University of Canberra

This thesis is available in print format from the University of Canberra Library.

If you are the author of this thesis and wish to have the whole thesis loaded here, please contact the University of Canberra Library at *e-theses@canberra.edu.au* Your thesis will then be available on the www providing greater access.

COLLEGE DEPOSIT

PLANT SPECIES RARITY AND DATA RESTRICTION INFLUENCE THE PREDICTION SUCCESS OF SPECIES DISTRIBUTION MODELS

James Mugodo B.Sc. (Hons.)

Applied Ecology Research Group

University of Canberra ACT 2601

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Applied Science at the University of Canberra

July 2002

ACKNOWLEDGEMENTS

I am grateful to my supervisors Dr David Williams and Associate Professor George Cho for their advice, encouragement and patience. I am equally thankful to the Commonwealth Scientific and Industrial Research Organisation (CSIRO) for kindly providing the data used in this study and to Dr Mike P. Austin for facilitating the collaboration and for his assistance throughout the study.

The statistical help from Alice Richardson (University of Canberra) in the early days of the project and from David Pederson (University of Canberra) in the latter days was most appreciated. My communications on model assessment with Beth Atkinson (Mayo Foundation for Medical Education and Research), Michael I. Westphal (University of Queensland) and Jane Elith (Melbourne University) were invaluable.

My studies were sponsored by the WWF Africa and Madagascar Programme in conjunction with WWF-US, and I am most thankful for their assistance.

A big thanks to my fellow students in the Applied Ecology Research Group and my house-mates for their friendship and unbelievable moral support that kept me going. Most of all, I thank my family, especially my Mom. Even though they were on the other side of the Indian Ocean, their support and encouragement were not diminished by the physical distance that separated us during my studies.

ABSTRACT

There is a growing need for accurate distribution data for both common and rare plant species for conservation planning and ecological research purposes. A database of more than 8 500 observations for nine tree species with different ecological and geographical distributions and a range of frequencies of occurrence in south-eastern New South Wales (Australia) was used to compare the predictive performance of logistic regression models, generalised additive models (GAMs) and classification tree models (CTMs) using different data restriction regimes and several model-building strategies. Environmental variables (mean annual rainfall, mean summer rainfall, mean winter rainfall, mean annual temperature, mean maximum summer temperature, mean minimum winter temperature, mean daily radiation, mean daily summer radiation, mean daily June radiation, lithology and topography) were used to model the distribution of each of the plant species in the study area.

Model predictive performance was measured as the area under the curve of a receiver operating characteristic (ROC) plot. The initial predictive performance of logistic regression models and generalised additive models (GAMs) using unrestricted, temperature restricted, major gradient restricted and climatic domain restricted data gave results that were contrary to current practice in species distribution modelling. Although climatic domain restriction has been used in other studies, it was found to produce models that had the lowest predictive performance. The performance of domain restricted models was significantly (p = 0.007) inferior to the performance of major gradient restricted models when the predictions of the models were confined to the climatic domain of the species. Furthermore, the effect of data restriction on model predictive performance was found to depend on the species as shown by a significant interaction between species and data restriction treatment (p = 0.013). As found in other studies however, the predictive performance of GAM was significantly (p = 0.003) better than that of logistic regression. The superiority of GAM over logistic regression was unaffected by different data restriction regimes and was not significantly different within species.

The logistic regression models used in the initial performance comparisons were based on models developed using the forward selection procedure in a rigorous-fitting model-building framework that was designed to produce parsimonious models. The rigorous-fitting model-

building framework involved testing for the significant reduction in model deviance (p = 0.05) and significance of the parameter estimates (p = 0.05). The size of the parameter estimates and their standard errors were inspected because large estimates and/or standard errors are an indication of model degradation from overfitting or effects such as multi-collinearity. For additional variables to be included in a model, they had to contribute significantly (p = 0.025) to the model predictive performance. An attempt to improve the performance of species distribution models using logistic regression models in a rigorous-fitting model-building framework, the backward elimination procedure was employed for model selection, but it yielded models with reduced performance.

A liberal-fitting model-building framework that used significant model deviance reduction at p = 0.05 (low significance models) and 0.00001 (high significance models) levels as the major criterion for variable selection was employed for the development of logistic regression models using the forward selection and backward elimination procedures. Liberal fitting yielded models that had a significantly greater predictive performance than the rigorous-fitting logistic regression models (p = 0.0006). The predictive performance of the former models was comparable to that of GAM and classification tree models (CTMs). The low significance liberal-fitting models had a much larger number of variables than the high significance liberal-fitting models, but with no significant increase in predictive performance. To develop liberal-fitting CTMs, the tree shrinking program in S-PLUS was used to produce a number of trees of different sizes (subtrees) by optimally reducing the size of a full CTM for a given species. The 10-fold cross-validated model deviance for the subtrees was plotted against the size of the subtree as a means of selecting an appropriate tree size. In contrast to liberal-fitting logistic regression, liberal-fitting CTMs had poor predictive performance.

Species geographical range and species prevalence within the study area were used to categorise the tree species into different distributional forms. These were then used to compare the effect of plant species rarity on the predictive performance of logistic regression models, GAMs and CTMs. The distributional forms included restricted and rare (RR) species (*Eucalyptus paliformis* and *Eucalyptus kybeanensis*), restricted and common (RC) species (*Eucalyptus delegatensis, Eucryphia moorei* and *Eucalyptus fraxinoides*), widespread and rare (WR) species (*Eucalyptus elata*) and widespread and common (WC) species (*Eucalyptus sieberi, Eucalyptus pauciflora* and *Eucalyptus fastigata*). There were

significant differences (p = 0.076) in predictive performance among the distributional forms for the logistic regression and GAM. The predictive performance for the WR distributional form was significantly lower than the performance for the other plant species distributional forms. The predictive performance for the RC and RR distributional forms was significantly greater than the performance for the WC distributional form. The trend in model predictive performance among plant species distributional forms was similar for CTMs except that the CTMs had poor predictive performance for the RR distributional form.

This study shows the importance of data restriction to model predictive performance with major gradient data restriction being recommended for consistently high performance. Given the appropriate model selection strategy, logistic regression, GAM and CTM have similar predictive performance. Logistic regression requires a high significance liberal-fitting strategy to both maximise its predictive performance and to select a relatively small model that could be useful for framing future ecological hypotheses about the distribution of individual plant species. The results for the modelling of plant species for conservation purposes were encouraging since logistic regression and GAM performed well for the restricted and rare species, which are usually of greater conservation concern.

TABLE OF CONTENTS

CERTIFICATE OF ORIGINALITY II

COPYRIGHT II

ACKNOWLEDGEMENTS III

ABSTRACT IV

LIST OF TABLES X

LIST OF FIGURES XII

CHAPTER 1 1

INTRODUCTION AND LITERATURE REVIEW 1

1.1 INTRODUCTION 1

1.2 LITERATURE REVIEW 5

1.2.1 Theory of plant species distribution modelling *5*
1.2.2 Assumptions of plant species distribution modelling *7*
1.2.3 Improvements in species distribution modelling *11*
1.2.4 Parsimonious models *12*
1.2.5 Variable selection procedures *12*
1.2.6 Choice of predictor variables for species distribution modelling *15*
1.2.7 Prediction methods *18*
1.2.7.1 Algorithms 18
1.2.7.2 Logistic regression model 20
1.2.7.3 Generalised additive models (GAMs) 20
1.2.7.4 Classification tree models (CTMs) 21
1.2.8 Data restriction *21*
1.2.9 Distribution modelling of common and rare plant species *23*
1.2.10 Measurement of model predictive performance *24*
1.2.11 Conclusion *28*

1.3 THE STUDY QUESTIONS 28

1.3.1 Research hypotheses and thesis structure *29*

CHAPTER 2 31

COMPARISON OF THE PREDICTIVE PERFORMANCE OF LOGISTIC REGRESSION AND GENERALISED ADDITIVE MODELS UNDER DIFFERENT REGIMES OF DATA RESTRICTION 31

2. 1 INTRODUCTION 31

2.1.1 Strategies for obtaining independent test data *32*
2.1.2 An explicit and more robust measure of model predictive performance *32*
2.1.3 Issues that may cause problems for model predictive performance comparisons *34*
2.1.4 Data restriction *35*
2.1.5 Aims *37*

2.2 METHODS 37

2.2.1 Study Area *37*
2.2.2 Biological data and predictor variables *39*
2.2.3 Analysis *42*
2.2.3.1 Data modifications 42
2.2.3.2 Data restriction 43

2.2.3.3 Comparison of model performance using the receiver operating characteristic (ROC) 47
2.2.3.4 Training and testing data sets 48
2.2.3.5 Model building 49
2.2.3.6 Fitting of logistic regression models 49
2.2.3.7 Fitting of GAMs 50
2.2.3.8 Analysis of model predictive performance 52
2.3 RESULTS 54
2.3.1 Correlation between independent variables 54
2.3.2 Species distribution models 57
2.4 DISCUSSION 70
2.5 CONCLUSION 73
CHAPTER 3 75
COMPARISON OF THE PREDICTIVE PERFORMANCE OF FORWARD SELECTION AND BACKWARD ELIMINATION LOGISTIC REGRESSION FOR PLANT SPECIES DISTRIBUTION MODELLING 75
3.1 INTRODUCTION 75
3.1.1 Model selection procedures 75
3.1.2 Criticisms of model selection procedures 78
3.1.3 Forward selection versus backward elimination 81
3.1.4 Other possible ways of improving predictive performance of logistic regression models 83
3.1.5 Aims of the chapter 83
3.2 METHODS 84
3.2.1 Data 84
3.2.2 Species distribution modelling 84
3.2.3 Rigorous modelling framework 84
3.2.4 Forward selection and backward elimination species distribution models 85
3.2.5 Model predictive performance comparisons 87
3.3 RESULTS 88
3.4 DISCUSSION 93
3.5 CONCLUSION 96
CHAPTER 4 98
EFFECT OF PLANT SPECIES RARITY ON THE PREDICTIVE PERFORMANCE OF DIFFERENT MODELLING METHODS 98
4.1 INTRODUCTION 98
4.1.1 Definitions of species rarity 98
4.1.2 Distribution modelling of plant species across different distributional forms 101
4.1.3 Current studies on the statistical modelling of rare plant species 103
4.1.4 Larger logistic regression models 104
4.1.5 Assessing model predictive performance on independent data 106
4.1.6 Aims 106
4.2 METHODS 108
4.2.1 Classification of tree species into distributional forms 108
4.2.2 Generation of training and testing sets 111
4.2.3 Species distribution modelling 111
4.2.4 Model predictive performance analysis 114
4.3 RESULTS 116

4.4 DISCUSSION 131

4.5 CONCLUSION 136

CHAPTER 5 138

SYNTHESIS 138

5.1 CONCLUSION 144

REFERENCES 145

APPENDICES 156

Appendix 1: Description of environmental variable estimates used as candidate predictor variables for the modelling of plant species distributions. 156

Appendix 2: The comparison of model predictive performance (AUC of ROC plot) for different plant species using logistic regression under different data restriction regimes. 160

Appendix 3: The comparison of model predictive performance (AUC of ROC) for different species using GAM under different data restriction regimes. 164

Appendix 4: The variables selected by GAM and the various model selection procedures for logistic regression models for the given plant species from each distributional form. 167

Appendix 5: The low significance level (0.05) liberal-fitting backward elimination model (lbls) for *E. Fraxinoides* that had a high predictive performance (AUC of ROC plot = 0.97) despite showing the undesirable model characteristics of outrageously high parameter estimates and standard errors for some its predictor variables. 169

Appendix 6: Major gradient restricted classification tree model for *E. delegatensis* that had a few nodes but showed high predictive performance (AUC of ROC plot = 0.98). 170

Appendix 7: The extent of occurrence (EOO) of *E. pauciflora* in the study area showing the patchy distribution of this common species among the sites sampled in it its range within the study area 171

LIST OF TABLES

Table 2.1: The environmental space within the study area as measured by the continuous rainfall (mm) and temperature (°C) and radiation (mj/m^2/day) climatic variables. 40

Table 2.2: Key for topography, lithology, and soil nutrient status 41

Table 2.3: The thresholds that were used for the application of the major gradient (M), temperature (T) and domain (D) restriction regimes. 45

Table 2.4: Spearman correlation matrix for environmental variables over the entire environmental space of the study area. 55

Table 2.5a) The variables chosen for the logistic regression models for *E. sieberi, E. pauciflora, E. fastigata, E. elata*, *Eucryphia moorei, E. delegatensis, E. fraxinoides, E. paliformis*, and *E. kybeanensis* under the unrestricted, major gradient, temperature and domain restriction regimes 58

Table 2.5b) The variables chosen for the generalised additive models for *E. sieberi, E. pauciflora, E. fastigata, E. elata*, *Eucryphia moorei, E. delegatensis, E. fraxinoides, E. paliformis*, and *E. kybeanensis* under the unrestricted, major gradient, temperature and domain restriction regimes 60

Table 2.6: Shows that the comparison of the predictive performance (AUC of a ROC plot) of logistic regression models based on different levels of data restriction gave different results depending on whether or not differences in data restriction were taken into account. 63

Table 2.7: ANOVA showing the effects of species, modelling methods, and data restriction on the predictive performance in tree species distribution modelling. 67

Table 3.1: Variables included in the logistic regression models for plant species distribution models developed using forward selection (p = 0.05) (fwd), full backward elimination (p = 0.05) (fbe) and constrained backward elimination (p-value = 0.05) (cbe) model selection procedures. 89

Table 3.2: ANOVA showing the effects of species and model selection procedures on the predictive performance of logistic regression species distribution models. 90

Table 3.3: The comparison of logistic regression model predictive performance for different species using forward selection (p = 0.05) (fwd), full backward elimination (p = 0.05) (fbe) and constrained backward elimination (p = 0.05) (cbe) model selection procedures. 91

Table 4.1: A typology of rare species based on three characteristics; geographic range, habitat specificity, and local population size. 100

Table 4.2: Shows the variety in the size of species geographic ranges (km^2) and species prevalence for the plant species that had their presence-absence distribution sampled over the study area of 56 160.61 km^2. 117

Table 4.3: A framework for defining distributional forms based on a species' geographic range as measured by the species extent of occurrence (EOO) and species' prevalence within the study area (56 160.61 km^2).. 118

Table 4.4: The number of variables used in logistic regression models (lbls, lbhs, lfls, lfhs and rf), GAMs (gam) and the different groups of classification tree models (ct, mct, shct, and shmct). 120

Table 4.5: ANOVA showing the significant effects of species and various model-fitting procedures for logistic regression, GAM and CTMs, on the predictive performance of tree species distribution models. 123

Table 4.6: ANOVA showing the significant effects of species and various model-fitting procedures for logistic regression and GAMs on the predictive performance of tree species distribution models. 123

Table 4.7: The means and p-values for the significant two-way ANOVA comparisons of the predictive performance of logistic regression and GAM models (lbls, lbhs, lfls, lfhs, rf and gam) with CTMs (shmct and shct) for plant species distribution. 126

Table 4.8: ANOVA showing the significant effects of species distributional form (SDF) and various model-fitting procedures for logistic regression, GAM and CTMs, on the predictive performance of tree species distribution models.................... 128

Table 4.9: The means and p-values for the two-way ANOVA comparisons of the predictive performance of CTMs for plant species with different distributional forms. 130

LIST OF FIGURES

Fig. 2.1: A map of the study area showing the location and boundary of the region in which tree species data were collected. 38

Fig. 2.2: Scatterplot matrix showing the pairwise relationships of the independent variables over the study area. 56

Fig. 2.3: The Tukey simultaneous 95% confidence intervals showing that the predictive performance (AUC of ROC plot = domrocs) of major gradient restriction models was significantly higher than that of domain restricted models (dom-resmdom). 68

Fig. 2.4: The effect of no restriction (resdom), temperature restriction (restdom), major gradient restriction (resmdom) and climatic domain restriction (dom) data restriction regimes on the predictive performance (AUC of ROC plot) of distribution models for different tree species. 69

Fig. 4.1: The extent of occurrence (EOO) of *E. sieberi* in the study area showing its high occurrence among the sites sampled in its range within the study area.......... 109

Fig. 4.2: Plot of the Dunnett simultaneous 95% confidence limits for the predictive performance (AUC of a ROC plot = ptroc) of species distribution models. 124

Fig. 4.3: Plot of the Tukey's simultaneous 95% confidence limits showing that the predictive performance (AUC of a ROC plot = ptroc) for WR species was significantly less than that for the other three species distributional forms. 129