing the usefulness of the hypothetical distribution investigation. The species investigation revealed that model design and sample size had a significant effect on model performance. These results suggest that, if presence and absence data are available, the LRSA model design should be selected in preference to the other designs. If only presence data are available, then the PCA model design should be selected, as it is likely to yield superior models more often than the CEM, FEM and LRPSA designs.

The PCA design did not differ significantly in performance from the LRSA design for six of the eight species, suggesting that profile techniques can produce equivalent results to group discrimination techniques under certain conditions. However, the sources of data that profile techniques typically rely on may be of poor quality, thus reducing their performance.

The results further suggest that optimal thresholds used to discriminate between presence and absence on continuous probability maps may differ among model designs and among species, suggesting that the meaning of the response surfaces produced by various model designs may be fundamentally different.

Introduction

Predictive modelling of species' distributions is becoming increasingly important in many fields of biology. A number of techniques are available for making predictions about the potential distribution of a species using distribution records and associated environmental predictor variables (Franklin 1995; Guisan and Zimmermann 2000; Chapter 3). The classification by Caithness (1995) of these techniques into group-discrimination and profile techniques is useful for separating models that make use of absence data from those that do not. Group-discrimination techniques are those that make use of distribution records that indicate localities where a target organism has been found to be present and localities where it has been found to be absent. Profile techniques are those that make use of presence locality records only.

Group discrimination techniques have been popular, especially Generalised Linear Models (GLM; Austin *et al.* 1984; Austin *et al.* 1990; Osborne and Tigar, 1992; Austin *et al.*, 1994; Ferrier and Watson, 1997; Guisan *et al.* 1998; Higgins *et al.*, 1999; Manel *et al.*, 1999 a & b; Cumming, 2000 a & b; Pearce and Ferrier, 2000) and Generalised Additive Models (GAM; Austin and Meyers 1996; Ferrier and Watson, 1997; Leathwick *et al.* 1996; Leathwick, 1998; Pearce and Ferrier, 2000; Hirzel *et al.*, 2001; Leathwick and Whitehead, 2001).

In order to produce reliable models using group discrimination techniques, ideally a survey should be conducted using a stratified sampling approach to reliably establish the presence or absence of a target species at a large number of sites throughout the map region by means of a field survey (Austin, 1998). These field surveys tend to expensive, labour-intensive and time-consuming. As a result, presence-only data may often be the only source of data available for predictive modelling. Sources of presence-only data typically include museum and herbarium collection records where data have been collected on an *ad hoc* basis rather than by means of a well design field survey. For data that have been collected on an *ad hoc* basis, absence data are often not available since usually only the presence and not absence of organisms is recorded (Margules and Austin, 1994; Stockwell and Peters, 1999; Hirzel *et al.*, 2001; Peterson, 2001). However, the presence data are often geographically biased (Margules and Austin, 1994; Austin, 1998; Freitag *et al.*, 1998;

Lawes and Piper, 1998; Funk and Richardson, 2002; Ferrier, 2002; Zaniewski *et al.* 2002).

Various profile modelling techniques have been developed so that presence data can be used when presence/absence data are either not available or are unreliable. Examples of these techniques include models developed by Palmer and Van Staden (1992), Erasmus *et al.* (2000), Robertson *et al.* (2001), Hirzel *et al.* (2002) and the approaches used in the modelling packages known as BIOCLIM (Nix, 1986; Busby, 1991) and DOMAIN (Carpenter *et al.*, 1993). Various studies have investigated the use of pseudo-absence data (Ferrier and Watson, 1997; Cumming, 2000 a & b; Zaniewski *et al.*, 2002) that can be used when absence data are either not available or unreliable.

In a recent review, Guisan and Zimmermann (2000) highlighted the need for comparisons to be made among modelling techniques. Several studies have compared the performance of various predictive modelling techniques (Carpenter *et al.* 1993; Franklin 1998; Manel *et al.* 1999 a & b, Özesmi and Özesmi, 1999; Cumming 2000 b, Hirzel *et al.*, 2001). Most comparisons have been made among group-discrimination techniques. Carpenter *et al.* (1993) compared two profile techniques with a group-discrimination technique but these comparisons were not quantitative. Ferrier and Watson (1997) made quantitative comparisons among various profile and group discrimination techniques. Recently, Hirzel *et al.* (2001) and Zaniewski *et al.* (2002) quantitatively compared the performance of profile and group discrimination techniques.

Although comparisons have been made among certain profile and group discrimination techniques, further comparisons among different techniques are needed. It also is not known whether more sophisticated profile techniques such as PCA (Robertson *et al.*, 2001) can yield models with significantly better performance than those built using less sophisticated techniques such as BIOCLIM (Nix, 1986; Busby, 1991).

This study attempts to determine: whether model design and sample size have a significant effect on model performance; which profile model design(s) produce models with the highest average performance; and whether profile model designs can produce models that perform equivalently on average to models produced from group

discrimination model designs. In addition, the validity of using a single threshold with which to assess model performance is also investigated.

The performance of three profile and two group discrimination techniques (five model designs) were compared over a range of sample sizes using hypothetical organisms and real target organisms (cicadas).

The profile techniques included the CEM algorithm implemented in the BIOCLIM modelling package (Nix, 1986; Busby, 1991), the Fuzzy Envelope Model (FEM, Chapter 4) and a PCA model (Robertson *et al.*, 2001; Chapter 5). As the algorithm used in BIOCLIM (recently renamed to "ANUCLIM") can be classified as an envelope technique (Chapter 3), I refer to this design as the "Crisp Envelope Model" (CEM). This design has also been referred to as a "boxcar" model since it is analogous to the "parallel-piped" or "boxcar" image classification algorithm used in remote sensing (Carpenter *et al.*, 1993).

The group discrimination models included two designs of logistic regression model, the Logistic Regression using Surveyed Absence (LRSA, Chapter 6) and the Logistic Regression using pseudo-absence (LRPSA, Chapter 6) models.

Methods

Quantitative model performance comparisons were made among five model designs (CEM, FEM, PCA, LRSA and LRPSA) and across five sample size categories (40, 80, 160, 320 and 640 presence records) using four hypothetical distributions (Fig. 1). In addition, comparisons were also made among the five model designs using locality records collected for eight cicada species (Table 1). Cicadas are true bugs (Homoptera: Hemiptera) of the family Cicadidae.

The hypothetical distributions

Four hypothetical distributions were produced from simple envelope models (Chapter 3) that used localities where four real organisms (cicadas) were recorded as being present, and the predictor variables given in Table 2. In the hypothetical distribution maps, grid-cells were coded 1 to represent presence of the hypothetical organism, and coded 0 to represent absence of the hypothetical organism. These

presence-absence maps were taken to represent the "known" distribution of the hypothetical target organisms. The hypothetical distributions were generated using locality records from the following cicada species: a) *Albanycada albigera* Walker, b) *Capicada decora* Germar, c) *Platypleura deusta* Thunberg, and d) *Platypleura haglundi* Stål (Fig. 1).

For each hypothetical distribution, samples consisting of 40, 80, 160, 320 and 640 locality records were randomly selected from the presence and the absence categories of the map region to represent surveyed presence and surveyed absence localities respectively. This selection was repeated seven times so that each sample size category for each hypothetical distribution had seven replicates. An algorithm was developed to ensure that there was a spacing of at least two grid-cells between any of the locality record grid-cells.

Distribution predictions were made using the values extracted from a set of environmental predictor variables (Table 2) associated with the samples of locality records for each of the five model designs.

The cicada distributions

Potential distribution models were produced for eight species of cicada (Table 1). Locality data were obtained from the Albany Museum (Grahamstown), Durban Museum (Durban), National Museum of South Africa (Bloemfontein), National Collection of Insects (Pretoria), Natal Museum (Pietermaritzburg), Natural History Museum (London), Museum für Naturkunde (Berlin), Transvaal Museum (Pretoria), Rhodes University (Grahamstown), Pretoria University (Pretoria), and the private collections of Isak Coetzer, Rudi Mijburgh, Renzo Perissinotto, Martin Villet, Richard Stephen, Tony Ewart and published records of Michelle Boulard. For each species the available locality records were partitioned into a training and an evaluation dataset (using a 3:1 ratio). For each species, partitioning was repeated 25 times so that different combinations (but the same number) of training and evaluation localities were used. The numbers of presence and absence localities in the training and evaluation sets are listed (Table 1). The models were developed using the localities from the training datasets and the performance of these models was calculated using the independent localities from the evaluation datasets.

These species were selected as target species for the following reasons. Their collection records were fairly easily accessed, there is a fair understanding of their biology, they are taxonomically distinct, they occupy a range of different habitats, they are indigenous and their ranges appeared in most cases to be contained within the map region. The number of available locality records (sample size) was also taken into account, to allow a range of sample sizes to be considered.

Model designs

In the case of the CEM, FEM and PCA model designs, only localities representing the presence of the hypothetical organisms were used to develop the models. In the case of the LRSA model design, samples of localities representing the presence and the absence of the hypothetical organisms were used to develop the models. The number of absence records selected was the same as the number of presence records. For the LRPSA design, a sample of localities representing the presence of the hypothetical organism represented the "surveyed presence" category and a sample of the remaining grid-cells in the map region were taken to represent "pseudo-absence" localities. In all cases the size of the pseudo-absence sample was the same as that of the surveyed presence sample.

CEM design

The CEM design is well known as the algorithm used by BIOCLIM (Nix, 1986; Busby, 1991) and has been used to predict distributions for a number of organisms (for examples see Chapters 3 and 4). To produce a prediction using the CEM design, core and marginal ranges have to be calculated for the target species based on the values in the training set. The marginal range was determined by reclassifying each predictor variable map using the maximum and minimum values of the training data for each predictor variable. The reclassified maps were superimposed using the intersection (AND) function in Boolean logic (Heuvelink and Burrough, 1993) to produce a map indicating the marginal range of the species. Similarly, a second map of the core range was produced by reclassifying each predictor variable map using the 10^{th} and 90^{th} percentiles of the training data as boundaries of the core range, as

defined by Lindenmayer *et al.* (1991). The core and marginal range maps were superimposed to produce a single map indicating the core and marginal ranges for each species.

FEM design

The FEM design (Chapter 4) represents a refinement of the CEM design. The FEM algorithm classifies the grid cells in each predictor variable map using an appropriate sigmoidal fuzzy membership function. The sigmoidal membership function can have symmetric, monotonically increasing or monotonically decreasing forms. Frequency histograms of the training data were examined for each variable (for each hypothetical organism and cicada species) to determine the appropriate function form (symmetric, monotonically increasing or monotonically decreasing) with which to classify each predictor variable.

The shape of the membership function is governed by four control points that are ordered from low to high on the measurement scale of the predictor variable axis. For the symmetric membership function, points "a" and "d" were assigned the minimum and maximum values respectively and points "b" and "c" were both assigned the median value from the training data set.

To produce the final distribution map, all of the fuzzily classified predictor variable maps are superimposed using a minimum overlay function.

PCA design

The PCA design uses the PCA-based modelling technique described by Robertson *et al.* (2001; see also Chapter 5). This technique constructs a hyperspace for the target species using principal components axes derived from a principal components analysis performed on the training dataset. The training dataset comprises the values of the predictor variables associated with presence records for the target species. The origin of this hyperspace is taken to represent the centre of the niche of the species. All the grid-cells in the map region are then fitted into this hyperspace using the values of the predictor variables at these grid-cells. The Euclidean distances from each of the localities to the origin of the hyperspace gives a

measure of the 'centrality' of those localities in the hyperspace and these distances are used to derive probabilities for each of the grid-cells in the map region.

The CEM, FEM and PCA algorithms were implemented in MATLAB and visualisation of distribution maps was done using IDRISI32.

LRSA and LRPSA designs

The LRSA and LRPSA designs were based on logistic regression, a form of Generalised Linear Model (GLM: McCullagh and Nelder, 1989) in which a binomial error distribution and a logistic link function are used (Guisan and Zimmermann, 2000). For the LRSA design, the dependent variable consisted of presence and surveyed absence localities and the LRPSA consisted of presence and pseudo-absence localities (see Chapter 6). The independent variables consisted of the predictor variables listed in Table 2 as well as the squares of selected variables. For each hypothetical distribution or species, scatter plots were examined to determine which predictor variables should be squared.

Model fitting was performed using GLMFIT, a generalised linear modelling function within MATLAB. A binomial distribution was used for the error function and a logit link function was used to calculate the values of the regression coefficients (β_i) of the linear predictor (Eq. 1).

A backwards elimination procedure was used to remove variables that did not significantly improve model fit. A predictor variable was removed if it did not result in a significant reduction in the deviance calculated from the full model (a model containing all the predictor variables). Significance tests were based on the assumption that the change in deviance followed a chi-square distribution. A threshold of 0.20 was used for exclusion of variables as several authors have suggested that a threshold of 0.05 may be too stringent (Pearce and Ferrier, 2000).

Probability values for each grid cell in the map region were calculated by substituting the values of the predictor variables associated with that cell into the linear predictor (Eq. 1) and then transforming these values using the inverse logistic transform (Eq. 2).

$$\beta_1 variable_1 + \beta_2 variable_2 + \ldots + \beta_8 variable_8 = \eta \quad (1)$$

$$P_{(y=1)} = \exp(\eta)/(1+\exp(\eta)) \quad (2)$$

Model evaluation

Most quantitative model performance tests are based on a confusion matrix in which the observed (actual) and predicted presence/absence patterns are cross-tabulated (Fielding and Bell, 1997). Various threshold-dependent and threshold-independent accuracy measures can be calculated from the confusion matrix (reviewed by Fielding and Bell, 1997). Threshold-independent measures (e.g. ROC curves) are considered to be more robust and more objective than threshold-dependent measures (e.g. Kappa statistics) since they do not rely on a single threshold to distinguish between predicted presence and predicted absence (Fielding and Bell, 1997). Since the CEM design does not produce a map of continuous values such as those produced by the other designs, threshold-independent accuracy measures based ROC curves could not be calculated for this design. As a result a threshold-dependent measure was required. Thus, the kappa statistic was calculated for the models of each design.

In order to calculate the value of the kappa statistic for quantifying model performance, the parameters of the confusion matrix have to be calculated (Fielding and Bell, 1997). These parameters are calculated using "observed" presence and absence localities as well as "predicted" presence and absence localities for a target species. For both the cicada predictions and those predictions produced using samples drawn from the hypothetical distributions (the hypothetical distribution predictions), those grid-cells in the predicted distribution maps with values greater or equal to the threshold represented the "predicted presence" category and the remaining grid-cells represented the "predicted absence" category. Kappa values were calculated for all probability or suitability values produced by the models, and then the maximum value of kappa (κmax) was selected as a measure of overall performance for that model. The threshold associated with κmax can be regarded as an optimum threshold for the model (Guisan and Zimmermann, 2000) as it is the threshold at which the highest

value of the performance measure can be calculated. This optimum threshold approach to model evaluation has been used elsewhere (Franklin, 1998; Guisan *et al.*, 1998) and is discussed in Chapter 2.

The value of κmax was calculated for the models of all the designs except for the CEM design. Since the CEM design does not produce a map of continuous values, kappa was calculated at the threshold defined by the marginal range of the model. For the other model designs, κ-values were calculated at all thresholds and then the maximum value of κ was selected as a measure of performance for that model (as outlined above).

For the cicada predictions, the presence and absence testing locality records (those records reserved for evaluation in the partition) were taken to represent the "observed presence" and "observed absence" localities respectively. For the models built using samples drawn from the hypothetical distributions, the presence (coded 1) and absence (coded 0) grid-cells of these hypothetical distribution maps represented the "observed presence" and "observed absence" localities respectively. This is equivalent to sampling every grid-cell in the map region to determine whether it falls into the "observed presence" or "observed absence" category. This means that predictions generated from the model can be compared with an independent set of records representing the "observed" distribution over the entire map region, resulting in one of the most objective tests of model performance possible (see Chapter 6).

Results

The ranges of agreement for the kappa statistic proposed by Monserud and Leemans (1992) are used here to describe the results. These ranges are: no agreement <0.05; very poor 0.05-0.20; poor 0.20-0.40; fair 0.40- 0.55; good 0.55-0.70; very good 0.70-0.85; excellent 0.85-0.99 and perfect 0.99-1.00 (Monserud and Leemans, 1992).

Hypothetical distributions

Model design had a significant effect on model performance for all four hypothetical distributions (Table 3). Sample size has a significant effect for three of

the four (B, C & D) and there were significant interactions between model design and sample size for two of the four distributions (A & B; Table 3).

The range in overall model performance for the five model designs varied among the hypothetical distributions (Fig. 1 & 2). Average model performance ranged from good to excellent for hypothetical distributions A and C, from poor to excellent for hypothetical distribution B, and from very poor to excellent for hypothetical distribution D (Fig. 1 & 2). The CEM design performed best for hypothetical distributions A and B (Fig. 1). For hypothetical distribution B, the CEM performed considerably better than the other designs across all sample size categories (Fig. 1). The CEM and FEM designs both performed better than the other designs (LRSA, LRPSA and PCA) for hypothetical distributions C and D. The PCA design performed poorly in relation to the other designs for hypothetical distributions B and D. The LRPSA design performed very poorly in relation to the other designs for hypothetical distribution D.

A comparison of model performance among model designs for each sample size category allows those models that differ significantly on average from one another can be identified. Those model designs that are members of the same group do not differ significantly from one another while those that do not appear in the same group are significantly different (Table 4). The average performance of CEM design was significantly higher than the remaining model designs across all sample size categories, except for sample size category of 40 records where it did not differ significantly from the FEM design. The FEM design did not differ on average in performance from the LRSA design across all sample size categories. The PCA and LRPSA model designs consistently appeared in the group of lowest average performance across all sample size categories.

Cicada distributions

The results of a two-way ANOVA performed on data pooled across all eight species indicated that model design and sample size had significant effects on model performance (Table 5). There was also a significant interaction between model design and sample size (Table 5).

The overall performance of the models produced for the cicada species (Fig. 3-6) was higher than those produced for the hypothetical distributions (Fig. 1 & 2). Average performance of the cicada distribution predictions ranged from good to excellent. In contrast, the performance of the hypothetical distribution predictions ranged from poor to excellent.

Average model performance ranged from good to excellent for *P. haglundi*; from very good to excellent for *A. albigera*, *C. decora*, *P. deusta,* and *P. divisa* (Fig. 3-6). For *P. mijburghi, P. capensis* and *Pycna semiclara* the performance of all designs was excellent.

There was a fair amount of variation in the relative performance among designs for each species (Fig. 3-6). No single design performed significantly better than all others across all eight species. Similarly, no single design performed significantly worse than all others across all eight species (Fig. 3-6). Tukey HSD tests were performed to determine which designs did not differ significantly from one another. Those designs in the same group do not differ significantly from one another but do differ significantly from those designs that are not in the same group. For example, in the case of *A. albigera* (Fig. 3) the CEM and FEM designs occur in the same group (G1) and thus the performance of these two designs does not differ significantly. However, the performance of these two designs is significantly lower than the performance of those designs that appear in the second group (LRPSA, PCA, LRSA).

For each species the Tukey HSD analysis revealed only two groups, with the exception of *P. deusta* where there were three groups. Table 6 indicates the frequency with which a particular model design appeared in the group of lowest performance and the frequency with which it appeared in the group of highest performance. It is possible for a design to appear in both groups e.g. in the case of *P. capensis* the FEM design appears in both the lower and the upper groups. Table 6 summarises the overall performance of the designs. If a particular design appears in the upper group more times than another design, then its overall performance can be considered to be better than the other design. The LRSA design appeared in the upper group eight times and never in the lower group (Table 6). This indicates that the LRSA was always in the upper group and that on average no other design performed significantly better. The PCA design appeared in the upper group six times and in the lower group three times. On average the performance of the PCA design was inferior to that of the

LRSA design but superior to that of the LRPSA design (Table 6). This is because the LRPSA design appeared five times in the upper group and four times in the lower group (Table 6). The CEM and FEM designs both appeared in the lower group eight times indicating that their performance was on average inferior to the other designs (Table 6).

A 2-way ANOVA indicated highly significant differences in the threshold at which the maximum kappa value (κmax) was calculated among model designs ($F=250.71$, d.f.=3, $p<0.001$) but not among species ($F=1.39$, d.f.=7, $p=0.206$). There was a significant interaction between model design and species ($F=2.24$, d.f.=21, $p<0.001$). The CEM design was excluded from this analysis since the κ statistic could only be calculated at one threshold for models of this design.

Discussion

Hypothetical distribution predictions

The use of a hypothetical distribution as a basis for building and evaluating models has been used previously (Cumming, 2000 b; Hirzel *et al.*, 2001; Hirzel and Guisan, 2002; Chapter 6). In the current study a hypothetical distribution was generated from a simple envelope model, which was built using a set of locality records for real organisms and a suite of environmental predictor variables. The same suite of predictor variables that was used to generate the hypothetical distribution was then also used to make predictions for the hypothetical organism, but using different samples of point records. The advantage of this approach is that environmental variables used as predictor variables can exhaustively describe the realised niche of the organism. As a result the suite of predictor variables represent the best possible set of variables for predicting the distribution of that hypothetical organism.

When a hypothetical distribution is used, then the entire distribution of the hypothetical organism is known. This allows model evaluation to be done using the entire hypothetical distribution.

The hypothetical distribution results suggest that both model design and sample size significantly affected model performance. For all four hypothetical distributions

model performance differed significantly with sample design and for three hypothetical distributions model performance differed significantly with sample size.

There was considerable variation in both the relative and absolute performance of the model designs among the four hypothetical distributions.

The CEM performed on average significantly better than all the other designs for each of the sample size categories, except at a sample size category of 40 where it did not differ significantly from the FEM design. This high relative performance of the CEM model design is unexpected. Similarly the FEM design also had an unexpectedly high performance. In contrast, Ferrier and Watson (1997), using data from real organisms, found that models generated using GLM and GAM performed significantly better than BIOCLIM models, which are the same as the CEM design in the current study.

These results can be explained as follows. Since both the CEM and FEM designs are envelope techniques, they had an advantage over the other designs because the hypothetical distributions were generated using a simple envelope technique.

The results from the species models suggest that the CEM and FEM designs perform worse on average than the other designs, which is in contrast to the results obtained from the hypothetical distribution predictions. The comparisons of model performance among model designs are thus probably more reliable from the cicada species investigation than from the hypothetical distribution investigation.

Hypothetical distributions may be useful for investigating certain data quality and model design issues (Cumming, 2000 b; Hirzel *et al.*, 2001; Hirzel and Guisan, 2002; Chapter 6), however conclusions drawn from these should be made with care. While hypothetical distributions may be similar to distributions of real organisms, there is no guarantee that these hypothetical distributions are reasonable surrogates for distributions of real organisms. Care should be taken to ensure that the method of generating the hypothetical distribution does not confer advantages to certain model designs over others (Hirzel *et al.*, 2001).

Cicada predictions

The results suggest that model design and sample size had significant effects on model performance. There was also a significant interaction between model design and sample size.

No single design performed significantly better than all other designs across all eight species. Similarly no single design performed significantly worse than all other designs across all eight species. For a given species, up to four model designs produced models whose performance did not differ significantly on average from one another. For seven of the eight species the performance of the model designs could be classified into two groups (an upper and a lower group), each containing models whose performance did not differ significantly from one another (using Tukey HSD multiple comparisons). The LRSA design most frequently produced models that occurred in the group with highest average performance (the upper group), followed by the PCA design and then the LRPSA design. The CEM and FEM designs produced models with the lowest average performance. Ferrier and Watson (1997) found that models generated from GLM and GAM performed significantly better than BIOCLIM models (equivalent to the CEM design). However, Hirzel *et al.* (2001) found that ENFA models (similar to the PCA design) performed better than GLM (with linear and quadratic terms) under certain data quality regimes.

The overall performance of a model depends on several factors including model design, sample size, data quality and the biology of the target species. These factors may explain the differences in performance among the model designs compared here. While the PCA and LRPSA designs performed fairly well relative to the LRSA design, this may not always be the case. In the comparison made in this study the PCA and LRPSA designs were built using the same data as the LRSA design. However, the sources of presence-only data that the PCA and LRPSA designs will typically rely on usually come from herbarium or museum collections. Data from these sources have a number of weaknesses (Funk and Richardson, 2002; Zaniewski *et al.*, 2002), the most serious of these is that samples obtained from these sources often contain bias (Margules and Austin, 1994; Austin, 1998; Freitag *et al.*, 1998; Lawes and Piper, 1998; Funk and Richardson, 2002; Ferrier, 2002; Zaniewski *et al.* 2002). This is likely to reduce the performance of models built using these data.

LRSA models are typically built from data that have been obtained by means of systematically designed surveys (Austin, 1998) where the presence and absence of the target organism is reliably established. Models built from these data sources are almost always likely to be superior to models built from data obtained from collections.

The cicada species models were built and evaluated using relatively small (<80 presence records) sets of locality records (Table 1) and different trends may emerge when a larger range of sample sizes is considered (e.g. see Chapter 6). Although the sizes of locality record sets used here are small, this is a real problem and probably represents the best dataset available for these cicada species (see list of data sources in the methods section). Small sample sizes such as these are probably representative of many other insect taxa, especially in the case of rare species. For example, in another study (Robertson and Villet, *in prep.*) all existing southern African records for three species of sliphid beetle (*Slipha punctulata*, *Thanatophilus micans* and *T. mutilatus*) have been assembled from museum collections and the literature. Despite considerable effort, very few records suitable for making predictions at minute-resolution are available (15, 73 and 85 presence records only respectively). Further studies using larger sample sizes, especially for evaluation, would complement the current study.

The overall performance of the models produced for the cicada species was higher than those produced for the hypothetical distributions, which suggests that this may be due to the way in which the hypothetical distributions were generated or because the evaluation of these models was more rigorous than the approach used for the species-based models. The evaluation of the hypothetical distribution predictions was based on the entire hypothetical distribution map, which was regarded as the "known" distribution in the evaluation. In contrast, the species models were evaluated using a relatively small sample of records that were probably not entirely independent from the training records.

The results of the optimal-threshold investigation suggest that optimal thresholds used to discriminate between presence and absence on continuous probability maps do differ among model designs. This suggests that the meaning of the response surfaces produced by various model designs may be fundamentally different, supporting similar conclusions drawn elsewhere (Chapter 8). Furthermore,

comparisons among model designs should not be done using a single threshold. Manel *et al.* (1999 a & b) used a single threshold (of 0.5) at which to define species presence for the purposes of comparing prediction success among models developed using discriminant analysis, logistic regression and artificial neural networks designs. The results obtained here suggest that this approach should be avoided. Instead, where possible, comparisons should be made using threshold-independent performance measures (e.g. ROC curves) or performance measures should be calculated using optimal thresholds (Guisan and Zimmermann, 2000), such as the approach used in this and other studies (Franklin, 1998; Guisan *et al.*, 1998; Collingham, *et al.*, 2000).

The use of pseudo-absence data

The use of pseudo-absence records in models for predicting potential distributions when true absence data are unavailable has been investigated previously using GLM and GAM (Ferrier and Watson, 1997; Cumming, 2000 a & b; Zaniewski *et al.* 2002). The success of this approach is likely to depend on the method used to define the pseudo-absence records.

The method used by Ferrier and Watson (1997) for generating pseudo-absence records was the same as that used in this study. In contrast, the methods used by Zaniewski *et al.* (2002) to generate pseudo-absence data differed from that used here. Zaniewski *et al.* (2002) used a group of 43 fern species for their study. Plots containing the target species were taken to represent species presence while all those plots that contained non-target species were available for selection as pseudo-absence records. The selection of plots to be used as pseudo-absence records was then done in two different ways. In the first method, plots were selected randomly. In the second method, plots with environmental attributes that were most similar to a set of plots representing true absence were selected to represent pseudo-absence.

In the current study, the records for non-target species were not used to define pseudo-absence. Instead a random sample of grid-cells where the target species was not recorded as being present were taken to represent pseudo-absence records. There was no reliance on a group of species having been surveyed, which is probably more realistic of situations in which true presence-only data are used.

Cumming (2000 a & b) successfully used a pseudo-absence approach in which all the grid-cells in the map region that did not contain the target species were taken to represent the absence of that species. This approach can be problematic when the number of presence records is small relative to the total number of grid-cells in the map region, as large differences in group sizes (prevalence) are likely to result in problems (see Chapter 6).

Conclusion

The results of this study suggest that model design and sample size significantly affect model performance. Significant interactions between model design and sample size may occur. For the hypothetical distributions the CEM and FEM model designs showed unexpectedly high performance relative to the other model designs. The way in which the hypothetical distributions were generated appeared to confer an unfair advantage on two of the designs (FEM and CEM), thus reducing the usefulness of the hypothetical distribution investigation. This suggests that care should be taken when using hypothetical distribution comparisons as there is no guarantee that these hypothetical distributions are realistic or that they do not confer an unfair advantage on certain model designs.

The species investigation found that model design and sample size had a significant effect on model performance. On average, the LRSA design most frequently produced models that occurred in the group (with other model designs whose performance did not differ significantly), that had the highest average performance, followed by the PCA and then the LRPSA designs. The CEM and FEM designs most frequently produced models with the lowest average performance.

These results suggest that if presence and absence data are available then the LRSA model design should be selected in preference to the other designs. If only presence data are available then the PCA model design should be selected, as it is likely yield superior models more often than the CEM, FEM and LRPSA designs.

The PCA design did not differ significantly in performance from the LRSA design for six of the eight species, suggesting that profile techniques can produce equivalent results to group discrimination techniques under certain conditions.

However, the sources of data that profile techniques typically rely on may be of poor quality, thus reducing their performance.

The results of the optimal-threshold investigation suggest that optimal thresholds used to discriminate between presence and absence on continuous probability maps can differ significantly among model designs. This suggests that the meaning of the response surfaces produced by various model designs may be fundamentally different, and that comparisons among model designs should not be done using performance measures that use a single threshold only.

This study is significant in that it quantitatively compares the performance of five correlative modelling techniques (three profile and two group discrimination) using real and simulated data. In their review, Guisan and Zimmermann (2000) noted the paucity of studies such as the current one, in which more than two statistical techniques were compared, and they further highlighted the need for such studies, a view that must be reiterated here.

Acknowledgements

I thank Clyde Mallinson of Geodatec GIS Consultants for use of computer facilities; Sarah Radloff for statistical advice; Mike Burton for advice with MATLAB software; Lesley Henderson at the Southern African Plant Invaders Atlas for locality data; Craig Peter for collecting some of the alien plant locality data; The National Botanical Institute for the use of data from the National Herbarium, Pretoria (PRE) Computerised Information System (PRECIS). I thank Fred Gess (Albany Museum), Clive Quickleberg (Durban Museum), Hamish Robertson (National Museum of South Africa), Ian Millar (National Collection of Insects), Jason Londt (Natal Museum), Mick Webb (Natural History Museum, London), Jeurgen Deckert (Museum Für Naturkunde, Berlin), Rob Toms (Transvaal Museum), Clarke Scholtz (Pretoria University), Isak Coetzer, Tony Ewart, Rudi Mijburgh, Renzo Perissinotto, Richard Stephen and Martin Villet for cicada distribution data. I thank the School of Bioresources Engineering and Environmental Hydrology (University of Natal), the Water Research Commission and the South African Country Study for Climate Change for the use of the climatic predictor variables. Funding from the National

Research Foundation and the Rhodes University Joint Research Council is gratefully acknowledged.

Table 1. The cicada species for which predictive distribution models were produced. The number of presence and absence localities of the training and evaluation sets are also indicated.

	Training		Evaluation	
Species	No. pres	No. abs	No. pres	No. abs
Albanycada albigera	27	27	5	5
Capicada decora	27	31	5	6
Platypleura capensis	78	80	16	16
P. deusta	23	34	5	7
P. divisa	42	42	8	9
P. haglundi	67	73	13	15
P. mijburghi	42	54	9	11
Pycna semiclara	67	65	14	13

Table 2. Predictor variables selected for building the distribution models.

No.	Predictor variable
1	Monthly potential evaporation - January
2	Monthly potential evaporation - July
3	Monthly maximum temperature - January
4	Monthly minimum temperature - July
5	Monthly rainfall - January
6	Monthly rainfall - July

Table 3. Results of 2-way ANOVAs conducted separately for each hypothetical organism.

Hypothetical distribution A					
	SS	df	MS	F	p
Intercept	116.232	1	116.232	17406.87	0.000
Model	1.826	4	0.457	68.38	0.000
Samp. Size	0.017	4	0.004	0.62	0.648
Model*size	0.256	16	0.016	2.40	0.003
Error	1.002	150	0.007		
Hypothetical distribution B					
	SS	df	MS	F	p
Intercept	62.396	1	62.396	12542.46	0.000
Model	8.523	4	2.131	428.30	0.000
Samp. Size	0.220	4	0.055	11.06	0.000
Model*size	0.779	16	0.049	9.79	0.000
Error	0.746	150	0.005		
Hypothetical distribution C					
	SS	df	MS	F	p
Intercept	133.122	1	133.122	74330.27	0.000
Model	1.143	4	0.286	159.62	0.000
Samp. Size	0.191	4	0.048	26.65	0.000
Model*size	0.033	16	0.002	1.15	0.317
Error	0.269	150	0.002		
Hypothetical distribution D					
	SS	df	MS	F	p
Intercept	83.388	1	83.388	31783.03	0.000
Model	11.523	4	2.881	1098.00	0.000
Samp. Size	0.143	4	0.036	13.65	0.000
Model*size	0.046	16	0.003	1.09	0.371
Error	0.394	150	0.003		

Table 4. Results of Tukey HSD multiple comparisons comparing model designs (model) across four hypothetical distributions at sample sizes of 40, 80, 160, 320 and 640 presence records (n). Those models that appear in the same group (e.g. G1, G2, G3 etc.) do not differ significantly in performance from one another. Group membership is indicated by **. The mean value of the kappa statistic is given for each model design.

n	Model	Mean	G1	G2	G3	G4
40	LRPSA	0.536	**			
	PCA	0.574	**	**		
	LRSA	0.683		**	**	
	FEM	0.787			**	**
	CEM	0.910				**
n	Model	Mean	G1	G2	G3	G4
80	PCA	0.578	**			
	LRPSA	0.580	**			
	LRSA	0.770		**		
	FEM	0.777		**		
	CEM	0.956			**	
n	Model	Mean	G1	G2	G3	G4
160	PCA	0.591	**			
	LRPSA	0.629	**			
	FEM	0.783		**		
	LRSA	0.806		**		
	CEM	0.979			**	

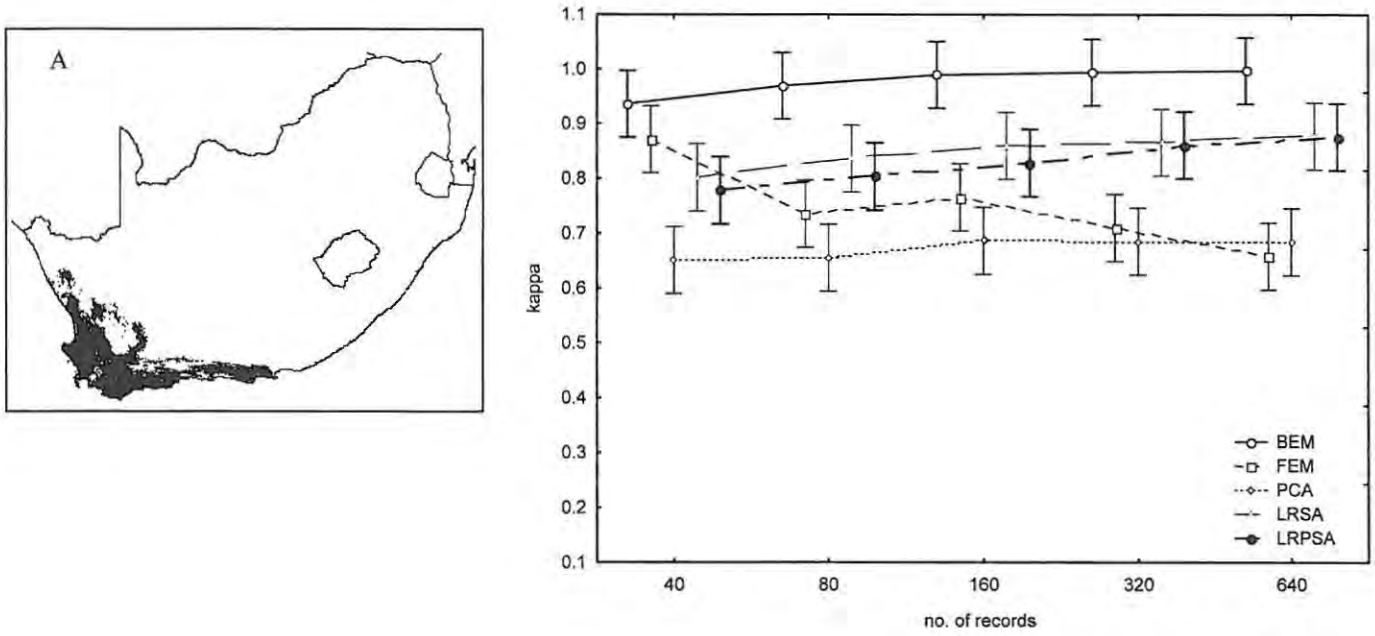
n	Model	Mean	G1	G2	G3	G4
640	PCA	0.583	**			
	LRPSA	0.659	**	**		
	FEM	0.767		**	**	
	LRSA	0.837			**	
	CEM	0.988				**
n	Model	Mean	G1	G2	G3	G4
320	PCA	0.591	**			
	LRPSA	0.652	**	**		
	FEM	0.762		**	**	
	LRSA	0.828			**	
	CEM	0.984				**

Table 5. Results of a 2-way ANOVA on model design (Model) and sample size (samp. size) performed on model performance data (based on the kappa statistic) for eight cicada species.

	SS	df	MS	F	p
Intercept	712.274	1	712.274	44799.26	0.000
Model	2.776	4	0.694	43.65	0.000
Samp. Size	2.242	4	0.560	35.25	0.000
Model*size	1.348	16	0.084	5.30	0.000
Error	16.138	1015	0.016		

Table 6. The frequency with which a particular model design appeared in the group of lowest performance and the frequency with which it appeared in the group of highest performance in Tukey multiple comparisons. For each of the cicada species, models were built using five different model designs (CEM, FEM, PCA, SA, PSA). Tukey tests were conducted to determine which designs performed significantly better than the rest (Figs. 3-6). For each species the Tukey tests revealed two or more groups containing models that did not differ significantly from one another. This allowed a group of models with lowest performance to be distinguished from a group of lowest performance (Figs. 3-6). The table indicates the frequency with which a particular model design appeared in each group.

Group	Model Design				
	CEM	FEM	PCA	SA	PSA
Highest	1	2	6	8	5
Lowest	8	8	3	0	4



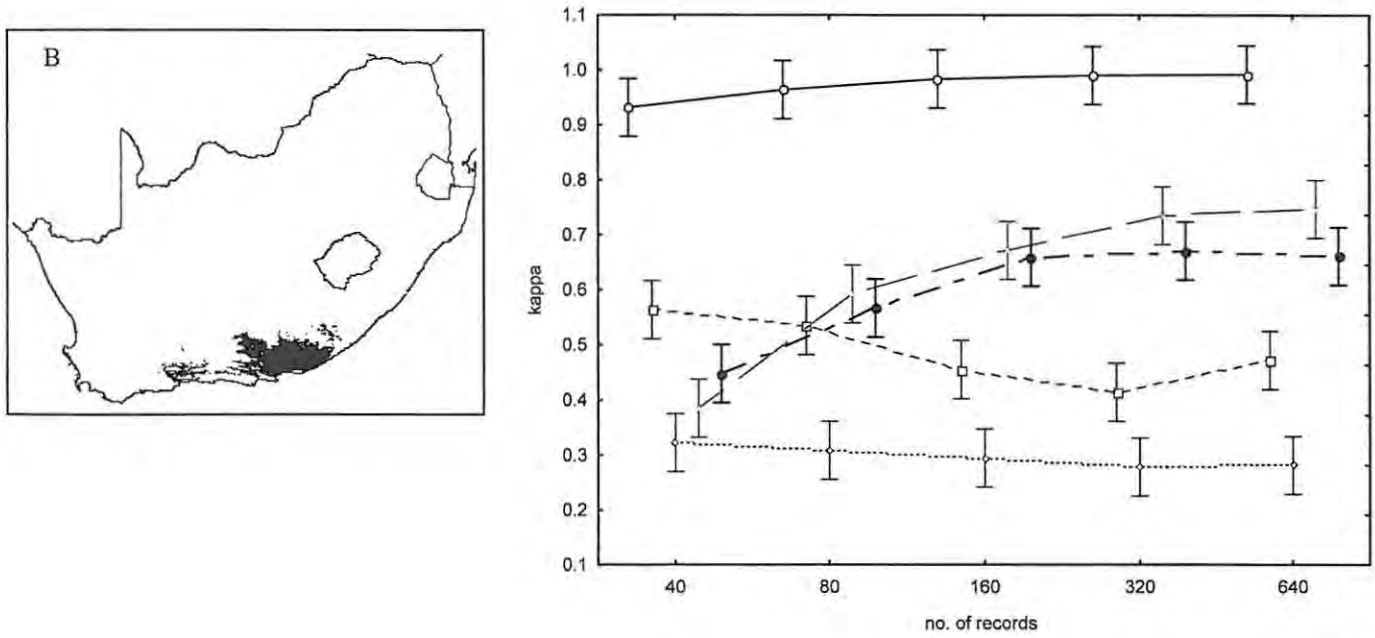
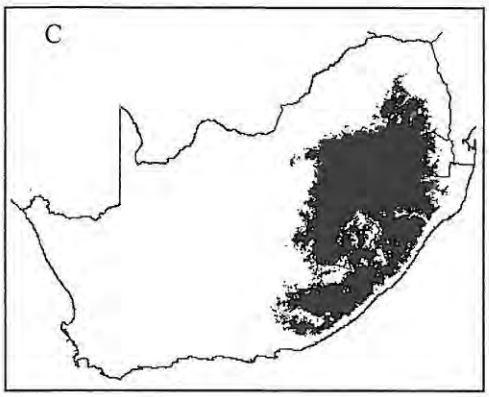
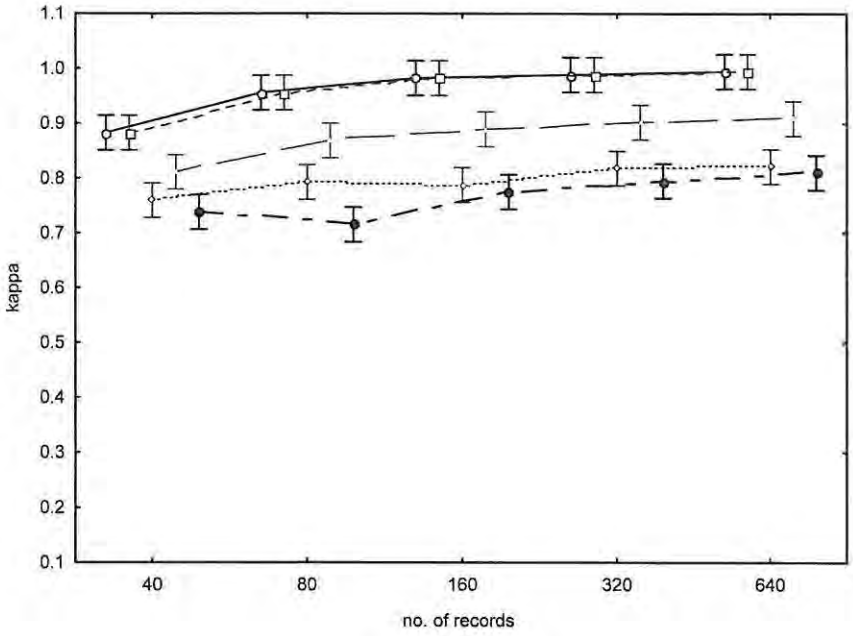
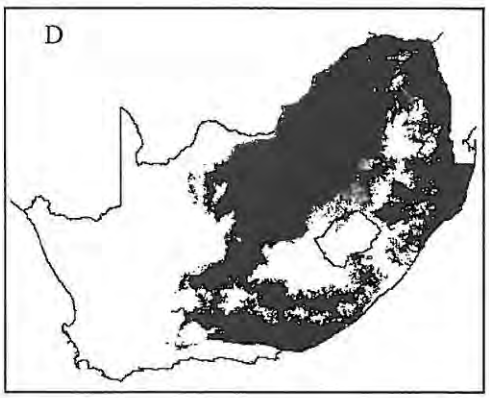


Figure 1a&b. Performance among model design and sample size for hypothetical distributions A and B. In the maps, black indicates the presence of the hypothetical organism.







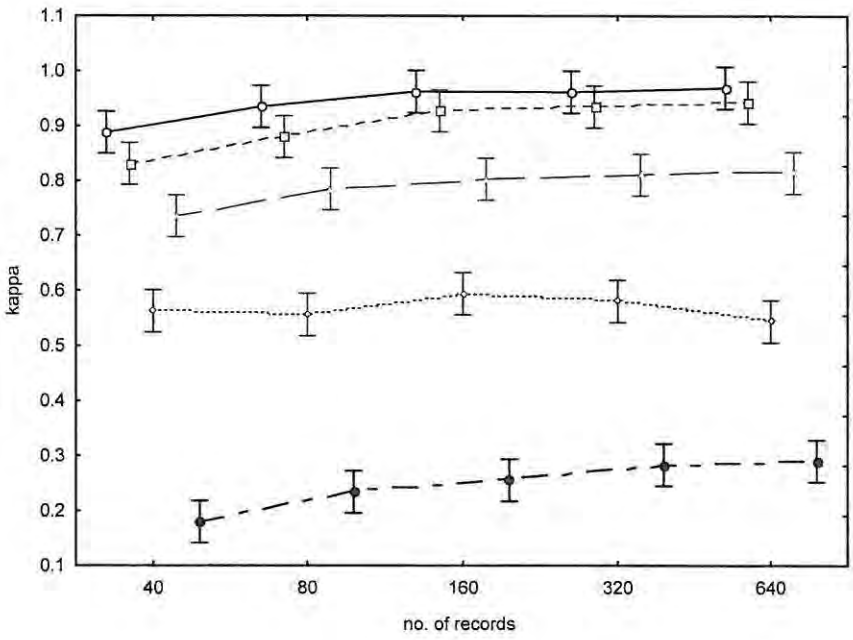
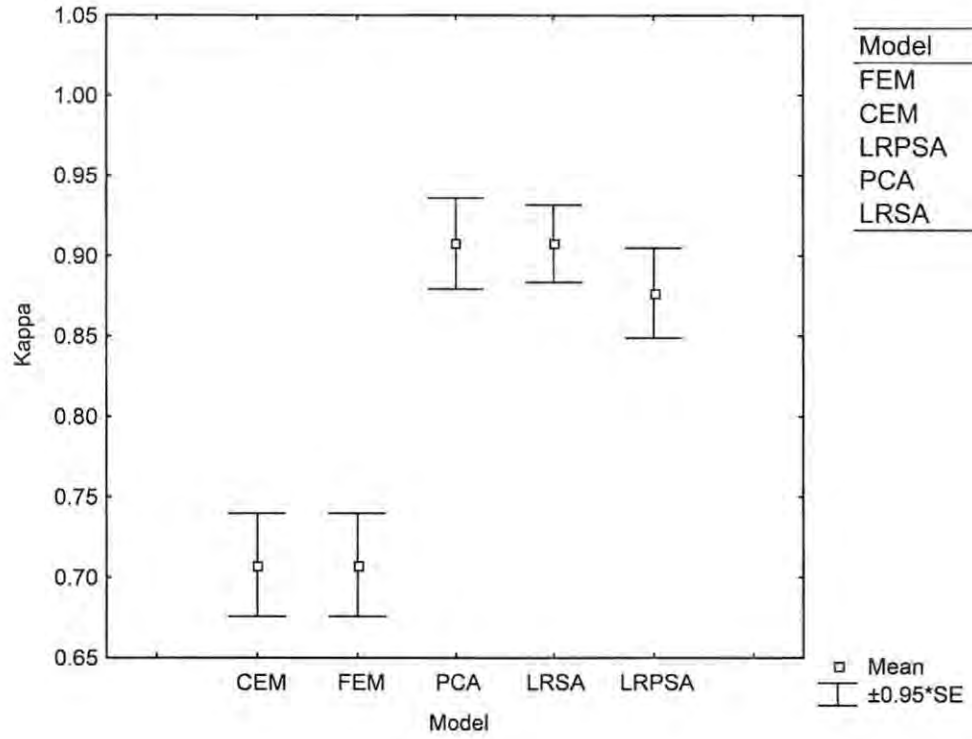
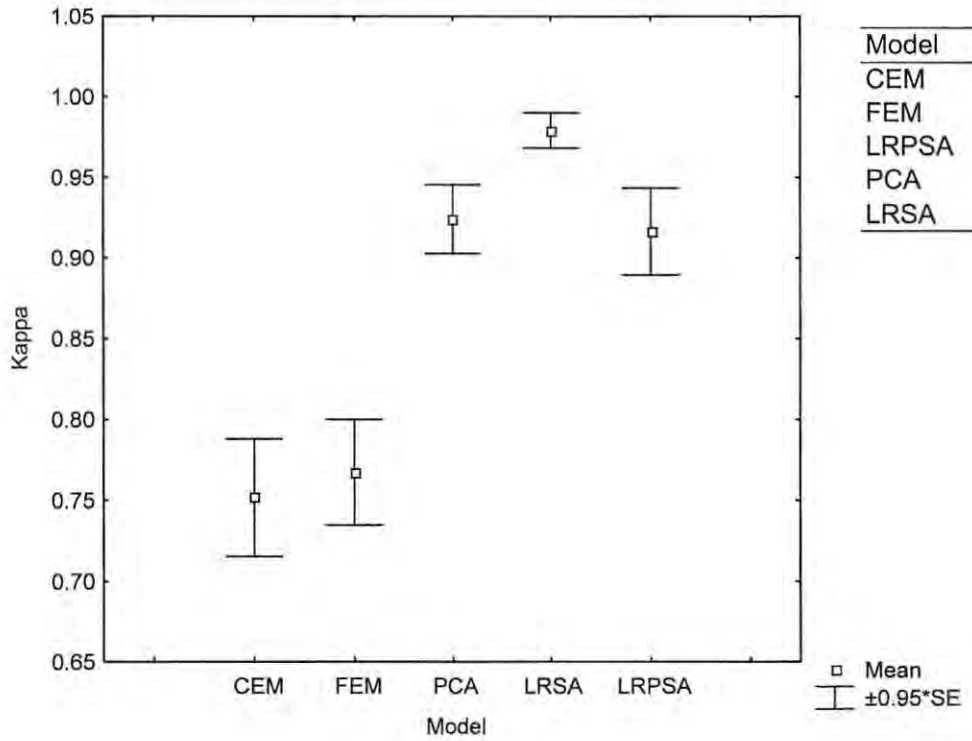


Figure 2c&d. Performance among model design and sample size for hypothetical distributions C and D. In the maps, black indicates the presence of the hypothetical organism.

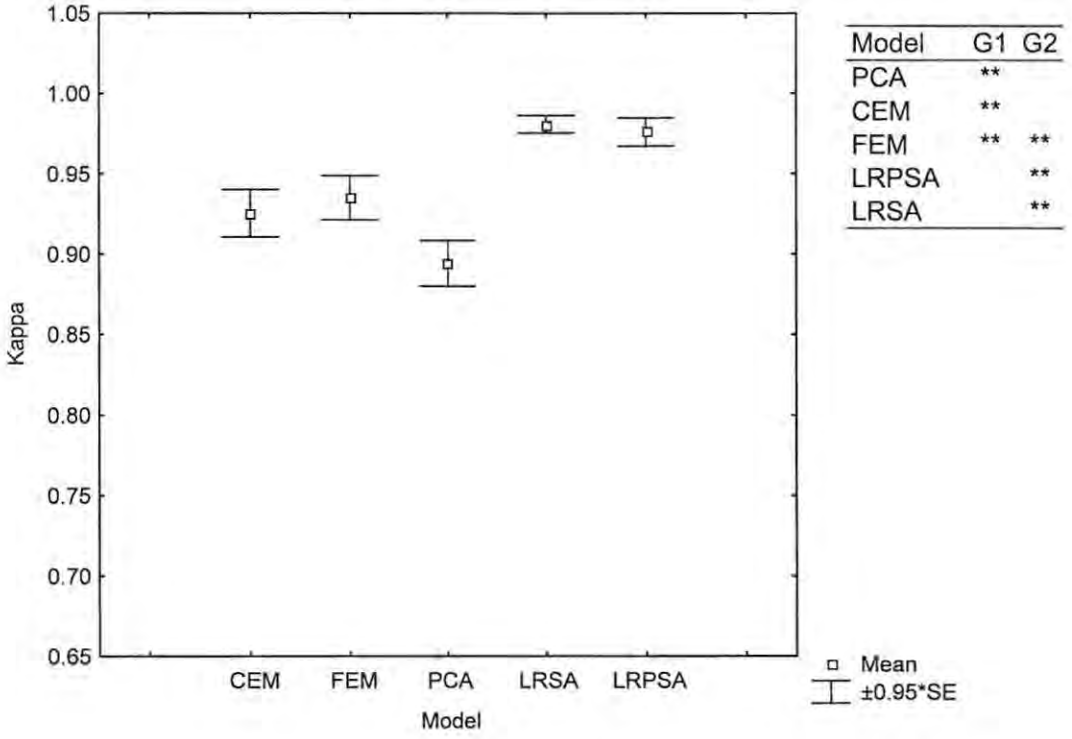


Model	G1	G2
FEM	**	
CEM	**	
LRPSA		**
PCA		**
LRSA		**

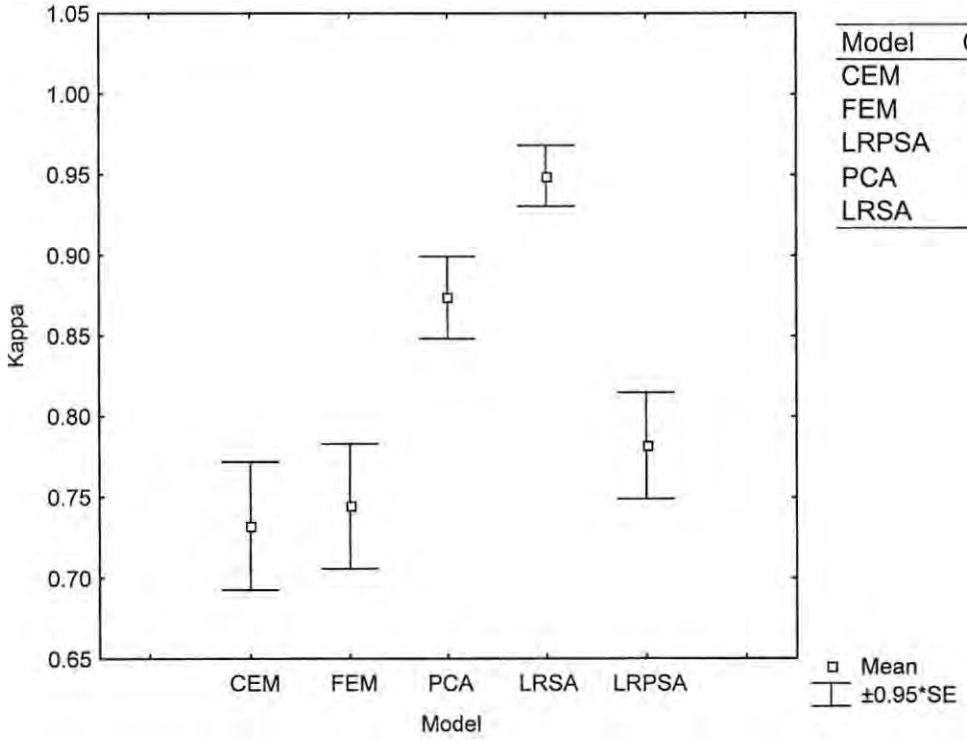


Model	G1	G2
CEM	**	
FEM	**	
LRPSA		**
PCA		**
LRSA		**

Figure 3. The average performance of five model designs for a) *A. albigera* and b) *C. decora* Means and standard deviations were calculated maximum kappa values from 25 partitions of the original data. Tukey HSD multiple range tests (using α=0.05) indicate those designs that do not differ significantly from on another. Those model designs that appear in the same group (G1, G2, G3) do not differ significantly from one another. Group membership is indicated by '**'. Model designs with the lowest average performance appear at the top of the table.

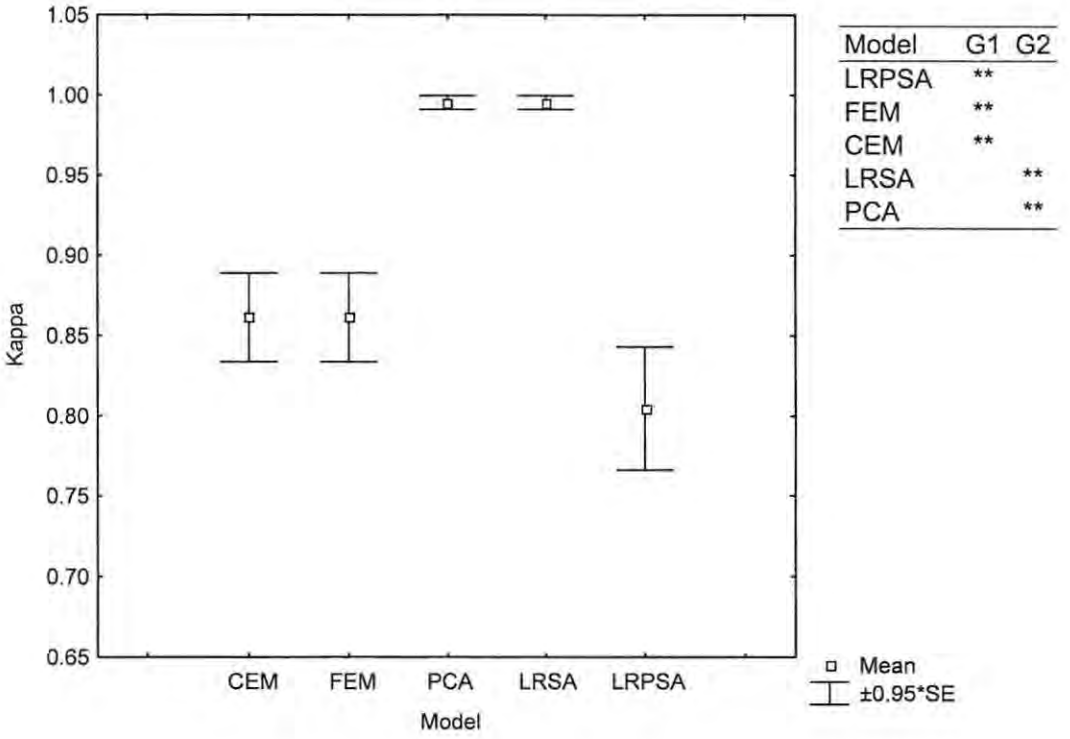


Model	G1	G2
PCA	**	
CEM	**	
FEM	**	**
LRPSA		**
LRSA		**

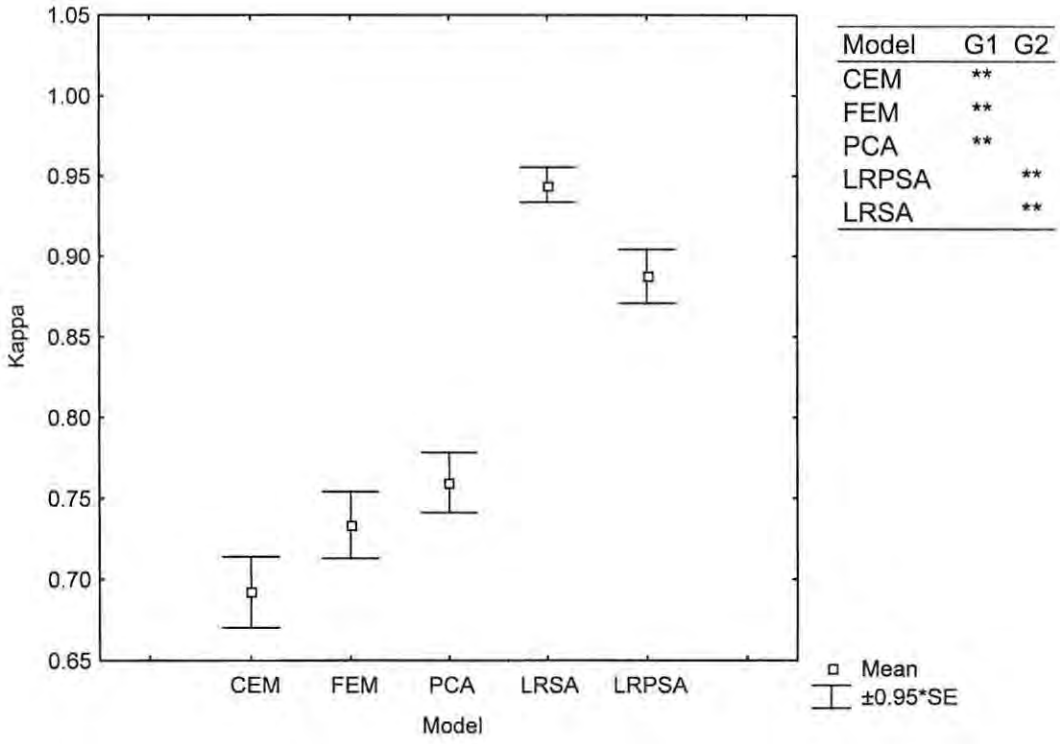


Model	G1	G2	G3
CEM	**		
FEM	**	**	
LRPSA	**	**	
PCA		**	**
LRSA			**

Figure 4. The average performance of five model designs for a) *P. capensis* and b) *P. deusta* Means and standard deviations were calculated maximum kappa values from 25 partitions of the original data. Tukey HSD multiple range tests (using α=0.05) indicate those designs that do not differ significantly from on another. Those model designs that appear in the same group (G1, G2, G3) do not differ significantly from one another. Group membership is indicated by '**'. Model designs with the lowest average performance appear at the top of the table.

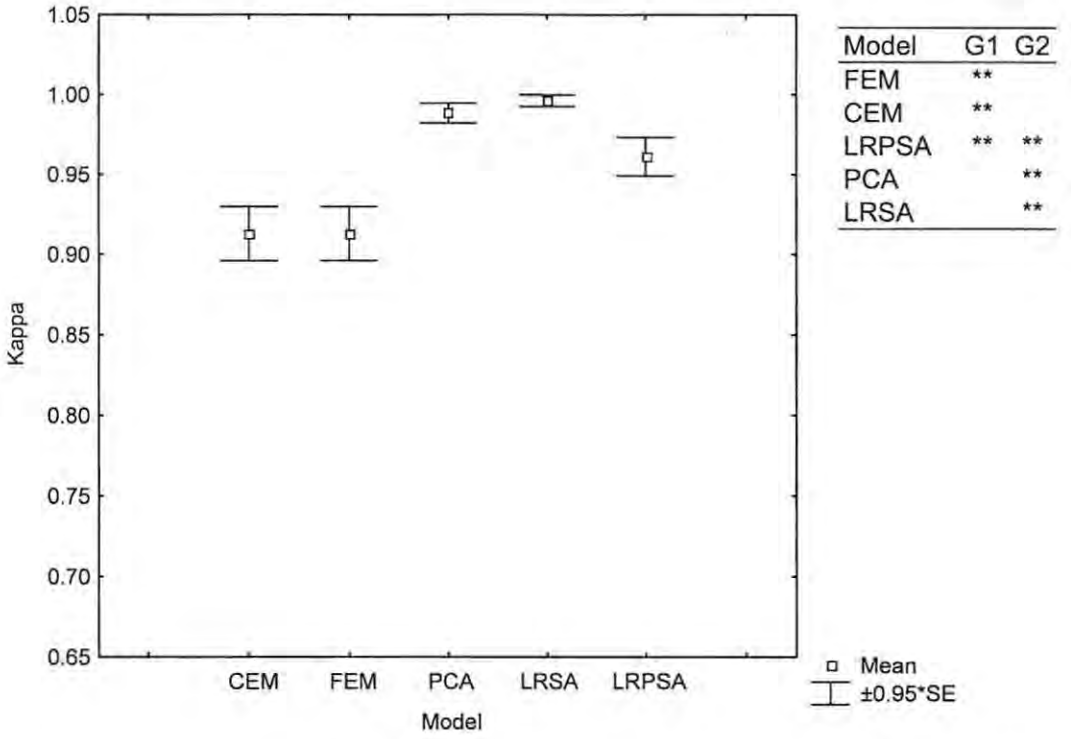


Model	G1	G2
LRPSA	**	
FEM	**	
CEM	**	
LRSA		**
PCA		**

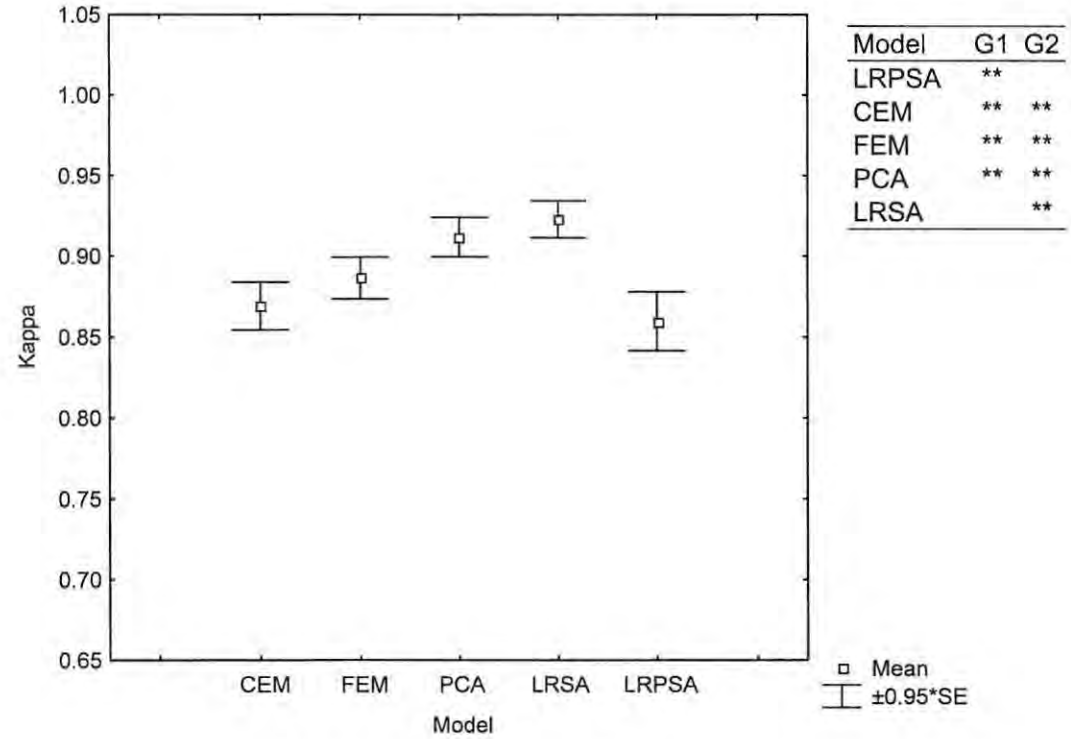


Model	G1	G2
CEM	**	
FEM	**	
PCA	**	
LRPSA		**
LRSA		**

Figure 5. The average performance of five model designs for a) *P. divisa* and b) *P. haglundi* Means and standard deviations were calculated maximum kappa values from 25 partitions of the original data. Tukey HSD multiple range tests (using α=0.05) indicate those designs that do not differ significantly from on another. Those model designs that appear in the same group (G1, G2, G3) do not differ significantly from one another. Group membership is indicated by '**'. Model designs with the lowest average performance appear at the top of the table.



Model	G1	G2
FEM	**	
CEM	**	
LRPSA	**	**
PCA		**
LRSA		**



Model	G1	G2
LRPSA	**	
CEM	**	**
FEM	**	**
PCA	**	**
LRSA		**

Figure 6. The average performance of five model designs for a) *P. mijburghi* and b) *Pycna semiclara* Means and standard deviations were calculated maximum kappa values from 25 partitions of the original data. Tukey HSD multiple range tests (using α=0.05) indicate those designs that do not differ significantly from on another. Those model designs that appear in the same group (G1, G2, G3) do not differ significantly from one another. Group membership is indicated by '**'. Model designs with the lowest average performance appear at the top of the table.

References

Austin, M.P. 1998. An ecological prespective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*. 85: 2-17.

Austin, M.P., Cunningham, R.B., Fleming, P.M., 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio*. 55: 11-27.

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

Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a beta-function. *Journal of Vegetation Science*. 5: 215-228.

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

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

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

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

Collingham, Y.C., Wadsworth, R.A., Huntley, B., Hulme, P.E., 2000. Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology*. 37: 13-27.

Cumming, G.S., 2000 a. Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*. 27: 425-440.

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

Erasmus, B.F.N., Kshatriya, M., Mansell, M.W., Chown, S.L., Van Jaarsveld, A.S., 2000. A modelling approach to antlion (Neuroptera: Myreleontidae) distribution patterns. *African Entomology*. 8: 157-168.

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*. 51: 331-363.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra, p. 193.

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

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*. 19: 474-499.

Franklin, J., 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*. 9: 733-748.

Freitag, S., Hobson, C., Biggs, H.C., Van Jaarsveld, A.S., 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*. 1: 119-127.

Funk, V.A., Richardson, K.S. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*. 51: 303-316.

Guisan, A., Theurillat, J-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*. 9: 65-74.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

Heuvelink, G.B.M., Burrough, P.A., 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*. 7: 231-246.

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

Hirzel, A., Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*. 157: 331-341.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N. 2002. Ecological-Niche Factor Analysis: how to compute habitat-suitability maps without absence data? *Ecology*. 83: 2027-2036.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*. 145: 111-121.

Lawes, M.J., Piper, S.E., 1998. There is less to binary maps than meets the eye: the use of species distribution data in the southern African sub-region. *South African Journal of Science*. 94: 207-210.

Leathwick, J.R. 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*. 9:719-732.

Leathwick, J.R., Whitehead, D. 2001. Soil and atmospheric water deficits and the distribution of New Zealand's indigenous tree species. *Functional Ecology*. 15: 233-242.

Leathwick, J.R., Whitehead, D. McLeod, M. 1996. Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach. *Environmental Software*. 11:81-90.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F., Tanton, M.T., 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*. 18: 371-383.

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

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

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

McCullagh, P., Nelder, J.A. 1989. Generalized Linear Models. Chapman and Hall, London. p. 511.

Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*. 62: 275-293.

Nix, H.A., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*, Australian Government Publishing Service, Canberra, pp. 4-15.

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

Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*. 116: 15-31.

Palmer, A.R., Van Staden, J.M., 1992. Predicting the distribution of plant communities using annual rainfall and elevation: an example from southern Africa. *Journal of Vegetation Science*. 3: 261-266.

Pearce, J., Ferrier, S., 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*. 128: 127-147.

Peterson, A.T., 2001. Predicting species' geographic distributions based on ecological niche modeling. *The Condor*. 103: 599-605.

Robertson, M.P., Caithness, N., Villet, M.H., 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*. 7: 15-27.

Robertson, M.P., Villet, M.H. Geographic distribution of Silphidae (Coleoptera) in southern Africa: a case study in predicting presence of necrophagous invertebrates in an area. In preparation for *Journal of Forensic Science*.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science*. 13: 143-158.

Zaniewski, A.E., Lehmann, A., Overton, J. McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*. 157: 261-280.

VIII

Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques

Preface

This chapter compares a mechanistic model with two correlative models of the distribution of a coastal sand dune plant. A modified version of this chapter has been submitted to *Ecological Modelling* for publication (Robertson, M.P., Peter, C.I., Villet, M.H., Ripley, B.S. Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques).

Abstract

Models used to predict species' potential distributions have been described as either correlative or mechanistic. An attempt was made to determine whether correlative models could perform as well as mechanistic models for predicting species potential distributions, using a case study. Potential distribution predictions made for a coastal dune plant (*Scaevola plumieri*) along the coast of South Africa, were compared using a mechanistic model based on summer water balance, and two correlative models (a profile and a group discrimination technique). The profile technique was based on Principal Components Analysis (PCA) and the group-discrimination technique was based on multiple logistic regression (LR). Kappa (κ) statistics were used to objectively assess model performance and model agreement. Model performance was calculated by measuring the levels of agreement (using κ) between a set of testing localities (distribution records not used for model building)

and each of the model predictions. Using published interpretive guidelines for the kappa statistic, model performance was "excellent" for the Summer Water Balance (SWB) model ($\kappa = 0.852$), perfect for the LR model ($\kappa = 1.000$), and "very good" for the PCA model ($\kappa = 0.721$). Model agreement was calculated by measuring the level of agreement between the mechanistic model and the two correlative models. There was "good" model agreement between the SWB and PCA models ($\kappa = 0.679$) and "very good" agreement between the SWB and LR models ($\kappa = 0.786$). The results suggest that correlative models can perform as well as or better than simple mechanistic models. The predictions generated from these three modelling designs are likely to generate different insights into the potential distribution and biology of the target organism and may be appropriate in different situations. The choice of model is likely to be influenced by the aims of the study, the biology of the target organism, the level of knowledge the target organism's biology, and data quality.

Introduction

Models for predicting species' potential distributions have been used in many fields of biology. Franklin (1995) and Guisan and Zimmermann (2000) have reviewed a number of these models and give examples of their application. A current question is which models perform best, given particular circumstances (Guisan and Zimmermann, 2000; Hirzel *et al.*, 2001; Zaniewski, *et al.*, 2002). This study attempts to address this question empirically.

Predictive modelling techniques have been described as static or dynamic models (Beerling *et al.*, 1995). Static models provide time-independent equilibrium predictions while dynamic models predict time-dependent dynamic responses to a changing environment (Beerling *et al.*, 1995). Both types of model have in turn been divided into two groups, namely *correlative* and *mechanistic* techniques (Beerling *et al.*, 1995). Correlative models rely on strong, often indirect, links between species distribution records and environmental predictor variables to make predictions (Beerling *et al.*, 1995). These models use values of a predictor variable or more commonly a set of predictor variables associated with distribution records to classify the predictor variable or predictor variable hyperspace into presence-absence regions,

suitability values or probabilities, which are visualized as maps. These predictor variables can be direct, resource or indirect gradients and tend to be distal rather than proximal. Austin (2002) defines proximal and distal as being to the position of the predictor in the chain of processes that link the predictor to its impact on the plant. Mechanistic models attempt to simulate the mechanisms considered to underlie the observed correlations with environmental attributes (Beerling *et al.*, 1995) by using a detailed knowledge of the target species' physiological responses to environmental variables as well as life-history attributes (Stephenson, 1998). Such models have also been referred to as *ecophysiological models* (Stephenson, 1998) and *process orientated* models (Carpenter *et al.*, 1993). In contrast to correlative models, mechanistic models do not use values of a predictor variable or predictor variables associated with distribution records to classify the predictor variable(s). The predictor variables used in mechanistic models tend to be resource or direct rather than indirect gradients. These predictor variables tend to be more proximal than those used in correlative models.

Stephenson (1998) maintains that the distinction between correlative and ecophysiological (mechanistic) models is often not clear. For example, he observes that ecophysiological studies of plants depend on empirical correlations to determine quantitative relationships between physiologically important factors and vegetation distribution. Similarly, correlative models have an ecophysiological basis when they employ predictor variables that are suspected to be of broad physiological importance to plants (Stephenson, 1998).

Predictive models have generally been used to predict the potential distribution of a target species under current climatic conditions or various climate change scenarios and to determine the importance of selected climatic variables on the distribution of the target species. Recent interest in the possible consequences of global climate change has resulted in a number of studies focusing on the climatic controls of vegetation distribution (Stephenson, 1998). Mechanistic models are considered to be more promising at successfully predicting climatically induced changes in the distribution of plant species (Stephenson, 1998), as these models will be more robust under changed climatic conditions than correlative models as certain correlations may cease to apply under changed conditions (Prentice *et al.*, 1992). While mechanistic models are likely to yield superior results to correlative models

(particularly under climate change scenarios) they are often extremely time-consuming and more difficult to build, relying on a greater knowledge of the biology of the target organism than correlative models.

Correlative models are particularly suited to cases where an initial estimate of the potential distribution of an organism is required, especially when the biology of the organism is not well known. Correlative models also can be used to obtain some insight into factors that may be responsible for limiting the distribution of the target organism when its biology is not well known. These insights can then be used to incorporate mechanistically more important predictor variables into the model, thus making it more mechanistic. Through an iterative process, a greater understanding of the target organism's biology can be developed and further insights into the factors that limit its distribution may be obtained. This may culminate in developing a mechanistic distribution model. Stephenson (1998) suggests that correlative approaches may play an important role in understanding of the relationships between climate and species distributions by identifying potentially significant and previously overlooked physiological mechanisms.

Correlative models that use both presence and absence locality records to make predictions have been referred to as *group discrimination techniques*, while those that use only presence locality records have been referred to as *profile techniques* (Caithness, 1995). Examples of group-discrimination techniques include those models based on discriminant analysis (Rogers and Randolph, 1993; Rogers and Williams, 1993; Rogers *et al.*, 1996), Generalised Linear Models (Austin *et al.* 1984; Austin *et al.* 1990; Osborne and Tigar, 1992; Austin *et al.*, 1994; Guisan *et al.* 1998; Higgins *et al.*, 1999; Manel *et al.*, 1999; Cumming, 2000 a & b), Generalised Additive Models (Austin and Meyers 1996; Leathwick *et al.* 1996; Leathwick, 1998; Pearce and Ferrier, 2000; Leathwick and Whitehead, 2001) and decision-tree-based methods (Walker, 1990; Lees, 1994; Michaelsen *et al.*, 1994; Williams *et al.*, 1994). Examples of profile techniques include models developed by Palmer and Van Staden (1992), Erasmus *et al.* (2000), Robertson *et al.* (2001), Hirzel *et al.* (2002) and the approaches used in the modelling packages known as BIOCLIM (Nix, 1986; Busby, 1991) and DOMAIN (Carpenter *et al.*, 1993). For examples of comparisons between group-discrimination and profile techniques see (Hirzel *et al.*, 2001; Zaniewski *et al.*, 2002).

The choice between using a mechanistic or correlative approach will depend largely on the purposes of the study and the current state of knowledge of the biology of the target organism. One of the central questions is whether correlative models can perform as well as mechanistic models for predicting species potential distributions.

Most of the mechanistic models in the literature have tended to make predictions at the continental scale and usually for several species or functional types e.g. Woodward and Williams (1987); Prentice *et al.* (1992) and Neilson (1995). However, Peter *et al.* (2002) recently developed a model based on water balance, temperature and plant phenology to predict the potential distribution of a single species (*Scaevola plumieri*) at a regional scale. The development of this water balance model presents an opportunity to compare quantitatively the success with which mechanistic and correlative approaches are able to predict the potential distribution of a target species.

This study compares potential distribution predictions made using three static modelling approaches for a coastal dune plant (*S. plumieri*) along the coast of South Africa. A comparison is made of the performance of a mechanistic approach based on water balance (described in Peter *et al.*, 2002) and two correlative models (a profile and a group discrimination technique). The profile technique is based on Principal Components Analysis (PCA), and is described and implemented by Robertson *et al.* (2001). Similar profile techniques include the approach used in the FloraMap package (Jones and Gladkov, 1999), the approach used by Erasmus *et al.* (2000) and Ecological Niche Factor Analysis (ENFA; Hirzel *et al.*, 2002).

The group-discrimination technique is based on multiple logistic regression (LR), a form of Generalised Linear Model (GLM; McCullagh and Nelder, 1983) that has been used frequently in biology (e.g. Austin *et al.* 1984; Nicholls 1989; Austin *et al.*, 1990; Leathwick and Mitchell, 1992; Osborne and Tigar, 1992; Austin *et al.* 1994; Austin and Meyers 1996; Guisan *et al.*, 1998, 1999; Higgins *et al.*, 1999; Manel *et al.*, 1999, Cumming, 2000 a & b; Pearce and Ferrier, 2000; Hirzel *et al.*, 2001).

Materials and Methods

The target species

S. plumieri represents a good test case for developing and comparing mechanistic and correlative models. Ecophysiological measurements could be made with relative ease since the plant is short, has broad leaves and stands of this plant were relatively easily accessible. Since it has an open, well ventilated canopy and it occurs on the same substrate (sand) throughout its range, certain simplifying assumptions could be made (Peter and Ripley, 2000). Making the necessary ecophysiological measurements that would be required to build a similar mechanistic model for other species (e.g. a large forest species), would be more challenging.

An important implicit assumption made by these models is that the target species is in equilibrium with its environment (in an ecological rather than a physiological sense), since these are static models (Guisan and Zimmermann, 2000; Austin, 2002). Since *S. plumieri* is indigenous to southern Africa, it is likely to be in equilibrium with its environment in the sense that it has had sufficient time to occupy all suitable sites (as opposed to an alien plant species that may not yet be in equilibrium with the environment). It is unlikely to compete directly with other species and thus is unlikely to be excluded from a particular site due to competition for space. Since it is confined to a narrow habitat, namely coastal sand dunes, it was possible to sample a large proportion of the plant's potential habitat along the South African coast. The chance of make false negative errors is low because the plant is conspicuous and thus is unlikely to have been recorded absent when it was present at a particular site. False positive errors are unlikely since *S. plumieri* is not easily confused with other species. This attribute of the plant has thus enabled accurate presence and absence locality data to be collected. This in turn has allowed reliable profile (PCA) and group-discrimination (LR) models to be built.

The data

The predictor variables used in these models consisted of various digital climatic variable maps and the response variable consisted of point distribution

records for *S. plumieri*. The digital climatic variable maps were developed by Schulze *et al.* (1997) for South Africa, Lesotho and Swaziland. Each of these climatic maps was interpolated from point data obtained from a network of weather recording stations distributed throughout South Africa, to produce continuous digital maps at a 1-minute spatial resolution (Schulze *et al.*, 1997). Monthly maps of mean maximum temperature, mean minimum temperature, mean relative humidity and median rainfall (Schulze *et al.*, 1997) were used in the water balance model (Peter *et al.*, 2002) while further pre-processing of these maps was done to yield a smaller subset of variables that were used to develop the correlative models.

Localities where *S. plumieri* was found to be present or absent (Fig. 1) were obtained by direct surveys using a GPS, from herbarium specimens and from historical photographs (for details see Peter *et al.*, 2002). Locality data were partitioned randomly into a set of training localities and a set of testing localities in a ratio of 3:1, based on Huberty's (1994) recommendations. The training localities consisted of 158 presence and 57 absence records and the testing localities consisted of 53 presence and 19 absence records. The training localities were used to calibrate the models and the testing localities were used for model evaluation. Locality data were used only to calibrate the PCA and LR models and not in the calibration phase of the water balance models.

The water balance model

An empirical relationship of transpiration (E) to atmospheric vapor pressure deficit (VPD) was calculated at the leaf level for *S. plumieri* (Peter and Ripley, 2000). VPD can be calculated from atmospheric temperature and relative humidity. Peter and Ripley (2000) also successfully scaled leaf level transpiration rates to the canopy level. Transpiration rates of *S. plumieri* were extrapolated from VPD which was in turn calculated from regional level values of temperature and relative humidity. Water balance was calculated by subtracting transpiration from rainfall for a given month (Peter *et al.*, 2002). Monthly maps of mean maximum temperature, mean minimum temperature, mean relative humidity and median rainfall were used to calculate 12 monthly water balance maps using the approach and equations described by Peter *et al.* (2002). These monthly water balance maps were cross-correlated to

investigate their relationships (Table 1). Based on these correlations maps for October, November, December January, February and March were summed to produce a map of summer water balance (SWB). Similarly maps for May, June, July and August were summed to produce a map of winter water balance (WWB). Seasonal, rather than annual, water balance was calculated because *S. plumieri* was found at sites that experienced water surpluses during the summer months, when the plants were most actively growing and reproducing (Peter *et al.*, 2002).

SWB model is referred to as a *mechanistic* model (although this term was not used to describe this model by Peter *et al.*, 2002). The SWB model is mechanistic for two reasons. Firstly, a predictor variable (summer water balance) was calculated using the physiological responses of the plant to environmental variables obtained by means of field measurements and through a knowledge of life history attributes (phenology). Secondly, the predictor variable was not classified using distribution records to produce a distribution map. This was unnecessary because the values of the predictor variable (SWB) had direct physiological significance to the plant. The plant should be absent from those sites experiencing summer water deficits and present at sites experiencing summer water surpluses (Peter *et al.*, 2002). This is based on the hypothesis that the plant is unable to survive at sites that experience water deficits during those periods when the plant was actively growing and reproducing (i.e. summer).

Predictor variable pre-processing

To reduce the dimensionality of available climatic variable data, four principal components analyses (PCA's) were performed on the 12 maps for each of the following: mean monthly maximum temperature, mean monthly minimum temperature, mean monthly relative humidity and median monthly rainfall. Each of the monthly maps was calculated from mean values for a calendar month e.g. January. PCA has previously been employed as a pre-analytical data reduction technique used in distribution modelling (Osborne and Tigar, 1992; Buckland and Elston, 1993; Robinson *et al.*, 1997; Guisan *et al.*, 1998). Those principal component axes whose eigenvalues were greater in magnitude than eigenvalues obtained from datasets of random numbers of the same sample size were retained as predictor variables. This

follows the "broken stick" stopping rule for PCA (Jackson, 1993). Predictor variables used in the PCA and logistic regression models are listed in Table 2.

Although variables such as potential evaporation or elevation (as used previously, Robertson *et al.*, 2001) could have been used in addition to those predictor variables listed for building the correlative models (Table 2), the same sets of predictor variables were used to build the mechanistic model and the two correlative models so that any differences in model performance could be attributed to model design and were not confounded by differences in the predictor variables used.

The PCA-model

A PCA-based modelling technique, described by Robertson *et al.* (2001), was used for predicting environmental suitability for a target organism from environmental predictor variables using only presence locality records. This technique constructs a hyperspace for the target species using principal component axes derived from a training data set. The training data set comprises the values of the predictor variables associated with those localities where the species has been recorded as present. The origin of this hyperspace is taken to characterise the centre of the niche of the organism. All the localities (grid-cells) in the map region are then fitted into this hyperspace using the values of the predictor variables at these localities (termed the prediction data set). The Euclidean distance from any locality to the origin of the hyperspace gives a measure of the "centrality" of that locality in the hyperspace. These distances are used to derive probability values for each grid cell in the map region. The probability values are taken to indicate the suitability of each grid cell in a map for the target species in terms of the suite of predictor variables. The approach taken in this paper is identical to that of Robertson *et al.* (2001) except that different predictor variables were used.

The logistic regression model

A logistic regression was performed using GLMFIT, a generalised linear modelling function within MATLAB. In order to calculate the values of the coefficients (β_i in Eq. 1), a binomial error distribution and a logit link function were

used. The response variable consisted of localities representing surveyed presence (coded 1) and surveyed absence (coded 0). The response variables consisted of the values of the environmental variables associated with the surveyed localities. Probability values for each grid cell in the map region were calculated by substituting the values of the predictor variables associated with that cell into the following equations:

$$\beta_1 variable_1 + \beta_2 variable_2 + \ldots + \beta_8 variable_8 = \eta \quad (1)$$

$$P_{(y=1)} = \exp(\eta)/(1+\exp(\eta)). \quad (2)$$

The first equation (Eq. 1) is known as the linear predictor and the second equation (Eq. 2) is the inverse logistic transformation. In order to constrain the values of the linear predictor between 0 and 1, the inverse logistic transformation has to be applied.

Model evaluation

There are several measures for assessing model performance (Fielding and Bell, 1997; Guisan and Zimmermann, 2000). A number of these measures are derived from a confusion matrix (Table 3). A reliable and well-known measure based on the confusion matrix is the Kappa (κ) statistic (Fielding and Bell, 1997). The κ statistic is dependent on a single threshold to distinguish between predicted presence and predicted absence and thus falls into the class of threshold-dependant measures (Fielding and Bell, 1997). Threshold-independent measures, such as Receiver Operating Characteristic (ROC) plots which are emerging as useful measures of model performance (Packer *et al.*, 1999; Cumming, 2000 a & b; Robertson *et al.*, 2001) are considered to be superior since they use a range of thresholds and are therefore less likely to introduce distortions (Fielding and Bell, 1997). Although the ROC measure is considered to be a superior measure, it could not be used on the data in this study since it requires values to be constrained between 0 and 1, and in the case of the water balance models, negative values were evident. As a result, the κ statistic was used to evaluate the models.

κ statistics were calculated from the parameters in the confusion matrix (Table 3) in two ways. Firstly, κ-values were calculated for all models using the presence and absence testing localities as *observed presence* and *observed absence*, respectively (Table 3). This gives a measure of *model performance*. Secondly, κ-values were calculated using the SWB model. Grid-cells with water balance values above zero were taken to represent *observed presence* and grid-cells with water balance values below zero were taken to represent *observed absence*. This gives a measure of *model agreement* between the SWB model and the PCA model, and between the SWB model and the LR model. κ statistics were calculated using thresholds that yielded maximum κ-values for the PCA and LR models, following recommendations of Guisan and Zimmermann (2000). In contrast κ-values for summer and winter water balances were calculated using thresholds of zero, which represent a biologically justifiable threshold. *S. plumieri* is unlikely to be able to survive water deficits (values below zero) for extended periods, particularly in the summer when it is actively growing (see Peter *et al.*, 2002).

Results

Scaevola plumieri was recorded as being present along the south and east coasts of South Africa (Figs. 1 & 2a). Arniston was the most westward locality at which S. plumieri was recorded present. All localities west of Arniston represent observed absence and all localities to the east represent observed presence (Fig. 1 & 2a). For simplicity, the coast to the north of Cape Columbine is referred to as west coast, between Cape Columbine and Arniston as the south west coast; between Arniston and Port Elizabeth as the south coast and between Port Elizabeth and the Mozambique border as the east coast (Fig. 2a).

Winter water balances (Fig. 2b) had low positive values to slight negative values on much of the east coast. Water balance values were particularly low between Port Alfred and Durban. On the south west coast between False Bay and Cape Columbine, relatively large water surpluses were evident. Localities on the west coast experienced relatively large water deficits.

In summer the south and east coasts were characterised by large water surpluses (Fig. 2c). Localities immediately to the east of Arniston (at the distribution limit) had small water surpluses or experienced small deficits. The west and south west coasts were characterised by large water deficits, particularly to the west of Cape Town.

On the south and east coasts, predicted suitability (from the PCA model) was generally high but variable (Fig. 2d). Suitability on the west and south west coasts was lower than that on the south and east coasts with the exception of two peaks in suitability occurring near Cape Columbine and at localities just west of Arniston. The trend of lower suitability on the west and south west coasts and higher suitability on the south and east coasts follows the trend observed for SWB (Fig. 2c). A trough of low predicted suitability just to the east of Mtunzini (in the region of grid-cell 100) corresponds with a peak in winter water surplus (Fig. 2b).

The results of the LR model have been reported as probabilities which can be interpreted as probability of occurrence (Fig. 2e). The east coast was characterised by consistently high probabilities. The south coast was characterised by greater variability in the probability values with more localities having low probabilities. On the south west coast probabilities were consistently very low or zero. On the west coast probabilities were mostly zero but increased towards the Namibian border (grid-cell 1440). The coefficients, their associated standard errors and Wald statistics are presented in Table 4 for the LR model. Only linear terms were included in the linear predictor as scatter plots indicated that there was no justification for including higher order terms.

Kappa statistics

Kappa statistics can be used to objectively assess the level of agreement between observed and predicted data. Monserud and Leemans (1992) suggested the following ranges of agreement for the κ statistic: no agreement < 0.05; very poor 0.05-0.20; poor 0.20-0.40; fair 0.40- 0.55; good 0.55-0.70; very good 0.70-0.85; excellent 0.85-0.99 and perfect 0.99-1.00. Negative values indicate extremely poor agreement (Monserud and Leemans, 1992). These ranges were used to describe the levels of agreement reported here using two tests.

The first test involved using the locality records reserved for testing (the testing localities). The WWB model had a negative κ-value of -0.150 (Table 5) indicating extremely poor agreement between observed and predicted values (Monserud and Leemans, 1992). In contrast, the SWB model had a high κ-value (0.852) indicating excellent agreement (Table 5). The LR model had a κ-value of 1.000, indicating perfect agreement. The PCA model had a lower value (0.721) than that of SWB model, although this falls into the category of very good agreement.

The second test involved using all of the predicted values from the PCA and LR models and measuring their agreement (using κ) with the SWB model. In order to test model agreement in this way, a set of "observed presence" and "observed absence" localities (grid-cells) had to be defined. Grid-cells in the SWB model with water surpluses (≥ 0 liters) were taken to represent "observed presence" and grid-cells with water deficits (< 0 liters) were taken to represent "observed absence" to assess the level of agreement between the SWB model and the PCA and LR models (Table 6). There was good agreement between the SWB and the PCA model (0.679), and very good agreement between the SWB and the LR model (0.786; Table 6).

Discussion

Interpretation of model predictions

I suggest that the predictions produced by each of these models may offer different insights into the potential distribution and biology of the target organism. The probabilities in the map generated from the logistic regression (LR) model are interpreted as the probability of occurrence for the target species (*S. plumieri*). In contrast the probabilities in the map generated from the PCA model are interpreted as environmental suitability values (Robertson *et al.*, 2001). The PCA model has no explicit "knowledge" of the conditions that exist at localities where the target organism is absent and thus the probabilities generated cannot be interpreted as probability of occurrence for the target organism. The term "suitability" is used to distinguish these from the probabilities produced using a LR model (Robertson *et al.*, 2001). Similarly, the term "suitability" is also used to describe the values produced

by another profile modelling technique based on Ecological Niche Factor Analysis (Hirzel *et al.*, 2001, 2002). The values of the water balance models are biologically meaningful since they empirically integrate energy and moisture levels.

Guisan and Zimmermann (2000) distinguish models that predict the fundamental niche from those that predict the realised niche of the target organism. Correlative models such as those presented here (PCA and LR) use actual distribution records to make predictions and these therefore must be drawn from the realised niche of that organism (Malanson *et al.*, 1992). Thus, although biotic interactions are not explicitly accounted for (Robertson *et al.*, 2001), their influence will be implicit by sampling the realised niche, and the result has been considered to be a prediction of the realised niche (Austin and Smith, 1989; Malanson *et al.*, 1992; Franklin, 1995; Guisan and Zimmermann, 2000). Austin (2002) has recently suggested that statistical models (correlative models) may not represent the realised niche but rather an amalgam of realised niche and sink areas. Sink areas are those areas where population growth is below replacement and populations are maintained by dispersal from source areas, where population growth is positive (Pulliam, 1988).

In contrast, mechanistic models that are based only on physiological constraints and that do not explicitly account for biotic interactions (such as the SWB) tend to predict the fundamental niche of the target organism (Austin and Smith, 1989; Guisan and Zimmermann, 2000). These can be refined to model the realised niche by adding rules to account for biotic interactions (see Prentice *et al.*, 1992).

In the case of *S. plumieri,* there is probably very little difference between its realised and fundamental niche because it has few predators or pathogens, and effectively no competitors. This may help to explain the close agreement between predictions made by the mechanistic and correlative models in this study.

Mechanistic models (that are fundamental niche models) are considered to be most promising at successfully predicting climatically induced changes in the distribution of plant species (Malanson *et al.*, 1992; Stephenson, 1998; Guisan and Zimmermann, 2000). These models will be more robust under changed climatic conditions than correlative models as certain correlations may cease to apply under changed conditions (Prentice *et al.*, 1992). In particular, correlative models have no way of handling the effects of climate change on the other organisms involved in the biotic interactions underlying realised niches. One solution is to incorporate certain biotic

interactions into predictive models. For example Leathwick *et al.* 1996 and Leathwick and Austin (2001) incorporated the effects of competition from a dominant species into models used to predict the spatial distribution of density of other species.

Model performance

There is fairly good visual agreement (Fig. 2) between the observed data in the form of presence and absence testing locality records and the SWB, the PCA and LR models. This was confirmed by κ statistics calculated using testing locality records (Table 5). In contrast, the κ statistic calculated for the WWB model was negative (Table 5), which indicates extremely poor agreement between the model and the testing localities (Monserud and Leemans, 1992). Using the scale of agreement proposed by (Monserud and Leemans, 1992), the κ statistics indicated "very good" agreement for the PCA and "excellent" agreement for the SWB model and "perfect" agreement for the LR model.

Although the κ-values calculated for the LR model indicated "perfect" agreement with the testing locality records, this probably represents an overestimate of the actual performance of the LR model. The LR model predicted high probabilities of *S. plumieri* being present along the west coast close to the Namibian border (between grid-cells 1470 and 1440; Fig. 2e). However, *S. plumieri* is unlikely to occur in this region. Unfortunately, it was not possible to evaluate the performance of the LR model quantitatively (using κ statistics) along this section of the west coast due to a lack of testing localities (Fig.2 a). Two sources of indirect evidence suggest that *S. plumieri* is unlikely to occur in this region. Firstly, the LR model suggests that *S. plumieri* is likely to occur at the Namibian border (grid-cell 1440 at the edge of the map region), which in turn suggests that it is also likely to occur just beyond the current map region, along the arid coast of Namibia. However, no herbarium records exist for this species from Namibia, and a distribution map produced by Tinley (1985) suggests that *S. plumieri* does not occur in Namibia but only occurs considerably further north along the northern coast of Angola. Finally, the SWB and PCA models also suggest that *S. plumieri* is unlikely to occur along this section of the South African coast. Although indirect evidence suggests that *S. plumieri* may be absent along this section of the coast, a survey is required to confirm this.

The LR model attained a higher κ-value than the SWB model, suggesting better agreement between the observed and predicted values for the LR model than for the SWB model. This may be at least partly influenced by the method of evaluation used. Chatfield (1995) maintains that splitting data into a training and a testing set is a poor substitute for true replication, as the two datasets are not completely independent. This is likely to be true in this case and this dependence (between the training and testing sets) could explain the higher κ-value calculated for the LR model than that calculated for the SWB model. The LR model was built using a training set of presence and absence localities, and then evaluated using a set of testing localities that were not completely independent of the training set. In contrast, the SWB model was based on empirical ecophysiological data rather than on presence and absence localities. This model was thus evaluated using data that were independent of the data used to build it. The LR model thus had a better chance of performing well using this model evaluation test than the SWB model. Similarly, the PCA model also had a better chance of performing well using this test than the SWB model – the only difference being that the PCA model was built using only presence data (rather than presence and absence data as was the case with the LR model). This is the reason for using the *model agreement* tests (Table 6) discussed below.

κ-values suggest that the LR model also performed better than the PCA model. This is possibly because the LR model had the advantage of being built using 57 absence localities in addition to the 158 presence localities used in the PCA model.

The performance of the mechanistic model would probably have been better had another mechanistic process been incorporated into this model. Although the LR and PCA models had a greater chance of performing well using these model evaluation tests than the SWB model, all three models demonstrated good performance.

Model agreement

Kappa values that were calculated using the reclassified SWB map were used to assess the level of agreement between the SWB and both the PCA and LR models. Kappa values indicated "good" agreement between the SWB model and the PCA model (Table 6). The agreement between the SWB model and the LR model was "very good" using the scale of agreement proposed by Monserud and Leemans

(1992). These values suggest very good correspondence between predictions made using a simple mechanistic model (the SWB model) and two correlative models (the PCA and LR models). Slightly better agreement between the LR model and the SWB model than between this model and the PCA model can again be explained by the LR model having 57 absence localities in addition to the 158 presence localities used to build the PCA model. The close agreement between the correlative (PCA and LR) models and the mechanistic model (SWB) may be because the realised niche of *S. plumieri* probably quite closely resembles its fundamental niche.

Water balance as a predictor variable in correlative models

The results of model performance tests suggest that the SWB model is a far better predictor of *S. plumieri* presence than the WWB model. This may be explained by considering the phenology of the plant (Peter *et al.*, 2002). The SWB model was calculated for those months which coincide with periods when the plant is actively growing and reproducing (Peter *et al.*, 2002). This suggests that water balance values calculated for these periods may be mechanistically more important, and as a result should be more important and useful for predicting plant distributions. This is particularly important when water balance is used as a predictor variable in correlative models (in contrast to the mechanistic approach adopted in this study). Water balance has been used as a predictor variable in various correlative studies (for example, Leathwick, 1995; Leathwick *et al.*, 1996; Leathwick, 1998; Leathwick and Whitehead, 2001). Stephenson (1998) suggested that site water balance should be used as a predictor variable for the purposes of building correlative models, even if it is only crudely calculated. Site water balance considers the interactions of energy and water which are important for predicting distributions (Stephenson, 1998). In addition, Guisan and Zimmermann (2000) have suggested that "physiology-based" parameters such as site water balance should be used in preference to physiographic predictors for developing static models that are more mechanistic. The data in this study suggest that the phenology of the target species should be taken into consideration when calculating water balance values in order to focus on those periods which are biologically significant.

Profile vs. group discrimination techniques

An issue that is likely to generate considerable debate in the future is the choice between profile and group discrimination techniques for predicting species distributions. Profile models are likely to be appropriate when absence data are not available or are unreliable. In a recent study, using simulated data, Hirzel *et al.* (2001) found that a profile technique (Ecological Niche Factor Analysis, ENFA) performed better than a group discrimination technique (GLM) in cases where the target organism was not in equilibrium with the environment e.g. alien invasive organisms. However, when reliable presence and absence data are available then group discrimination techniques are likely to perform best (Ferrier and Watson, 1997; Hirzel *et al.*, 2001). This is confirmed by the slightly better agreement between the SWB and LR models than the agreement between the SWB and PCA models.

When a target organism is not in equilibrium with its environment, absence data collected for this organism are likely to contain a proportion of false absence records, which may adversely affect model performance (Hirzel *et al.*, 2001). However, logistic regression has been successfully applied using "unsurveyed absence" records (e.g. Cumming 2000 a & b). Recently, Zaniewski *et al.* (2002) successfully used "pseudo-absence" data to apply generalised additive models (GAM) for predicting the distribution of ferns.

There appears to be a need for further comparative studies that compare the performance of profile and group discrimination techniques using data of varying quality. In particular, an important question is whether group discrimination techniques such as GLM or GAM can be applied without using surveyed absence data (e.g. Cumming, 2000 a & b; Zaniewski *et al.*, 2002) to produce models that perform better than those produced by profile techniques (e.g. Robertson *et al.*, 2001; Hirzel *et al.*, 2002), given data of the same quality.

Conclusion

The results suggest that correlative models can perform as well or better than simple mechanistic models. However, the generality of this statement requires testing. These models have different requirements in terms of input data and prior knowledge

of the target organism's biology. The predictions generated from these models are each likely to offer slightly different insights into the potential distribution and biology of the target organism and may be appropriate for different purposes. Mechanistic models such as those used here (the SWB model) require a greater knowledge of the biology of the organism and require making time-consuming ecophysiological measurements. These models may in turn yield greater insights into the biology of the organism that are mechanistic in nature. In contrast, correlative models are easier and less time-consuming to build and require less prior knowledge of the target organism's biology. These models will yield different insights into the biology and potential distribution of the target organism from mechanistic models. The choice of model is likely to be influenced by several factors, such as the aims of the study, the biology of the target organism, the level of knowledge the target organism's biology and data quality. The type of target organism and the level of knowledge of the target organism's biology is likely to play a central role in the model development process. Simple correlative models may be used initially when the biology and distribution of the organism is not well known. These models can then be iteratively refined (see Chatfield, 1995) to produce models that are more mechanistic in nature. This is equivalent to incorporating more knowledge of "ecological process" (Austin, 2002) into the model. This may be followed (if feasible) by the development of ecophysiological models that simulate the mechanisms considered to underlie the correlations between the distribution records and the predictor variables observed in developing the correlative models.

Acknowledgements

I thank Ursula Hertling for absence locality data collected on the west coast; Sarah Radloff for statistical advice; Mike Burton for assistance with MATLAB software; the School of Bioresources Engineering and Environmental Hydrology (University of Natal), the Water Research Commission and the South African Country Study for Climate Change for the use of the climatic predictor variables. Funding from the Rhodes University Joint Research Council and the National Research Foundation is gratefully acknowledged.

Table 1: Correlation matrix of monthly water balance values for the entire coast (n=1439). All significant correlations ($p < 0.05$) are indicated with *.

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.
Feb.	0.992*										
March	0.979*	0.981*									
April	0.830*	0.839*	0.862*								
May	0.216*	0.208*	0.269*	0.676*							
June	-0.200*	-0.210*	-0.160*	0.323*	0.900*						
July	-0.190*	-0.200*	-0.150*	0.326*	0.901*	0.989*					
Aug.	0.039	0.019	0.088*	0.502*	0.958*	0.939*	0.948*				
Sept.	0.735*	0.723*	0.783*	0.915*	0.762*	0.440*	0.449*	0.661*			
Oct.	0.927*	0.921*	0.959*	0.906*	0.464*	0.056*	0.068*	0.317*	0.909*		
Nov.	0.986*	0.984*	0.991*	0.861*	0.283*	-0.140*	-0.130*	0.113*	0.796*	0.965*	
Dec.	0.996*	0.987*	0.980*	0.829*	0.230*	-0.180*	-0.170*	0.064*	0.756*	0.939*	0.990*

Table 2: Predictor variables and their abbreviations (Abbrev.) used in the PCA and LR models.

Abbrev.	Predictor variable
RH1	Component axis 1 of a PCA on 12 monthly mean relative humidity map
RH2	Component axis 2 of a PCA on 12 monthly mean relative humidity map
MXT1	Component axis 1 of a PCA on 12 monthly maximum temperature map
MXT2	Component axis 2 of a PCA on 12 monthly maximum temperature map
MNT1	Component axis 1 of a PCA on 12 monthly minimum temperature maps
MNT2	Component axis 2 of a PCA on 12 monthly minimum temperature maps
RN1	Component axis 1 of a PCA on 12 monthly rainfall maps
RN2	Component axis 2 of a PCA on 12 monthly rainfall maps

Table 3: A confusion matrix used to calculate kappa statistics (Fielding & Bell, 1997). Where + indicates presence and – indicates absence. The parameters a, b, c and d represent counts rather than percentages.

		Observed	
		+	-
Predicted	+	a	b
	-	c	d

Table 4: Coefficients (B) with their associated standard errors (SE) and Wald statistics for the LR model. The full names of the variables can be found in Table 2.

Variable	B	SE	Wald
RN1	3.893	2.60E-08	2.25E+16
RN2	6.382	4.66E-08	1.88E+16
MNT1	-3.981	3.09E-08	1.66E+16
MXT1	3.723	3.53E-08	1.11E+16
RH1	-2.547	4.23E-08	3.63E+15
MXT2	-1.597	4.73E-08	1.14E+15
RH2	1.688	5.08E-08	1.10E+15
MNT2	-1.631	4.94E-08	1.09E+15
Constant	7.517	2.62E-07	8.23E+14

Table 5: Tests of model performance using κ statistics and confusion matrix parameters. Kappa statistics were calculated using thresholds that yielded maximum values for each of the models, with the exception of summer (SWB) and winter water balances (WWB) which were calculated using thresholds of zero. Kappa statistics were calculated using only those grid-cells in which *S. plumieri* was observed to be either present or absent (surveyed grid-cells).

Model	Threshold	κ	a	b	c	d	N
WWB	0	-0.150	40	17	13	2	72
SWB	0	0.852	52	3	1	16	72
PCA	0.30	0.721	44	0	9	19	72
LR	0.13	1.000	53	0	0	19	72

Table 6: Tests of model agreement using κ statistics and confusion matrix parameters calculated for the PCA and LR models using all the grid-cells along the entire coast. Calculations were performed as follows: those grid-cells with summer water balance values greater or equal to zero were taken to represent "observed presence" while grid-cells with water balance values below zero were taken to represent "observed absence". Thresholds that yielded maximum kappa (κmax) values were selected.

Model	Threshold	κmax	a	b	c	d	N
PCA – all grid-cells	0.04	0.679	891	147	52	349	1439
LR – all grid-cells	0.89	0.786	827	29	116	467	1439

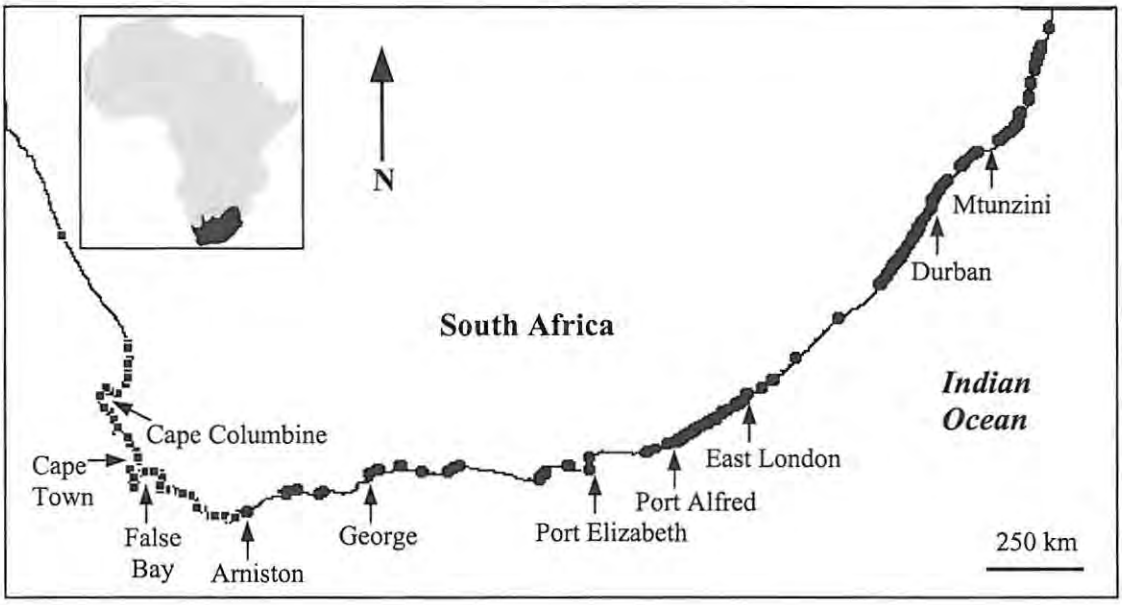


Fig. 1. The coast of South Africa indicating localities where *S. plumieri* was recorded as being present (filled circles) and where it was recorded as being absent (filled squares). In the inset, black indicates the position of South Africa relative to Africa.

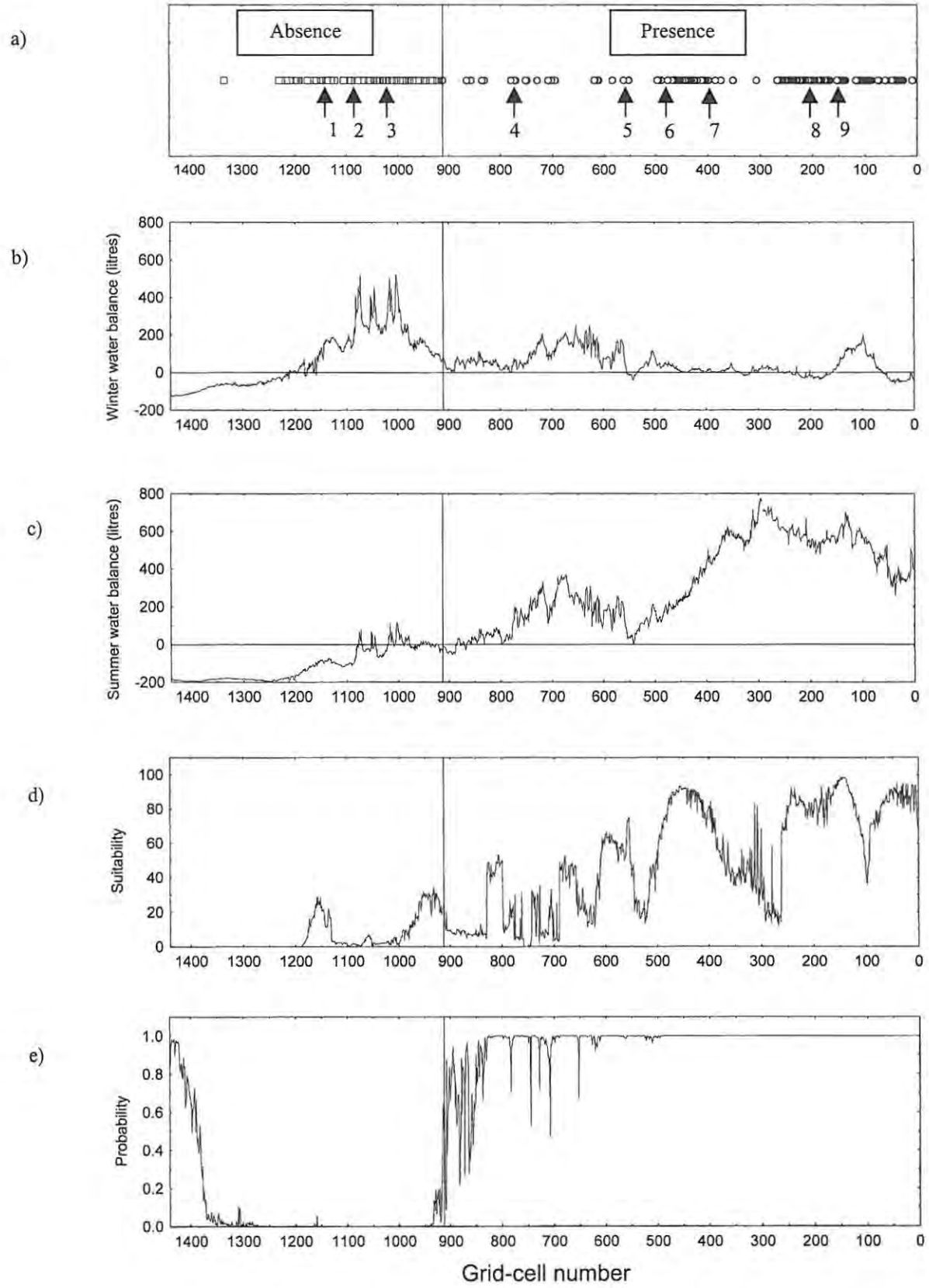


Fig. 2. (a) The observed presence (n=211) and absence (n=76) of *S. plumieri* at localities along the South African coast from the Mozambique border (grid-cell 0) to the Namibian border (grid-cell 1439). The position of major towns and geographical features along the coast: 1- Cape Columbine; 2- Cape Town; 3- False Bay; 4- George; 5- Port Elizabeth; 6- Port Alfred; 7- East London; 8- Durban; 9- Mtunzini. The vertical line in all the figures indicates the position of Arniston, the most westward locality at which *S. plumieri* was observed present. The results of four predictive models: (b) Winter water balance; (c) Summer water balance; (d) Suitability calculated using the PCA model; (e) Probability calculated using the LR model.

References

Austin, M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*. 157: 101-118.

Austin, M.P., Cunningham, R.B., Fleming, P.M., 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio*. 55: 11-27.

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

Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a beta-function. *Journal of Vegetation Science*. 5: 215-228.

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

Austin, M.P., Smith, T.M., 1989. A new model for the continuum concept. *Vegetatio*. 83: 35-47.

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

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

Busby, J.R., 1991. BIOCLIM - a bioclimatic analysis and prediction tool. In: Margules, C.R., Austin, M.P. (Eds.), Nature conservation: cost effective biological surveys and data analysis. CSIRO, Melbourne, pp. 64-68.

Caithness, N. 1995. Pattern, process and the evolution of the African antelope (Mammalia: Bovidae). Ph.D. Dissertation, University of the Witwatersrand, Johannesburg, 210 pp.

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

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A*. 158: 419-446.

Cumming, G.S., 2000 a. Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*. 27: 425-440.

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

Erasmus, B.F.N., Kshatriya, M., Mansell, M.W., Chown, S.L., Van Jaarsveld, A.S., 2000. A modelling approach to antlion (Neuroptera: Myreleontidae) distribution patterns. *African Entomology*. 8: 157-168.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra, p. 193.

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

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*. 19: 474-499.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

Guisan, A., Theurillat, J.-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*. 9: 65-74.

Guisan, A., Weiss, S.B., Weiss, A.D., 1999. GLM versus CCA spatial modelling of plant species distribution. *Plant Ecology*. 143: 107-122.

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

Hirzel, A.H., Helfer, V., Métral, F., 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*. 145: 111-121.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-Niche Factor Analysis: how to compute habitat-suitability maps without absence data? *Ecology*. 83: 2027-2036.

Huberty, C.J., 1994. Applied discriminant analysis. Wiley Interscience, New York, 466 pp.

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

Jones, P.G., Gladkov, A., 1999. FloraMap - a computer tool for predicting the distribution of plants and other organisms in the wild. International Center for Tropical Agriculture, Cali, Columbia, p. 99.

Leathwick, J.R. 1995. Climatic relationships of some New Zealand forest tree species. *Journal of Vegetation Science*. 6: 237-248.

Leathwick, J.R. 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*. 9:719-732.

Leathwick, J.R., Austin, M.P. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*. 82: 2560-2573.

Leathwick, J.R., Mitchell, N.D. 1992. Forest pattern, climate and vulcanism in central North Island, New Zealand. *Journal of Vegetation Science*. 3: 603-616.

Leathwick, J.R., Whitehead, D. 2001. Soil and atmospheric water deficits and the distribution of New Zealand's indigenous tree species. *Functional Ecology*. 15: 233-242.

Leathwick, J.R., Whitehead, D. McLeod, M. 1996. Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach. *Environmental Software*. 11:81-90.

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

Malanson, G.P., Westman, W.E., Yan, Y-L., 1992. Realized versus fundamental niche functions in a model of chaparral response to climatic change. *Ecological Modelling*. 64: 261-277.

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

McCullagh, P., Nelder, J.A. 1983. Generalized Linear Models. Chapman and Hall, London. p. 261.

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

Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*. 62: 275-293.

Neilson, R.P., 1995. A model for predicting continental-scale vegetation distribution and water balance. *Ecological Applications*. 5: 362-385.

Nicholls, A.O. 1989. How to make biological surveys go further with generalised linear models. *Biological Conservation*. 50: 51-75.

Nix, H.A. 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R. (Ed.), Atlas of Elapid Snakes of Australia. Australian Government Publishing Service, Canberra, pp. 4-15.

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

Packer, M.J., Canney, S.M., McWilliam, N.C., Abdallah, R. 1999. Ecological mapping of a semi-arid savanna. In: Coe, M.J., McWilliam, N.C., Stone, G.N., Packer, M.J. (Eds.), Mkomazi: the Ecology, Biodiversity and Conservation of a Tanzanian Savanna. Royal Geographical Society (with the Institute of British Geographers), London, pp. 43-68.

Palmer, A.R., Van Staden, J.M. 1992. Predicting the distribution of plant communities using annual rainfall and elevation: an example from southern Africa. *Journal of Vegetation Science*. 3: 261-266.

Pearce, J., Ferrier, S. 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*. 128: 127-147.

Peter, C.I., Ripley, B.S., 2000. An empirical formula for estimating the water use of *Scaevola plumieri*. *South African Journal of Science*. 96: 1-4.

Peter, C.I., Ripley, B.S., Robertson, M.P. 2002. The distribution of *Scaevola plumieri* along the South African coast is limited by seasonal water balance and temperature. *Journal of Vegetation Science*. (in press).

Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A., Solomon, A.M., 1992. A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*. 19: 117-134.

Pulliam, H.R., 1988. Sources, sinks, and population regulation. *American Naturalist*. 132: 652-661.

Robertson, M.P., Caithness, N., Villet, M.H., 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*. 7: 15-27.

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

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

Rogers, D.J., Randolph, S.E., 1993. Distribution of Tsetse and Ticks in Africa: past, present and future. *Parasitology Today*. 9: 266-271.

Rogers, D.J., Williams, B.G. (1993): Tsetse distribution in Africa: seeing the wood and the trees. In: Edwards, P.J., May, R. (Eds.), Large-scale ecology and conservation biology. Blackwell Scientific Publications, Oxford, pp. 247-271.

Schulze, R.E., Maharaj, M., Lynch, S.D., Howe, B.J., Melvil-Thomson, B., 1997. South African Atlas of agrohydrology and climatology, 1st Edition. Water Research Commission, Pretoria.

Stephenson, N.L., 1998. Actual evapotranspiration and deficit: biologically meaningful correlates of vegetation distribution across spatial scales. *Journal of Biogeography.* 25: 855-870.

Tinley, K.L. 1985. Coastal dunes of South Africa. South African National Scientific Programmes, Report no. 109.

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

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

Woodward, F.K., Williams, B.G., 1987. Climate and plant distribution at global and local scales. *Vegetatio.* 69: 189-197.

Zaniewski, A.E., Lehmann, A., Overton, J. McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling.* 157: 261-280.

IX

General discussion

Preface

The aim of this chapter is to discuss the major findings of the thesis and to place them within a broader context. In order to do this, the major findings of the thesis are discussed and issues that most urgently require further study are highlighted. Some of these findings are also discussed in the context of invasive alien plant management.

A wide variety of models have been produced to address a large number of questions in applied ecology (Chapter 1). These models have different designs, require different input data, make different assumptions about these data and differ in the results that they produce and the way in which these results can be realised and applied (Chapters 2 and 3). In predictive biogeography, a good understanding of data quality issues is essential for making the necessary decisions in order to predict the potential distribution for a target organism. A number of decisions will be influenced by the quality of available input data and the quality required of the predictions. This highlights the need for documenting data quality issues and in particular the need to understand and minimise sources of error in data (Chapter 2). Another important need is to understand the effects that these data quality issues have on model performance (Chapters 6 and 7) and how to deal with these.

These issues are currently of central importance and are likely to become even more important as correlative predictive modelling is more widely used. In the last few years the number of publications in the literature on predictive modelling techniques and their application has increased rapidly (Chapter 3) and will probably continue, especially considering current concerns of climate change. Interest in this field is likely to increase rapidly as data and predictive modelling software becomes

more easily accessible to a larger number of users (Kaiser, 1999; Stockwell and Peters, 1999; Lehmann *et al.*, 2002). Integrated spatial analysis systems for predicting distributions make both data and software accessible to users (Kaiser, 1999; Stockwell and Peters, 1999). These systems link databases of species location records to predictive modelling algorithms to make predictions and display the results in a World Wide Web browser. An example of this sort of system is the GARP Modelling System (Stockwell and Peters, 1999) an implementation of which can be found in the Biodiversity Species Workshop at the web site http://biodic.sdsc.edu. Another example is The Species Analyst, which can be found at the web site http://tsadev.speciesanalyst.net/.

Lehmann *et al.* (2002) describe GRASP (http://www.cscf.ch), which is a software package for predicting spatial distributions of species using generalised regression. It incorporates all of the necessary processes required to make predictions using point records and environmental predictor variables.

These systems will put predictive modelling tools in the hands of a large number of users who are not necessarily familiar with predictive modelling techniques and data quality issues relating to these predictions. One of the major challenges will be to provide guidance on acceptable use of these techniques (Stockwell and Peters, 1999), and to make suggestions on possible limitations of the data available for making these predictions, especially presence-only data. While some of the limitations of the techniques and the data is already known, much research is still required.

In particular, there is a need for studies to investigate the influence of sampling bias on the predictive performance of models. This is a major factor that may limit the usefulness of predictions based on presence-only data. In particular, we need to know how biased collections data are in general and how this bias influences model performance and the conclusions drawn from these models. While the study of Funk and Richardson (2002) is an important first step, further studies are required. Hypothetical distributions show potential in this regard but studies should also examine real data, preferably for a range of organisms.

The major part of the thesis is devoted to data quality issues, their influence on model performance and the choice of modelling technique selected. Specifically, one of the issues addressed was that of absence data availability and reliability. In response to the limited number of reliable profile techniques available for predicting species distributions at the commencement of the thesis, two profile techniques were implemented and evaluated.

The Fuzzy Envelope Model (FEM, Chapter 4) is an envelope technique that uses fuzzy logic to classify a set of predictor variable maps based on the values associated with presence records to produce a potential distribution map for a target species. This technique represents several refinements of the crisp envelope approach used in the BIOCLIM modelling package (the CEM design). These refinements are related to the way in which FEMs deal with uncertainty, the way in which this uncertainty is represented in the resultant potential distribution maps, and the way that these maps can be interpreted and applied. Moreover, the FEM technique predicted the potential distribution of three alien invasive plant species and three cicada species with an average performance that was significantly higher than similar models built using the CEM design. The FEM technique shows promise as a profile technique although further studies are required to establish its reliability under a range of different data quality conditions. In the multiple comparison of modelling techniques (Chapter 7; species comparison) the PCA technique appeared to perform better than the FEM technique but there were cases where the performance of the FEM technique was equivalent or better than that of the PCA.

The PCA-based modelling technique (Chapter 5) uses a fundamentally different approach from the FEM technique. It is based on a hyperspace defined by principal component axes obtained from a PCA on the training set. Each of the grid-cells in the map region is fitted into this hyperspace and its distance from the origin is a measure of centrality in this hyperspace, which is expressed as a probability. Other techniques that are similar to the PCA technique implemented here (Chapter 5) have been described and implemented (Jones and Gladkov, 1999; Erasmus *et al.*, 2000; Hirzel, 2001). The algorithms and procedures used by these techniques have been described in relation to the PCA technique (Chapter 3). Quantitative comparisons among these alternative profile techniques would be useful to determine their domains of application and to establish whether techniques that are similar can perform equally

well. Comparisons such as those used in Chapter 7 and elsewhere (Ferrier and Watson, 1997; Hirzel *et al.*, 2001) would be particularly useful. In the multiple comparison (Chapter 7; species results) the PCA technique performed better on average than the BEM and FEM techniques and in certain cases demonstrated equivalent performance to the LRSA (logistic regression surveyed absence) models.

The results of investigations into the effects of sample size, prevalence and false absence records on the performance of logistic regression models (Chapter 6) suggest that false absence records and sample size have a significant effect on model performance. However, logistic regression appears to be robust to a certain proportion of false absence records. The use of pseudo-absence data appears to be viable in certain cases, however this depends on the extent of the range of the target species and the method used to select pseudo-absence records. If false absences can be kept to a minimum then the pseudo-absence design appears to be viable. However, the bias likely to be present in presence-only data are will pose a problem.

Prevalence was found to significantly effect model performance. Samples with very low (10%) or very high (90%) prevalence produced models that were significantly lower in performance than those built using samples with less extreme prevalence (30%-80%). Prevalence did not appear to have a negative effect on model performance when smaller samples of records were used (160-320 records) but were more serious when large samples were used (2560-5120 records). These results suggest that it may be possible to reduce the uncertainly in pseudo-absence approaches by selecting fewer pseudo-absence records than there are presence records, for modelling.

A more general point that this study raises is that the AUC measurement scale has a much smaller range than more commonly used measurement scales such as percentages or probabilities. AUC values can range between 0.5 for no agreement and 1.0 for perfect agreement between a predicted distribution and an observed distribution (Zwieg and Campbell, 1993). Therefore, care should be taken when interpreting these values. Ranges of agreement proposed by Pearce and Ferrier (2000) may be useful here. In addition, significant differences in model performance can occur within a very narrow range of AUC values.

This study illustrates the utility of hypothetical distribution approaches for evaluating the effects of data quality issues on model performance and for testing

hypotheses. Although these hypothetical distributions are useful, they are unlikely to accurately represent real organisms in every respect, even when precautions are taken to make them as realistic as possible (Hirzel *et al.*, 2001). Therefore, hypothetical distribution studies should be complemented by suitable examples using real data, where possible (e.g. Chapter 7).

Comparisons among profile and group discrimination techniques were performed using hypothetical distributions and data from real organisms. For the hypothetical distributions the CEM and FEM model designs showed unexpectedly high performance relative to the other model designs. The way in which the hypothetical distributions were generated appeared to confer an unfair advantage on two of the designs (FEM and CEM), thus reducing the usefulness of the hypothetical distribution investigation. This suggests that care should be taken when using hypothetical distribution comparisons as there is no guarantee that these hypothetical distributions are realistic or that they do not confer an unfair advantage on certain model designs.

The species investigation found that model design and sample size had a significant effect on model performance. On average, the LRSA design most frequently produced models that occurred in the group (with other model designs whose performance did not differ significantly), that had the highest average performance, followed by the PCA and then the LRPSA designs. The CEM and FEM designs most frequently produced models with the lowest average performance.

These results suggest that if presence and absence data are available then the LRSA model design should be selected in preference to the other designs. If only presence data are available then the PCA model design should be selected, as it is likely yield superior models more often than the CEM, FEM and LRPSA designs.

The PCA design did not differ significantly in performance from the LRSA design for six of the eight species, suggesting that profile techniques can produce equivalent results to group discrimination techniques under certain conditions. However, the sources of data that profile techniques typically rely on may be of poor quality, thus reducing their performance.

An important finding is that optimal thresholds (Franklin, 1998; Guisan *et al.*, 1998) used to discriminate between presence and absence on continuous probability maps can differ significantly among model designs. This suggests that the meaning of

the response surfaces produced by various model designs may be fundamentally different, and that comparisons among model designs should not be done using performance measures that use a single threshold only.

This study is significant in that it quantitatively compares the performance of five correlative modelling techniques. Studies such as this one and that of Ferrier and Watson (1997) can make major contributions by quantitatively comparing several techniques simultaneously.

Comparisons among mechanistic and correlative techniques suggest that correlative models can perform as well as, or better than, simple mechanistic models (Chapter 8). The predictions generated from these three modelling designs are likely to generate different insights into the potential distribution and biology of the target organism and may be appropriate in different situations.

Quantitative comparative studies

One of the important needs that has been highlighted in this thesis and elsewhere (Guisan and Zimmermann, 2000) is for quantitative comparative studies in which the performance of more than two techniques is assessed using the same dataset. In addition, the influence of several data quality issues on these techniques needs to be evaluated. The number of comparisons and factors that need to be assessed is extremely large. As a result, managing the data and modelling software required to make these predictions will be a challenge. One of the ways in which these comparisons can be made effectively and efficiently is by making a variety of competing techniques and various datasets available to users by means of integrated spatial analysis systems (e.g. Kaiser, 1999; Stockwell and Peters, 1999; Lehmann *et al.*, 2002). Studies have suggested that different predictive modelling techniques may be appropriate to different species, types of distribution or quality of data (Hirzel *et al.*, 2001; Chapters 6 and 7). This suggests that a choice among several different predictive techniques will probably be better than trying to provide only a single robust technique (e.g. GARP, Stockwell and Peters, 1999).

Model evaluation

The basis for making quantitative comparisons among models is a reliable accuracy measure for quantifying model performance. Several measures are available (e.g. Kappa, ROC curves) but these measures evaluate the overall performance of the model and do not give any details of the spatial distribution of errors in the predictions (Guisan and Zimmermann, 2000). Cowley *et al.* (2000) acknowledged this problem and presented maps of predicted and observed distributions in addition to statistics of agreement, in an attempt to address the problem. Although this approach may partially overcome this problem, it is likely to be quite subjective and probably not rigorous enough.

Viewed in another way, these measures do not give any indication for which part of the hyperspace the predictions are unsuccessful. Perhaps consideration should be given to developing accuracy measures that assess model performance by partitioning the hyperspace into subsets and assessing model performance within these.

Another important issue is to quantitatively compare optimal threshold accuracy measures (such as the maximum kappa value measure) with more popular threshold independent measures (such as ROC curves) to determine whether the optimal threshold measures can produce equivalent results.

Collaboration between biologists and modellers

Making potential distribution predictions for a target organism may involve collaboration between a biologist who is familiar with the biology and systematics of the target organism and a modeller who is familiar with predictive modelling techniques and spatial data management (e.g. Robertson *et al.*, 2000; Peter *et al.*, 2002; Baars and Robertson, *in prep*). In practice, the modeller usually doesn't know much about the biology of the organism and the biologist doesn't know much about predictive modelling. In order to predict the potential distribution of the target organism and to answer further questions that may arise from this prediction, a mutual learning process must occur. In this process, the modeller needs to describe the type of result that can be achieved using available techniques, while the biologist needs to

describe what is known about the biology of the target organism and the type of data that is available for this organism. The process will probably begin with the development of a preliminary model, which is then iteratively refined as new insights into the biology of the target organism are gained, or as new hypotheses regarding the distribution of the target organism are framed.

The success of these interactions between the modeller and the biologist will largely determine the success of the modelling effort. These types of interactions should be documented in an attempt to find ways of making interactive modelling more efficient. This type of data would be particularly useful for designing and developing integrated modelling systems (e.g. Kaiser, 1999; Stockwell and Peters, 1999).

Pearce *et al.* (2001) investigated the possibility of incorporating expert opinion into faunal distribution models. They identified particular aspects of the modelling process where experts could make the greatest positive impact but highlighted the need for a more thorough evaluation, particularly at different scales (Pearce *et al.*, 2001). Further studies such as this would also be extremely beneficial to the development of integrated spatial analysis systems and to predictive modelling in general.

Predictive modelling and invasive alien plants

The FEM (Chapter 4) and PCA-based (Chapter 5) profile modelling techniques described and implemented in the thesis have considerable potential for predicting invasive alien plant distributions, especially as absence data are often unreliable (Chapter 2). The pseudo-absence design of logistic regression (Chapters 6 and 7) also shows potential for predicting alien plant distributions, without the need for surveyed absence data. Although these techniques show considerable potential, they require further evaluation under a range of data quality conditions and for various types of distribution. The success of these techniques will rely on the sample of presence data being an unbiased representative sample.

In South Africa (and elsewhere), considerable benefit would be derived from developing integrated spatial analysis systems designed specifically for invasive alien plants. Such a system would be particularly useful for management of invasive alien

plants, especially to the government's Working for Water Programme (www-dwaf.pwv.gov.za/wfw/). This system could be linked to existing atlas projects e.g. southern African plant invaders atlas (SAPIA: Henderson, 1998; Henderson, 1999). The system could also incorporate an online version of the prioritisation system (see Appendix). Some of the criteria in the prioritisation system could be based on potential distribution predictions of the plants.

Potential distribution predictions can be used as inputs for other models, such as those used to investigate the consequences of various control programmes on invasive alien plants (Wadsworth *et al.*, 2000).

Another area of considerable promise for predictive modelling is in assessing the potential distribution and likely impact of insect biological control agents introduced to control invasive alien plants (Baars, 2002; Baars and Robertson, *in prep*).

Conclusions

Predictive models show considerable promise in a number of areas of applied ecology, but Rogers *et al.* (1996) have warned that although statistical models describing the habitat or distribution of a target species may be useful, they are not a substitute for an understanding of the biology of these phenomena. This thesis contributes to our understanding of some of the important considerations about the type and quality of the data used to calibrate distribution models.

References

Baars, J-R, 2002. Biological control initiatives against *Lantana camara* L. (Verbenaceae) in South Africa: an assessment of the present status of the programme, and an evaluation of *Coelocephalapion camarae* Kissinger (Coleoptera: Brentidae) and *Falconia intermedia* (Distant) (Hemiptera: Miridae), two new candidate natural enemies for release on the weed. Ph.D. Thesis, Rhodes University, Grahamstown.

Baars, J-R., Robertson, M.P. The potential distribution of selected biocontrol insects introduced into South Africa for the control of *Lantana camara* L. In preparation for *Biological Control*.

Cowley, M.J.R., Wilson, R.J., Leon-Cortes, J.L., Gutierrez, D., Bulman, C.R., Thomas, C.D., 2000. Habitat-based statistical models for predicting the spatial distribution of butterflies and day-flying moths in a fragmented landscape. *Journal of Applied Ecology*. 37: 60-72.

Erasmus, B.F.N., Kshatriya, M., Mansell, M.W., Chown, S.L., Van Jaarsveld, A.S., 2000. A modelling approach to antlion (Neuroptera: Myreleontidae) distribution patterns. *African Entomology*. 8: 157-168.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra, p. 193.

Franklin, J., 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*. 9: 733-748.

Funk, V.A., Richardson, K.S. 2002. Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*. 51: 303-316.

Guisan, A., Theurillat, J-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*. 9: 65-74.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*. 135: 147-186.

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

Henderson, L., 1999. The Southern African Plant Invaders Atlas (SAPIA) and its contribution to biological weed control. *African Entomology Memoir* 1:159-163.

Hirzel, A., 2001. When GIS come to life. Linking landscape and population ecology for large population management modelling: the case of Ibex (*Capra ibex*) in Switzerland. Ph.D. Thesis, Institute of Ecology, Laboratory for Conservation Biology, University of Lausanne.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*. 145: 111-121.

Jones, P.G., Gladkov, A., 1999. FloraMap - a computer tool for predicting the distribution of plants and other organisms in the wild. International Center for Tropical Agriculture, Cali, Columbia, pp. 99.

Kaiser, J., 1999. Searching museums from your desktop. Science. 284: 888.

Lehmann, A., Overton, J. McC., Leathwick, J.R. 2002. GRASP: generalised regression analysis and spatial prediction. *Ecological Modelling*. 157: 189-207.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*. 133: 225-245.

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

Peter, C.I., Ripley, B.S., Robertson, M.P. 2002. The distribution of *Scaevola plumieri* along the South African coast is limited by seasonal water balance and temperature. *Journal of Vegetation Science*. (in press).

Robertson, M.P., Herbert, D., Villet, M.H., 2000. Predictive modelling of invasive snail distributions in South Africa- *Theba pisana* Müller (1774). *Molluscs 2000 conference*, Sydney, Australia, December 2000.

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

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science*. 13: 143-158.

Wadsworth, R.A., Collingham, Y.C., Willis, S.G., Huntley, B., Hulme, P.E., 2000. Simulating the spread and management of alien riparian weeds: are they out of control? *Journal of Applied Ecology*. 37: 28-38.

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

Appendix

A proposed prioritisation system for the management of weeds in South Africa

Preface

The manuscript presented here describes a proposed prioritisation system for the management of weeds in South Africa. This manuscript is currently in preparation for *The South African Journal of Science* (Robertson, M.P., Villet, M.H., Palmer, A.R., Fairbanks, D.H.K., Henderson, L., Higgins, S., Hoffmann, J.H., Le Maitre, D.M., Riggs, I., Shackleton, C.M., Zimmermann, H.G. A proposed prioritization system for the management of weeds in South Africa). This system represents a potentially important and useful means of making policy decisions for managing invasive plants. It is relevant here as at least one of the criteria in the system may make use of potential distribution predictions. In addition, the system was also used in the process of selecting some of the target species (invasive alien plants).

Abstract

In any country a number of weed species occur which in some way conflict with human management objectives and needs. When resources for research and control are limited, priority should be given to species that are most problematic or undesirable. A prioritisation system was designed to objectively assess research and control priority of alien and indigenous weeds at a national scale in South Africa. The system consists of eighteen criteria, grouped into five modules which assess invasiveness; spatial characteristics; potential impact; potential for control and conflicts of interest for each plant species under consideration. Total prioritisation scores, calculated from criterion and module scores, are used to assess the priority for a species. Prioritisation scores are calculated by combining independent assessments provided by several experts, thus increasing the reliability of the prioritisation. The total confidence score, a separate index, indicates the reliability and availability of data used to make an assessment. Candidate species for assessment were identified and these species assessed by a number of experts using the prioritisation system. The

system has a multi-assessor, modular design with criterion and module score standardisation, which offer a number of advantages over existing systems.

Introduction

A number of publications contain lists of plant taxa which are described as "problem plants", "declared weeds", "declared invaders" or "alien invaders" in South Africa (Wells *et al.*, 1986; Henderson, M. *et al.*, 1987; Richardson *et al.*, 1997). These taxa have been described in this way because they possess at least some characteristics that bring them into conflict with human interests or because they are ecologically harmful under certain circumstances. The lists include many species for example Wells *et al.* (1986) produced a catalogue of problem plants in southern Africa which contains 1653 taxa. This represents a vast number of undesirable species, which in many cases require the application of some form of control measure. Given the limited availability of resources for research and control there is a need to focus attention on the control, study and monitoring of the most problematic and thus most undesirable species, rather than less problematic species.

In South Africa, and very likely other developing countries, there is currently no formal means of identifying those species that are most problematic and most warranting attention for intervention measures. In view of this, a system was designed to prioritise research and control efforts against alien invasive plant species, which have already become established in South Africa, and for indigenous invasive plant species.

Indigenous invasive species referred to in this manuscript are typically those that are involved in a phenomenon known as bush encroachment, whereby trees and shrubs increase in density in savanna situations (Smit *et al.*, 1996). This process often results in adverse effects on a natural community or changes in a natural community which conflict in some way with human activities or management objectives. Bush encroachment is considered to be a major problem in South Africa (Grossman and Gandar, 1989; Trollope, 1992), being most detrimental in arid savannas (Tainton, 1984).

Alien invasive species have numerous deleterious impacts on the environment; as summarised by Macdonald *et al.* (1986) for impacts in southern Africa. An alien species is defined as one that is remote from its centre of origin, usually from a different continent or subcontinent.

This system is different to other systems which have been designed largely for preventing species invasions at the quarantine stage (Navarantham and Catley, 1986; Smallwood and Salmon, 1992; Pheloung, 1995; Tucker and Richardson, 1995). To distinguish this system from the abovementioned systems, this system is known as a "prioritisation system", rather than a ranking or rating system. Prioritisation systems, also known as rating systems (Smallwood and Salmon, 1992), or ranking systems (Macdonald and Jarman, 1984; Hiebert, 1997) generally consist of a set of criteria and some sort of scoring system against which the threat posed by a species can be assessed.

Methods

In describing this prioritisation system, the design of the system is first outlined followed by the criteria against which the organism is assessed.

System design

The design of the system refers to the manner in which the scores obtained from the assessment criteria are arranged, combined or weighted to produce a total score for a given species. A modular approach was followed in which the 18 assessment criteria were grouped into five modules (Table 1), with each module being dedicated to a particular aspect or issue. A similar approach has been used previously (Smallwood and Salmon, 1992) although the definition of terms differs from those used here.

It is undesirable to have a situation where the potential maximum scores for each criterion can have different values because it often leads to arbitrarily uneven weighting of criteria. To overcome this problem, the score for each criterion is scaled so that the potential maximum score for each criterion is 1 (Smallwood and Salmon, 1992). Similarly, the score for each module is divided by the number of criteria in

that module. The modules can be differentially weighted according to the needs of users, to emphasise certain aspects of interest (Moran, 1983; Macdonald and Jarman, 1984; Smallwood and Salmon, 1992).

An effective means of obtaining a reliable assessment for a number of plants is to use a multi-assessor approach. In a multi-assessor approach, the opinions of several assessors are considered to be better than the opinion of a single assessor (for discussion see Hiebert, 1997). This type of approach is less sensitive to biases or individual experience of assessors. Assessors may be able to provide data for some criteria and not for others, depending on their field of specialisation. Each assessor's field of specialisation is likely to be slightly different and thus a multi-assessor approach can make optimal use of available data.

Hiebert (1997) outlines two decision-making techniques, namely nominal group techniques (NGTs) and the Delphi method. Nominal group techniques involve an interactive group structure while the Delphi method uses the individual opinions of experts with no "face-to-face" interaction (Hiebert, 1997). For this reason, the Delphi method has a number of advantages over NGTs. The Delphi method allows a group of individuals to reach consensus without ever meeting. This helps to reduce certain factors such as dominance of a discussion by one or more participants, the bandwagon effect, and unwillingness to abandon a previous opinion (Hiebert, 1997). Apart from the effects of group dynamics and human interaction on responses, it is often difficult to get a panel of experts together at the same time. The prioritisation system was designed so that independent assessments from several different assessors could be made using the Delphi method.

An issue of concern is how to deal with uncertain or missing data. One might omit the criterion (and alter the module scaling factor accordingly), or add a small penalty score that makes the system err on the conservative side (Smallwood and Salmon, 1992; Tucker and Richardson, 1995). The disadvantage of these approaches is that the total score is artificially inflated by uncertain or missing data. It is thus impossible to determine whether the species has a high score because of uncertain or missing data or because the species is genuinely problematic.

To overcome this limitation, a separate score known as a *Confidence Score* was used, which gives an indication of uncertainty and availability of data for each criterion. The lower the confidence score the greater the uncertainty and amount of

missing data for that criterion. This system has the advantage that it explicitly indicates a level of confidence in the *Total Prioritisation Score* assigned to a species, i.e. it can be used as a measure of how much faith and further research should be placed in a given prioritisation score. In addition, the confidence score can be used as a measure of the state of knowledge of a given species.

Using the prioritisation system

To apply the prioritisation system, a species is scored for each of the criteria (Table 1). These scores are then added up for each module, and the total divided by the number of criteria for which assessments were made in the respective module (criteria for which no score is provided are ignored). The result is a set of module scores that can be weighted according to the needs and emphasis of particular users. For general purposes, modules can be given equal weight. The system is designed so that the user can customise the system by weighting the modules, but the criteria within those modules have fixed weightings that cannot be altered by the user. A final, user-specific prioritisation of the candidate species can be made by summing the (un)weighted module scores for each species (this is the equivalent of weighted averaging of modules) and ranking the taxa by this index.

To calculate the total confidence score for a species, a similar approach is taken to that described above. In cases where data are missing, a confidence score of zero is assigned; where data are uncertain, a confidence score of 0.5 and where data are certain, a confidence score of 1 is assigned; for a given criterion. The confidence scores are re-scaled for each criterion and module to produce a total confidence score as described above for prioritisation scores.

The criteria

The criteria used in the prioritisation system were selected by the authors at a workshop convened specifically to design a system for ranking plant species at a national scale in South Africa. The best assessment criteria may not necessarily be those for which data are available (Macdonald and Jarman, 1984), or for which data can easily be acquired. These constraints as well as ease-of-use considerations

influenced the criterion selection process. A detailed account of the selection process and a justification for the inclusion of each criterion can be found elsewhere (Robertson and Palmer, 1999).

Prioritising a candidate list of species

The list of candidate species to be assessed was compiled by including those species for which: herbicides have been registered (Henderson, L., 1995; Vermeulen *et al.*, 1996); biocontrol agents have been released (Wells *et al.*, 1986; Hoffmann, 1991; Henderson, L., 1995); legislation has been passed (Wells *et al.*, 1986; Henderson, L., 1995) or for which legislation is proposed (Henderson, L., 1995). A list of alien invasive species compiled by Richardson *et al.* (1997) was also included. Unpublished sources include a list of taxa targeted by the South African Working for Water Programme (Department of Water Affairs and Forestry; unpublished), and a list of important encroachment species (Trollope, *pers comm.*, 1997).

Independent assessments for the list of candidate species were then provided by a number of assessors, using the Delphi method. Scores for each criterion were obtained by calculating the median of the individual scores assigned by each of the assessors for that criterion. Each criterion score was standardised by dividing the score obtained by the potential maximum for that criterion. These criterion scores were summed to produce the module score which was then also standardised by dividing that score by the number of criteria used to obtain that score. The total prioritisation score was obtained by summing the standardised module scores for that species. The module scores were each given a weighting of one.

The species were ranked according to the product of the total prioritisation score and the total confidence score. This provides a single meaningful index by which the species can be ranked objectively. This approach produces a more realistic ranking of the species than rankings produced by first sorting the list using prioritisation score and then sorting by the confidence score. The interpretation of the ranking should still be done using the total prioritisation and total confidence scores as well as the number of assessors.

Results

Sixty two weedy species were ranked according to the product of their total prioritisation scores and total confidence scores (Table 2). This list is not intended to be a national, prioritised list of South African weeds, but merely demonstrates the prioritisation system and illustrates ways in which the results can be analysed.

Those species that have high prioritisation scores and high confidence scores are of most concern for example *Lantana camara*, *Chromolaena odorata* and *Opuntia ficus-indica* (ranked 1, 2 and 3 respectively). Those species that have low prioritisation scores and high confidence scores are of least concern, for example *Harrisia martinii, Opuntia spinulifera* and *Opuntia exaltata* (ranked 59, 60 and 61 respectively).

The highest ranking species (*Lantana camara*) obtained a score of 3.33 and the lowest ranking species (*Opuntia rosea*) obtained a score of 1.26 (Table 2). Total confidence scores ranged from 5 to 3.97 and the number of assessors ranged from 3 to 11 (Table 2).

Discussion

Prioritisation Scores

The maximum possible total prioritisation score attainable in the prioritisation system is 5 and the lowest is zero. A number of species, which may be highly desirable, are likely to obtain a total prioritisation score which is very close to zero, but it is unlikely that any weeds will obtain a total prioritisation score of 5 (Table 2).

Some of the results should however be treated as preliminary results due to the low numbers of assessors (e.g. 3 assessors) used to assess many of the species. Care should be taken when comparing total prioritisation scores calculated from data provided by different numbers of assessors, as is the case here. Total prioritisation scores calculated from a greater number of assessors are likely to be more reliable. For example the score calculated for *Acacia mearnsii* (ranked 9th) is much more reliable because it was calculated using data from 10 assessors as opposed to that of

Tamarix ramossima (ranked 8th) which was calculated using data from only 3 assessors. Based on the experience of the authors, *Tamarix ramossima* appears to have been rated too highly by the system, possibly because it was only assessed by three assessors. Comparing total prioritisation scores which were calculated using different numbers of assessments can be likened to comparing the results of experiments which were conducted using different sample sizes. The results indicate the need for assessments to be made using at least 10 assessors, although further work is needed to establish minimum number of opinions to canvass for a dependable result. The system can be used to identify gaps in expertise by identifying those species which had low assessor numbers.

The reliability of the total prioritisation score (and hence of the rank) is clearly dependent on the number of criteria assessed, the quality of the data available to the assessor and the number of assessors involved in the assessment. When the results of a species prioritisation are interpreted, then each of these factors has to be taken into account.

In an ideal situation all of the criteria would be used to calculate the total prioritisation score for a species. This is not always possible due either to a genuine lack of reliable data or a lack of access to these data. If the total prioritisation score is calculated using all the criteria then one can be sure that all the relevant factors were taken into account. If only some of the criteria were taken into account then the species may obtain an artificially high or low score due to its unique set of undesirable attributes. Caution should be exercised when comparing total prioritisation scores that were calculated using different numbers of criteria. The total confidence score gives an indication of the perceived quality and availability of data used by an assessor. This score is extremely important for evaluating the assessments provided by an individual assessor for a given species as well as for assessing the rank that a species has been assigned.

Confidence scores

Species that had confidence scores with values of less than 3.5 were not reported by assessors. There is some quality control in the system through "self-

censorship" by the assessors themselves, as they appeared to be unwilling to assess a species about which they had little confidence.

Species that have both high total prioritisation scores and high total confidence scores are most likely to be highly problematic weeds (Table 2). *Lantana camara*, *Chromolaena odorata* and *Opuntia ficus-indica* are good examples which have high prioritisation scores (3.33, 3.06 and 2.96 respectively), and high confidence scores (4.52, 4.86 and 4.77 respectively). These species are ranked 1, 2 and 3 respectively (Table 2). These are species for which the greatest number of resources, human and economic, can confidently be invested.

Species which have high total prioritisation scores (TP) but low total confidence scores indicate an urgent need for further investigation or research to obtain more confidence in their prioritisation score and hence their rank. In order to be conservative, these species should be treated as serious until evidence to the contrary is provided. It is more difficult to justify large investments of resources in these species than in those with high prioritisation scores and high confidence scores. These species should be carefully monitored and researched.

Species with low total prioritisation scores and high total confidence are of little cause for concern to weed managers for the foreseeable future. Species with low total prioritisation scores and low total confidence scores are of more concern that the former group and should be monitored. These species could be more problematic than their prioritisation scores suggest, due to the uncertainty of the data used to obtain these scores, as indicated by low confidence scores.

Ranks

The highest ranking species in the list is *Lantana camara* (Table 2). *L. camara* is described as one of the most serious invader species in South Africa and considered to be one of the world's ten worst weeds (Bromilow, 1995). Other species with high total prioritisation scores include *Chromolaena odorata*, *Opuntia ficus-indica*, *Acacia saligna*, *Cestrum laevigatum*, *Acacia mearnsii* and *Prosopis* spp. (Table 2) which are also considered to be problem species in South Africa (Richardson *et al.*, 1997). Based on the experience of the authors, a number of species appear to have been ranked surprisingly low, these include: *Acacia mearnsii, Harrissia martinii, Hakea*

sericea, Acacia cyclops, Melia azdedarach, Acacia dealbata, Pinus pinaster, Psidium gujava, Opuntia stricta and *Opuntia rosea*. Species which appear to have been ranked unexpectedly high include: *Opuntia ficus indica, Cinnamum caphoratus, Datura stramonium, Cestrum laevigatum, Schinus molle, Ricinus communis, Myriophyllum aquaticum, Salvinia molesta, Tamarix ramossima* and *Arundo donax*.

Despite the outliers, the prioritisation system appears to have delivered credible results with a meaningful decrease in the status of the species being observed as one moves from highest to lowest rank on the list (Table 2). For example, a species which is known to be highly problematic (Bromilow, 1995) such as *Lantana camara* is ranked at the top of the list while a species such as *Acacia karroo*, which is an indigenous invasive involved in bush encroachment and is comparatively much less problematic, is ranked near the bottom of the list (Table 2). The system has heuristic value because it challenges preconceptions about the ranking of species. In addition, the system can be used to provide sound reasons for assigning a particular rank to a species. These reasons can be found by performing a detailed examination of the criteria used in the system and the criterion scores that the species attained.

The results of any prioritisation should be treated with caution because the rank which is assigned to a given species is at least in part affected by the other species included in the list of candidate species. For example, if a number of potentially high-ranking species (i.e. serious or highly undesirable weeds) are excluded from a list then other, less serious weeds will have higher ranks than they would if these species were included (although the total prioritisation scores will be unaffected). This illustrates the need for thoughtful decisions regarding criteria for selection of candidate species for assessment. While any plant could potentially be assessed using the prioritisation system, it would be wasteful to assess species at random from a national species list since most of them would not be weeds and would thus have a prioritisation score close to zero.

A ranking of plant species based on prioritisation scores is valid for a limited period of time because the status of these plants can change due to successful intervention strategies, introduction of new species, or rapid population increases of certain plant (McLaren *et al.*, 1998).

The scale at which the prioritisation is performed is also likely to influence the rank of the species. Species rankings produced at a provincial scale (local scale) are

likely to be different from rankings produced at a national scale (regional scale). This is because a different list of candidate species will almost certainly be used at the provincial scale because not all of the species included in the national list are likely to occur within the particular province of interest. Rankings will also be influenced by the land area covered by the province and the climatic suitability of the area.

In the species list, a given species may have a higher prioritisation score than another, indicating a higher rank and therefore a higher priority status. If the prioritisation scores and confidence scores used to assign the ranks are very similar, then this ranking becomes arbitrary. For example *Melia azedarach* is ranked 28 in our list, based on a prioritisation score of 2.57, while *Caesalpinia decapetala* is ranked 29, based on a prioritisation score of 2.56 (Table 2). The confidence scores are 4.48 and 4.49 respectively (Table 2). These scores are not appreciably different, indicating that these ranks should be treated with discretion.

Conclusions

The prioritisation system presented here is a useful decision support system for weed-control and research organisations, not only for South Africa but also in other countries and at various spatial scales. This system can be customised according to the needs of these organisations by altering the module weightings or altering some of the criteria used for assessment. In addition, the system can also be used to assess the state of knowledge of weeds by determining: (i) gaps in expertise by identifying species with low assessor numbers; (ii) gaps in knowledge by identifying species for which few criterion questions were answered; and (iii) gaps in insight by examining those species with surprisingly high or low ranks.

The design of the system in terms of standardisation of criterion and module scores, the confidence scoring system and the Delphi assessment approach have many advantages as outlined above. As a result, these design features should be incorporated into future prioritisation systems. These design features are also applicable to prioritisation systems used to assess other organisms (Smallwood and Salmon, 1992) and to systems designed for different management purposes such as risk assessment at the quarantine (Navarantham and Catley, 1986; Tucker and Richardson, 1995).

Acknowledgements

We thank Estelle Brink (Selmar Schonland Herbarium, Grahamstown), Clive Bromilow (Bayer Chemicals), Tony Gordon (ARC - Plant Protection Research Institute), Martin Hill (ARC - Plant Protection Research Institute), Pat Hulley (Dept. Zoology and Entomology, Rhodes University), Hildegard Klein (ARC - Plant Protection Research Institute), Dinemari Nel (Western Cape Department of Agriculture), Geoff Nichols (Geoff Nichols Horticultural Services), Carl Stoltz (ARC - Plant Protection Research Institute) and Costas Zachariades (ARC - Plant Protection Research Institute) for providing additional species assessments. This work was funded by the National Department of Agriculture and the Agricultural Research Council - Range and Forage Institute. Financial assistance from the Foundation for Research Development, towards the workshop, is gratefully acknowledged. The Agricultural Research Council - Plant Protection Research Institute is thanked for hosting the workshop.

Table 1. The prioritisation system outlined above, consists of five modules: Potential Invasiveness, Actual Spatial Extent, Potential Impacts, Potential for Control and Conflicts of Interest. Each module consists of a number of criteria, with their weightings appearing in square brackets on the right of each criterion. Highest weightings are associated with the most undesirable characteristics. The score for each criterion divided by the maximum possible score gives the Criterion Score (denoted by a lower-case letter). The sum of the Criterion Scores divided by the number of criteria in the module gives the Module Score (denoted by an upper-case letter). Module can be weighted according to the requirements of the user, by applying a module weighting (Wa, Wb, Wc, Wd, We) to the module score. The sum of the Module Scores gives the Total Score. Confidence scores are assigned to each criterion based on the uncertainty and availability of data. For a given criterion, the following confidence scores are assigned: 0 where data are missing, 0.5 where data are uncertain, and 1 where data are certain. For indigenous invasives, the *Impact on Water Resources* and the *Invasive Elsewhere* criteria, are not applicable. This should be taken into account when calculating the Module Score and Total Score of an indigenous invasive.

Criteria and Modules		Module Score	Confidence Score
MODULE A: POTENTIAL INVASIVENESS			
a) Long-distance dispersal			
There is:			
1) no known long-distance dispersal mechanism	[0]		
2) a known long-distance dispersal mechanism (dispersal >5 km)	[1] __/1=		
b) Invasive elsewhere:			
The species is invasive elsewhere, outside of South Africa?			
1) Yes	[1]		
2) No	[0] __/1=		
		__/2 =	__/2 =
MODULE B: SPATIAL CHARACTERISTICS			
c) Distribution:			
The current percentage of 15' (quarter degree) grid squares in the entire country (approx. 2000) occupied by the species is:			
1) 1 - 2% (up to 40 quarter degree squares) e.g. *Hakea drupacea*	[0]		
2) 3 - 5% (up to 100 quarter degree squares) e.g. *Cereus jamacaru* & *Chromolaena odorata*	[1]		
3) 6 - 10% (up to 200 quarter degree squares) e.g. *Jacaranda mimosifolia*	[2]		
4) 11 - 20% (up to 400 quarter degree squares) e.g. *Prosopis* spp. & *Acacia dealbata*	[3]		
5) 21 - 40% (up to 800 quarter degree squares) e.g. *Acacia mearnsii* & *Melia azedarach*	[4]		
6) > 40% (over 800 quarter degree squares) e.g. *Opuntia ficus-indica*	[5] __/5 =		
d) Density:			
The species occurs predominantly as:			
1) individual plants	[0]		
2) small clumps	[1]		
3) vast monospecific stands	[2]		
4) mixed stands with other invasives	[3] __/3 =		
		__/2 =	__/2 =
MODULE C: POTENTIAL IMPACTS			
e) Biodiversity:			
Reduction in biodiversity where the species occurs is:			
1) none	[0]		
2) minor (1-30%)	[1]		
3) moderate (31-80%)	[2]		
4) profound (>80%)	[3] __/3 =		

Table 1 contd.

f) Water resources:

The species' impact on water resources is:

1) no impact	[0]
2) reduction of stream flow by 10-30%	[1]
3) reduction of stream flow by > 30%	[2]
4) flow eradicated	[3] __/3 =

g) Negative economic impact:

The negative economic impact of the species is:

1) no negative impact	[0]
2) < 10% reduction in profit	[1]
3) 11 - 30% reduction in profit	[2]
4) > 30% reduction in profit	[3]
5) land unusable	[4] __/4 =

h) Positive economic impact:

The positive economic impact of the species is:

1) none	[4]
2) informal	[3]
3) small business	[2]
4) commercial (industrial)	[1]
5) any two or more of the above	[0] __/4 =

i) Poison status:

The species is poisonous to stock or humans

1) yes	[1]
2) no	[0] __/1 =

__/5 = __/5 =

MODULE D: POTENTIAL FOR CONTROL

j) Chemical control:

The options for realistic chemical control of the species are:

1) not available	[3]
2) impractical in most situations	[2]
3) partially successful	[1]
4) effective and practical	[0] __/3 =

k) Biological control:

The options for biological control of the species are:

1) complete control	[0]
2) substantial control	[1]
3) negligible control	[2]
4) no agents released yet	[3] __/3 =

l) Mechanical control:

The options for mechanical control of the species are:

1) not available	[3]
2) impractical in most situations	[2]
3) partially successful	[1]
4) effective and practical	[0] __/3 =

n) Legislation:

Legislation to assist in the control of the species
(e.g. classification as a declared weed or declared invader) is:

1) absent	[1]
2) in place	[0] __/1 =

Table 1 contd.

o) Accountability:

Can any agency be held accountable for the introduction or proliferation of an invasive species in South Africa?

1) No	[1]		
2) Yes	[0]	__/1 =	
		__/6 =	__/6 =

MODULE E: CONFLICTS OF INTEREST

p) Commercial sector:

Possible conflicts of interest at the commercial sector level are:

1) No conflict	[0]		
2) Possible resolution to conflict	[1]		
3) Biological control precluded	[2]		

q) Informal sector:

Possible conflicts of interest at the informal sector level are:

1) None	[0]		
2) in cases where rural households harvest plants to meet their daily needs of food or energy	[1]		
3) in cases where rural households sell plants or plant products as a source of income on a supplementary or full-time basis	[2]	__/2 =	

r) Cost/ benefit analysis:

The species has:

1) substantial economic value (including informal sector & commercial markets)	[0]		
2) some economic value (e.g. building material or windbreaks)	[1]		
3) limited value (e.g. ornamental or horticultural value)	[2]		
4) no apparent commercial, ornamental or horticultural value	[3]	__/3 =	
		__/3 =	__/3 =

	A __x Wa	A __
	B __x Wb	B __
	C __x Wc	C __
	D __x Wd	D __
	E __x We	E __
TOTAL		

Table 2. A list of 62 species ranked according to the product of total prioritisation score and the total confidence score for each species. Total Criteria refers to the total criterion score and total confidence to the total confidence score. Critxconf refers to the product of the total criterion score and the total confidence score. The number of assessors refers to the number of assessors who provided scores for one or more of the criteria for a given species. Indigenous species are indicated by *.

Rank	Species	Total Criteria	Total Confidence	Critxconf	No. of Assessors
-	Maximum score	5.00	5.00	-	-
1	*Lantana camara*	3.33	4.52	15.05	11
2	*Chromolaena odorata*	3.06	4.86	14.87	9
3	*Opuntia ficus-indica*	2.96	4.77	14.12	11
4	*Acacia saligna*	2.75	5.00	13.75	3
5	*Cestrum laevigatum*	2.78	4.90	13.62	3
6	*Prosopis spp.*	2.98	4.57	13.62	7
7	*Hakea gibbosa*	2.79	4.88	13.62	3
8	*Tamarix ramossima*	3.02	4.44	13.41	3
9	*Acacia mearnsii*	2.99	4.48	13.40	10
10	*Azolla filiculoides*	2.92	4.57	13.34	10
11	*Solanum mauritianum*	2.95	4.52	13.33	11
12	*Myriophyllum aquacticum*	2.66	4.96	13.19	4
13	*Acacia cyclops*	2.82	4.61	13.00	3
14	*Pinus patula*	2.80	4.63	12.96	3
15	*Salvinia molesta*	2.62	4.90	12.84	3
16	*Pinus elliottii*	2.75	4.63	12.73	3
17	*Pereskia aculeata*	2.58	4.91	12.67	4
18	*Acacia melanoxylon*	2.75	4.58	12.60	4
19	*Cinnamomum camphoratus*	2.65	4.75	12.59	3
20	*Datura stramonium*	2.59	4.83	12.51	3
21	*Arundo donax*	3.08	4.03	12.44	3
22	*Leptospermum laevigatum*	2.85	4.29	12.26	4
23	*Eichhornia crassipes*	2.63	4.65	12.23	5
24	*Ricinus communis*	2.62	4.56	11.95	8
25	*Acacia baileyana*	2.94	3.97	11.67	3
26	*Pinus radiata*	2.83	4.09	11.60	3
27	*Acacia dealbata*	2.70	4.28	11.56	9
28	*Melia azedarach*	2.57	4.49	11.54	8
29	*Caesalpinia decapetala*	2.56	4.48	11.47	10
30	*Acacia decurrens*	2.60	4.37	11.36	3
31	*Pinus halepensis*	2.64	4.29	11.33	3
32	*Pinus canariensis*	2.61	4.29	11.20	3
33	*Nassella trichhotoma*	2.44	4.49	10.96	3
34	*Pinus pinea*	2.50	4.38	10.95	3
35	*Cirsium vulgare*	2.37	4.61	10.93	3
36	*Rubus cuneifolius*	2.47	4.37	10.79	3
37	*Solanum elaeagnifolium*	2.38	4.44	10.57	4
38	*Psidium guajava*	2.39	4.41	10.54	9
39	*Pinus pinaster*	2.46	4.28	10.53	9
40	*Paraserianthes lophantha*	2.30	4.48	10.30	4
41	*Litsea glutinosa*	2.50	4.11	10.23	3
42	*Schinus molle*	2.28	4.47	10.19	3
43	*Opuntia aurantiaca*	2.11	4.75	10.02	3
44	*Hakea sericea*	2.23	4.46	9.95	10
45	*Acacia longifolia*	2.24	4.28	9.59	11
46	*Opuntia stricta*	1.97	4.73	9.32	3

Table 2. Contd.

47	*Pennisetum clandestinum*	2.04	4.51	9.20	3
48	*Acacia pycnantha*	2.06	4.38	9.02	3
49	*Passiflora edulis*	2.01	4.36	8.76	3
50	*Pistia stratiotes*	1.98	4.35	8.61	3
51	*Opuntia imbricata*	1.86	4.62	8.59	3
52	*Opuntia monocantha*	1.81	4.62	8.36	3
53	*Hypericum perforatum*	2.01	4.07	8.18	3
54	*Solanum sisymbrifolium*	1.76	4.61	8.11	3
55	*Ipomoea purpurea*	1.77	4.47	7.91	3
56	*Opuntia lindheimeri*	1.49	5.00	7.45	3
57	*Acacia karroo* *	1.52	4.90	7.45	3
58	*Jacaranda mimosifolia*	1.47	4.90	7.20	3
59	*Harrisia martinii*	1.42	4.54	6.45	3
60	*Opuntia spinulifera*	1.26	4.83	6.09	3
61	*Opuntia exaltata*	1.28	4.45	5.62	3
62	*Opuntia rosea*	1.26	4.07	5.13	3

References

Bromilow, C., 1995. Problem plants of South Africa, 1st Edition. Briza Publications, Arcadia, pp. 315.

Grossman, D., Gandar, M.V., 1989. Land transformation in South African Savanna regions. *South African Geographical Journal*. 71: 38-45.

Henderson, L., 1995. Plant invaders of southern Africa. Plant Protection Research Institute Handbook no.5. Agricultural Research Council, Pretoria, pp. 177.

Henderson, M., Fourie, D.M.C., Wells, M.J., Henderson, L., 1987. Declared weeds and alien invader plants in South Africa. Department of Agriculture and Water Supply, Pretoria, pp. 167.

Hiebert, R.D., 1997. Prioritizing invasive plants and planning for management. In: Luken, J.O., Thieret, J.W. (Eds.), *Assessment and management of plant invasions*, Springer, New York, pp. 195-212.

Hoffmann, J.H., 1991. Biological control of weeds in South Africa. *Agriculture, Ecosystems and Environment*, Vol. 37 Special Issue, Elsevier Science Publishers, Amsterdam.

Macdonald, I.A.W., Jarman, M.L., 1984. Invasive alien organisms in the terrestrial ecosystems of the fynbos biome, South Africa. South African National Scientific Programmes Report no. 85.

Macdonald, I.A.W., Kruger, F.J., Ferrar, A.A., 1986. The ecology and management of biological invasions in southern Africa. Oxford University Press, Cape Town.

McLaren, D.A., Stajsic, V., Gardener, M.R., 1998. The distribution and impact of South/North American stipoid grasses (Poaceae: Stipeae) in Australia. *Plant Protection Quarterly*. 13: 62-70.

Moran, V.C., 1983. The phytophagous insects and mites of cultivated plants in South Africa: patterns and pest status. *Journal of Applied Ecology*. 20: 439-450.

Navarantham, S.J., Catley, A., 1986. Quarantine measures to exclude plant pests. In: Groves, R.H., Burdon, J.J. (Eds.), *Ecology of Biological Invasions*, Cambridge University Press, Cambridge, pp. 106-112.

Pheloung, P.C., 1995. Determining the weed potential of new plant introductions to Australia. Draft report to the Australian Weeds Commitee and the Plant Industries Committee. Agriculture Protection Board, Western Australia.

Richardson, D.M., Macdonald, I.A.W., Hoffmann, J.H., Henderson, L., 1997. Alien plant invasions. In: Cowling, R.M., Richardson, D.M., Pierce, S.M. (Eds.), *Vegetation of southern Africa*, Cambridge University Press, Cambridge, pp. 535-570.

Robertson, M.P., Palmer, A.R., 1999. Bioclimatic modelling of invasive alien plants. A report to the National Department of Agriculture (Directorate of Resource Conservation). Agricultural Research Council - Range and Forage Institute, Grahamstown.

Smallwood, K.S., Salmon, T.P., 1992. A rating system for potential exotic bird and mammal pests. *Biological Conservation*. 62: 149-159.

Smit, G.N., Rethman, N.F.G., Moore, A., 1996. Vegetative growth, reproduction, browse production and response to tree clearing of woody plants in African savanna. *African Journal of Range and Forage Science*. 13: 78-88.

Tainton, N.M., 1984. Veld and pasture management in South Africa. Shuter and Shooter, Pietermaritzburg.

Trollope, W.S.W., 1992. Control of bush encroachment with fire in the savanna areas of South Africa. In: Hurt, C.R., Zacharias, P.J.K. (Eds.), *Prestige Farmers Day Proceedings 1991-1992*, Grassland Society of Southern Africa, Pietermaritzburg, pp. 8-11.

Tucker, K.C., Richardson, D.M., 1995. An expert system for screening potentially invasive alien plants in South African fynbos. *Journal of Environmental Management*. 44: 309-338.

Vermeulen, J.B., Dreyer, M., Grobler, H., van Zyl, K., 1996. A guide to the use of herbicides, 15th Edition. Government Printer, Pretoria, pp. 153.

Wells, M.J., Balsinhas, A.A., Joffe, H., Engelbrecht, V.M., Harding, G., Stirton, C.H., 1986. A catalogue of problem plants in southern Africa. *Memoirs of the Botanical Survey of South Africa*. No. 53.