



<http://researchspace.auckland.ac.nz>

ResearchSpace@Auckland

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.

<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form.

**THE ANALYSIS OF BINARY
DATA IN QUANTITATIVE
PLANT ECOLOGY**

by

THOMAS WILLIAM YEE

A thesis submitted for the
degree of Doctor of Philosophy
at the University of Auckland.

Department of Mathematics & Statistics
Department of Botany
June 1993

UNIVERSITY OF AUCKLAND LIBRARY
SCIENCE
THESIS

93-285

Abstract

The analysis of presence/absence data of plant species by regression analysis is the subject of this thesis. A nonparametric approach is emphasized, and methods which take into account correlations between species are also considered. In particular, generalized additive models (GAMs) are used, and these are applied to model species' responses to greenhouse scenarios and to examine multispecies interactions. Parametric models are used to estimate optimal conditions for the presence of species and to test several niche theory hypotheses.

An extension of GAMs called vector GAMs is proposed, and they provide a means for proposing nonparametric versions of the following models: multivariate regression, the proportional and nonproportional odds model, the multiple logistic regression model, and bivariate binary regression models such as the bivariate probit model and the bivariate logistic model. Some theoretical properties of vector GAMs are deduced from those pertaining to ordinary GAMs, and its relationship with the generalized estimating equations (GEE) approach elucidated.

Acknowledgements

I wish to express my gratitude to my supervisors, Drs Neil Mitchell and Chris Wild, for their guidance and encouragement throughout the course of this study. Special thanks to Chris for doing the bulk of the proofreading. Thanks to a number of people with whom I had discussions—Drs Alan Lee and Brian McArdle, John Leathwick, and members of the Statistics Unit. I wish to thank Professors Gauld and Scott for granting me a part-time lecturing position over the last five years, and for the environment conducive to study provided by the Department of Mathematics & Statistics to work in.

I wish to thank Professor R. J. Tibshirani and Dr T. J. Hastie for sending me copies of GAIM during the early stages of this work, and Dr J. A. Fessler for VSPLINE; all three kindly answered all my e-mail enquiries. I also wish to thank Dr M. F. Hutchinson for referring me to vector splines and for CUBGCV.

My family, especially my mother, have been very supportive over many years and I wish to thank them too.

Finally, I thank God for giving me the grace and strength to complete this work.

I gratefully acknowledge the financial assistance provided from the following scholarships: U.G.C. Postgraduate Scholarship, William Georgetti Scholarship, C. Alma Baker Postgraduate Scholarship.

Contents

Abstract	iii
Acknowledgements	v
Contents	vii
List of Tables	xiii
List of Figures	xv
1 INTRODUCTION	1
1.1 Notation	4
1.2 The North Island data set	6
1.2.1 Data Sources	6
1.2.2 Climate data	7
1.2.3 The Plant Species	8
1.3 GLMs	9
1.3.1 Logistic Regression	11
1.3.2 Extending the Gaussian Logit Model	13
2 GENERALIZED ADDITIVE MODELS	15
2.1 Introduction	15
2.2 Generalized Additive Models	16

2.2.1	Technical Details	20
2.3	An Analysis of Several North Island Species	25
2.4	Discussion	41
3	MODELLING THE GREENHOUSE EFFECT	45
3.1	Introduction	45
3.2	A Climate Change Scenario	47
3.3	Salinger & Hicks' Scenario	49
3.4	Discussion	52
4	MODELLING MULTISPECIES INTERACTIONS	57
4.1	Preliminaries	57
4.1.1	Conditional and Joint Probabilities	59
4.2	Data Analysis	61
4.3	Two Competing Species	64
4.3.1	The 2-Species Competition Model	67
4.3.2	Mutualism	73
4.3.3	Other Effects of Competition on Response Curves	75
4.4	S Competing Species	79
4.5	Discussion	82
5	ESTIMATION AND INFERENCE FOR OPTIMA	85
5.1	Introduction	85
5.2	GLMs	87
5.2.1	1 Dimension	87
5.2.2	2 Dimensions	95
5.2.3	Designs for Gaussian Response Curves	98

5.3	Weighted Averaging	102
6	TESTING SEVERAL NICHE THEORY HYPOTHESES	107
6.1	Introduction	107
6.2	Testing Equality of Response Surfaces	109
6.3	Testing $H_0 : \mathbf{u} = \mathbf{c}$	110
6.4	Equal Optima	111
6.4.1	1 Dimensional Case	112
6.4.2	2 Dimensional Case	113
6.5	Equally Spaced Optima	115
6.6	Equal Tolerances	116
6.7	Correlated Data	117
7	VECTOR GAMs	121
7.1	Introduction	121
7.2	Vector Additive Models	123
7.2.1	Preliminaries	123
7.2.2	Vector Splines	124
7.2.3	Degrees of Freedom for Linear Vector Smoothers	129
7.2.4	Standard Errors	133
7.2.5	Consistency and Convergence	133
7.3	Vector GAMs	135
7.3.1	Justification via Penalized Likelihood	138
7.3.2	Degrees of Freedom and Standard Errors	139
7.3.3	Smoothing Parameter and Variable Selection	140
7.4	Other Topics	141

7.4.1	Some Parameters as Constants	141
7.4.2	Computational Aspects	142
8	NONPARAMETRIC METHODS FOR BIVARIATE RESPONSES	145
8.1	Introduction	145
8.2	Parametric Bivariate Probit Models	146
8.2.1	The Usual Bivariate Probit Model	146
8.2.2	Separate Correlations For Each Group	150
8.2.3	Correlation as a Function of x	151
8.3	A Nonparametric Bivariate Probit Model	152
8.4	The Bivariate Logistic Model	156
8.5	MacKay's Model	160
8.6	The 2-Species Competition Model	160
9	OTHER VECTOR GAMs	163
9.1	Introduction	163
9.2	The Proportional Odds Model	164
9.2.1	A Nonparametric Nonproportional Odds Model	164
9.2.2	A Nonparametric Proportional Odds Model	165
9.2.3	Coalminers Example	166
9.2.4	<i>Hieracium</i> Example	169
9.2.5	Discussion	171
9.3	The Multiple Logistic Regression Model	173
10	NONPARAMETRIC GEE AND VECTOR SMOOTHING	175
10.1	Introduction	175
10.2	Original Formulation	177

10.3 Extensions of Liang & Zeger (1986)	183
10.4 GEE2	186
Appendix A: Variable Site Sizes	189
Appendix B: Weighted Averaging Proofs	191
References	195

NOTE: Errata sheet enclosed at back

List of Tables

1.1	A portion of the North Island data set	5
1.2	Climate variables and their abbreviations for the North Island data set. . .	8
2.1	GAM fitted on <i>Vitex lucens</i>	34
2.2	2-dimensional Gaussian logit model with an interaction term fitted to <i>Vitex lucens</i>	35
2.3	Final GLM for <i>Vitex lucens</i> , applied to the complete North Island data set	36
2.4	Final GLM for <i>Knightia excelsa</i> , applied to the complete North Island data set	38
2.5	Stages of the stepwise regression procedure in fitting the model for <i>Agathis australis</i>	39
4.1	Fitted competition estimates	71
5.1	Simulation results for 95% confidence intervals for u	90
5.2	Optimal design points for Gaussian logit curves. Equal sample sizes at each design point is assumed	102
5.3	Optimal $D = 3$ designs for Gaussian logit curves	103
8.1	The coalminers data set	148
8.2	Fitted regression coefficients for the coalminers data in the usual bivariate probit model	150
8.3	Fitted regression coefficients for the coalminers data in the bivariate probit model where ρ is a function of the covariate age	152

9.1	Period of exposure and severity of pneumoconiosis amongst a group of coalminers	167
9.2	The <i>Hieracium</i> data set	169
10.1	\hat{R} for 16 species in the Hunua and Waitakere Ranges	182

List of Figures

2.1	Case-control sampled site locations used in the <i>Agathis australis</i> study . . .	26
2.2	Estimated presence-absence response curves for three contrasting species against annual mean temperature	28
2.3	The estimated contribution of annual mean temperature and solar radiation in the wettest quarter to <i>Vitex lucens</i>	30
2.4	Contour plots for <i>Vitex lucens</i>	32
2.5	Two convex hulls	34
2.6	Fitted function values for <i>Vitex lucens</i>	35
2.7	Fitted function values for <i>Knightia excelsa</i>	37
2.8	Fitted function values for <i>Agathis australis</i>	37
2.9	Grey-scale plot for <i>Agathis australis</i> overlaid upon the North Island	40
2.10	Case sites used in the <i>Agathis australis</i> study	42
3.1	Probability of presence for <i>Agathis australis</i> in the 1.5°C self-developed scenario	50
3.2	1.5°C self-developed scenario for <i>Agathis australis</i> : areas of probability increase and decrease	51
3.3	1.5°C scenario of Salinger & Hicks (1990)	53
3.4	Probability of presence for <i>Agathis australis</i> in the 1.5°C scenario of Salinger & Hicks (1990)	54
3.5	1.5°C scenario of Salinger & Hicks (1990) for <i>Agathis australis</i> : areas of probability increase and decrease	55

4.1	Fitted response curves for <i>Agathis australis</i> and <i>Knightia excelsa</i> against minimum monthly temperature	62
4.2	Fitted joint probabilities of <i>Agathis australis</i> and <i>Knightia excelsa</i>	64
4.3	Grey-scale plots of four joint probabilities over New Zealand	65
4.4	Areas of maximum probability	66
4.5	Effects of competition on one species with another along a gradient x . . .	69
4.6	Approximate upper and lower bounds of competition intensity between <i>Agathis australis</i> and <i>Knightia excelsa</i>	70
4.7	Inferences assuming $c = p_1/(p_1 + p_2)$	78
4.8	The 3-species competition model	81
5.1	Gaussian logit curves, where $a = \pm 1$; $t = 1$; $u = 0$	89
5.2	An estimated bimodal response curve $\hat{p}(x)$	94
5.3	95% confidence regions for <i>Knightia excelsa</i>	98
5.4	$\left\{ \frac{2}{3} x^2 p(x) q(x) \right\}^{-1}$ versus x for various p_{\max} and the asymptotic relative efficiency of the $D = 3$ design versus designs where the design points are optimally equally spaced about the origin	101
6.1	Location of the Waitakere and Hunua Ranges near Auckland, New Zealand	110
6.2	Fitted GAMs to three species from the Waitakere Ranges	111
6.3	GAMs fitted to <i>Agathis australis</i> against altitude (m) collected in the Waitakere and Hunua Ranges	114
6.4	Three species' response curves from data collected in the Waitakere Ranges	115
7.1	Equivalent kernels	127
7.2	$df_{(m)} - \text{tr}(\mathbf{S}(\lambda_m))$ for various ρ and λ_m	131
8.1	$\hat{\rho}_i$ for the coalminers data	151
8.2	Fitted functions \hat{f}_m with ± 2 standard error bands, for the coalminers data	154
8.3	Fitted probabilities for the coalminers data	155

8.4	Functions used for simulating a two variable bivariate probit model	156
8.5	Fitted functions for a two variable bivariate probit model fitted to simulated data. Each function has 5 degrees of freedom	157
8.6	Fitted functions for a two variable bivariate probit model fitted to simulated data. Each function has 3 degrees of freedom	158
8.7	Results on the <i>Agathis australis</i> - <i>Knightia excelsa</i> data	159
8.8	Fitted functions for the two variable bivariate probit model applied to the North Island data set	160
8.9	Marginal probabilities $P(Y_i = 1 \text{minimum monthly temperature, minimum monthly solar radiation})$ of the fitted two covariate nonparametric bivariate probit model	161
9.1	Fitted nonparametric proportional odds model for pneumoconiosis data . .	167
9.2	Fitted nonparametric nonproportional odds model for pneumoconiosis data	168
9.3	Fitted functions and probabilities of a nonparametric nonproportional odds model fitted to the <i>Hieracium</i> data	170
9.4	Fitted functions and probabilities of a nonparametric proportional odds model for the <i>Hieracium</i> data	172
9.5	Fitted nonparametric multiple logistic regression model for pneumoconiosis data	174
10.1	Probability of presence response curves for 16 species in the Hunua Ranges	181
10.2	Probability of presence response curves for three species in the Waitakere Ranges	183