

Department of Spatial Sciences

**Spatial Decision Support for Selecting
Tropical Crops and Forages in Uncertain Environments**

Rachel Anne O'Brien

**This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University of Technology**

June 2004

ABSTRACT

Farmers in the developing world frequently find themselves in uncertain and risky environments, often having to make decisions based on very little information. Functional models are needed to support farmers' tactical decisions. In order to develop an appropriate model, a comparison is carried out of potential modelling approaches to address the question of what to grow where. A probabilistic GIS model is identified in this research as a suitable model for this purpose. This model is implemented as the stand-alone Spatial Decision Support System (SDSS) CaNaSTA, based on trial data and expert knowledge available for Central America and forage crops. The processes and methods used address many of the problems encountered with other agricultural DSS and SDSS. CaNaSTA shows significant overlap with recommendations from other sources. In addition, CaNaSTA provides details on the likely adaptation distribution of each species at each location, as well as measures of sensitivity and certainty. The combination of data and expert knowledge in a spatial environment allows spatial and aspatial uncertainty to be explicitly modelled. This is an original approach to the problem of helping farmers decide what to plant where.

ACKNOWLEDGEMENTS

Thanks are due first and foremost to my supervisors, Dr Robert Corner (Department of Spatial Sciences, Curtin University of Technology) and Dr Simon Cook (Land Use, International Centre for Tropical Agriculture [CIAT]). Both provided excellent guidance and encouragement, even when physically on the other side of the world. I particularly appreciate Simon's big picture approach and Rob's attention to detail.

This project was made possible primarily through the efforts of Dr Michael Peters (Tropical Forages, CIAT). His constant support and enthusiasm have been invaluable.

The research was funded by the *Bundesministerium für Wirtschaftliche Zusammenarbeit und Entwicklung* (German Federal Ministry for Economic Cooperation and Development, BMZ), with support from the *Deutsche Gesellschaft für Technische Zusammenarbeit* (German Agency for Technical Cooperation, GTZ). Support was also received through a Curtin University Postgraduate Scholarship (CUPS).

Many at CIAT have been closely involved in this project. I would like to thank Luis Horacio Franco, Belisario Hincapie (both Tropical Forages) and Arturo Franco (Database) especially for assistance with RIEPT data and other forages information. William Diaz (Land Use) provided valuable technical assistance with Delphi and MapObjects LT programming. Dr Carlos Lascano (Tropical Forages), Dr Thomas Oberthür, Dr Douglas White, Dr Peter Jones, and Dr Glenn Hyman (all Land Use) all provided feedback and support at various stages of the research.

Outposted CIAT staff not only provided feedback on, and assistance with, my project, but also made me superbly welcome in various corners of the globe. Many thanks go to Dr Axel Schmidt and the CIAT team in Nicaragua, Rein van der Hoek, Heraldo Cruz and the rest of the CIAT team in Honduras, Dr Pedro Argel in Costa Rica and Dr Peter Horne and the rest of the CIAT team in Laos. Of course, I am

much indebted also to the many farmers and technicians in these countries who offered their time and knowledge.

Thanks to Dr Bruce Pengelly (CSIRO) and the rest of SoFT team for giving me access to the SoFT knowledge base, and to Dr Robert Hijmans (MVZ, UC Berkeley) for help with Delphi and MapObjects LT programming and, in particular, code for grid mapping. Thanks also go to other colleagues around the world in the fields of forages, spatial modelling and psychology of knowledge, for useful discussions and feedback.

During the course of my research I spent just over half my time based at CIAT in Cali, Colombia, and the rest of my time at Curtin University in Perth, Western Australia. Both places became home bases for me, and provided excellent working and living environments.

As well as those mentioned above, I would like to thank the entire Land Use and Forages teams at CIAT for their support and friendship, *muchas gracias*. Other colleagues at CIAT also provided support and friendship, both professionally and personally. In particular I would like to acknowledge Harriet Menter, Nadine Saad, Meike Andersson and Karen Tscherning.

At Curtin University I would like to thank all staff and research students in the Department of Spatial Sciences for providing such a warm welcome on each return (and for finding desk space for me). Thanks go to the Spatial Sciences Research Group and in particular to Lesley Arnold, Georgina Warren and Deavi Purnomo for both professional and personal support and friendship.

Thank you to those who proofread various drafts and gave valuable feedback, particularly Michael Peters, Lester O'Brien, Jennifer O'Brien, Sara Hatton, Gabriela Pracilio, Tim Chalke and Lori Patterson.

Thanks to all the wonderful friends I've met in Cali and Perth, and special thanks to my Cottesloe housemates for being understanding and supportive these last few months. And finally, a big thank you to friends and family in New Zealand, USA,

Bangkok and London, for providing accommodation and entertainment throughout my travels.

Rachel O'Brien

June 2004

A NOTE ON TERMINOLOGY

The acronym GIS refers both to Geographic Information Systems and Geographic Information Science, depending on the context.

In some places the text refers to a farmer or an expert in the singular, followed by the gender-neutral pronoun ‘they’ or possessive pronoun ‘their’. This is because the author prefers to err in number rather than in gender, and finds the constructs ‘s/he’, ‘he or she’ or alternating between ‘he’ and ‘she’ cumbersome. Although arguably the majority of farmers and forage experts are indeed male, using the singular pronoun ‘he’ would render invisible those who are not.

The term ‘data’ is used throughout the text as a singular noun, similar to the uncountable noun ‘information’. ‘Data’ is used to refer to entries in databases, usually directly observed from trials. ‘Knowledge’ is used to refer to knowledge in the heads of some kind of experts, ‘Information’ is used to denote both data and knowledge.

TERMS AND ACRONYMS

Below are some terms and acronyms used in the thesis.

Accession	Unique identification for entries in germplasm collections
ACIAR	Australian Centre for International Agricultural Research
ANN	Artificial Neural Networks
CaNaSTA	Crop Niche Selection in Tropical Agriculture
CART	Classification and Regression Trees
CI	Conditional Independence
CIAT	International Center for Tropical Agriculture (in Spanish, <i>Centro Internacional de Agricultura Tropical</i>)
CPT	Conditional Probability Table
DEM	Digital Elevation Model
DSS	Decision Support System
ES	Expert System
FAO	Food and Agriculture Organisation
Forage	Field vegetation which can be used for pasture, hay and other uses
GAM	Generalised Additive Models
Germplasm	Genetic material of a plant
GIS	Geographical Information Systems <i>or</i> Science
GLM	Generalised Linear Models
GUI	Graphic User Interface
manzana, mz	Unit of area commonly used in Central America, equivalent to 0.6988 ha
masl	Metres above sea level
NGO	Non-Government Organisation
PCA	Principal Components Analysis
RABAOC	Network for Research on Livestock Feed in West and Central Africa (in French, <i>Réseau de Recherches en Alimentation du Bétail en Afrique Occidentale et Centrale</i>)
RIEPT	International Network for the Evaluation of Tropical Pastures (in Spanish, <i>Red Internacional de Evaluación de Pastos Tropicales</i>)
SDSS	Spatial Decision Support System
SoFT	Selection of Forages for the Tropics

TABLE OF CONTENTS

	Page
ABSTRACT.....	i
ACKNOWLEDGEMENTS.....	ii
A NOTE ON TERMINOLOGY.....	v
TERMS AND ACRONYMS.....	vi
TABLE OF CONTENTS.....	vii
LIST OF FIGURES.....	xiv
LIST OF TABLES.....	xvii

Chapter	Page
1 INTRODUCTION	1
1.1 Introduction.....	1
1.2 The Decision Problem.....	2
1.3 Decision Support.....	3
1.4 Methodology	4
1.4.1 Research Methodology	4
1.4.2 Modelling the Decision.....	6
1.4.3 Developing Spatial Decision Support Systems.....	6
1.4.4 Testing and Validating Decision Support Systems.....	7
1.4.5 Delivering Decision Support Systems	7
1.5 Aims of the Thesis	8
1.6 Conceptual Model and Thesis Structure	8
2 THE DECISION PROBLEM	11
2.1 Agricultural Development	11
2.1.1 Developing Countries in the Tropics.....	11
2.1.2 Agricultural Development in Developing Countries	19
2.2 The Case Study	21
2.3 Role of Forages in Tropical Agriculture.....	24
2.3.1 Benefits of Forages	24
2.3.2 Adoption of Improved Forages.....	26
2.4 Illustrative Examples	27
2.4.1 Juan Gea López.....	28

Chapter	Page
2.4.2 Tomas Banegas Rosales.....	31
2.4.3 Unique Biophysical, Socio-economic and Management Environments	31
2.4.4 Selecting Suitable Forage Species	33
2.5 Summary	35
3 RISK, UNCERTAINTY AND KNOWLEDGE.....	36
3.1 Risk and Uncertainty.....	36
3.1.1 Definitions of Risk and Uncertainty	36
3.1.2 Spatial Uncertainty.....	38
3.2 Knowledge	40
3.2.1 Knowledge to Reduce Spatial Uncertainty	40
3.2.2 Types of Knowledge	40
3.2.3 Formalisation of Knowledge.....	41
3.2.4 Eliciting Expert Knowledge.....	42
3.2.5 Issues with Expert Knowledge.....	43
3.3 Models to Support Decision-Making.....	45
3.3.1 Steps in Decision-Making.....	47
3.3.2 Functional Modelling.....	48
3.3.3 Addressing Spatial Uncertainty	49
3.4 Summary	49
4 ADDRESSING THE DECISION PROBLEM.....	51
4.1 Risk and Uncertainty in Forage Selection	51
4.1.1 Decision-Making under Risk and Uncertainty	51
4.1.2 Risk Perception	52
4.1.3 Sources of Uncertainty.....	54
4.1.4 Reducing Uncertainty Through Knowledge	55
4.2 Modelling the Forage Selection Decision.....	56
4.3 Decision Support Systems	56
4.3.1 Types of Decision Support Systems	56
4.3.2 Decision Support Systems in Agriculture.....	57
4.3.3 Spatial Decision Support Systems	58
4.3.4 Expert Systems.....	60
4.3.5 The Case for a DSS for Selecting Forage Species.....	60

Chapter	Page
4. 4 Information Sources for a Forages DSS	61
4.4.1 Forage Databases	62
4.4.2 Expert Knowledge	62
4.4.3 Spatial Reference Data.....	64
4.4.4 Spatial Biophysical Data.....	65
4.4.5 Spatial Socio-Economic Data	67
4.4.6 Other Spatial Data.....	69
4. 5 Summary	69
5 CRITERIA FOR MODEL SELECTION	71
5. 1 Modelling The Decision Problem.....	71
5. 2 Model Selection Criteria.....	73
5. 3 Data for Modelling.....	75
5. 4 Model Validation	77
5. 5 Summary	79
6 MODELLING APPROACHES.....	80
6. 1 Logistic Regression.....	80
6.1.1 Description.....	80
6.1.2 Strengths and Weaknesses	81
6.1.3 Applicability to Tropical Forage Selection.....	82
6. 2 Generalised Linear Models and Generalised Additive Models	82
6.2.1 Description.....	82
6.2.2 Strengths and Weaknesses	83
6.2.3 Applicability to Tropical Forage Selection.....	84
6. 3 Artificial Neural Networks	84
6.3.1 Description.....	84
6.3.2 Strengths and Weaknesses	86
6.3.3 Applicability to Tropical Forage Selection.....	86
6. 4 Classification and Regression Trees	86
6.4.1 Description.....	86
6.4.2 Strengths and Weaknesses	89
6.4.3 Applicability to Tropical Forage Selection.....	90
6. 5 Environmental Envelopes	90
6.5.1 Description.....	90

Chapter	Page
6.5.2	Strengths and Weaknesses 92
6.5.3	Applicability to Tropical Forage Selection..... 92
6. 6	Fuzzy Rule-Based Methods 93
6.6.1	Description..... 93
6.6.2	Strengths and Weaknesses 95
6.6.3	Applicability to Tropical Forage Selection..... 95
6. 7	Bayesian Probability Models 96
6.7.1	Description..... 96
6.7.2	Strengths and Weaknesses 100
6.7.3	Applicability to Tropical Forage Selection..... 101
6. 8	Other Methods 101
6.8.1	Description..... 101
6. 9	Conclusions..... 102
6. 10	Summary 102
7	PROBABILISTIC GIS MODEL 104
7. 1	Fuzzy Envelopes 104
7. 2	Bayesian Models..... 110
7.2.1	Formulation..... 110
7.2.2	Calculating Posterior Probabilities Under Conditional Independence 113
7.2.3	Testing for Conditional Independence..... 116
7.2.4	Dealing with Violations of Conditional Independence..... 120
7.2.5	Causality 122
7.2.6	Uncertainty Measures 123
7.2.7	Sensitivity Analysis 124
7. 3	Modelling in a GIS Context..... 125
7. 4	Proposed Modelling Approach 126
7. 5	Summary 129
8	DATA AND KNOWLEDGE SELECTION 131
8. 1	Predictor Variables..... 131
8.1.1	RIEPT Database..... 132
8.1.2	Representativeness of RIEPT Database..... 136
8.1.3	Accuracy of RIEPT Attribute Data..... 141

Chapter	Page
8.1.4 Correlation Analysis of Potential Predictor Variables.....	143
8.1.5 Specifying Predictor Variables	145
8.1.6 Expert Knowledge	153
8.2 Response Variables.....	155
8.3 Summary of Predictor and Response Variables.....	157
8.4 Summary	158
9 SPATIAL DECISION SUPPORT SYSTEM.....	159
9.1 Overcoming Potential Issues.....	159
9.2 Conceptual Model.....	160
9.2.1 Nomenclature.....	160
9.2.2 SDSS Design.....	161
9.3 Building the Knowledge Base	162
9.3.1 Joint and Conditional Probability Distributions	162
9.3.2 Potential Uses and Filter Variable Thresholds.....	163
9.3.3 Joint and Conditional Probabilities from RIEPT.....	165
9.3.4 Prior Probability Distributions.....	166
9.3.5 Joint and Conditional Probabilities from SoFT	167
9.3.6 Transforming SoFT Categories to CaNaSTA Categories.....	169
9.3.7 Calculating Prior Probabilities for Adaptation	171
9.3.8 Calculating Prior Probabilities for Predictor Variables	175
9.3.9 Combining Data Sources	176
9.4 Model Calculations	177
9.4.1 Calculating Posterior Probability Distributions	177
9.4.2 Suitability and Sensitivity	178
9.5 Model Inputs and Outputs.....	179
9.5.1 User Inputs.....	179
9.5.2 Model Outputs	180
9.6 Summary	182
10 IMPLEMENTATION.....	184
10.1 Approach and Objectives.....	184
10.2 Software Design.....	184
10.3 Libraries	186

Chapter	Page
10.3.1 Map Routine Library.....	186
10.3.2 Grid Routine Library.....	186
10.3.3 Probability Routine Library	187
10.4 Screens	187
10.4.1 Location Selection Screen.....	187
10.4.2 Location Characteristics Screen.....	189
10.4.3 Data Updating Screen	190
10.4.4 Results Screen.....	191
10.4.5 CaNaSTA Manager Tool	192
10.5 Using CaNaSTA	193
10.5.1 Selecting Species for a Location.....	193
10.5.2 Selecting Locations for a Species	199
10.5.3 Updating Species Data.....	202
10.6 Using CaNaSTA Manager	206
10.7 Summary	207
11 RESULTS AND DISCUSSION	208
11.1 Accuracy of Model	208
11.1.1 Selecting Species for a Location.....	209
11.1.2 Selecting Locations for a Species	217
11.2 Appropriateness of Method.....	231
11.2.1 Comparison with Other Models.....	231
11.2.2 Feedback From Users	232
11.3 Achievement of Stated Aims	233
11.3.1 Ability to Work with Small Datasets	233
11.3.2 Ability to Work with Expert Knowledge.....	234
11.3.3 Ability to Predict a Range of Species Responses	234
11.3.4 Low Structural Complexity.....	234
11.3.5 Ease of Communication.....	235
11.3.6 Ability to Implement Spatially.....	235
11.3.7 Appropriateness of Agricultural SDSS	235
11.4 Summary	235
12 CONCLUSIONS.....	237
12.1 The Decision Problem.....	237

Chapter	Page
12. 2 Addressing the Decision Problem.....	237
12. 3 SDSS Development and Implementation	238
12. 4 Further Research and Development	239
12. 5 Applicability to Other Fields	241
12. 6 Lessons Learned.....	241
12. 7 Summary	242
REFERENCES	243
APPENDIX A – Forage Database Analysis	261
APPENDIX B – Farmer Surveys.....	267
APPENDIX C – Forage Expert Questionnaire	272

LIST OF FIGURES

	Page
Figure 1.1 Total meat and milk production 1963 – 2003	1
Figure 1.2 Conceptual model and structure of research	9
Figure 2.1 World bioclimatic soils regions.....	13
Figure 2.2 Bimodal rainfall patterns in Nicaragua.....	14
Figure 2.3 Monthly average rainfall and temperature for two locations in Nicaragua.	15
Figure 2.4 Percentage of problem soils in three continents.	16
Figure 2.5 Map of world hunger.	17
Figure 2.6 A family farm system.	20
Figure 2.7 GDP per capita and percentage of population on less than US\$1 per day for Central American countries compared to the developing world.	23
Figure 2.8 Matagalpa district, Nicaragua and Yoro district, Honduras.	28
Figure 2.9 Decision process for selecting a forage species	34
Figure 3.1 Decrease in uncertainty as knowledge increases.....	46
Figure 3.2 Relationship between decreasing uncertainty and increasing ability to manage risk.....	46
Figure 3.3 Interaction between expert decision-making and farmer decision-making	48
Figure 4.1 A spatial forage decision support system	61
Figure 5.1 Openshaw's (1996) model of system complexity vs. scientific precision.....	75
Figure 6.1 ANN with four input nodes, three hidden nodes in one intermediate level and two output nodes.	85
Figure 6.2 Classification and Regression Tree	87
Figure 6.3 Environmental envelope for two factors	91
Figure 6.4 Typical node in a Bayesian schema.....	96
Figure 7.1 Environmental envelope for <i>S. guianensis</i> related to soil pH and rainfall	104
Figure 7.2 Fuzzy suitability of soil pH	108
Figure 7.3 Stepped suitability of soil pH	109
Figure 7.4 Probability distribution with three states.....	111

	Page
Figure 7.5 Populating a CPT for 6 conditionally independent variables X^* with 5 states each and a dependent variable Y with 4 states.....	115
Figure 7.6 Populating a CPT for 6 independent variables X^* with 5 states each and a dependent variable Y with 4 states.....	122
Figure 7.7 Simple Bayesian network.....	127
Figure 7.8 Simple Bayesian network with map purity	128
Figure 7.9 Model implementation diagram	130
Figure 8.1 RIEPT trial locations in Central America	133
Figure 8.2 Adaptation trials for all species in RIEPT.....	134
Figure 8.3 Adaptation trials for selected species in RIEPT	135
Figure 8.4 Comparison of cumulative frequency of location elevations.....	137
Figure 8.5 Comparison of cumulative frequency of rainfall.....	138
Figure 8.6 Comparison of percentage of area or locations with soil pH in classes shown.....	139
Figure 8.7 Comparison of percentage of area or locations with soil texture in classes shown.....	140
Figure 8.8 FAO soil classification	140
Figure 8.9 Comparison of percentage of area or location with soil fertility in classes shown.....	141
Figure 8.10 Soil texture classification based on USDA triangle	148
Figure 8.11 RIEPT data in model	158
Figure 9.1 Defining conditional probability distributions	176
Figure 9.2 Defining CPT for each species	177
Figure 10.1 Overview of CaNaSTA	185
Figure 10.2 Three main modules in CaNaSTA	185
Figure 10.3 Views at six scales implemented in CaNaSTA	188
Figure 10.4 Locations characteristics screen	190
Figure 10.5 Data updating screen	191
Figure 10.6 Results screen	192
Figure 10.7 Empty map when CaNaSTA is first launched.....	193
Figure 10.8 Initial ‘Select Location’ screen.....	194
Figure 10.9 ‘Select Location’ screen with location near Luquigüe selected.....	195
Figure 10.10 Elevation classes for Luquigüe.....	196

	Page
Figure 10.11 Location characteristics	196
Figure 10.12 Ranked list of suitable species.....	197
Figure 10.13 Suitability, certainty and sensitivity values for top five species	198
Figure 10.14 Combined suitability score for top selected species.....	199
Figure 10.15 Suitability and stability details	199
Figure 10.16 Selecting locations for a species.....	200
Figure 10.17 Score for selected species in selected location.....	201
Figure 10.18 Access to market.....	201
Figure 10.19 New user prompt	202
Figure 10.20 Species and variable selection	203
Figure 10.21 Updating probabilities	203
Figure 10.22 Most likely adaptation for <i>S. guianensis</i> based on rainfall.....	204
Figure 10.23 Combining multiple variables	205
Figure 10.24 Changing value for soil pH.....	205
Figure 10.25 Different views of adaptation distributions	206
Figure 11.1 Elevation, annual rainfall and dry months for Central America and San Dionisio region.....	225
Figure 11.2 Probability of adaptation of <i>Stylosanthes guianensis</i>	226
Figure 11.3 Probability of adaptation of <i>Arachis pintoii</i>	227
Figure 11.4 Probability of adaptation of <i>Cratylia argentea</i>	228
Figure 11.5 Probability of adaptation of <i>Centrosema pubescens</i>	229
Figure 11.6 Probability of adaptation of <i>Brachiaria brizantha</i>	230

LIST OF TABLES

	Page
Table 2.1 Selected socio-economic indicators in tropical developing regions	18
Table 2.2 Selected socio-economic indicators in developed regions.	18
Table 2.3 Potential forage uses.	26
Table 2.4 Biophysical, socio-economic and management factors for some Central American farmers.....	29
Table 3.1 Classifications of uncertainty.....	38
Table 3.2 Types of spatial uncertainty.....	39
Table 4.1 Type I and II errors as possible outcomes of a decision.....	52
Table 4.2 Steps in DSS development.....	58
Table 4.3 Description of existing spatial reference data.....	65
Table 4.4 Description of existing spatial biophysical data	66
Table 4.5 Description of existing spatial socio-economic data	68
Table 4.6 Census data for Central America	68
Table 5.1 Decision problem formulations	71
Table 5.2 Model evaluation criteria	75
Table 5.3 Confusion matrix of predicted classification vs. observed classification	78
Table 7.1 Number of values required to populate the full CPT.....	121
Table 8.1 Variables relating to location included in the RIEPT database	137
Table 8.2 R^2 between RIEPT data and GIS data.....	142
Table 8.3 R^2 values for all Central America / all RIEPT locations.....	143
Table 8.4 Ecosystem classification used in RIEPT	146
Table 8.5 Holdridge lifezones.....	147
Table 8.6 Ecosystem classification based on Holdridge lifezones	148
Table 8.7 Soil texture classification based on USDA texture triangle	148
Table 8.8 Soil fertility classification based on organic matter and phosphorus	150
Table 8.9 Variable categories selected for SDSS development.....	151
Table 8.10 Joint information uncertainty for GIS data / RIEPT data (RIEPT data only for soil factors).....	152
Table 8.11 Cohen's kappa comparing RIEPT and GIS data for three variables	152

	Page
Table 8.12 Variables in SoFT database	154
Table 8.13 Trial variables in the RIEPT database	155
Table 8.14 Predictor variables and filter variables for SDSS implementation	157
Table 9.1 Notations for predictor variables	161
Table 9.2 Frequency counts / joint probability values for rainfall class against adaptation class	163
Table 9.3 Potential forage uses defined at CIAT and in SoFT	164
Table 9.4 Filter variables in SoFT	165
Table 9.5 SoFT variables related to predictor variables	168
Table 9.6 Transforming SoFT categories to predictor variable categories	169
Table 9.7 Probability distributions and certainty values for rainfall ranges in SoFT	171
Table 9.8 Transforming SoFT rainfall probability distributions to predefined classes	171
Table 9.9 Conditional probabilities $P(A x_i)$ for <i>S. guianensis</i>	174
Table 9.10 Prior probabilities derived from spatial data for predictor variables across all agricultural land in Central America	175
Table 9.11 Ranking calculation example for five species based on the score value	178
Table 10.1 Predefined scales in location selection screen in CaNaSTA	187
Table 11.1 Data and knowledge for functional model assessment	209
Table 11.2 Summary of selected locations	209
Table 11.3 Comparison of query variables for different sources	210
Table 11.4 Species suggested for farmer in Luquigüe	211
Table 11.5 Species suggested for farmer in San Dionisio-Wibuse	212
Table 11.6 Species suggested for farmer in El Corozo	213
Table 11.7 Species suggested for farmer near Flores	214
Table 11.8 Species suggested for farmer in Esparza	215
Table 11.9 Recommendations by CaNaSTA for species recommended by experts	217
Table 11.10 Records in RIEPT Adaptation for selected species	217
Table 11.11 Weighted kappa (κ_w) for each pair of experts	218

	Page
Table 11.12 Suitability of selected species for Luquigüe	218
Table 11.13 Suitability of selected species for San Dionisio-Wibuse	219
Table 11.14 Suitability of selected species for El Corozo	219
Table 11.15 Suitability of selected species for near Flores	220
Table 11.16 Suitability of selected species for Esparza.....	220
Table 11.17 Confusion matrix for expert assessment vs. CaNaSTA.....	220
Table 11.18 Weighted kappa for CaNaSTA recommendations against individual experts.....	221
Table 11.19 Weighted kappa (κ_w) for recommendations from all sources.....	221
Table 11.20 Classes for maps from different sources.....	223
Table 11.21 Joint information uncertainty for map comparisons	224

CHAPTER 1. INTRODUCTION

1.1 Introduction

Over 800 million of the world's people, including 200 million children, suffer from chronic undernutrition. While food production has more than kept pace with global population growth in recent years, agriculture faces enormous challenges to meet the food needs of a projected additional 1.7 billion people over the next 20 years (WRI, 2001). In the last 40 years, livestock production worldwide has more than doubled (FAOSTAT, 2004) (Figure 1.1). Most of this is meat production in the developing world, which has increased six-fold. Over the same time period, total meat exports from the developing world only increased four-fold (FAOSTAT, 2004), confirming that meat (and milk) consumption in the developing world is increasing.

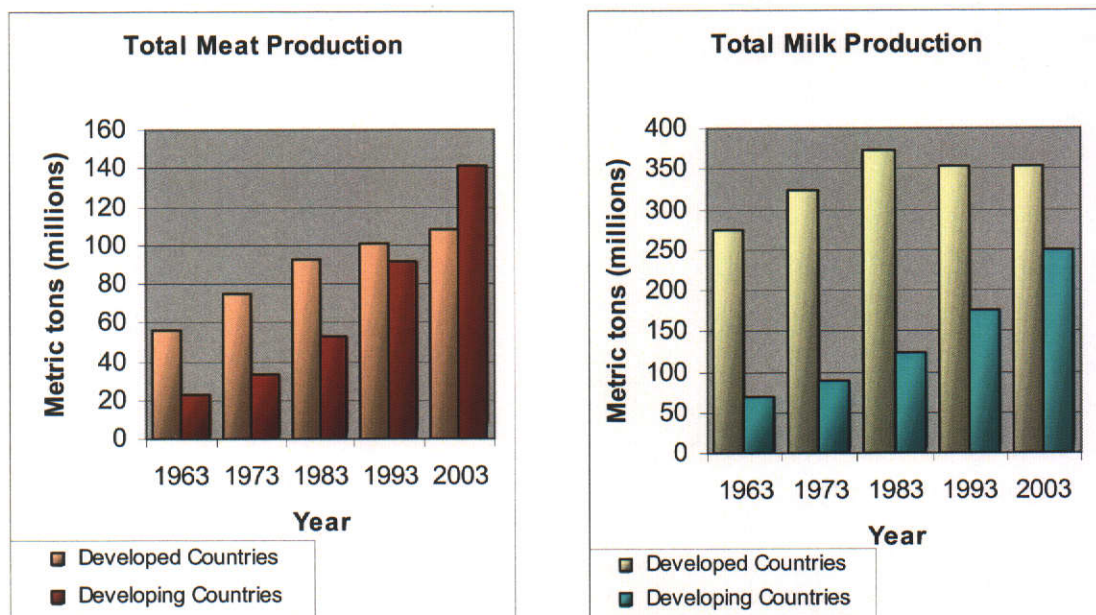


Figure 1.1 Total meat and milk production 1963 – 2003. Source: FAOSTAT, 2004.

Experience in Latin America and Asia has demonstrated the effectiveness of new forage-based technologies for intensifying meat and milk production on small farms (Peters *et al.*, 2003a). Whilst forage grasses have been widely adopted in Latin America for pasture improvement, adoption of forage legumes has been fairly limited. This is because increased livestock production during recent decades has been achieved by expanding the area used for cattle, rather than by intensifying

production. In addition, many farmers do not recognise the long-term benefits that forage legumes bring by enhancing the soil and making production more sustainable.

1.2 The Decision Problem

Although the potential benefits of forages have been demonstrated, uptake remains low. Farmers in the developing world frequently find themselves in uncertain and risky environments, often having to make decisions based on very little information. In the developing world, risks for smallholder farmers are often critical because of their poverty. In addition, in the tropics, the natural environment is spatially and temporally variable and often harsh, thereby increasing the uncertainty faced by these farmers. This thesis examines the nature of spatial decision support for decision problems in tropical agriculture. Risky and uncertain decision problems, such as those faced by smallholder farmers in the developing world, are investigated using the example of the decision of what forage species to plant where. It is argued here that farmers' decision problems can be reduced by providing information and that delivery of information can be improved through the use of computer tools and Geographical Information Systems (GIS).

The decision problem is a complex spatial decision problem. A problem is considered complex when it entails a web of related problems, covering many disciplines and interacting on various scales (van Asselt, 2000). Spatial decision problems refer to situations where location impacts on the problem in some way. Examples of complex spatial decision-making include urban planning, landscape planning, retail site selection and agricultural management. Uncertainty and risk are integral components of complex decision-making. Spatial decision support can assist in these types of decision problems by using appropriate modelling techniques to handle complexity and uncertainty.

Poor farmers are necessarily risk-averse and require confidence that a potential improvement will work in their situation. The decision of which forage species to adopt contains a measure of uncertainty. Reducing this uncertainty provides opportunities for forages to be targeted successfully. However, representing the function of forages in farming systems is difficult, because livestock systems are

complex and influenced by many factors, both biophysical and socio-economic. In addition, there are issues surrounding data availability and data uncertainty.

Methods are required to represent the essential variations of farming systems and to make available existing information on the system in a way that reduces uncertainty and assists with robust decision-making. The objectives of this research support the goal of improving forage adoption decisions for smallholder farmers in the developing world, thereby increasing sustainable intensification and ultimately contributing to increased sustainable world food production and the alleviation of undernutrition.

1.3 Decision Support

Decision support can help reduce uncertainty and improve the decision-making process. In the case of selecting forages, this can be achieved by combining diverse biophysical and socio-economic data and knowledge. A formal way of offering decision support is through a Decision Support System (DSS), providing access to data, procedures and analytical capability.

Decision support in agriculture has a chequered history. While agricultural DSS have been available to farmers for decades, the use of DSS in agriculture is declining (McCown *et al.*, 2002). Various barriers have been identified in the adoption of agricultural DSS, in particular poor understanding among researchers of how agricultural DSS are actually used by practitioners (Cox, 1996). If a DSS is designed in conjunction with targeted users, then it is more likely to have a positive impact on addressing agricultural decision problems (Cox, 1996; Stephens and Middleton, 2002; Walker, 2002).

Spatial Decision Support Systems (SDSS) work with explicitly spatial data, and outputs usually include maps. Spatial decision support for tropical agriculture can aid decision-making in a number of ways. Firstly, it can make information available to farmers and their advisors that may otherwise not be accessible. Secondly, through the use of GIS, inherent spatial uncertainties can be addressed. The use of GIS also allows for visual interpretation of results through the use of maps. Thirdly,

a well-designed SDSS will give reliable results in a consistent and timely manner, which, over time, should allow farmers and their advisors to have confidence in the results. Finally, if appropriate methods are used, farmer feedback and knowledge can be incorporated, thereby making this knowledge available to other farmers.

1.4 Methodology

1.4.1 Research Methodology

In the ‘physical’ sciences, research is traditionally assumed to be a linear process. Science is about creating new knowledge, through objective, experimental discovery of facts. Scientific methods are assumed to be objective and value-free.

Even within ‘physical’ science, this positivist view has been challenged by various philosophers, since facts can be shown to be theory-dependent and fallible, and deriving theories from facts is not always straightforward (Chalmers, 1999). Feyerabend (1975) went so far as to claim there is no such thing as a ‘scientific method’ and therefore ‘science’ is not necessarily superior to other forms of knowledge. There is debate as to whether human and social sciences can in fact be categorised as ‘science’ and whether the scientific methods of physical science can be legitimately transferred to social sciences (Chalmers, 1999).

Research into decision-making in tropical agriculture is cross-disciplinary in nature, drawing on both physical science (sometimes termed ‘hard’) and social science (sometimes termed ‘soft’). Physical sciences tend to rely more on quantitative research, while social science traditionally draws more on qualitative research. Historically, qualitative research was defined within the positivist paradigm (Denzin and Lincoln, 2000). More recently, a postpositivist perspective has emerged, arguing that reality can only be approximated. Postpositivism uses multiple methods to capture as much of reality as possible, including quantitative methods and qualitative procedures that lend themselves to structured analysis. It is argued that positivist methods are “just one way of telling stories about society”, and as such they are no better or worse than other methods (Denzin and Lincoln, 2000).

The distinction between hard and soft sciences is, in many cases, artificial. Many research questions, such as global warming, poverty and world hunger, span both hard and soft disciplines. Wilbanks (1986) states simply that, when it comes to questions such as to how to assure that people get enough to eat, such subdivisions become meaningless.

Debate on the nature of science, on the validity of the ‘scientific method’ and difficulty in clearly distinguishing physical and social sciences suggest a more holistic approach is warranted, particularly in cross-disciplinary research. Therefore, the methodologies employed in this research draw on both physical and social science research methods. Although the current research is largely quantitative, the postpositivist perspective is accepted, and therefore both quantitative and qualitative methods are incorporated in the research. These include data collection and analysis, literature review, farmer interviews, questionnaires and participatory methods. The main attribute of participatory research is shared ownership of research projects; it is usually also community based, with an orientation towards community action (Kemmis and McTaggart, 2000).

Literature was reviewed on decision support, tropical agriculture, tropical forages, farmer technology adoption, risk and uncertainty in decision-making, knowledge representation, expert knowledge, spatial habitat and classification models, DSS, SDSS, and software development. Literature reviews for each topic are included in the relevant chapters.

In the beginning of 2002, scientists, extension workers and farmers working with tropical forages were visited in Honduras, Nicaragua and Costa Rica. Informal interviews with some of these farmers are included as case studies in the research. The purpose of the trip was also partly participatory in nature, with the aim of attempting to recognize what farmers and researchers in the field required in order to support their decision-making processes.

Throughout the project, potential users of an SDSS for tropical agriculture were consulted, in particular forage experts associated with the International Center for Tropical Agriculture (in Spanish, *Centro Internacional de Agricultura Tropical*

[CIAT]) in Cali, Colombia. They were involved in all steps of the development, and continue to guide the process in a participatory manner. Although this research is perhaps not participatory in regards to farmers and extension workers, as they were not involved in defining the research, it is, however, participatory with regards to forage experts.

1.4.2 Modelling the Decision

The decision problem is illustrated by the selection of which forage species to plant, on the assumption that a strategic decision has already been made, that is, identifying the need for forage technology. The decision is informed by biophysical data, socio-economic data and expert and farmer knowledge. The modelling of the decision-making process is based on how an expert would be expected to make a decision, given information about the biophysical characteristics of the field and the purposes for adopting the forage.

Models are representations of the real world, with the purpose of describing or predicting attributes that are not directly observed. The modelling process introduces a number of uncertainties. Reducing and describing these uncertainties allows risks associated with the decision to be managed by the decision-maker. Therefore it is necessary to select a suitable model to address the decision problem.

Models relevant to the decision problem are reviewed and compared, with the aim of selecting a model that best reflects this process, whilst at the same time using available data and knowledge, dealing with uncertainty and remaining as transparent as possible. This process leads to development of a probabilistic GIS model to support decision-making in tropical agriculture. Relevant data and knowledge are then assembled, analysed and finally incorporated into the model.

1.4.3 Developing Spatial Decision Support Systems

The model is implemented as an SDSS, based on the probabilistic GIS model. The data and knowledge identified form the inputs to the SDSS, along with direct inputs

from the user. The outputs of the SDSS include maps, graphs and tables, designed to support the decision-maker in selecting suitable species to trial.

The implementation allows for sparse and uncertain data, works with expert knowledge and deals with uncertainty. The SDSS attempts to predict accurate results in a structurally uncomplicated model, providing results that are straightforward in their interpretation.

1.4.4 Testing and Validating Decision Support Systems

Testing and validation of the probabilistic GIS model is carried out using a number of validation sources, including independent data and expert knowledge. Validation of the model shows how well the SDSS works, that is, how accurate the results are. Validation is also required to test the effectiveness of the SDSS and how it contributes to addressing the decision problem. Because the SDSS developed has not yet been released in the field at the time of writing, this validation is currently limited to qualitative assessment by potential users.

1.4.5 Delivering Decision Support Systems

Effective delivery of DSS requires implementation that facilitates adoption and use of the DSS by the intended audience. In the case of an SDSS for tropical agriculture, the intended users are extension workers, Non-Government Organizations (NGOs), national research institutions, development agencies and international agricultural research institutions involved in tropical agriculture.

Guiding principles for design include ease of use, flexibility and transparency. Software was developed based on the model and data identified in the research. The software is called CaNaSTA (Crop Niche Selection in Tropical Agriculture) (*canasta* is Spanish for basket, and the tool aims to offer a basket of options to farmers, primarily in Spanish-speaking Central America) and was developed using Borland Delphi 6 (Borland Software Corporation, 2002) and ESRI MapObjects LT (ESRI, 2000).

1.5 Aims of the Thesis

The overall aim of the thesis is to investigate ways of providing decision support in uncertain and risky environments. Tropical agriculture is such an environment, particularly for smallholder beef and dairy farmers. Increasing demand for their products in the developing world requires intensification, which in turn can be achieved using forage-based technologies. However, the decision to adopt these technologies is both uncertain and risky.

Decision support can facilitate the decision process by making available relevant data and knowledge. In order to provide this decision support, an appropriate model needs to be developed, and this is a supporting aim of the thesis. Implementation of the model as a SDSS supports the aim of delivering decision support to smallholder farmers in the tropics, and is evaluated in this context.

1.6 Conceptual Model and Thesis Structure

A conceptual model of the research process and thesis structure is presented in Figure 1.2. The decision problem of what to grow where is influenced by uncertainty and risk and also available data and knowledge. These subjects in turn inform the model that is developed to address the decision problem. Subject to validation, the model is implemented as an SDSS. The SDSS is then reviewed, and feeds back into the decision problem. The entire process is illustrated using a case study.

In Chapter 2, the decision problem is introduced, namely, which actions to take in the face of uncertainty in tropical agriculture. In this chapter, the case study of tropical forages in Central America is introduced, and the role of forages in tropical agriculture is examined.

Chapter 3 examines the concepts identified as relevant to the decision-making process, namely, uncertainty, risk, data and knowledge. Tactics for incorporating these in models of decision-making are investigated.

Chapter 8 discusses the data and information available for developing the SDSS, incorporating the methods identified in the previous chapters. Variables and their categories are defined based on statistical analysis, functional equations and expert opinion.

Chapter 9 then discusses development of the SDSS, pulling together methods, data and information discussed in the research so far. It is shown how prior and conditional probability distributions can be derived from existing databases and expert knowledge, as well as calculations to provide appropriate outputs for the SDSS.

The implementation of the SDSS is continued with Chapter 10 describing the software developed. The SDSS 'CaNaSTA' recommends species for a given location and situation, and recommends locations for a given species. In addition, users can update data interactively and examine results through maps, tables and graphs. This process is described and illustrated.

Chapter 11 presents results and discussion, including model output and an analysis of whether the aims of the SDSS were achieved. Accuracy of the model is checked by comparing results from CaNaSTA with results from a number of other sources. Spatial comparisons are also made for selected species by visually inspecting maps produced from different sources.

The thesis concludes in Chapter 12 with a summary of the research, lessons learned and conclusions.

CHAPTER 2. THE DECISION PROBLEM

Tropical agriculture faces a unique set of problems defined by the biophysical and socio-economic environment of the tropics. Poverty and sometimes harsh biophysical conditions force many farmers to take risky actions in the face of uncertainty. When risk is perceived as too great, it is often avoided completely, and no action is taken. Some farmers therefore find themselves unable to act, and are not able to leave the vicious cycle of poverty.

Agricultural development is an important factor in alleviating poverty in tropical developing countries. Research in this area includes approaches to mitigating the risks and reducing the uncertainties that farmers contend with. The decision problem discussed here is what actions can and should be taken in the face of uncertainty in tropical agriculture. A case study – that of forages in Central America – is then introduced in general terms.

2.1 Agricultural Development

Agricultural development plays a significant role in the economies of developing countries. Natsios (2001) stresses the importance of agricultural development as a means to alleviating poverty in developing countries. A relatively large proportion of the population in these countries is smallholder farmers. Before examining the role of agricultural development in more depth, a general description is put forward of developing countries in the tropics.

2.1.1 Developing Countries in the Tropics

The tropics may be defined in a number of ways, based on climatic values, maps of plant geography and physiognomy, or economic and cultural criteria (Manshard, 1968). The most straightforward definition is geographic, with the tropics bounded by the Tropics of Cancer (23° 27' N) and Capricorn (23° 27' S). There are also no established definitions of a 'developing' country, however, in general it is accepted that most are within the tropics, although some developing nations do lie outside of

the tropics (e.g., some countries of the former Soviet republic). Therefore *tropical* agriculture is generally also *developing* agriculture, which means that the issues are not just biophysical in nature but also socio-economic.

Biophysically, the tropics are often characterised by mean temperature, generally limited by the 18°C isotherm of the coldest month (Köppen, 1923). The boundary between the tropics and subtropics can be established in a number of ways. Köppen (1923) places the subtropics between 20° and 40° latitude. Other classifications depend on average temperature of the coldest and warmest months, and therefore subtropical climates can be found within the Tropics of Cancer and Capricorn. Rainfall is also a distinguishing characteristic of the tropics and subtropics, with considerably different patterns from those found in temperate climates. Quantity, seasonal distribution, regional variability, temporal variability and intensity of tropical rainfall are all of profound ecological importance in the tropics and subtropics (Weischet and Caviedes, 1993). Another characteristic of the tropics is the fact that the largest variations in temperature are found with elevation – not only are elevation and temperature highly correlated, but elevation is often also a proxy for other limiting factors in tropical agriculture.

Upton (1996) classifies the tropical environment into broad climatic regions, namely, the humid tropics, the sub-humid tropics and semi-arid environments. In the humid tropics close to the equator, the dry season is short (less than five months, and usually between one and three months). In the sub-humid tropics, there is seasonal rainfall, with either one longer dry season or two short dry seasons. Semi-arid regions have long dry seasons. Therefore, there is great variation between the tropics and the subtropics particularly in terms of humidity and aridity (Figure 2.1).

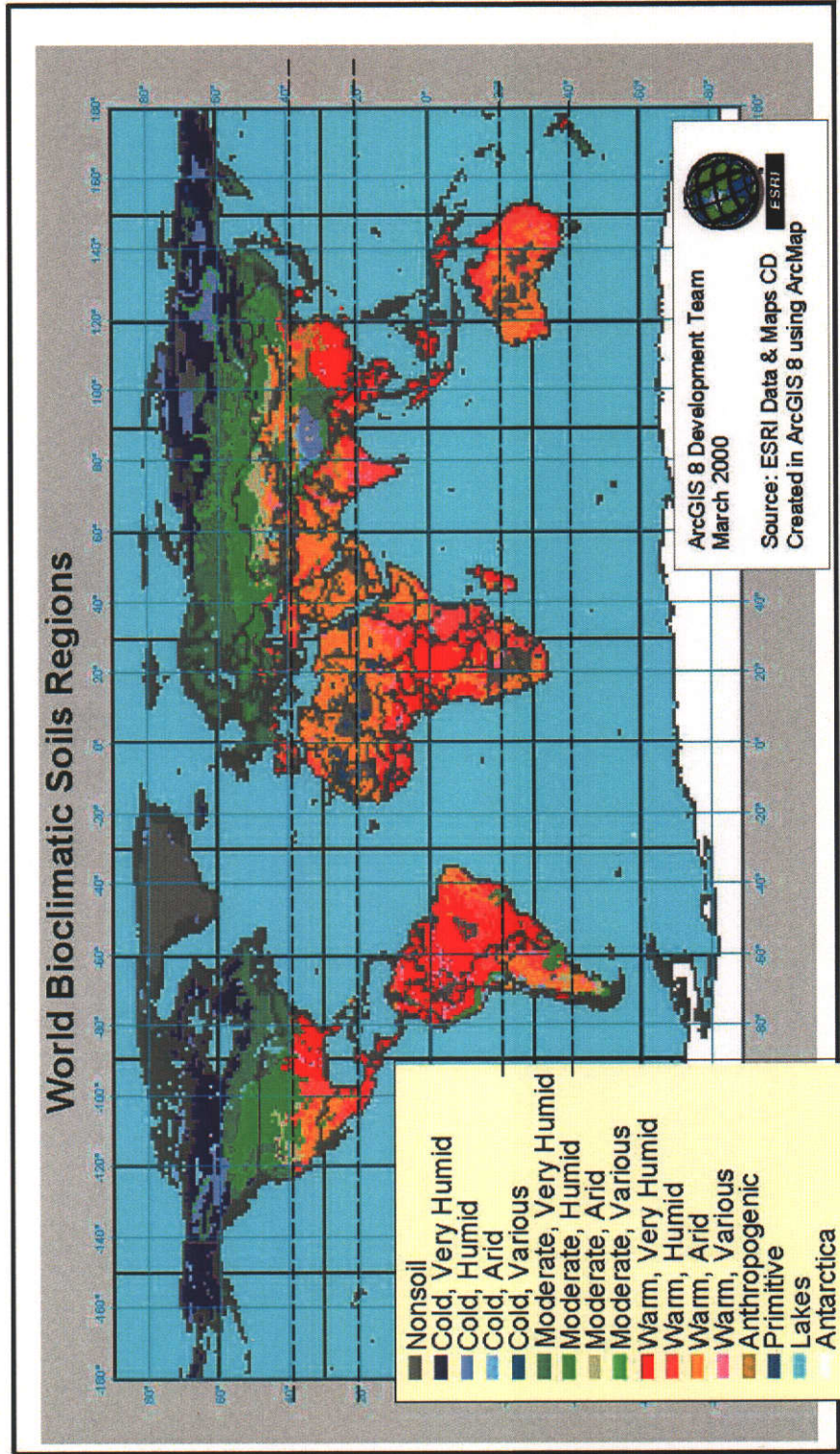


Figure 2.1 World bioclimatic soils regions. Source ESRI (1997).
 Dotted lines denote delimitations of the tropics and subtropics according to Köppen (1923).

Bimodal rainfall patterns (with two wet seasons and two dry seasons) typify some of the tropics. As an example, in Nicaragua, while there is no bimodal rainfall pattern in the central region or on the Atlantic coast, on the Pacific coast the pattern is quite defined, and sometimes severe (Figure 2.2). In Central America, the bimodal rainfall pattern is usually a long dry period from December to April and a short dry period in August, known locally as the *canícula*.

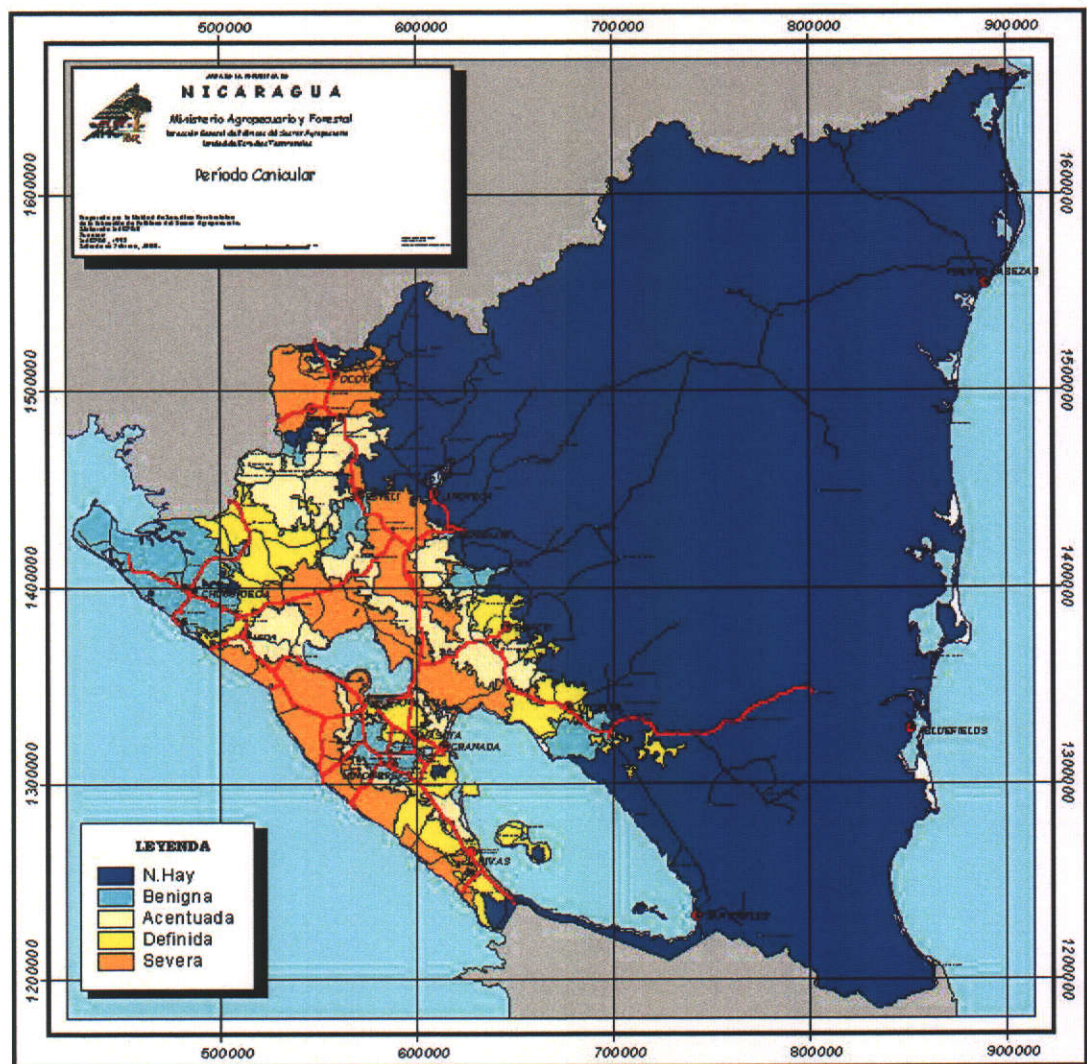


Figure 2.2 Bimodal rainfall patterns in Nicaragua. Source: *Atlas Rural de Nicaragua*, MAGFOR *et al.* (2001). Title: *Período Canicular* is equivalent to "Presence of Bimodal Rainfall Pattern". Legend: (dark blue) none, (light blue) benign, (light yellow) noticeable, (dark yellow) defined, (orange) severe (*own translation*).

This means that some locations with relatively low rainfall may in fact have multiple short dry seasons, with a quite different impact on agriculture than the same total rainfall at locations with one long dry season. Figure 2.3 shows average monthly rainfall and temperature for two locations in Nicaragua, within 150km of each other, but with markedly different rainfall patterns.

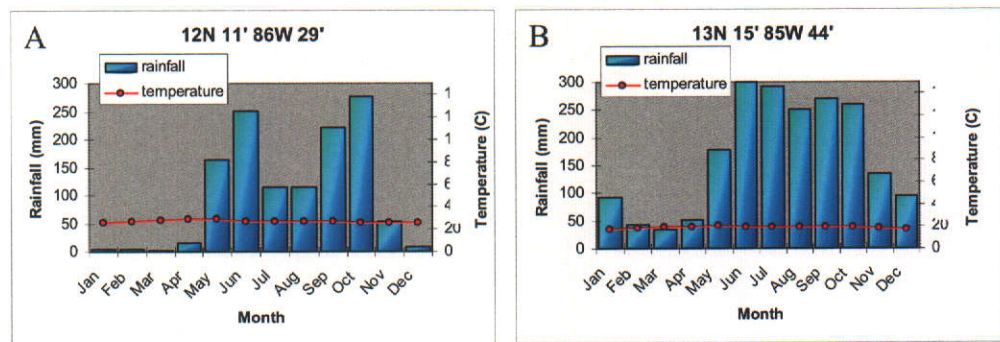


Figure 2.3 Monthly average rainfall and temperature for two locations in Nicaragua.
A. Two dry seasons (November – April and July – August). B. One dry season (December – April).

Tropical soils are also significantly different from their temperate counterparts. Well into the 1930s, it was assumed that tropical lowlands were enormously fertile with high potential for agricultural production (Weischet and Caviedes, 1993). However, in the 1960s it was discovered that soil conditions in the tropics and subtropics were, in fact, limiting factors for agricultural productivity, with many tropical soils (e.g., xanthic and orthic ferralsols) being extremely poor in nutrients. The Food and Agriculture Organisation of the United Nations (FAO) publishes some statistics on problem soils at continental level (FAO, 2004a). Low fertility acid soils are a particular problem in Latin America, whereas in Africa infertile sandy soils are the most represented problem soil. South and South East Asia has a lower percentage of problem soils, but of those problem soils, acid and low-fertility calcareous soils predominate (Figure 2.4).

Socio-economically speaking, developing countries tend to have higher levels of food insecurity. As Figure 2.5 shows, food deprivation is, on the whole, both more prevalent and deeper (meaning there is a higher degree of food deprivation) in tropical countries. Other indicators of poverty and food insecurity, such as

malnutrition, high infant mortality, low calorie consumption and insufficient grain production versus demand, all typify tropical regions (FAO *et al.*, 2004b).

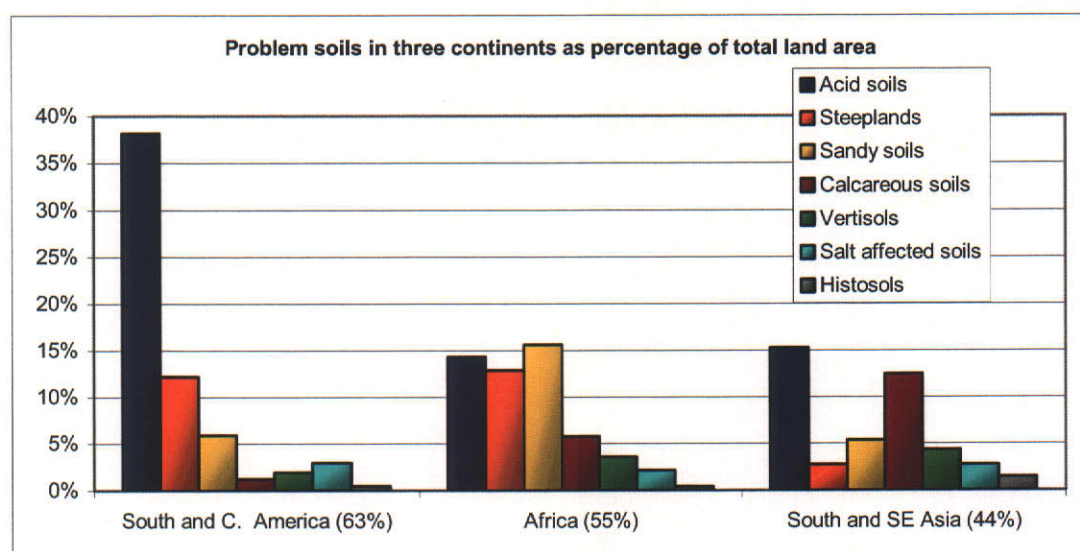


Figure 2.4 Percentage of problem soils in three continents. Acid soils include acrisols, ferralsols and podzols. Steeplands are classified as leptosols. Sandy soils are arenosols. Calcareous soils are calcisols. Salt-affected soils include solanchaks and solonetz. Source: FAO (2004a).

In Table 2.1, an analysis of all tropical developing countries shows that the population is largely rural – between 21 percent rural (South America) and 69 percent rural (Asia) – and with low GDP per capita – between US\$525 (Africa) and US\$4,980 (Central America). If Mexico's relatively high GDP (US\$6,144) is excluded, then Central America's GDP drops to US\$1,898. A relatively large proportion of the population survives on less than US\$1 per day – between 12 percent (South America) and 46 percent (Africa) (United Nations, 2004). The same figures are shown in Table 2.2 for North America, Western Europe and Australia, for comparison purposes. In addition, in both tables agricultural land is presented as percentage of total land, a statistic derived from FAO statistical databases (FAOSTAT, 2004). As Table 2.1 shows, this statistic ranges from 31 percent in South America to 64 percent in the Middle East.

World Hunger

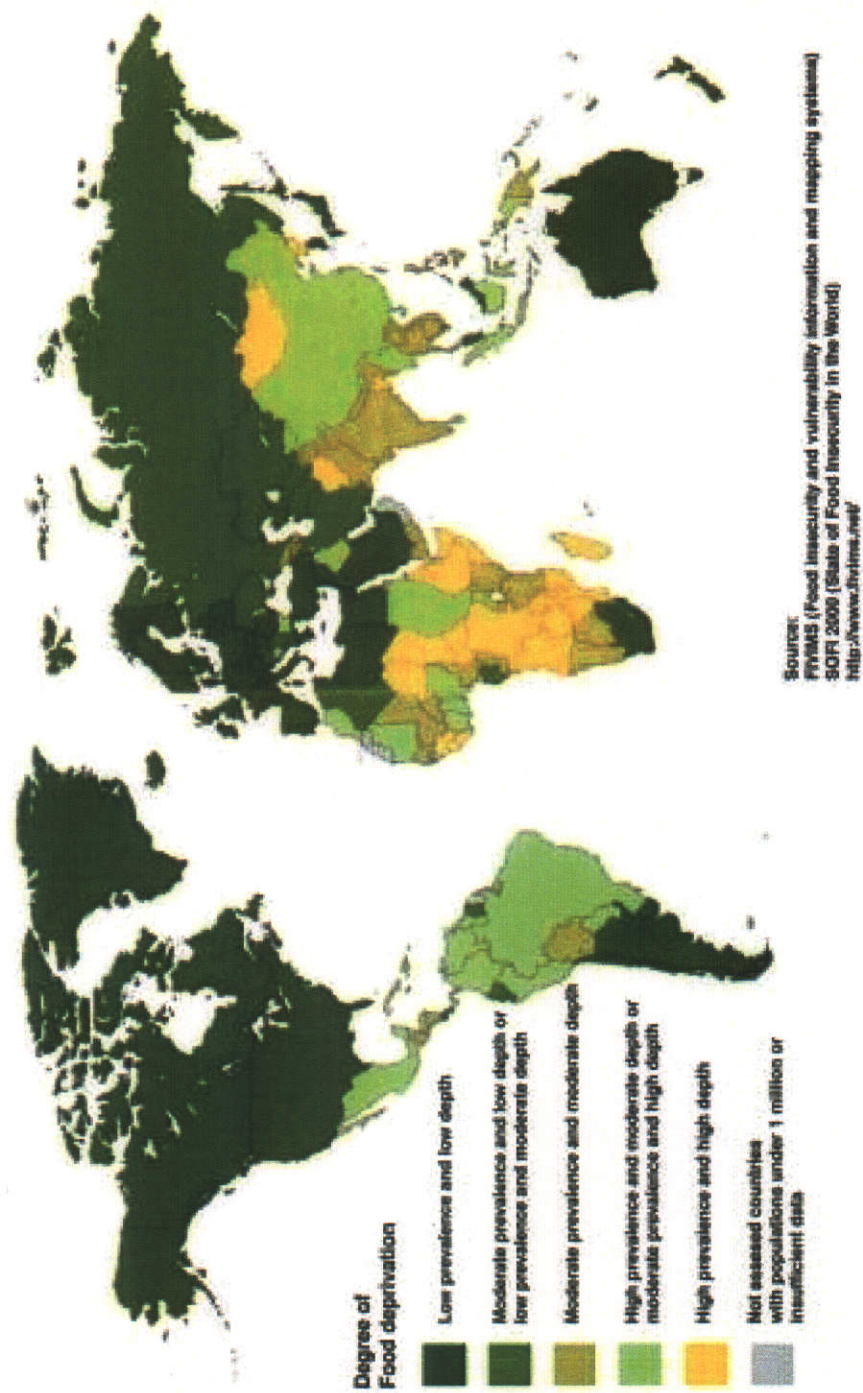


Figure 2.5 Map of world hunger. Source: SOFI (2000).

Region ^a	% Rural population ^b	GDP per capita ^b	% Pop < US\$1 per day ^c	% Agricultural population ^d	% Agricultural area ^d
Africa	65%	US\$ 525	46%	59%	34%
Asia	69%	US\$ 605	28%	52%	36%
Central America (CA)	31%	US\$ 4,980	13%	25%	52%
CA excluding Mexico	48%	US\$ 1,898	36%	32%	41%
Middle East	40%	US\$ 4,835	16%	27%	64%
South America	21%	US\$ 2,572	12%	18%	31%

Table 2.1 Selected socio-economic indicators in tropical developing regions. Source: United Nations (2004) and FAOSTAT (2004).

^a Only countries identified as developing by FAO, and that fall at least partly within the tropics.

^b United Nations 2001 figures.

^c United Nations 1989-2000 figures. Not all countries in region, only countries selected as Millennium Indicators by United Nations.

^d FAO 2002 figures.

Region	% Rural population ^b	GDP per capita ^b	% Pop < US\$1 per day	% Agricultural population ^d	% Agricultural area ^d
North America	23%	US\$33,588	No data	2%	25%
Western Europe	20%	US\$21,657		4%	42%
Australia	8%	US\$19,056		4%	59%

Table 2.2 Selected socio-economic indicators in developed regions. Source: United Nations (2004) and FAOSTAT (2004).

^{b, d} As in Table 2.1.

Although percentage of agricultural land is similarly high in Western Europe, Australia and USA (North America's figure is lower because of the vast unpopulated regions of Canada), the relatively much larger rural population in developing countries means there are many more people sharing this agricultural land. Much of the rural population is smallholder farmers and landless farm workers, as evidenced by the relatively large agricultural population in developing countries, compared to developed countries.

The majority of people in developing countries are poor (United Nations, 2004), and many are rural. Those who work in agriculture need to be self-sufficient, and often a large proportion of farm produce is destined for household consumption, with very little income from sales to local markets. Although many farmers interact with commercial markets, distance, poor road quality and lack of transport may make access to these markets problematic.

2.1.2 Agricultural Development in Developing Countries

In developing countries, farmers are often poor and farm very small areas using rudimentary techniques. Solutions that have the potential to improve farmers' situations are likely to be complex, consisting of a mix of technology, improved crop varieties, technical assistance, financial support and education.

In the past, it has been assumed that problems in tropical agriculture could be ameliorated by the adoption of technological advances from developed countries. However, as Weischet and Caviedes (1993) point out, shifting cultivation and field/fallow rotation practices are often necessary on tropical soils, and attempts to impose other agricultural techniques from developed nations have often been unsuccessful for this reason.

An example of an attempt to improve farmers' situations is the so-called Green Revolution, after high-yielding varieties of major cereal crops were developed in the 1960s. As a result, food production in many areas increased, although the anticipated global decrease in poverty did not eventuate. However, Lipton and Longhurst (1989) point out that modern plant science has mitigated what would have been increased poverty brought about by increasing population, and Upton (1996) claims the Green Revolution as a dramatic illustration of the benefits of new technology.

FAO predicts that one of the challenges for agriculture as we move further into the 21st century will be the need for precision agriculture technology in the widest sense of the phrase (Fresco, 2001). Research into site-specific development is needed, alongside biotechnology, to address yield issues. Fresco asserts that the ultimate challenge lies in two organising principles.

“First, to guarantee and facilitate access of poor countries and poor people to markets, to technology and to knowledge... Second, to maintain and enhance diversity options – diversity of products and production technology allows consumers and producers to make informed choices rather than having a blueprint approach pushed down their throats. In all this, openness about production processes and their scientific underpinnings is essential” (Fresco, 2001).

In order to alleviate poverty and improve food and income security in the developing world, sustainable production systems are necessary, balancing environmental protection with social and economic sustainability. Intensification of production may be the only solution for resource-poor farmers (Peters et al, 2001).

An overview of a farm household system, adapted from Upton (1996), is presented in Figure 2.6. In general, labour is provided by the family, although income is increasingly supplemented with off-farm work. Labour is also hired in some cases.

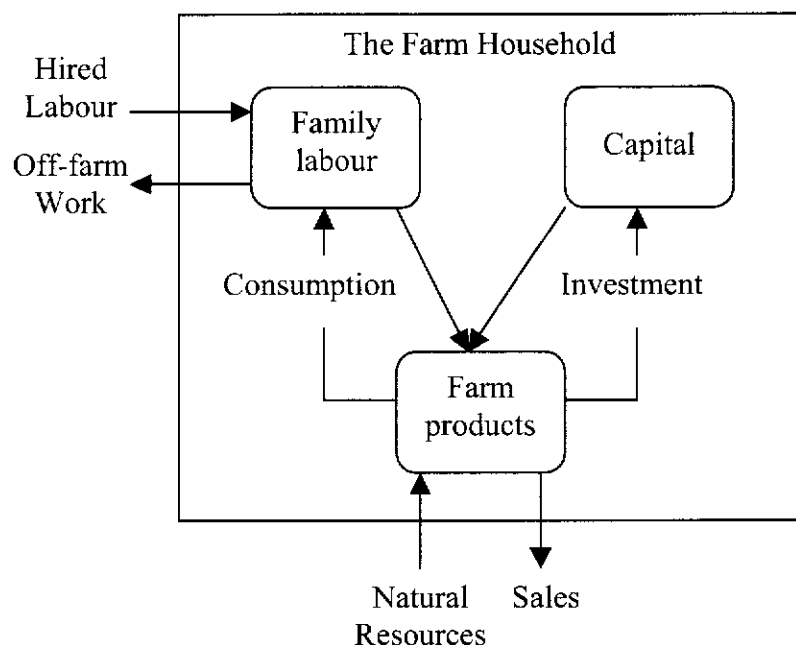


Figure 2.6 A family farm system. Adapted from Upton (1996, p 21).

Agricultural research concerning the impacts of adopting new technologies (in the sense of improved plant species and their management) on smallholder farmers, should consider them not only as farmers, but also as potential consumers and employees. Increased adoption of technology impacts on each role in different ways. Farmers are affected, for example, by changes in seed prices, management techniques, required inputs, yield, and, if the crop is sold, market prices. Consumers are affected by changes in quantity, price and quality of crops and meat. Finally, employees may be farm labourers, in which case changes in the economics of

running a farm will affect their working conditions and pay. Technology adoption means less labour is required, which will affect farm labourers, but may also mean family members are free to seek employment elsewhere (Upton, 1996).

The production environment for smallholder farmers in the tropics is characterised by uncertainty and risk. It is clear that farming systems are complex, and decisions are influenced by biophysical, social, economic and market factors. Furthermore, any decisions made will, in turn, impact on biophysical, social, economic and market factors. Lipton and Longhurst (1989) stress the need for applied agricultural research to examine the effects of adoption in specific political and demographic circumstances, rather than just concentrating on biophysical and, sometimes, economic circumstances.

Much research in agricultural development is focussed on how to reduce uncertainty and mitigate risk. Seasonality of crop production, caused by rainfall fluctuations (both within and between seasons), contributes to this uncertainty and risk. This, in turn, contributes to labour excess and shortage at different times, as well as seasonal food shortages for humans and livestock (Upton, 1996). Strategies to even out these fluctuations could reduce uncertainty in the system and allow farmers to better manage risk.

2.2 The Case Study

Central America, with a focus on Honduras and Nicaragua, has been chosen as the case study region. For the remainder of this discussion, Central America is defined as Central American countries wholly within the tropics, namely, Belize, Guatemala, El Salvador, Honduras, Nicaragua, Costa Rica and Panama. Mexico is excluded because it does not lie wholly within the tropics, and because it is significantly more developed than the rest of Central America. Biophysically, Central America encompasses a wide range of tropical and subtropical environments, representing a fair subset of the developing world (see for example Figure 2.1). Of course, not all biophysical environments in the tropics are represented in Central America, and certainly not in the same proportions. Any methodology developed concentrating on this area must be validated and adjusted for use in other environments. However, it

is also the purpose of this research to show that information that exists for a certain part of the world can, to a large extent, be extrapolated to other regions. If it is known that a crop species thrives in certain conditions in Central America, then there is no reason to believe it will perform differently in Sub-Saharan Africa or South East Asia under similar conditions, if these conditions exist. There will, of course, be other factors to take into account, such as vulnerability to disease and pests in a new environment, differing management practices and cultural preferences.

Central America is also representative of the developing world socio-economically, as was shown in Table 2.2. From Figure 2.7, it can be seen that Central America as a whole has a higher GDP than the rest of the developing world, but also a higher percentage of population living on less than US\$1 per day. Within Central America there is a wide range in both indicators, with Nicaragua and Honduras by far poorer than Panama and Costa Rica. This pattern is repeated for other socio-economic factors. In Latin America (Central America, South America and the Caribbean), there is wider variation in farm sizes and systems than in other developing countries, where most farmers are smallholders. Also, Latin America is less intensively cultivated than Africa and Asia, but because of the wider variation many smaller farmers are comparable to those in Africa and Asia (Upton, 1996). Soil degradation is also higher in Latin America. Socio-economic factors, and certainly management factors, can be expected to show a greater spatial heterogeneity throughout the developing world than biophysical factors (with the possible exception of soil, which also is highly heterogeneous [Burrough *et al.*, 1997]). Despite this spatial variability, methods can still be extrapolated to other regions, if suitable conditions can be identified.

In tropical agriculture, many strategies exist for development. One strategy is improved forages to give livestock farmers better options for animal feed. As livestock and dairy farming becomes more important in tropical agriculture, the potential for forages increases (Peters *et al.*, 2003a). It has been shown that within a given production system, livestock are particularly important for poorer households, especially to those with limited access to land (FAO, 2004b). In addition, a great proportion of the rural poor in developing countries are livestock producers (Dixon, 2003).

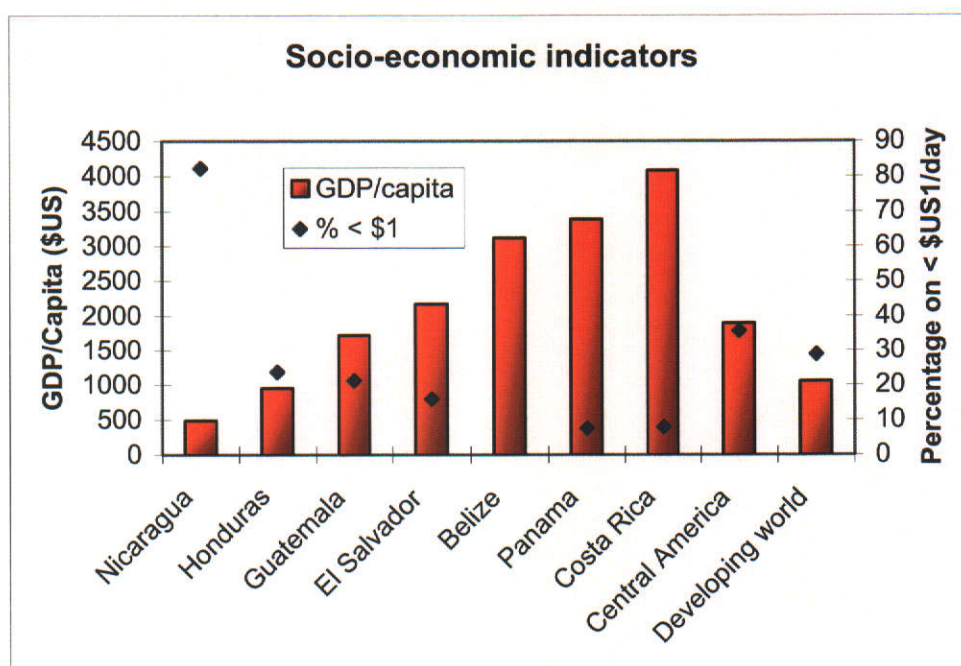


Figure 2.7 GDP per capita and percentage of population on less than US\$1 per day for Central American countries compared to the developing world. There is no figure for Belize for % < \$1. Source: United Nations (2004).

The importance of livestock farming in developing agriculture is underscored by FAO's Pro-Poor Livestock Policy Initiative (FAO, 2003). In their first steering committee report, they state that:

“livestock contribute to the livelihoods of an estimated 70% of the world's rural poor. For many of these rural poor, livestock provide a small but steady stream of food and income, help raise whole farm productivity and are often the only way of increasing assets and diversifying risks. In addition, livestock have an important role in improving the nutritional status of low-income households, confer status, are of cultural importance and create employment opportunities within and beyond the immediate household” (FAO, 2003).

Therefore, while it could be argued that human food crops are a more efficient means of addressing hunger, there are a number of reasons why livestock production is also a valid strategy. In addition, in some areas, land that is unsuitable for food crops may still be used for livestock grazing. As Upton (2004) points out, diversifying, by adding livestock to cropping systems, adds to total farm production and household income, and may also alleviate risk.

The research presented here will focus on decisions surrounding the selection of tropical forages as livestock feed, but the methodologies developed can easily be extended to decisions on selection of other tropical crops.

2.3 Role of Forages in Tropical Agriculture

Despite the potential of improved forages, adoption has been slow, of legumes in particular. One barrier to adoption is lack of information on which forage species are suited to a farmer's unique environment and why (Schultze-Kraft and Peters, 1997). As with other tropical crops, many uncertainties exist in these environments, and decision-making is often risky.

2.3.1 Benefits of Forages

Improved forages can have an important function in intensification of livestock systems. Forages are grasses, herbaceous legumes and shrub or tree legumes that are used primarily as animal feed (Horne and Stür, 1999). Peters *et al.* (2001) report that while adoption of all improved forages is low, this is particularly the case for legumes. In addition they assert that legumes have the highest potential of forages to improve smallholder farming systems. Forages play a central role in what Delgado *et al.* (1999) term the Livestock Revolution, led by increasing worldwide demand for livestock products. They contend that in the next 20 years, cattle and other livestock will play an increasingly important role in agriculture, particularly in developing countries.

Central America has 93.5 million ha of grazing land, supporting 41.4 million head of cattle (FAOSTAT, 2004). It is estimated that in tropical America, 90 percent of grazing lands are still in native pastures. The definition of smallholder farmer varies from country to country, and in Central America may be taken as meaning a farmer with around ten cows (Staal, 2003). Resource-poor livestock keepers form an extremely diverse group. Attempts to categorise them by the number of animals owned may therefore be misleading (Chipeta *et al.*, 2003), and other socio-economic factors should be taken into account. With some exceptions, a high agricultural population per km² of agricultural land indicates a large percentage of smallholder

farmers and landless farm labourers. The agricultural population per km² of agricultural land is particularly high in Guatemala (126/km²), El Salvador (122/km²) and Honduras (76/km²), with an overall figure of 59 agricultural population per km² of agricultural land in Central America. This is lower than in Asia, and many farmers do exist in Central America who are not smallholders. However, those who are often have very small landholdings and are very poor. In Central America, the percentage of rural population living below the poverty line ranges from 41 percent (Costa Rica) to 76 percent (Nicaragua) (World Development Report 2000/2001, cited in Thornton *et al.*, 2002). This translates to just under half of Central America's rural population of 43.4 million (2001 estimate (FAOSTAT, 2004)) living below the poverty line. It is reasonable to expect that the majority of smallholder livestock farmers will belong to this group.

As farmers move towards beef and dairy farming, flow-on benefits include increased purchasing power for the poor, alleviation of protein and micronutrient deficiencies in the community and increased fertiliser and draft power for the farmer. There are also risks and adverse side effects, such as increased deforestation, soil erosion, animal borne diseases and degradation of grazing areas. These benefits and risks are all direct effects of a larger number and proportion of cattle in agriculture. However, the selective introduction of forages into the agriculture system has additional benefits and can mitigate some of the adverse effects mentioned above.

The primary use of forages is as animal feed, however, they can also be used for such diverse purposes as human nutrition and natural resource management (Table 2.3). The selective introduction of forages into smallholder farming systems can help reduce soil erosion and aid sustainable intensification by regenerating degraded soils and replenishing nitrogen through N-fixation. In addition, they can help control weeds, allow the farmer to be less dependent on external inputs and are suited to diverse production systems (Humphreys, 1994; Schultze-Kraft and Peters, 1997). Furthermore, intensification may prevent deforestation in areas where farmers may otherwise have been forced to clear trees to access more fertile land. Common forage systems include cut and carry plots, grazed plots, living fences, hedgerows, improved fallows, cover crops in annual crops and cover crops under trees, ground

covers for erosion control, legume supplementation for the dry season and legume leaf meal (Horne and Stür, 1999).

Therefore the adoption of improved forages can improve a farming system in a number of ways. Forages introduce diversity into the system, reducing the risk of the impacts of crop failure. High nutritional value can increase the return on investment for animal products. The additional benefits, such as N-fixation of many legumes and weed control, can free up resources otherwise used for purchasing fertiliser and herbicides and reduce labour input. The adoption of improved forages can increase the farmer's ability to maintain livestock, and in many cases livestock are the only means of capital accumulation available to farmers.

Uses	Examples
Human nutrition	Root crops, pulses, fruits and leaves, shoots, pods as vegetables
Animal nutrition	Pasture, protein banks, hay
Technical uses	Wood, fuel, paper, luxury timbers, gum/resin, fibre, dye, tannery
Improvement and maintenance of soil productivity	Mulch, green manure and planted fallows, soil cover, erosion control, soil stabilisation, windbreaks, weed control
Other uses	Folk medicine, fish poison, rat poison, shade trees, living fences, ornamentals

Table 2.3 Potential forage uses. Source: Schultze-Kraft and Peters (1997).

2.3.2 Adoption of Improved Forages

In the tropics, and in particular in Central America, the adoption of improved forages, especially legumes, has been low (Peters et al, 2001). Stür et al (2002) and Peters et al (2001) identify some reasons for limited and slow adoption of forages. These include unfamiliarity with the technology, the fact that longer-term benefits may not be immediately obvious, unavailability of planting material, unfavourable policies and lack of participation by farmers in research and development. Adoption may also be limited by restrictions imposed by climate, soil and other biophysical factors. In addition, cultural traditions and preferences may also play a role. Farmers may be risk-averse for a number of reasons, and uncertainty surrounding the likelihood of a new forage species being successful will also be a factor. As both

biophysical and socio-economic factors can be heterogeneous, even at very fine scales, these uncertainties are inevitably site-specific.

Often farmers and extension workers are aware of constraints and opportunities in their production system, but have limited access to information on potential solutions. This is particularly true as the poverty level increases, and in environments far away from institutional support. In such a situation, adoption of a new technology may be facilitated when an expert is called upon to recommend a forage species. In a typical case, once a potential need for improved forages has been identified, a farmer or their adviser will contact an extension worker, a seed provider or an institution knowledgeable in forage technology, with a request for a forage recommendation. The expert will attempt to identify constraints and problems in the whole farming system to see where forages fit in the production system. For example, issues with production levels, erosion or degradation may all call for a different approach. Once the need is identified, the niche, or site-specific environment, is then considered and described in both biophysical and socio-economic terms. For example, the niche may be constrained by climate and soils and also by the current cropping system, farmer resources and management practices. Suitable forage options can then be suggested, matched to the farmer's unique niche. The farmer may then choose to trial a number of forages, before making an adoption decision. Experts able to make these recommendations may, however, be scarce and hence not be easily accessible, particularly in remote rural areas. Adoption may also occur when seed becomes available or is promoted and when neighbours adopt.

2.4 Illustrative Examples

As part of this research, a field trip was undertaken to Honduras, Nicaragua and Costa Rica in early 2002. During the field trip, a number of farmers were visited and some completed a short survey (see Appendix B) which, in conjunction with informal discussions, form the basis of the illustrative examples described below.

As previously mentioned, the type of farmers and size of landholdings varies greatly throughout Central America. Biophysical, socio-economic and management characteristics differ from farmer to farmer. Table 2.4 shows these characteristics for

six farmers in Honduras and Nicaragua. Even though some of these farmers are close to each other in both physical and biophysical space, their specific situation, or niche, can still vary greatly.

2.4.1 Juan Gea López

To illustrate the decision process involved in the selection of forages, two typical smallholder farmers in Central America are described. The first, Juan Gea López (Farmer 5 in Table 2.4), lives near Esquipulas in the central region of Nicaragua (Figure 2.8).

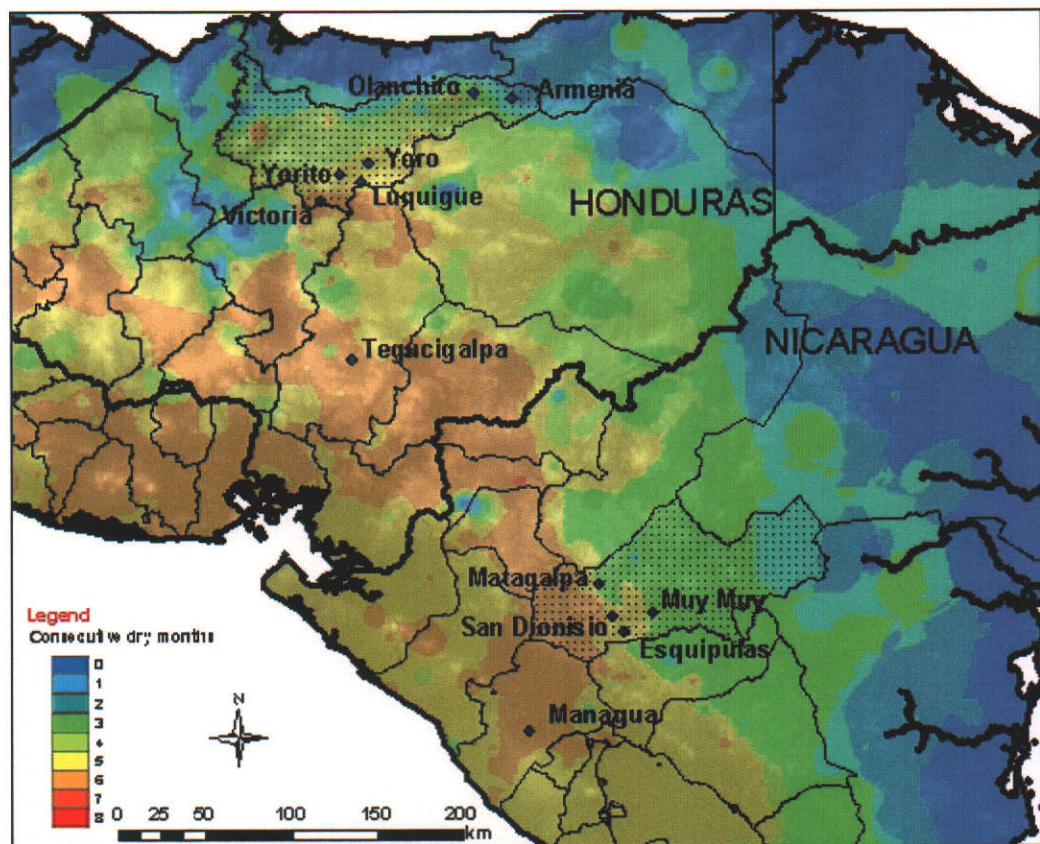


Figure 2.8 Matagalpa district, Nicaragua and Yoro district, Honduras (stippled areas). Background of map is length of dry season.

	Farmer 1	Farmer 2	Farmer 3	Farmer 4	Farmer 5	Farmer 6
Location	Luquigüe, Yoro, Honduras 15°2'N 87°10'W	Luquigüe, Yoro, Honduras 15°2'N 87°10'W	Near Flores, Atlántida, Honduras 15°36'N 87°15'W	Susuli, Matagalpa, Nicaragua 12°49'N, 85°51'W	Esquipulas, Matagalpa, Nicaragua 12°40'N, 85°47'W	Muy Muy, Matagalpa, Nicaragua 12°46'N, 85°38'W
Elevation	1514 masl	1514 masl	70 masl	554 masl	425 masl	313 masl
Annual rainfall	1230mm	1230mm	2608mm	1387mm	1206mm	1533mm
Dry season	5 months	5 months	None	3 months	5 months	4 months
Soil	Neutral, clay loam, low fertility	Moderately acid, clay loam, low fertility	Acid, loam, high fertility	Moderately acid, loam, high fertility	Moderately acid, loam, high fertility	Moderately acid, clay, high fertility
Crops	2 mz [†] coffee 2 mz maize	5.5 mz maize (first planting) 2 mz beans (second planting, personal consumption) 1 mz coffee	None	2 mz maize (personal consumption) 5 mz beans (first planting) 3 mz sorghum (second planting) 2 mz beans (second planting, personal consumption)		
Forages	8 mz pasture (trees, <i>Hypparrhenia rufa</i> and native pasture)	14 mz native pasture and pine 0.5 mz improved forages	40 mz pasture, 80% in improved forages	25 mz pasture 1 ½ mz improved forages	84 mz pasture (10 mz improved forages)	100 mz pasture (5 ½ mz improved forages)
Animals	3 horses, 1 bull, 3 pigs, 35 hens	15 cattle, 6 horses, 20 hens	230 cows	22 cattle 2 donkeys	172 cattle 3 horses	57 cattle
Family labour	4 adult children				2 permanent farm hands, temporal staff equivalent to 5 permanent positions	
Dependents	5	4				

Table 2.4 Biophysical, socio-economic and management factors for some Central American farmers.
[†] 1 mz = manzana, a unit of measure commonly used in Central America. 1 mz = 0.6988 ha.

Twenty years ago he owned one cow and seven ha of native pasture. Over time he sold his old truck, bought more land and built a house. Now he owns 59ha and 172 head of cattle (of which about 50 are descendants of his original cow), and a new four wheel drive. He milks 24 cows, producing on average 8.7l/cow/day, in a region where the average is as low as 2.4l/cow/day. He has an artificial insemination programme, sells a number of cattle each year and provides the equivalent of five permanent farmhand positions. He has put his three children through university. He is considering moving to two milkings a day, but at this stage cannot justify the cost of mechanised milking, irrigation or electric fences.

His farm is situated in a part of the country where the dry season lasts six or seven months, and Hurricane Mitch destroyed a couple of hectares of his land in 1998. Gea López has created a thriving business (where other, more traditional, farmers have failed) through good management, careful research and by chasing opportunities. During the 1980s, when monetary assistance was by preference given to cooperatives, he struggled as an independent farmer to gain access to loans and technical assistance. Slowly but surely, he increased his land and his stock, experimenting along the way with cattle breeds and fodder crops. Although he can no longer be classed as a smallholder farmer, he certainly started out as one, and improved forages played a part in his development. He planted the grasses Taiwan (*Pennisetum purpureum*) and King Grass (*Pennisetum purpureum* x *Pennisetum typhoides* hybrid) and sugar cane (*Saccharum officinarum*), but decided the work required to process these for cattle fodder could not be justified. In 1993 he was offered *Brachiaria brizantha* seeds, an improved forage pasture introduced into Central America by CIAT. He planted just a little in places where other pastures were failing, gradually increasing to seven ha of *Brachiaria brizantha* mixed with *Arachis pintoi*, a wild peanut providing quality protein, also introduced by CIAT. The cattle graze the pastures in the wet season and are given hay cut from the pastures in the dry season, supplemented with his own concentrate of chicken manure, *Mucuna pruriens* (Velvet Bean) and mineral salt mixtures.

Although Gea López's success can mostly be attributed to good management practices and a willingness to experiment, the introduction of improved forages into his pastures played an important role in increasing both milk production and milk

quality. He is planning on planting a larger area in improved forages, and would like to know if there are other species able to be used for both grazing and hay that are better adapted to the climate and soil conditions of his farm. Without ready access to expert opinion, he is unlikely to find out about potentially excellent options for his situation that could greatly increase his return on investment.

2.4.2 Tomas Banegas Rosales

Tomas Banegas Rosales (Farmer 1 in Table 2.4) also lives in a region of Central America with a long dry season. His 8.5 ha of pasture and forest, maize and coffee lie in Luquigüe, in the Yoro district of Honduras (Figure 2.8). He had four cows, but had to sell them when the coffee price dropped in 2001, to pay for his children's primary education. Four of his adult children work on the farm, and five of his children are students. He has three horses, three pigs and 35 hens, which provide ten eggs a day on average. He also still has a bull. Banegas' paddocks consist of trees, *Hyparrhenia rufa* (a pasture species naturalised in much of Latin America and locally known as *jaragua*) and native pasture. His grazing land is not high quality, but as he has no cows his resources are focused on production of maize and coffee.

He would like to buy more cows to provide milk for household consumption. With cows to feed, investment in improved pasture or legumes could pay off through higher yields and less input requirements. Better quality milk production is not a requirement in this case, because the milk is strictly for domestic consumption. However, intensifying with improved forages could eventually allow Banegas to purchase enough cows to also sell some milk. On the other hand, Banegas is risk-averse, and he cannot afford to invest in a forage species, unless he knows for certain the investment will pay off. The uncertainties in this case are related to the success of selected forages in his farm's environment, the monetary and labour costs in planting them, the benefits received from feeding them to his current livestock.

2.4.3 Unique Biophysical, Socio-economic and Management Environments

Gea López and Banegas's situations are vastly different, yet they share the challenge of making a living from the land in the dry region of Central America. Within the

region, some farmers have twice-daily mechanised milking, irrigation, a large area of improved forages and silage and hay production for the dry season. Others milk two cows by hand for only part of the year, producing a couple of litres of milk for household consumption. Some drive their cattle large distances to reach higher, slightly greener pastures in the dry season; others build a second milking shed to save the cows a walk of 1km. One farmer near Victoria in Yoro, Honduras, trained as a veterinarian, has a lot of experience and good land near a river. Yet he receives a low price for his milk, therefore lacking incentive and means to make improvements. In a pattern common to the region, this means he continues to milk by hand and graze his cattle on native pastures, which in turn means that milk quality and quantity remain the same. Therefore his income remains unchanged, even though a small investment in improved forages, matched to his unique situation, could improve his livelihood in a number of ways. He is intelligent and knowledgeable, but there are a number of barriers, including lack of confidence that investment in improved forages will pay off.

This lack of confidence is often warranted in Central America, where improved forage species may be offered based on seed availability or on latest releases, often disregarding climatic and other pertinent factors. For example, in Armenia and La Ceiba in the humid Atlantic region of Honduras, pastures such as *Brachiaria brizantha*, which are purported to be widely adapted throughout Central America, are considered poor quality in these locations because of problems with waterlogging. In reality the problem is that *B. brizantha* has low tolerance to poor drainage and hence is poorly adapted to this environment.

Just as biophysical conditions vary within Central America, so do socio-economic factors and management factors. Poor farmers with small landholdings are less likely to be able to take on risk than less poor farmers with slightly more land. Farmers with only cattle are more likely to see benefits from adopting improved forages, whilst farmers with a mixture of food crops and pasture are more likely to make investments in their food crops. Farmers with good market access are more likely to benefit from increase in milk quality (higher protein content due to improved forages), as long as this is rewarded by the purchaser. For cheese-makers, protein and fat content are especially important. Smallholder farmers with poor

market access are more likely to produce milk for household consumption, therefore low production cost may be more important.

Furthermore, different farmers prefer different management techniques. Some have sufficient land in pasture to allow grazing and to produce hay for dry season feed. Others prefer to plant smaller areas intensively in cut and carry forages, although this may not be an option if labour is unavailable. Infertile soils may benefit from legumes providing N-fixation, reducing the need for fertilisers. Hillsides may benefit from legumes that help prevent soil erosion. These and the many other options for forage use will also determine whether a farmer deems a forage crop a good investment. Knowing which forages are suitable for which management purposes is obviously valuable information for farmers selecting improved forages.

2.4.4 Selecting Suitable Forage Species

A large number of improved tropical forage species exist that are well suited to Central America. CIAT and its partners have tested these species, evaluating adaptability, establishment and production in various locations throughout Latin America. One such location is at Las Minas, a few kilometres from Luquigüe, Honduras. If Banegas decides he wants to try a forage species, he could visit this site and see for himself how well these forages grow under conditions in the immediate vicinity of his farm. He can evaluate the different possible uses of grasses, legumes and shrubs, and he can receive technical advice on sowing and management practices. Of course, farmers can only trial species if they have access to seed, which is not always the case.

The closest trial site to Gea López is about 20 km away in San Dionisio. Over this distance there is already variation in elevation, soils and rainfall patterns. Other farmers live a great deal further from forage trial sites, making it difficult for them to benefit from this research. Although CIAT has compiled a comprehensive database with data from the trial sites (Barco *et al.*, 2002), it is not an easy task to translate the results to other locations.

An approach to both generalising results to other locations and making this information available is the development of a Decision Support System (DSS). The process of selecting forage species consists of a series of decisions (Figure 2.9), firstly deciding whether change is needed, if so, what type of change, and, if that type of change is to plant a forage species, then which one.

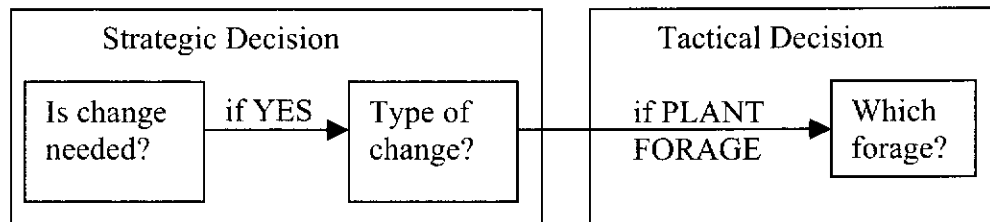


Figure 2.9 Decision process for selecting a forage species.

At the strategic level, the decision on whether or not to plant an improved forage species depends on whether or not a need for change is identified, and if so, what type of change. This could be a number of options, one of which might be the decision to plant an improved forage. This decision will depend on a number of management and socio-economic factors, including current crops, current management practices, number and type of animals, land availability, experience, education, access to extension and other socio-economic factors, as well as biophysical factors. Staal *et al.* (2002) found, in the decision of Napier grass uptake in Kenya, that level of education, access to formal milk outlets, rainfall, temperature and market access were particularly significant.

At the tactical level, the decision is which improved forage species to select, and assumes the strategic decision has already been made. This decision depends mostly on biophysical factors and management practices. However, some socio-economic factors may still be important at this level of decision-making, in particular farmers' level of risk-aversion.

In considering the development of a decision support system to aid the selection of forages, it is assumed that an improved forage species is required and the problem is to select which one(s), i.e. the decision at the tactical level. A DSS can assist the

decision-making process by making information available to farmers (increasing familiarity and describing long-term benefits) and by matching forage species to unique biophysical and socio-economic niches. Decision support systems will be discussed in more detail in subsequent chapters.

2.5 Summary

Agricultural development is an important means of poverty alleviation in tropical developing countries. In tropical agriculture, many strategies exist for development, and one strategy is the introduction of improved forages.

Improved forages are often a suitable option for smallholder farmers seeking to sustainably improve livelihoods. However, for a number of reasons, forage adoption is low, particularly in the case of legumes. One of these reasons is uncertainty on the part of the farmer about how these forages will perform in specific environments. Providing information on forages and their suitability to particular biophysical and socio-economic niches can equip farmers with the ability to make better-informed decisions.

The decision problem revolves around what information to provide and how to provide it in order to best assist smallholder farmers. In particular, issues arise around inherent uncertainty in many facets of the adoption process. It is important to acknowledge and work with risk and uncertainty in developing solutions to the decision problem. In addition, knowledge about the domain, and how this can be incorporated in the solution, is important. These topics are examined in the following chapter.

CHAPTER 3. RISK, UNCERTAINTY AND KNOWLEDGE

The previous chapter introduced the decision problem, namely the problem of supporting selected decisions made by smallholder farmers. It was shown that there are many risks inherent in this process, specifically related to uncertainties about the unique biophysical and socio-economic environment of farmers, as well as their management practices. The role of information, and in particular knowledge, was highlighted as a potential means of managing risk through reducing uncertainty.

3.1 Risk and Uncertainty

3.1.1 Definitions of Risk and Uncertainty

Risk and uncertainty are integral issues in agricultural decision-making. They are distinct concepts, but they are related. Risk relates to the utility of the outcome of a decision of uncertain outcome. In the case of agriculture, utility for the farmer may be a function of multiple objectives, such as ability to provide food for household consumption, profit, and providing employment for family members (Upton, 1996). Hardaker *et al.* (1997) define uncertainty as imperfect knowledge and risk as uncertain consequences.

Some researchers use the terms ‘uncertainty’ and ‘risk’ interchangeably (e.g., Pannell *et al.*, 2000). Norman and Shimer (1994), discussing complexity theory, claim that some theorists insist there is no valid distinction between risk and uncertainty. Van Asselt (2000) sees risk and uncertainty as two sides of the same coin, both being a consequence of limited predictability as a result of complexity. In this discussion, it is recognised that while they are interrelated, they should be understood as two separate concepts.

Risk is usually defined as a function of two factors: the utility of an event and its likelihood. Low risk could imply small probability of an event with negative utility or large probability of an event with positive utility. Conversely, a high risk implies large probability of an event with negative outcome (high negative utility). The

complete definition of risk includes all factors that define the likelihood of the event and its utility to the decision-maker. Risk cannot always be defined objectively, because in the context of decision-making, subjective (sometimes inaccurate) risk estimates, as perceived by the decision-maker, may play a much greater role than objective risk.

Uncertainty in the context of this discussion will be examined for its influence on the decision-making process. If all conditions are certain, decision-making becomes an optimisation problem – the algorithm can be complex, but the best solution is clearly demonstrable. Since practical problems are often unbounded, decision-makers will expand definition of the problem to include uncertain factors.

Uncertainty arises from ignorance or variability in the decision-making process (Ferson and Ginzburg, 1996), where ignorance describes uncertainty caused by factors that are not considered in a decision, and variability describes the uncertainty that is caused by factors of undefined degree that are known to exist.

Rowe (1994) further classifies uncertainty and variability into four main classes, namely, metrical, temporal, structural and translational (Table 3.1). Metrical uncertainty relates to uncertainty and variability in measurement. Temporal uncertainty stems from uncertainty in future and past states. Structural uncertainty involves complexity, including model structure, and translational uncertainty stems from explaining uncertain results (Rowe, 1994). In this classification, risk is related specifically to temporal uncertainty in the future. All four types stem from both variability and ignorance, however, metrical and temporal uncertainties are, arguably, more likely to have their sources in variability, and structural and translational uncertainties are more likely to stem from ignorance.


Source	Type	Description
Variability  Ignorance	Metrical	Variability in measurement
	Temporal	Uncertainty in future states Uncertainty in past states
	Structural	Uncertainty in model structure due to complexity
	Translational	Uncertainty in explaining uncertain results

Table 3.1 Classifications of uncertainty, adapted from Rowe (1994).

3.1.2 Spatial Uncertainty

The decision problem in agriculture is inherently spatial, since crops and forages are produced within a spatially variable environment. The environment displays heterogeneity at different scales, for example, soils tend to vary at a finer scale than climate. Farmers are interested in species that will thrive in their particular location. Hence much of the uncertainty surrounding the selection of species is also spatial (and in fact spatio-temporal) in nature.

Spatial information in the real world is complex, and in order to use it to facilitate decision-making, it must be simplified in some way. This simplification can consist of abstraction and discretisation, both spatially and in other dimensions. For example, consider the problem of representing ecosystems. In the first place, the classifications (such as ‘humid tropics’) are an abstraction of climatic factors, grouped with potentially arbitrary cut-off points. Secondly, climate data is usually summed and averaged over time, for example, average monthly rainfall. Finally, when representing ecosystems spatially, the space itself is discretised into polygons or rasters, each assigned an ecosystem classification. Spatial representations of data are therefore models of reality, and uncertainty can be introduced during the modelling process.

This simplification necessarily introduces an element of uncertainty. A body of literature exists on spatial uncertainty and how to deal with it (see for example Agumya and Hunter, 2002; Crosetto and Tarantola, 2001; Davis and Keller, 1997; Heuvelink, 1998; Zhang and Goodchild, 2002).

Error is a special case of spatial uncertainty, although some researchers (e.g. Crosetto and Tarantola, 2001) claim they are equivalent. The term ‘error’ implies that true values are definable and obtainable by removing inaccuracy and imprecision (Zhang and Goodchild, 2002). This is only the case, however, with metrical spatial uncertainty. Most spatial uncertainty is likely to be structural, introduced through abstraction and discretisation. Uncertainty may result from lack of information and also from vagueness, randomness, heterogeneity and spatial dependence inherent in much geographical information (Zhang and Goodchild, 2002).

Spatial uncertainty can arise from a variety of sources (Table 3.2). Metrical uncertainty can be caused by errors in positional accuracy during data acquisition. Natural variation also impacts on metrical uncertainty, depending on data collection techniques. Spatial data is normally processed and reduced (e.g., through classification), introducing structural uncertainty. Temporal uncertainty is introduced by dynamic change within biophysical and socio-economic dimensions. Future temporal uncertainty can be exacerbated because of nature’s ability to change over time (Davis and Keller, 1997). Elith *et al.* (2002) also define linguistic uncertainty, including vagueness, ambiguity, underspecificity and compounded uncertainty.

Source	Type	Example
Positional inaccuracy	Metrical	Rounding of latitude and longitude
Data acquisition	Metrical	Measurement accuracy of satellites
Natural variation	Metrical	Temperature ranges at a given location
Geoprocessing	Structural	Estimations of rainfall at locations with no data collection
Classification	Structural	Soil type classification
Currency	Temporal	Infrequent census collection Climate change

Table 3.2 Types of spatial uncertainty. Sources: Zhang and Goodchild (2002); Davis and Keller (1997).

Agumya and Hunter (2002) explore uncertainty in geographic data in light of how this impacts on risk assessment. They recognise that the presence of uncertainty in spatial data may increase the risks associated with using this data. Therefore, it is not sufficient to simply describe and visualise uncertainty, and it is crucial to examine the impacts of spatial uncertainty in a given decision-making task. The

classifications of Rowe (1994) and of Ferson and Ginzburg (1996) provide a useful framework for this.

3.2 Knowledge

3.2.1 Knowledge to Reduce Spatial Uncertainty

Decision-makers aim to reduce risk and improve the likelihood of return on investment. They do so by removing the uncertainty associated with specific decisions. Using the ignorance/variability classification of Ferson and Ginzburg (1996), uncertainty that is introduced by ignorance may be reduced by increasing knowledge, and uncertainty introduced by variation may be reduced through spatial information describing spatial variability.

Spatial information can be classed as data, as in numerical databases, maps and images, or as knowledge. Knowledge implies information that is meaningful in a specific context, and it can be derived from formalised data, or from less tangible sources, such as personal experience, ideas or impressions. Here, the discussion focuses on knowledge and how it can be formalised and applied to the decision problem.

3.2.2 Types of Knowledge

Expert knowledge is specialised knowledge about a specific domain, based on the experience of the expert. Farmers are experts in the specific types of agriculture they have experience in. In this discussion, farmer knowledge is treated separately to other types of expert knowledge, generally relating to scientists. Hence the term 'expert' here means scientists and technicians with specialised knowledge about forage selection.

A large amount of expert knowledge concerning tropical forages exists in aspatial knowledge bases. Much knowledge appears in the literature, but it is often not easily accessible to non-scientists, and therefore difficult to assemble and summarise. Even more has never been recorded, and resides in the heads of experts. Trying to fit

expert knowledge into a database format may be counterproductive and cause much of the knowledge to be lost. It is important, however, to somehow capture the existing knowledge about how forages respond to varying conditions.

Farmers possess a great deal of knowledge, but usually very specific to conditions of their own land. They are experts about their specific conditions. Chambers (1980) states that:

“the most difficult thing for an educated expert to accept is that poor farmers may often understand their situations better than he does... It is difficult for some professions to accept that they have anything to learn from rural people, or to recognise that there is a parallel system of knowledge to their own which is complementary, that is usually valid and in some aspects superior”.

Studies where farmer knowledge is collected are often ad-hoc and difficult to collate. When the decision-maker is a farmer, it is appropriate to incorporate that farmer's own knowledge into the decision process. Ideally, different farmers' knowledge should be incrementally incorporated into a decision support system, so that other farmers can benefit from this knowledge. However, farmers themselves, and in particular poor farmers in the developing world, often have limited access to information and data from any sources.

3.2.3 Formalisation of Knowledge

Knowledge invariably includes an element of subjectivity. Arguably, human experts will always possess more knowledge than a machine-based system is able to discover (see Gödel's incompleteness theorem, which proved that every formal system is incomplete [Jongeneel and Koppelaar, 1999]). Two different experts will never completely agree on a decision, and often experts disagree vastly. This is because experts make decisions under conditions of uncertainty and are biased in various ways. In addition, experts sometimes make inconsistent decisions (Jungermann, 1983). When experts make decisions, they consciously or unconsciously do so based on models of the real world. In this process, both structural and translational uncertainty may be present.

Uncertainty in decisions can arise from incomplete information or knowledge about a situation or because the probability of occurrence of possible alternatives is unknown. Uncertainty may be internally or externally attributed (Scholz, 1983), that is, due to ignorance or due to variability. Shafer and Pearl (1990) point out that most everyday decision-making is uncertain, with most actions based on guesses.

Many researchers (e.g., Shafer and Pearl, 1990; Tversky and Kahneman, 1974) have argued that decision-makers think in terms of the likelihood of different outcomes occurring, and that therefore the mathematical theory of probability should be used for the formulation of reasoning under uncertainty. Shafer and Pearl (1990) claim that, historically, it has provided the most successful approach in a wide variety of fields. Therefore much of the literature on the theory of expert knowledge and decision-making refers to probabilistic decision-making.

Probability can be defined as either objective or subjective. Objective, or frequentist, probability is the frequency with which an event occurs if an experiment is performed repeatedly. Subjective probability is the degree of belief an expert has in an outcome. Subjective probability may be based on objective probability, as when an expert recalls previous occurrences of an event to estimate the probable outcome of the event. Degrees of belief may, however, also be based on other processes, in particular where no previous knowledge exists.

3.2.4 Eliciting Expert Knowledge

Knowledge elicitation refers to the process of obtaining knowledge from experts, usually in order to address a decision problem. Girard and Hubert (1999) define knowledge acquisition as the “collection, elicitation, and interpretation of data on the functioning of expertise in some domain in order to design, build, extend, adapt or modify” an expert system. Standard methods for eliciting knowledge are interviews and surveys.

A common difficulty encountered in knowledge elicitation is disagreement between experts. Approaches range from taking the mean of experts' responses (Seidel *et al.*, 2003) to sophisticated probabilistic risk assessment techniques (Edwards and Fasolo,

2001). The Delphi method (Linstone and Turoff, 1975) facilitates the convergence of expert knowledge from a number of sources.

Seidel *et al.* (2003) hypothesise that even for very narrow diagnostic problems, different human experts will come to different conclusions. They conclude that in creating a knowledge base, as well as in validating an expert system, the knowledge of several experts should be amalgamated, rather than depending on just one expert.

It is not always desirable to resolve conflicts, as this may cause information loss (Messing, 1997). In particular, where two experts disagree completely, it may be more useful to retain this conflicting information than to average their assessments. Messing's approach is to provide a formal framework that allows inconsistencies. In essence, inconsistencies and uncertainties are preserved and remain transparent to the decision-maker.

Expert knowledge elicitation can be facilitated through the use of appropriate tools. Graphic User Interfaces (GUIs) can be instrumental in organising and displaying knowledge. Where the knowledge is spatial, a GIS-based approach can be valuable for quantitative knowledge elicitation and qualitative assessment (Yamada *et al.*, 2003).

3.2.5 Issues with Expert Knowledge

A number of issues can arise when attempting to elicit expert knowledge in the form of subjective probabilities. Tversky and Kahneman (1974) consider a number of misconceptions and biases which can occur due to the subjective nature of belief assessment. Some of these are discussed below in the context of the decision problem.

Consider a situation where an expert is asked to assess whether a species will thrive under certain conditions. This is a purely subjective assessment – this species has not been trialed under these conditions previously. However, the expert may be aware of trials under slightly different conditions, or of similar species, and hence can be expected to make an 'educated guess'.

The first bias discussed by Tversky and Kahneman (1974) is ‘insensitivity to prior probability of outcomes’, which can have a major effect on probability assessments. Prior probability is the same as the base-rate frequency, or the frequency with which an outcome is expected to occur in the absence of any information. In the situation described above, the base-rate frequency is the frequency with which the species survives across all trials, regardless of conditions. If the species has a very high probability overall of thriving, then conditions would have to be very poor to significantly change the probability distribution. Bar-Hillel (1983) calls this the base-rate fallacy.

Another bias is termed the ‘illusion of validity’, which is the unwarranted confidence produced by a good fit between the predicted outcome and the input information (Tversky and Kahneman, 1974). This bias can present itself when input information is highly correlated. For example, if the location in question is in the humid tropics, the expert may assess that the species is unlikely to thrive. If, in addition, they are told that rainfall is very high, this might increase their confidence in their assessment, even though humidity and high rainfall are obviously correlated.

A further bias is that due to the ‘retrievability of instances’ (Tversky and Kahneman, 1974). Instances which are easily recalled will appear more numerous than those which are less easily recalled, even if they in fact occur with the same frequency. In particular, recent occurrences are likely to be more easily retrieved. In the example given above, trials on similar species may have been carried out over a period of time. The expert is more likely to assess the likelihood of the new species thriving based on the most easily recalled trials of the first species – either more recent trials or trials with which the expert had direct involvement.

Another possible source of bias is ‘insufficient adjustment from anchoring’ (Tversky and Kahneman, 1974). An assessment of probability may be made by starting from an initial value and adjusting this value. Say the expert makes an initial guess that the species has a 50 percent chance of thriving. Upon inspection, the conditions may prove to be very favourable. However the principle of insufficient adjustment suggests that the new estimate is likely to be much lower than if the expert had started with an initial estimate of, say, 80 percent.

Bar-Hillel (1983) also suggests that experts can make erroneous assessments when they confuse predictive accuracy and retrospective accuracy. Predictive accuracy relates to the probability of an event based on a hypothesis, while retrospective accuracy relates to the probability of a hypothesis being true, given that the event is observed. Say most trial sites for a given species are at locations with acidic soils. Therefore most instances where the species is known to thrive are characterised by acidic soils, and the retrospective accuracy of acidic soils being present when the species thrives is very high. This might be quite different from the predictive accuracy of whether the species will thrive given acidic soils.

These examples of biases in expert knowledge illustrate some of the difficulties in working with expert knowledge. In addition, there are considerations of how uncertain expert knowledge is communicated. Although it may be convenient to represent expert knowledge as numerical probabilities, people are much more likely to assess relative qualities than absolute quantities (Pearl, 1988). Bordogna and Pasi (2000) examine linguistic qualifiers of uncertainty, including vagueness of terms such as 'high' and linguistic qualifiers such as 'fairly reliable' and 'almost certain'. They show that it is possible to create mathematical models of the probability distribution denoted by these qualifiers.

Despite these problems with expert knowledge potentially introducing greater uncertainty, decisions based on qualitative and sometimes incomplete knowledge are still better than making decisions based on no knowledge (Mackinson, 2001).

3.3 Models to Support Decision-Making

As shown in Table 3.1, structural uncertainty is the uncertainty introduced in the model structure. A model is any representation of the real world, and structural uncertainty will always be present to some extent. Models to support decision-making describe and predict outcomes of various stages of the process. In the case of deciding which forage species to adopt, a model is needed to describe the likely performance of the species under certain conditions, without the need for the farmer to actually trial all candidates.

The reduction of uncertainty is an important component of supporting decision-making, although it is unrealistic, and usually impossible, to remove all uncertainty (Clark, 2002). Knowledge and appropriate modelling can reduce some uncertainty, but it is expected that there will be a point where the addition of more knowledge will no longer reduce uncertainty substantially, or possibly at all (Figure 3.1).

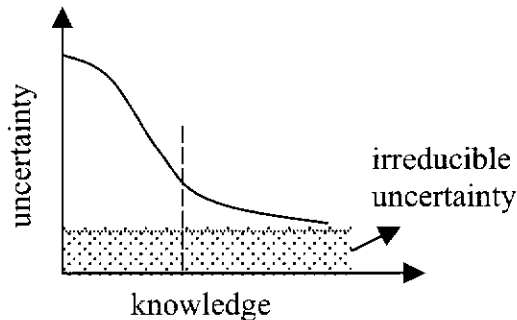


Figure 3.1 Decrease in uncertainty as knowledge increases.

Similarly, the reduction of uncertainty will not have a linear relationship with the ability to handle risk. As uncertainty is reduced, there will be a point at which reducing further uncertainty does not necessarily increase the ability to handle risk (Figure 3.2). It is important to attempt to define in the modelling process, the points at which increasing knowledge and decreasing uncertainty are no longer beneficial to the aim of the model, which is to support decision-making by making risk manageable.

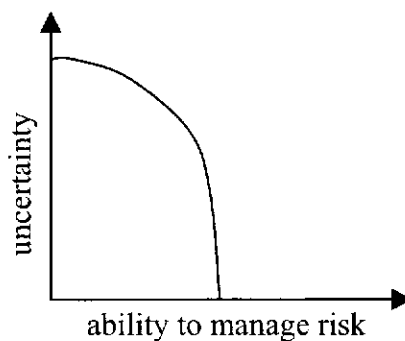


Figure 3.2 Relationship between decreasing uncertainty and increasing ability to manage risk.

Probability-based methods are not only useful for representing knowledge, but also for dealing with uncertainty by addressing uncertainty in model quantities. However, van Asselt (2000) points out that this approach still ignores uncertainty in model structure. Structural uncertainty is more challenging to address than metrical uncertainty, but it is clearly important to attempt to describe the structural uncertainty present in a model.

Decision support models are intended to represent the aspects of complex or vague decision problems in formats that are useful to decision makers. Models of decision-making simplify a complex process, making it easier to describe and analyse.

The purpose of a model may be descriptive, predictive, mechanistic or empirical. Many theories exist on how decisions are made and how they should be made (see for example Edwards and Fasolo, 2001). However, in the current research, the focus is on investigating models of reality which support the decision-making process, in addition to modelling the process of decision-making itself.

3.3.1 Steps in Decision-Making

The decision-making process can be divided into the decisions facing the farmer and the decisions facing the expert. In Figure 2.9, the farmer's decision was defined as a strategic decision (whether a change is needed and, if so, what type of change) followed by a tactical decision (what should be planted). Expert knowledge, either directly or through modelling, can be used to inform different stages of the farmer's decision (Figure 3.3).

The role of the expert in this process is to provide advice and options to the farmer. The knowledge and information available to the expert can be modelled to provide suitable advice and options, both at the strategic level and the tactical level.

The ways in which the models are defined identifies concepts that explain factors that are significant. Getting choice of model right reduces both structural and translational uncertainties.

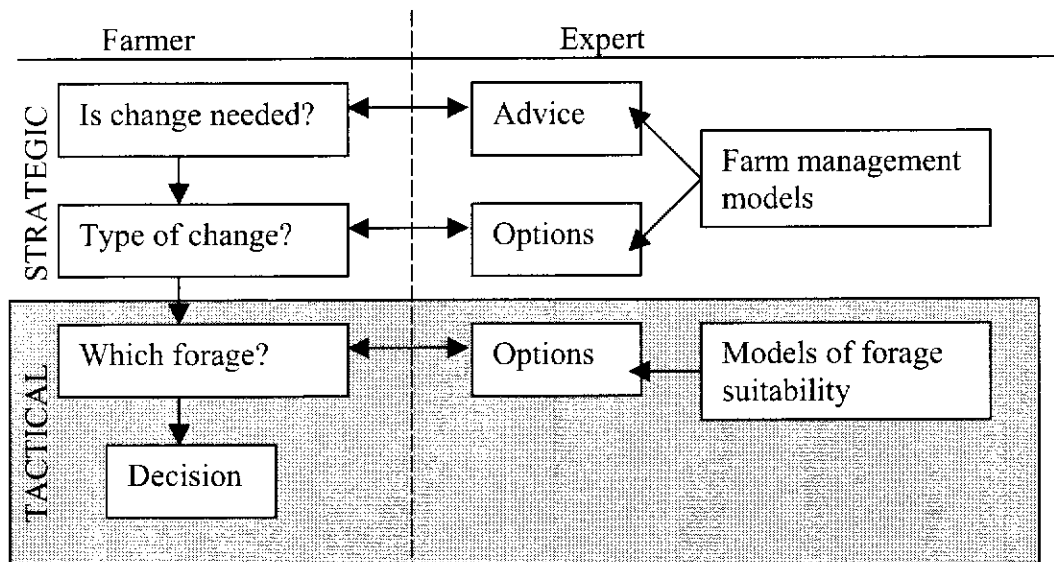


Figure 3.3 Interaction between expert decision-making and farmer decision-making.

3.3.2 Functional Modelling

The purpose of functional modelling in this case is to predict the success of forage species in specific locations with specific biophysical, socio-economic and management characteristics.

A pertinent question is how to combine both data from databases and expert knowledge in modelling. Although data may still be uncertain, it is likely to be more objective than knowledge, and represented in different ways. Data is likely to be directly measured, or derived from direct measurements, and lend itself more readily to statistical analysis than knowledge.

Data can be used to inform knowledge, in particular where knowledge is lacking or uncertain. Conversely, where data is missing or in error, knowledge can complement and correct the data where necessary.

Functional modelling relies on information about variables to reduce metrical and temporal uncertainty. A great deal of this information is spatial in the context of the decision problem.

3.3.3 Addressing Spatial Uncertainty

Recognising the spatial context of modelling is not only desirable, but also central to effectively modelling processes which are spatially heterogeneous. Spatial analyses can discover otherwise undiscernible patterns in data, and spatial visualisation can greatly assist in conveying model results.

The inclusion of spatial data can introduce uncertainty and sometimes error, as discussed in Section 3.1.2 above. Crosetto and Tarantola (2001) point out that often in spatial modelling, both the data and model are treated as error-free, and output uncertainty is completely disregarded. It is important therefore to recognise these sources of uncertainty, describing them and reducing them where possible. Davis and Keller (1997), for example, explore ways of visualising spatially-variable uncertainty.

However, the inclusion of spatial aspects in modelling can also reduce uncertainty. If a spatial model better represents the real world than an aspatial model, then structural uncertainty is reduced. In addition, translational uncertainty may be greatly reduced by presenting results visually in the form of maps.

In the decision problem, spatial information and knowledge relate to how species perform at different locations and the characteristics of those locations. In many cases, approaches are available to reduce or remove uncertainty. However, the incorporation of uncertainty measures allows data and knowledge to be utilised even when uncertainty cannot effectively be removed.

3.4 Summary

Risk and uncertainty are factors in most decision-making, especially when the decision-making process has spatial aspects. In the case of supporting farmers' decisions about forage selection, there are a number of sources of risk and uncertainty.

Approaches to reducing uncertainty depend on its source and its type. In order to incorporate knowledge, uncertainty and risk in the modelling process, appropriate forms of modelling must be selected.

Although many researchers have discussed uncertainty in spatial modelling, often only metrical uncertainty is considered. However, almost all decision problems will also contain temporal, structural and translational uncertainty. It is important therefore to attempt to account for all types of uncertainty in the modelling process.

The following chapter will examine ways to address the decision problem, taking into account sources of risk and uncertainty and sources of knowledge and data.

CHAPTER 4. ADDRESSING THE DECISION PROBLEM

In the previous chapter, risk, uncertainty and knowledge were discussed in general terms. This chapter explores ways of addressing the decision problem, taking into account uncertainty and risk in the system. It also examines the information sources available for the task of forage selection.

4.1 Risk and Uncertainty in Forage Selection

4.1.1 Decision-Making Under Risk and Uncertainty

Decision theory analyses human decision behaviour and draws on mathematics, economics and philosophy, as well as on concepts from political and social sciences (Scholz, 1983). Decision theory's principal rule is to maximise expected utility (Heckermann, 1995). Decision-making becomes a function of the decision-maker's alternatives, their beliefs and their preferences.

There are several alternative models of decision-making under risk, including the maximin decision criterion, maximising expected utility, minimising variation, and trade-off between expected utility and variation (Upton, 1996). These can all be applied to the problem of agricultural decision-making in the tropics.

In coming to a choice on how to deal with uncertainty in decision-making, the implications of uncertainty need to be explored. Lindner (1987) describes the possible outcomes of a decision (Table 4.1). The potential magnitude of the lost opportunity, or the loss caused, will influence the need to reduce uncertainty around the decision. Another important factor lies in defining how much uncertainty needs to be reduced in a particular decision in order to select the correct action (or inaction).

		Objective	
		Benefit occurs	No benefit occurs
Subjective	Adopt	Correct action	Type II error Loss caused
	Reject	Type I error Lost opportunity	Correct inaction

Table 4.1 Type I and II errors as possible outcomes of a decision. Adapted from Lindner (1987) and Abadi Ghadim (2000).

4.1.2 Risk Perception

Decision-makers can be classed as risk-taking (sometimes termed risk-seeking), risk-neutral or risk-averse, depending on the level of risk they are willing, or able, to accept. Abadi Ghadim (2000) makes an important distinction between perceived riskiness and attitudes to risk. Perceived riskiness is a subjective judgment about how risky a decision is, whereas attitude to risk refers to the decision maker's personal preferences about how to handle risk (Abadi Ghadim, 2000).

Attitudes to risk affect adoption of new crops, technology and practices. The term adoption refers to the process of deciding whether or not to use a new production technique (Lindner, 1987). Depending on the research context, adoption may be considered to have occurred based on different criteria. For example, this could be as soon as a farmer starts using new production technologies or, alternatively, only after a farmer has continuously used the innovation for a number of years.

Many researchers (Abadi Ghadim, 2000; Marra *et al.*, 2003; Pannell *et al.*, 2000) argue that farmers want risk to be better defined so they can choose how to respond to it. One way to better define risk is by reducing both spatial and aspatial uncertainty in the decision to adopt. It is often assumed that farmers wish to avoid risk, however, Pannell *et al.* (2000) point out that farmers do not necessarily want to avoid risk but rather to define the risk so they can respond tactically and dynamically.

Farmers have a reputation for making type I errors (incorrect rejection) (see Table 4.1) due to risk aversion (Antle, 1987; Kingwell, 1994). However, type II errors (incorrect adoption) also occur frequently in the developing world, for example,

when farmers adopt land use changes that have severe environmental effects. Information that identifies areas of likely benefit reduces type I errors, while information that identifies where likely loss would occur reduces type II errors.

Pannell *et al.* (2000) found that whilst risk aversion is a factor in farm management, it is only one of a number of factors that affect farm management decisions. For the farmer, it is more important to “solve the whole problem roughly than to attempt to solve part of the problem extremely well” (Pannell *et al.*, 2000).

Studies have shown that farmers’ perceived risk is modified as more information is acquired (Marra *et al.*, 2003). Seemingly contradictory studies on the relationship between risk-aversion and adoption can be explained as follows. A risk-averse farmer will take the option with the lowest perceived risk. Sometimes this is adoption and sometimes non-adoption. The idea that risk-seeking farmers will adopt sooner than risk-averse farmers assumes that adoption is always subjectively riskier than the status quo and that adoption is always the better option (Marra *et al.*, 2003). However, as the above analysis shows, this is not always the case.

Marra *et al.* (2003) review the role of uncertainty and risk in agricultural adoption. They found that learning reduces uncertainties and improves decision-making and that attitudes to risk affect adoption of new technologies. Risk-aversion does not always mean that farmers will be slower to adopt, but generally that the farmer wants more certainty around the benefits of adoption.

Abadi Ghadim (2000) investigated how risk perceptions and risk attitudes affect adoption behaviour for farmers considering adopting chickpeas and other legumes in Western Australia. He concluded that risk has an important influence on adoption decisions and that learning is an integral part of adoption under risk and uncertainty. While farmers in Western Australia vary widely in their levels of risk-aversion, for all farmers it is important to reduce uncertainty about an innovation as quickly as possible (Abadi Ghadim, 2000).

Although the preceding results come from studies in developed agriculture, it is assumed that farmers in developing agriculture approach adoption decisions in a

similar fashion. These farmers are, however, likely to be more risk-averse, because one negative outcome could severely threaten their livelihood.

Cramb (2000) argues that the decision to adopt or not is really a range of responses, depending on the farmer's goals and circumstances. A number of authors have researched adoption decisions in tropical agriculture, often from a farming systems research or participatory research approach (e.g., Peters *et al.*, 2003a; Staal *et al.*, 2002; Stür *et al.*, 1999; Thomas and Sumberg, 1995). These authors generally agree that the decision to adopt is complex and that participatory, farmer-centred models are needed to fully appreciate farmers' decision-making practices. Neill and Lee (2001) found levels of knowledge and understanding of the decision-makers to be fundamental to adoption, along with agronomic characteristics and economic contexts. Shively (1997) found risk to be an important factor in adoption decisions for low-income farmers in the Philippines. Perceptions of risk in the adoption decision are very important, and managing these risks appropriately is particularly significant for resource-poor farmers in the developing world.

4.1.3 Sources of Uncertainty

Uncertainty in the decision problem comes from a number of sources. Firstly, there is uncertainty for the farmer about which forage species exist, and what their properties are. Furthermore, a farmer may be uncertain about how to obtain the seeds and at what price – certainly availability and cost will be subject to variability. Also, uncertainty exists about forage species' management and environmental requirements. This uncertainty is due both to variability and ignorance.

There will also be uncertainty due to climate variability and soil variability. Rainfall and temperature values for a given location will exhibit both metrical and structural uncertainty (see Table 3.2), but arguably the most difficult to manage is temporal uncertainty caused by ignorance about future climate. In particular, extreme climate events, such as droughts or hurricanes, expose farmers to greater risk. Soil characteristics, such as fertility, may change over time, depending on which crops are planted and how they are managed.

Uncertainty may also be present regarding the impact of a forage species on livestock and milk production. A final source of uncertainty is translational uncertainty, introduced when an expert or extension worker communicates recommendations to a farmer. Note that this uncertainty can exist on both sides of the communication, that is, there may also be uncertainty for the expert regarding the farmer's requirements due to translational uncertainty.

Hardaker *et al.* (1997) classify the sources of risk in agriculture as financial risk, resulting from the method of financing, and business risk. The latter is comprised of production risk, market risk, institutional risk and personal risk. Production risk relates to the sources of uncertainty outlined above, namely, the unpredictable nature of weather and uncertainty about performance of crops and livestock (Hardaker *et al.*, 1997). Institutional risk relates to government changes in laws and rules, and personal risks are those that arise from major life crises or illness of the farmer or their household. See Hardaker *et al.* (1997) for a comprehensive review of risk and decision analysis in developing agriculture.

4.1.4 Reducing Uncertainty Through Knowledge

Some, but not all, of the uncertainty mentioned above can be reduced through information. Uncertainty about forage characteristics and uses can be reduced by providing information to the farmer via extension agents and publications. Information on availability and cost can also be provided by local extension agencies. Similarly, information on the impact of improved forages on livestock and milk production can be provided in this way.

Temporal uncertainty relating to climate is more difficult to reduce. However, information regarding which forages better withstand certain extreme events can help farmers to manage the associated risk. Also, information on the likelihood of extreme events occurring could be very useful. Various researchers have focussed on the temporal uncertainty associated with future climate change in agriculture (e.g., Gu *et al.*, 1996; Jones and Thornton, 2003), and deal with it using techniques ranging from conventional simulation modelling to belief network approaches (Gu *et al.*, 1996).

Trials play an important role in information gathering to reduce uncertainty. Farmers typically use trials to decide whether to adopt or reject a new technology. Both improved skills and information can be learned from trialing, both of which can reduce uncertainty and therefore assist decision-making around adoption or rejection.

Translational uncertainty can be addressed by investigating how information is provided to farmers and how information is gathered from farmers.

As previously discussed, where uncertainty cannot be reduced, it should be described. Digital maps provide an excellent opportunity for visualising spatial uncertainty.

4.2 Modelling the Forage Selection Decision

As discussed in the previous chapter, the purpose of modelling in this case is to predict the success of forage species in specific niches, with particular biophysical, socio-economic and management characteristics. Functional models support the tactical decision of the farmer regarding which forage species to select.

A model is only useful to the degree that it represents reality in a way that supports decision-making. Therefore a model needs to be consistent and accurate and, at the same time, reduce the complexities of the decision-making process to make them manageable.

Simply developing a model of species' suitability is only part of the process. The results of the model need to be made available to the decision-maker. An appropriate way of delivering this information is by means of a Decision Support System (DSS).

4.3 Decision Support Systems

4.3.1 Types of Decision Support Systems

The purpose of a DSS is to provide data, procedures and analytical capability leading to better-informed decisions. Typically, a DSS consists, therefore, of data, a rule-base, algorithms for combining these and a user interface. A DSS is not necessarily

computer-based, but in this discussion it is assumed a DSS is implemented in a personal computer environment.

Spatial Decision Support Systems (SDSS) are a special case of DSS, with an explicitly spatial component, usually formed by integrating GIS technology.

Expert Systems (ES) are a subset of DSS, developed in the field of Artificial Intelligence (AI). An ES usually consists of a knowledge base and an inference engine (Jennings and Wattam, 1994). An ES is usually not static, and new rules in the knowledge base can be ‘learned’ by the system over time.

DSS may comprise a combination of data and knowledge as a hybrid of data-based DSS and knowledge-based ES. The rule base and algorithms are developed in this case to work with both sources of information.

4.3.2 Decision Support Systems in Agriculture

Agricultural DSS have been in existence since at least the mid 1970s (e.g., SIRATAC, a cotton production decision system, and EPIPRE, a European wheat DSS, were both begun in 1976 [McCown *et al.*, 2002]). For farmers in the developed world, there are hundreds of DSS available, covering production decisions relating to crops such as cotton, wheat and pasture (McCown *et al.*, 2002).

A body of research is emerging on the ‘crisis’ of DSS research in agriculture, that is, DSS use in agriculture is declining and not living up to its apparent early promise (see for example Cox, 1996; McCown *et al.*, 2002; Walker, 2002). The journal ‘Agricultural Systems’ devoted a special issue (Issue 74, 2002) to reviewing the low adoption and frequent abandonment of DSS by farmers.

Barriers identified in the adoption of agricultural DSS in the developed world include complex design and presentation of DSS, unrealistic requirements for monitoring data and the need for the farmer to own a computer (Cox, 1996). Walker (2002) adds irrelevance and inflexibility of many DSS, lack of user confidence and institutional and political barriers, among others.

Less research has been done on the adoption of agricultural DSS in the developing world, but it can be expected that issues related to data requirements and computer ownership would be intensified. Hall *et al.* (1997) researched the implementation of GIS-based DSS for facility planning in developing countries, and found barriers including inadequate computing skills, poor computing facilities and poor data availability and quality. The same barriers would presumably exist for agricultural DSS adoption in developing countries.

The development of a DSS therefore requires more than just the implementation of the DSS itself. Walker (2002) lists six essential steps (Table 4.2), starting with needs analysis and followed by design and implementation. Development often stops after implementation, and, as Cox (1996) points out, there is poor understanding among researchers of what an agricultural DSS does when farmers or other practitioners use it. Walker's final three steps are capacity building, fostering uptake and monitoring and evaluation. Cox (1996) also argues that for a DSS to be successful, it needs to be embedded in a wider social process.

1	Needs analysis
2	Design
3	Implementation
4	Capacity building
5	Fostering uptake
6	Monitoring and evaluation

Table 4.2 Steps in DSS development.

4.3.3 Spatial Decision Support Systems

A spatially enabled DSS requires spatial input data, spatial analysis capabilities and/or spatial output. Because a considerable amount of uncertainty stems from spatial variation, a spatial DSS (SDSS) can provide the information necessary to manage some of this uncertainty. The spatial component is usually implemented using GIS technology.

A large number of SDSS exist within agriculture (see for example Berger, 2001; Booth, 1995; Hill, 2000), and even more in wider fields that could be applied to

agricultural decisions. As spatial data becomes more available and reliable in the developing world, SDSS are emerging that are specifically targeted to these environments, with encouraging results. Crossland *et al.* (1995) found unequivocal evidence that addition of GIS technology to the decision environment for a spatially referenced decision task reduced the decision time and increased the accuracy of individual decision-makers.

Staal *et al.* (2002) found that the inclusion of spatial analysis improved their models of technology uptake on smallholder dairy farms in Kenya. However, Hall *et al.* (1997) found that although in some developing countries (e.g., Chile and Costa Rica) GIS infrastructure is in place, expertise had not yet evolved to the point where practitioners had been able to develop their own DSS.

A number of SDSS have been developed for use in agriculture. GRAZPLAN (Donnelly *et al.*, 2002) was developed by Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia for consultants and farmers to improve the profitability and environmental sustainability of grazing enterprises. KEOPS (Girard and Hubert, 1999), a knowledge-based model to characterise the strategic pattern of farms, is used for grazing management on French sheep farms. PROMOD (Mummery and Battaglia, 2001) is a eucalyptus plantation growth model that can be used to produce site productivity and suitability maps. Agri-FACTS (Thomas *et al.*, 1999) is a web-based spatial application providing data to support suitability analysis of cropping systems alternatives.

FloraMap (Jones and Gladkov, 1999) was developed for predicting the distribution of plants in the wild, but can be applied to agricultural contexts. GEOMOD2 (Pontius *et al.*, 2001) models the spatial pattern of land-use change over time and was developed specifically for application in the tropics. ASSESS (Veitch and Bowyer, 1996) is a system designed to select suitable sites for waste disposal assessment, but could conceivably be used for other land use decisions. ILUDSS (Zhu *et al.*, 1996; 1998) is a knowledge-based spatial decision support system for strategic land use planning in rural areas. The assumption is that spatial land-use models could be adapted for agricultural decision-making.

Many other non-spatial agricultural DSS exist. In most cases, incorporating spatial analysis into non-spatial agricultural DSS would enhance these tools. White *et al.* (2002) assessed the use of GIS in agronomic research and conclude that GIS is under-utilised for analysing and presenting results.

4.3.4 Expert Systems

There is often an overlap between DSS and ES, with many DSS in reality depending on expert knowledge. Some of the agricultural DSS mentioned above could therefore also be classed as agricultural ES. A large number of ES exist, many for medical diagnosis. Here, the focus is more on ES related to natural resources, particularly those with a spatial component.

The Automated Land Evaluation System (ALES) is based on a decision tree built using expert knowledge (Rossiter, 1996). PROSPECTOR (Hart *et al.*, 1978) provides expert consultation on mineral exploration. The inference engine is based on Bayesian probabilistic reasoning. PROSPECTOR has been converted into an SDSS by coding the decision rules into a GIS framework (Katz, 1991). ArcWofE (Desnoyers, 2001) is an add-on to ArcView (ESRI, 2001) designed to incorporate weights of evidence (WofE), an essentially Bayesian approach in a spatial context. Expector (Corner *et al.*, 2002) is a spatially enabled expert knowledge system used in soil mapping.

Common evaluation criteria to compare expert systems have not yet been developed, but Seidel *et al.* (2003) suggest examining criteria including accuracy, comprehensiveness, usefulness and user-friendliness. A method of validation often applied is to validate cases where a group of experts have previously agreed upon a diagnosis or outcome (Seidel *et al.*, 2003).

4.3.5 The Case for an SDSS for Selecting Forage Species

As reported in Chapter 2, barriers to adoption of forages include unfamiliarity, lack of understanding of longer-term benefits, biophysical limitations and possibly cultural and social limitations. An SDSS can address these barriers, firstly by

making information available to farmers (increasing familiarity and describing long-term benefits) and secondly by matching forage species to unique biophysical and socio-economic niches.

A spatial forage DSS would therefore combine data, knowledge and appropriate algorithms to provide recommendations (Figure 4.1). However, rather than just providing recommendations, it is important that the SDSS support decisions appropriately by providing as much relevant information as possible. This information can include maps and graphs of species' suitability and species factsheets. In addition, uncertainties can be explicitly shown in both the maps and the graphs.

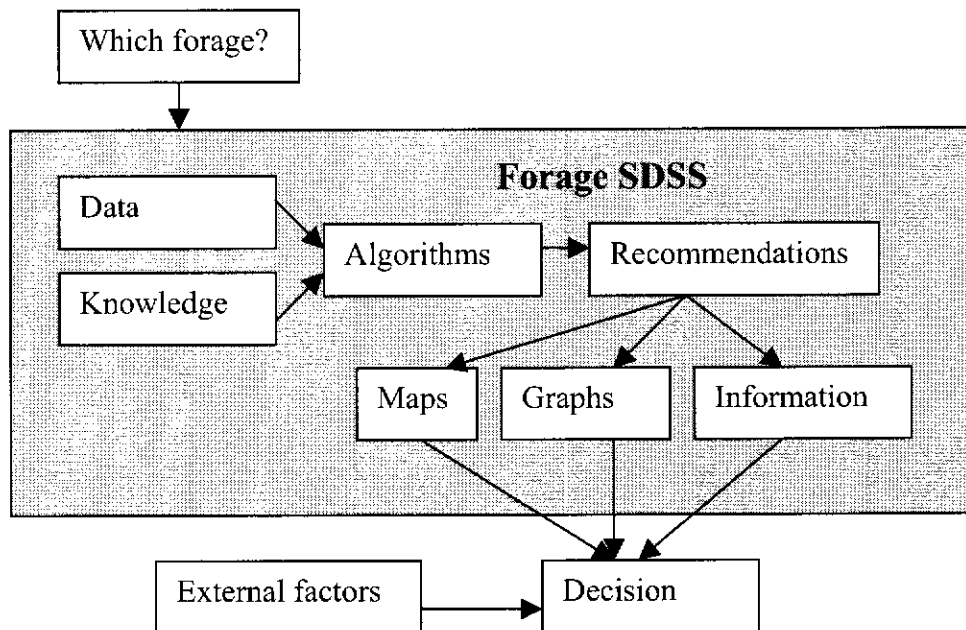


Figure 4.1 A spatial forage decision support system.

4.4 Information Sources for a Forages DSS

A substantial amount of data exists in trials databases, spatial databases and literature, as well as expert knowledge which may or may not be formalised in knowledge bases.

4.4.1 Forage Databases

Forage databases contain data on trials carried out on various species at various locations over time. Information expected to be recorded for each trial would be a measure of the success of the trial and characteristics of the trial conditions. Success of a trial can be measured in a number of ways, such as how well the plants establish and how much they produce. Information on the trial conditions could be as basic as location coordinates, but could also include climate data, soil information and details of the management of the trial site.

Databases also exist containing ‘passport data’ (Barco *et al.*, 2002), that is, data on the locations where species accessions have been collected in the wild. This provides information on where a species can survive, however, passport data does not allow any conclusions to be drawn about where the species will thrive.

Existing forage databases include the International Network for the Evaluation of Tropical Pastures (in Spanish, *Red Internacional de Evaluación de Pastos Tropicales* [RIEPT]) and the Network for Research on Livestock Feed in West and Central Africa (in French, *Réseau de Recherches en Alimentation du Bétail en Afrique Occidentale et Centrale* [RABAOC]), both maintained by CIAT (Barco *et al.*, 2002).

4.4.2 Expert Knowledge

Most expert knowledge on forage species is not formalised. However, apart from the experts (including farmers) themselves, a number of sources of expert forage knowledge exist. Although many booklets aimed at assisting tropical forage selection have been produced (e.g. Argel and Villareal, 1998; Argel *et al.*, 2000; Argel *et al.*, 2001; Lobo and Acuña, 2001; Sandoval *et al.*, 2001), each publication tends to focus on only one species.

Publications that cover a number of species are generally aimed at a specific environment, such as Central America (Peters *et al.*, 2003b), Costa Rica (Lobo and Solano, 1997) or South East Asia (PROSEA, 2001; Stür and Horne, 2001; Stür *et al.*, 2002), or at a specific management system (Holmann and Lascano, 2001).

There is inevitably overlap between databases and knowledge bases, as knowledge is often derived from data. Publicly available databases and knowledge bases usually both have inbuilt search mechanisms. The main distinction drawn here is that a database contains objective, measurable data (hence usually from trials), whereas a knowledge base contains more subjective data, often compiled from a number of sources and experts. One of the benefits of a DSS is the ability to combine both data and knowledge as input information.

The Tropical Grasslands Society of Australia has developed 'Pasture Picker' (Partridge, 2003), a knowledge base of tropical and subtropical grasses and legumes, searchable by selecting average rainfall, soil type and tolerance to frost, drought, waterlogging and heavy grazing. Fact sheets exist for each species, with further information on adaptation and production, and on varieties within the species.

FAO's Ecocrop (FAO, 2000) contains climate and soil requirements for 1,710 plant species, and permits the identification of plant species for defined uses, including some forages. The Animal Feed Resources Information System (AFRIS) (FAO *et al.*, 2004a) contains information on a variety of feed resources including grasses and legumes.

The International Institute of Tropical Agriculture's (IITA) database of legume characteristics, LEXSYS, has been converted to Lexsys KBS (University of Wales, 2003), a decision support system for the integration of legumes into tropical farming systems. It contains information on 91 parameters for 125 legume species, including detailed citations.

The CAB International compendia (CABI, 2003) include a forestry compendium and a crop protection compendium, incorporating information on species and countries. Some of the knowledge base in the compendia includes some tropical forage species.

NSW Agriculture's (2001) 'Pasture Planner' includes detailed information on around 25 tropical grasses and legumes, alongside temperate pastures suitable for New South Wales (Australia).

In addition, a number of other plant knowledge bases include some tropical forage species, such the United States Department of Agriculture Plants Database (USDA, 2002).

A recent development is the Selection of Forages for the Tropics (SoFT) project, an ACIAR-funded collaborative initiative between CSIRO, the Queensland Department of Primary Industries (QDPI), the International Livestock Research Institute (ILRI) and CIAT (Peters *et al.*, 2003a).

4.4.3 Spatial Reference Data

Spatial reference data is not necessarily used for modelling per se, but is useful in referencing the location of other data. Biophysical data is any data on the biological and physical components of the environment and socio-economic data is any data involving social as well as economic factors. Both biophysical and socio-economic data, as well as spatial reference data, are required to inform the decision support system. To this end, existing sources of spatial data are examined here.

A number of sources exist for spatial data worldwide, at country level and at subregional level. The spatial data used for this project has primarily been sourced for Central America, with the caveat that it should be available for the entire developing world. The data should also be ready to use for modelling purposes, without a significant processing overhead.

A selection of spatial reference data readily available for Central America is summarised in Table 4.3. Scale refers to the scale at which the data is intended to be viewed or used. A scale of 1:1,000,000 means that 1mm on a printed map corresponds to 1,000,000mm, or 1km, on the ground. In digital mapping, view scale is not statically confined, but the smaller the scale the less detailed the spatial data will be.

Data	Scale	Extent	Source
Country boundaries	1:1,000,000	Worldwide	ESRI, 1992
Department, province, municipality, district, canton	1:120,000 – 1:1,500,000	Central America	Winograd <i>et al.</i> , 2000
Roads	1:200,000 – 1:1,000,000	Central America	Winograd <i>et al.</i> , 2000
Rivers	1:1,000,000	Worldwide	ESRI, 1992
Populated places		Worldwide	NGA, 2004

Table 4.3 Description of existing spatial reference data.

Administrative boundaries are readily available in GIS format. The Digital Chart of the World (ESRI, 1992) comprises a number of thematic layers, including country boundaries, roads and rivers. However, roads data is also available at larger scales from the *Indicadores de Sustentabilidad Rural* CD-ROM (Winograd *et al.*, 2000) for Central America (hereafter referred to as the *Indicadores Atlas*). This CD-ROM is also the source for department, province, municipality, district and canton boundaries for Central American countries, compiled from a variety of sources. The US National Geospatial-Intelligence Agency maintains a database of foreign geographic feature names and locations, including all populated places (NGA, 2004).

4.4.4 Spatial Biophysical Data

Biophysical data (Table 4.4) includes 1km² resolution surfaces for elevation, temperature and precipitation. At the time of commencing the project in 2001, this was the smallest resolution available for all surfaces. More recently, 90x90m resolution surfaces have become available, but have not yet been incorporated in this research. When data is presented in raster (grid) format, rather than vector format (see for example Bonham-Carter [1994] for a discussion on vector and raster formats), it is convenient to refer to resolution rather than scale. Resolution refers to the size of one gridcell (or pixel) on the ground, and is commonly measured in km or degrees (or subdivisions of these). The length of one degree varies with latitude, and is around 111km at the equator. A 30 arc second grid is often referred to as a 1km² grid, and 3 arc second resolution is approximately equivalent to 90m. As with scale, grids can be visualised at larger resolution than intended but the level of detail will

remain at the original resolution. In the following tables, resolution is given in distance units.

Data	Resolution / Scale	Extent	Source
Elevation (m)	1km x 1km	Worldwide	NOAA, 1999
	100m x 100m	Central America	Jones, 2001
	90m x 90m	Worldwide	USGS, 2003
Temperature (monthly minimum, mean and maximum) (degrees Celsius)	1km x 1km	Central America	Jones, 2001
		Worldwide	Hijmans <i>et al.</i> , 2004a
Rainfall (monthly mean), (mm)	1km x 1km	Central America	Jones, 2001
		Worldwide	Hijmans <i>et al.</i> , 2004a
Ecosystems	1:250,000	Central America	World Bank <i>et al.</i> , 2003
Holdridge lifezones ^a	55km x 55km (30 minute)	Worldwide	Leemans, 1990
FAO soil units	1:1,000,000	Worldwide	FAO, 2002
Derived soil properties	55km x 55km (30 minute)	Worldwide	Batjes, 1997

Table 4.4 Description of existing spatial biophysical data.

^a Holdridge lifezones are ecosystem classifications developed for the tropics and subtropics (Holdridge, 1967).

The 1km² digital elevation model (DEM) was developed at CIAT, derived from the US National Oceanic and Atmospheric Administration's DEM at 30 arc seconds (NOAA, 1999). Climate surfaces were also developed at CIAT, based on the DEM and around 18,500 climate stations in the tropical world. These were produced at various scales, including a 1km² surface for Central America (Jones, 2001). A 90 x 90m DEM has also recently become available (USGS, 2003). Hijmans *et al.* (2004a) are developing worldwide 1km² grids of monthly total precipitation, monthly mean, minimum and maximum temperature and 19 derived bioclimatic variables.

Various classifications of ecosystems are available. The World Bank *et al.* (2003) have produced a dataset of 140 ecosystem classes for Central America. Holdridge Lifezones (Holdridge, 1967) are ecosystem classifications developed for the tropics and subtropics. A global spatial dataset exists for Holdridge Lifezones (Leemans, 1990).

In addition, soil maps exist at various scales for various areas within Central America. At a worldwide level, FAO publishes digital soil maps with derived soil properties (FAO, 2002). They are based on the FAO/UNESCO soil map of the world at an original scale of 1:5,000,000. The International Soil Reference and Information Centre (ISRIC) has used these to derive 30 x 30 minute grids of soil properties, including soil pH and soil organic carbon density (Batjes, 1997).

Soil data is problematic for a number of reasons. Firstly, it is generally not available at a large enough resolution to represent its heterogeneity. In the case of FAO data, 30 minute resolution equates to approximately 55km. Soils generally vary at a much finer scale than this. In addition, deriving soil properties from soil types is not always a straightforward process. Although various characteristics such as pH and fertility have been derived, there is not necessarily always a direct correspondence between soil type and soil characteristic. Finally, soils are often mixed types, and therefore attempting to classify soil type at any given location will be a complex process. The distribution of soils over space is complex and no single soil classification can be used at all locations and at all levels of resolution (Burrough *et al.*, 1997).

Soil classifications can be derived with a number of different purposes in mind. If the classification process has no relationship to plant growth variability, then it is unlikely to be useful for the decision problem. Despite these issues with soil data, it is clear that soils have a strong influence on forages, as with all plants, and soils must be considered in any approach to recommending forages to farmers. Appropriate ways of incorporating soils information are discussed in subsequent chapters.

4.4.5 Spatial Socio-Economic Data

Socio-economic data that could be useful includes population, market access and, specifically where forages are concerned, other factors such as locations of milk processing plants and livestock density (Table 4.5). Census data is available at municipality level and has been summarised in a number of different digital publications, including the *Indicadores Atlas* (Winograd *et al.*, 2000), the *Atlas de Honduras* (CIAT, 1999) and the *Atlas Rural de Nicaragua* (MAGFOR *et al.*, 2001).

Censuses vary from country to country (Table 4.6). The Center for International Earth Science Information Network (CIESIN), the International Food Policy Research Institute (IFPRI) and the World Resources Institute (WRI) have created a Gridded Population of the World (GPW) spatial database (CIESIN *et al.*, 2000) based on censuses around the world.

Data	Scale	Extent	Source
Population density	2.5 x 2.5 minute	Worldwide	CIESIN <i>et al.</i> , 2000
Population census	Municipality	Central America	Various, 1984 - 1995
Access to market	1:2,500,000	Central America	Winograd <i>et al.</i> , 2000
Livestock density	Municipality	Honduras	CIAT, 1999
		Nicaragua	MAGFOR <i>et al.</i> , 2001

Table 4.5 Description of existing spatial socio-economic data.

Country	Year	Source
Costa Rica	1984	<i>Ministerio de Economía, Industria y Comercio, Dirección General de Estadísticas y Censos</i>
El Salvador	1992	<i>Ministerio de Economía, Dirección General de Estadísticas y Censos (DIGESTYC)</i>
Guatemala	1994	<i>Instituto Nacional de Estadística (INE)</i>
Honduras	1988	<i>Dirección General de Estadísticas y Censos, Secretaría de Coordinación y Presupuesto (SECPLAN)</i>
Nicaragua	1995	<i>Instituto Nacional de Estadísticas y Censos (INEC)</i>
Panama	1990	<i>Contraloría General de la República, Dirección de Estadísticas y Censo</i>

Table 4.6 Census Data for Central America. Source: Winograd *et al.* (2000).

CIAT has developed a methodology for creating accessibility surfaces, based on a cost-distance algorithm. Access to market for Central America was included in the *Indicadores Atlas* (Winograd *et al.*, 2000), and an ArcView add-on is now available to create other accessibility surfaces (Eade *et al.*, 2000). Access to market could be useful information in determining the value of increasing milk and meat production through forages. In particular, milk needs to be delivered to market quickly. In the case of cash crops, access to market becomes even more relevant. Other market information, such as market price over time, also exists that can be spatially referenced (CIAT, 2003).

FAO's Agriculture Department (2004) publishes maps and data at country level on livestock population and production. However, country level data is not so useful for analysis at subregional level. Some countries carry out agricultural censuses, so some livestock data is available at municipality level. Two examples of this are the data available in the *Atlas de Honduras* (CIAT, 1999) and the *Atlas Rural de Nicaragua* (MAGFOR *et al.*, 2001).

4.4.6 Other Spatial Data

The two atlases mentioned above contain a wealth of GIS data, but, obviously, only for Honduras and Nicaragua respectively. The *Atlas de Honduras* (CIAT, 1999) contains biophysical, socio-economic and agricultural data at various scales, with the aim of supporting analysis after Hurricane Mitch in 1998. This atlas also contains digitised FAO-derived soils, as well as soils digitised from a Honduras soils map at 1:500,000. This classification includes more detail than the FAO-derived maps, including soil depth, drainage, pH, slope, colour, texture and soil name. The *Atlas Rural de Nicaragua* (MAGFOR *et al.*, 2001) also contains biophysical and socio-economic data, including soil data, climate, population, economic activity, illiteracy, education, housing, water, sanitation and agriculture, all at municipality level.

The *Indicadores Atlas* (Winograd *et al.*, 2000) contains environmental and sustainability data for Latin America and the Caribbean, and includes datasets on climate, soil, land use and population, to name a few. In total the *Indicadores Atlas* covers over 100 indicators of sustainability.

4.5 Summary

This chapter has explored the various elements required to address the decision problem. Firstly, risk and uncertainty in agricultural decision-making were discussed, concluding that risk is very important for farmers, and that attitudes to risk can be affected by increasing knowledge and reducing uncertainty. Sources of uncertainty in agricultural decision-making were then considered, as well as where and how these uncertainties can be reduced and described.

Functional models are needed to support farmers' tactical decisions, but the results of the model also need to be made available to the decision-maker. A DSS, and in particular an SDSS, was identified as a well developed method for achieving this aim. Because information sources include not only data but also knowledge, it is appropriate to develop an SDSS incorporating expert system concepts.

Examples have been given of available information for addressing the decision problem, including trials databases, expert knowledge bases, literature and a wealth of spatial data.

This thesis now turns to the problem of developing a functional model to address the decision problem, concentrating on the issues of uncertainty and knowledge. The availability and appropriateness of data mentioned above is also held in consideration.

CHAPTER 5. CRITERIA FOR MODEL SELECTION

In the previous chapters, the decision problem was described and explored. The problem can be split into two components, namely, the strategic decision and the tactical decision. In this and the following chapter, approaches to modelling the tactical decision using functional models are discussed.

A number of significant issues have already been identified. These include the importance of dealing appropriately with uncertainty and knowledge in developing a functional model. Also key is the ability to present model output in such a way that decisions are facilitated.

5.1 Modelling the Decision Problem

The question of which species are suitable where can be phrased in a number of ways (Table 5.1). The problem can be interpreted from an ecology perspective (natural occurrence of species) or from a niche modelling perspective (suitable cultivation of species). Essentially, all reduce to the same decision problem: “What is the likelihood that species α is suitable at location β ?”

Ecological modelling	Niche modelling
Will species α be found at location β ?	Will species α be successful at location β ?
What is the likelihood that species α will be found at location β ?	What is the likelihood that species α will be successful at location β ?
Which species will be found at location β ?	Which species are likely to be successful at location β ?
Where will species α be found?	Where is species α likely to be successful?

Table 5.1 Decision problem formulations.

If the characteristics of location β are denoted by biophysical variables X_1, X_2, \dots, X_n , and a measure of success of species α , or measure of whether species α is present, is denoted by Y , then the relationship between these is given by:

$$Y = f(X_1, X_2, \dots, X_n) \quad (5.1)$$

where f is some function.

In terms of the decision problem, Y is called the *response variable* (also known as dependent variable), and X_1, X_2, \dots, X_n are the *predictor variables* (also known as independent variables).

Although the purpose of ecological modelling is different from that of niche modelling, it is clear that there is significant overlap in their approaches. There is an extensive body of literature on ecological modelling that examines the question of where a plant or animal species might be found, based on environmental factors. The resulting models are known as habitat models or distribution models, and they relate the geographical distribution of species to their environment (Guisan and Zimmermann, 2000). Clearly, the decision of what to plant where is related to habitat modelling.

There is also some overlap with multivariate classification techniques, such as techniques for classifying remotely sensed data. Remotely sensed data refers to imagery acquired from terrestrial, aircraft and satellite sensors (Civco, 1993). The techniques used in both types of modelling are largely similar; the distinction is drawn here because there tend to be two separate bodies of literature. In subsequent discussion, only habitat models will be referred to, but it is understood that the discussion applies equally to classification techniques.

In this chapter, a number of selection criteria are proposed, against which models may be evaluated. These include the input data requirements of the model, the type of output and the complexity of the model. Critical for the decision problem is the ability of the model to include both data and knowledge. Complexity of the model affects both structural uncertainty and translational uncertainty. Complexity relates directly to the model structure (structural uncertainty), and reducing complexity of the model structure reduces associated uncertainty. Because the structure of the model is then also easier to communicate, translational uncertainty is reduced.

The types of data required for modelling are then discussed, followed by a discussion on model validation. In the following chapter, a number of models will then be reviewed, based on the criteria identified in this chapter.

5.2 Model Selection Criteria

As discussed in Chapter 3, the purpose of a model is to represent a complex real world system. Models can improve understanding of systems and can be used to predict or describe an outcome within a system. A model usually begins as a concept of how a real world system functions (conceptual model) and is then developed into a structure (structural model). Once the structure of the model has been defined, parameters are specified. The model specification process usually assumes that the overall model structure is correct, and it attempts to calculate the best parameters for the model.

Passioura (1996) distinguishes between two types of models in the field of agronomy. ‘Scientific models’ aim to improve understanding of a system. In the case of agronomy, this is an understanding of the physiology and environmental interactions of crops. In contrast, ‘engineering models’ aim to be functional, providing sound predictions and sound management advice to farmers. Passioura argues that ‘scientific models’ are qualitative and chiefly educational in nature. However, the aim of agronomic simulation models should be accurate prediction on which to base sound advice. Models should be as simple as possible, require as little data as possible and be based on simple robust empirical relationships between variables (Passioura, 1996). The approach is still scientific, but in order to be applied to practical problems, it also needs to be functional.

Although the models considered here are not agronomic simulation models, the goal is for the model to be used by farmers and those advising them. Therefore, while the model should be scientifically and statistically robust, it is clear that the approach should lean towards ‘engineering models’, as defined by Passioura. Hence two of the criteria for model selection are to keep model complexity low and to minimise the amount of input data required, both in physical space and in attribute space.

Guisan and Zimmermann (2000) point out that some models are better suited to reflect theoretical findings and therefore model selection should not depend solely on statistical considerations. They consider different types of models and, in particular, the differences between empirical models and causal models. Empirical models are also known as phenomenological or statistical models. Causal models have also been termed mechanistic, physiological or process-based (Guisan and Zimmermann, 2000). Empirical models are statistical data-driven models. Conversely, causal models base predictions on biologically functional relationships and are therefore knowledge-driven.

Austin (2002) stresses the importance of ecological knowledge in habitat modelling. When ecological theory is inadequate, incorrect assumptions can be made about the structure of the model. In addition, when using spatial data in habitat modelling, classical statistical methods are often inappropriate because of spatial uncertainties and biases. According to Mugglin *et al.* (1999), spatial data is usually multivariate, multilevel, misaligned and often non-randomly missing (i.e., systematically missing). Both of these observations strengthen the case for requiring expert knowledge to be able to be incorporated in the model.

Openshaw (1996) shows how different methods are needed depending on the complexity of a system and scientific precision required (Figure 5.1). Recalling Rowe's (1994) classification of uncertainty, scientific precision relates to metrical uncertainty and system complexity relates to structural uncertainty. Conventional statistical and mathematical models are appropriate for simple systems with little complexity or uncertainty. However, if complexity is high then structural uncertainty will remain. Where a large amount of data is available but little is known about the structure, 'model-free' methods can be more suitable. 'Model-free' methods include Artificial Neural Networks (ANN), genetic programming and automated model-design systems (Openshaw, 1996). As complexity and imprecision increase and less numerical data is available, fuzzy systems may be more appropriate. These methods may not reduce metrical uncertainty as well as mathematical models do, but by better reflecting system complexity, they may reduce more structural uncertainty.

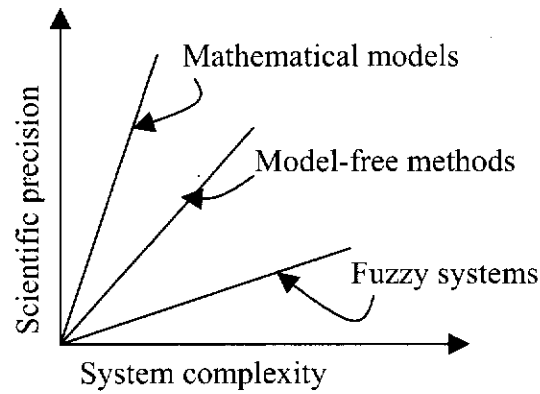


Figure 5.1 Openshaw's (1996) model of system complexity vs. scientific precision.

Even though Openshaw (1996) applies the word 'model' to mathematical methods only, 'model-free' methods and fuzzy systems are, arguably, still models, as they attempt to model the real world.

Guisan and Zimmermann (2000) state that most ecological models need to find a trade-off between precision (empirical models) and generality (causal models), a view that agrees with Openshaw's (1996) treatment.

The main criteria for model evaluation for this decision problem are summarised in Table 5.2.

Ability to work with small datasets
Ability to work with expert knowledge
Ability to predict range of species' responses
Low structural complexity
Easy to communicate
Easily implemented spatially

Table 5.2 Model evaluation criteria.

5.3 Data for Modelling

The types of data that are needed for modelling should be considered before the modelling approach is chosen. All models require response and predictor variables, but the nature of the variables, as well as quantity and quality of available data, influence model selection.

In GIS, representation of spatial data can be vector (points, lines and polygons) or raster (uniformly sized grid cells). In the kinds of models under discussion here, the area being modelled is generally raster-based, as this representation lends itself more readily to map algebra, and is, as Tomlin (1990) puts it, “better suited to interpretations of *where*”. However, data may sometimes be represented by polygons of varying shapes and sizes. Predictor variables are derived from the characteristics of factors within each cell or polygon, and the response variable is a value representing the predicted value or classification for that location.

Habitat and classification models are usually biophysical in nature, as opposed to social or economic. Biophysical here indicates that the model describes a biophysical outcome, but the system itself is not necessarily entirely biophysical. For example, species’ habitats may be influenced by human activity, such as road construction. In particular, in deciding what crop or forage to grow where, it is evident that socio-economic factors become important, such as distance to market and risk-aversion of the farmer. Therefore, in biophysical models the response variables are biophysical in nature, but the predictor variables may also be social and economic.

Data relating to the response variable in habitat modelling usually consists of a number of geo-referenced locations where the species in question is known to exist, i.e., presence data. In addition, sometimes data exists on where a species is known to not exist, i.e., absence data. However, absence data is much less reliable than presence data. In ecological habitat modelling, when a species is not found in a location, it may be true absence or it may be that the species was present but not found (false absence). It is also possible that the environmental conditions are favourable for the species, but the species is not present for other reasons. In niche modelling, absence of trials is often not because the species is unsuitable, but for any one of a number of other reasons. It is critical, therefore, to discriminate between absence of evidence and evidence of absence.

If the question is not only whether the species will *survive* but how well it will *thrive*, then more is needed than binary presence/absence data. A gradient is needed ranging from unsuitable, through marginally suitable and suitable, to highly suitable.

Data for response variables may also be represented in a knowledge base, in addition to, or instead of, geo-referenced data. In this case, information about conditions under which the species will be present or thrive is known, usually formalised in a rule base. This information may be based on geo-referenced data, but not necessarily from the area under investigation. Although this implies extrapolation to locations outside of the geographic area, the attributes (such as climate) of locations in the area under investigation should lie within the range represented in the database.

Independent data is the data that is known, or suspected, to be related to the habitat of a species (or the classification of land-use). Depending on the species these may include climatic, topographical, geological, edaphic and hydrological factors, as well as socio-economic factors.

5.4 Model Validation

An important step in model selection and development is model validation. Validation is necessary to evaluate how well the model performs and to determine where perhaps further investigation is necessary. Model validation is classically performed by retaining a portion of the dependent data as a testing or validation set. By applying the model to this data, an estimation can be made of how well the model is performing.

When there is not enough data available to withhold a portion for validation, techniques such as jack-knifing can be used. In a set of n data points, jack-knifing uses $n - 1$ data points to predict the value of the remaining one point. This process is iterated n times, resulting in a validation data set the same size as the original data set.

In classical statistical analysis, model performance is usually measured with statistics such as the correlation coefficient (R), the chi-squared statistic (χ^2), or, if sample sizes are small, binomial probabilities (Anderson *et al.*, 2003). When the model output is binary or discrete, other measures of performance are necessary. With presence/absence data, model performance is often analysed using a confusion matrix (Table 5.3).

	Observed	
Predicted	Present	Absent
Present	True presence	False presence
Absent	False absence	True absence

Table 5.3 Confusion matrix of predicted classification vs. observed classification.

The confusion matrix can be interpreted in a number of ways, including overall performance (total correct classification rate), sensitivity (in the sense of percentage where presence is correctly predicted), specificity (percentage where absence is correctly predicted), omission error (false negative rate), commission index (false positive rate), Cohen's kappa statistic, the odds ratio and the normalised mutual information statistic (NMI) (Anderson *et al.*, 2003; Manel *et al.*, 2001).

The validation techniques described above assume that there is enough data to allow robust statistical analysis. Guisan and Zimmermann (2000) suggest that in most statistical modelling, for each predictor variable used there should be at least ten observations in the least represented category of the response variable. Therefore, if we consider a simple binary classification model with only two predictor variables, then at least 40 observations are required for statistical robustness, and more if the data is unbalanced.

When this is not the case, validation becomes problematic. Coenen *et al.* (2001) recommend carrying out a number of test cases and comparing the results to those suggested by a domain expert. In the decision problem, this is the situation encountered. For each forage species, a limited amount of data is available for model specification, and even less for validation. Development of a causal model should reflect expert knowledge, so validating in this way will indicate whether this is in fact the case.

It is important to bear in mind that the model that correctly classifies the largest proportion of the test set is not necessarily the *best* model. In such a model, metrical uncertainty may be reduced at the expense of structural uncertainty. Reducing some, but not all, metrical uncertainty may be sufficient to allow the decision-maker to reach a valid conclusion. Reducing structural uncertainty may be more valuable to

the decision-maker, and this might be better achieved by considering a model that draws on expert knowledge. Translational uncertainty may be reduced by ensuring the model structure and output are transparent and easy to communicate.

5.5 Summary

In this chapter, criteria for model selection were introduced. For the functional model, the decision problem can be stated as follows: “What is the likelihood that species α is suitable at location β ?” In selecting a model to address the decision problem, a number of criteria should be considered. The first criteria are the ability to work with small datasets and expert knowledge and the ability to predict a range of species’ responses, rather than just ‘presence’ and ‘absence’. In addition, the model must display low structural complexity and must be easy to communicate and to implement spatially.

In the following chapter, a number of models will be considered that have been applied to habitat distribution modelling and classification modelling. Each model will be assessed based on the criteria given above. In addition, consideration will be given to how the model can be validated and the applicability of the model to tropical forage selection.

CHAPTER 6 MODELLING APPROACHES

The previous chapter discussed the purpose of modelling and identified a number of criteria to be considered in selecting a modelling approach.

The discussion will now turn to a description and comparison of potential modelling approaches. Techniques discussed are those most commonly used in environmental habitat modelling. Some methods are more empirical in nature and others are more causal. Each method is described, followed by some examples of how the method has been applied in the literature. Relative strengths and weaknesses are considered, followed by a brief discussion of possible applicability of the method to forage (and crop) selection.

6.1 Logistic Regression

6.1.1 Description

Traditional statistical approaches have been shown to be appropriate methods for modelling habitat distribution in many cases, especially when a large amount of data is available. However, it is generally recognised that environmental functions tend not to be linear, and therefore multiple linear regression is not often used. As Austin (2002) points out, agreement has not yet been reached on the expected shape of a response curve to an environmental gradient, but the shape is unlikely to be linear.

Logistic regression, however, is a popular technique in habitat modelling and is capable of producing good results, providing a number of assumptions are met. Logistic regression is useful for predicting a binary response from either continuous or categorical predictors. In the case of habitats, the binary response is typically presence/absence.

The logistic regression equation is of the form:

$$\text{logit}(P) = \beta_0 + \sum_i \beta_i X_i \quad (6.1)$$

where β_i are the regression constants and X_i are the regression variables. *Logit* is the natural logarithm of odds, written as:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) \quad (6.2)$$

If $Y = \text{logit}(P)$, then, from Equations 6.1 and 6.2, P can be written as:

$$P = \frac{e^Y}{1 + e^Y} \quad (6.3)$$

P is therefore a transformation of a multiple linear regression and is generally interpreted as ‘presence’ if $P \geq 0.5$ and ‘absence’ if $P < 0.5$.

Examples of models that have been developed using logistic regression are the distribution of otters (Barbosa *et al.*, 2003), swamp antechinus (Gibson *et al.*, 2003), wolves (Glenz *et al.*, 2001), red-crown cranes (Li *et al.*, 1997) and buzzard nests (Austin *et al.*, 1996). In addition, Aspinall (2002) used logistic regression to validate classification of vegetation species from remotely sensed data, and Brooker *et al.* (2002) applied the technique to infectious disease prediction. The predictors used in these studies range from two variables and 57 sites (Gibson *et al.*, 2003) to 25 independent variables and 6,187 sites (Barbosa *et al.*, 2003). Some of these methods also employ discriminant analysis to develop the predictive models. Some also use Principal Components Analysis (PCA) to reduce the number of factors used as dependent data (for example, Li *et al.*, 1997).

6.1.2 Strengths and Weaknesses

Strengths of logistic regression are that the method is well defined and statistically robust and various statistical measures exist for describing how well the model performs. However, a drawback of statistical models is their dependence on a large amount of data, both for specification and validation.

As with all regression problems, variables must also be uncorrelated. Incomplete attribute data cannot easily be accounted for in logistic regression, and instead, techniques must be utilised to complete missing data. Schafer (1997), for example, discusses various techniques for estimating and imputing missing values in databases. An additional weakness of logistic regression is that assumptions are made about the statistical distribution of the data (*logit* transformation of linearity).

As the output of logistic regression models is binary, they are generally validated by interpreting the confusion matrix of observations correctly and falsely classified (Table 5.3).

6.1.3 Applicability to Tropical Forage Selection

Although logistic regression is a robust method, it is dependent on relatively large datasets. In selecting suitable species for farmers in the tropics, a method is needed that works with relatively sparse datasets. The output of these models could feasibly be interpreted as a continuum from unsuitable to suitable, rather than as binary presence/absence. This would be necessary in a model used to predict species' success in tropical agriculture.

6.2 Generalised Linear Models and Generalised Additive Models

6.2.1 Description

Generalised Linear Models (GLM) and Generalised Additive Models (GAM) were first developed in the 1960s and have since been used extensively in ecological research (Guisan *et al.*, 2002). GLMs are extensions of linear models, allowing for non-linearity and non-constant variance structures in data. GAMs are a further extension of GLMs, where the only underlying assumption is that the functions are additive and the components are smooth (Guisan *et al.*, 2002). These models are particularly useful in ecology modelling because underlying data is usually highly non-linear and may take on many different distribution forms.

The equation for a GLM can be expressed as follows:

$$g(E(Y)) = LP = \beta_0 + \sum_i \beta_i X_i \quad (6.4)$$

where β_i are the regression constants, X_i are the regression variables and Y is the response variable. LP is a linear predictor, $E(Y)$ is the expected value of Y and g is a link function. The distribution of Y may be any of the exponential family of distributions, and the link function may be any monotonic differentiable function (Guisan *et al.*, 2002). When a binomial distribution is used with a logistic link then this approach is equivalent to logistic regression (Hirzel and Guisan, 2002).

GAMs are data-driven rather than knowledge-driven, with the shape of the response curve determined directly from the data (Lehmann, 1998). Therefore, they cannot be easily described using equations, but in general take the form:

$$Y \propto f(X_1) + g(X_2) + \dots + h(X_n) \quad (6.5)$$

where Y is the response variable, X_i are the predictor variables and f , g and h are various functions.

Both GLM and GAM have been used successfully in habitat modelling for a number of species. Laurance (1997) applied GLM with Poisson regression to predictions of bettong abundance. Carey and Brown (1994) used GLM with a log link function to identify suitable sites for a rare orchid in a future changed climate. Guisan *et al.* (1999) used GLM to predict multiple plant distributions. Lehmann (1998) applied GAM to submerged macrophyte distribution and Lehmann *et al.* (2002) used GAM to predict natural distribution and species abundance of ferns. Seoane *et al.* (2003) modelled the distribution of breeding birds with GAM.

6.2.2 Strengths and Weaknesses

GLMs and GAMs have the advantage over logistic regression that assumptions on the distribution of the data are relaxed, and ecologists tend to prefer them for this reason. Higher-order interactions can be accounted for and, with GAM, any response curve shape is possible. In addition, unlike regression models, GLM will

always yield predictions within the limits of observed values (Guisan and Zimmermann, 2000). However, these models still require relatively large amounts of data for specification. Guisan *et al.* (1999) found that GLM performs well with large amounts of data, but not so well for rare or uncommon species with only a small number of 'presence' records. Welsh *et al.* (1996) also found the fit to be poor for GLM for rare species. Some techniques have been developed to deal with this problem, such as zero-inflated regression (Pearce and Ferrier, 2001; Welsh *et al.*, 1996).

Elith *et al.* (2002) report that few ecological GLM present predictions with an indication of the uncertainty involved. Where uncertainty is considered, the focus is usually on metrical uncertainty. However, their work shows that attempts can be made to deal with other types of uncertainty, such as considering multiple GLM models to help describe and reduce structural uncertainty.

6.2.3 Applicability to Tropical Forage Selection

In essence, GLM and GAM are extensions of logistic regression and therefore face many of the same issues regarding applicability to forage selection. The methods are well suited to ecological presence/absence data, in particular where little is known about the relationship between predictor variables and species' presence. In the current decision problem, expert knowledge is available to help define these relationships and large amounts of species data often are not, making all data-driven methods less likely to be suitable.

6.3 Artificial Neural Networks

6.3.1 Description

Artificial Neural Networks (ANN) is an artificial intelligence technique based on a representation of the neural interactions in the human brain. Information is passed through a number of nodes, resulting in values or classifications. The model is initially assumed to be completely unspecified. The model learns how to classify data based on a training data set. Information flow can be in both directions (feed

forward and back propagation), and any number of levels of intermediate nodes can be present, although in practice most models use one or two (Figure 6.1).

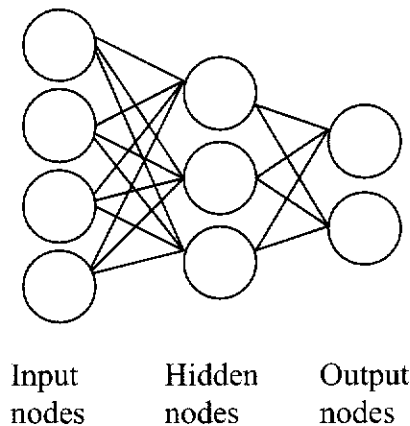


Figure 6.1 ANN with four input nodes, three hidden nodes in one intermediate level and two output nodes.

ANN has only recently been applied to the problem of ecological modelling. Bradshaw *et al.* (2002) used an ANN to model suitability of habitats for fur seal breeding and Drumm *et al.* (1999) modelled the habitat preference of the sea cucumber using ANN. Neural nets are often used more successfully for automated classification problems, for example, Skidmore *et al.* (1997) mapped forests using GIS and remotely sensed data based on a neural network and Pijanowski *et al.* (2002) looked at land use change. De la Rosa *et al.* (1999) evaluated land vulnerability using decision trees and neural networks. Civco (1993) used an ANN for land cover classification and mapping and Mas *et al.* (2003) modelled tropical deforestation using an ANN.

SPECIES (SPatial Evaluation of Climate Impact on the Envelope of Species) (Pearson *et al.*, 2002) couples an ANN with a climate-hydrological process model, and was used to model plant habitats under future climate scenarios in Great Britain. Antić *et al.* (2003) modelled the spatial distribution of soil groups in Croatia, using feed forward neural networks.

6.3.2 Strengths and Weaknesses

Neural nets have been shown to deal well with non-linear dynamic systems, discriminating between actual data and noise and processing previously unencountered patterns (de la Rosa *et al.*, 1999). ANN can therefore be a good approach when there is sufficient data for training and where little knowledge exists about the biologically functional relationships in the data.

Mas *et al.* (2003) found that ANN can easily be overspecified, reducing generalisation capabilities, and also that ANN tell us nothing about the functional form of the relationships between variables.

Therefore, it can be problematic to interpret the results of ANN modelling, precisely because of the lack of causal structure. The method offers no discernable benefits over other methods when data is sparse and the system is biologically fairly well understood. In fact, Manel *et al.* (1999) found some clear disadvantages when comparing ANN to conventional statistical methods, such as increased processing time and lack of identification of causal relationships.

6.3.3 Applicability to Tropical Forage Selection

In tropical forage selection, sparseness of data and the requirement of being able to incorporate expert knowledge make ANN a less promising method. For crops or forages for which little is known about their ecological processes, but where a large dataset is available, ANN could be useful, but this situation is unlikely to realistically occur.

6.4 Classification and Regression Trees

6.4.1 Description

Classification and Regression Trees (CART) (Figure 6.2) are derived from the concept of decision trees. A decision tree is a representation of all scenarios that can occur, depending on a sequence of decisions (Pearl, 1988). The purpose of decision trees is to analyse the utility of a given sequence of decisions, starting with a parent

node, and choosing an option at each decision node. This leads to branch nodes, where either new decisions are required or an outcome and its associated utility are given. In the latter case, the node is known as a leaf node. CART are similar in that a decision is taken at each node depending on the observation value. Leaves of the tree represent resulting classifications.

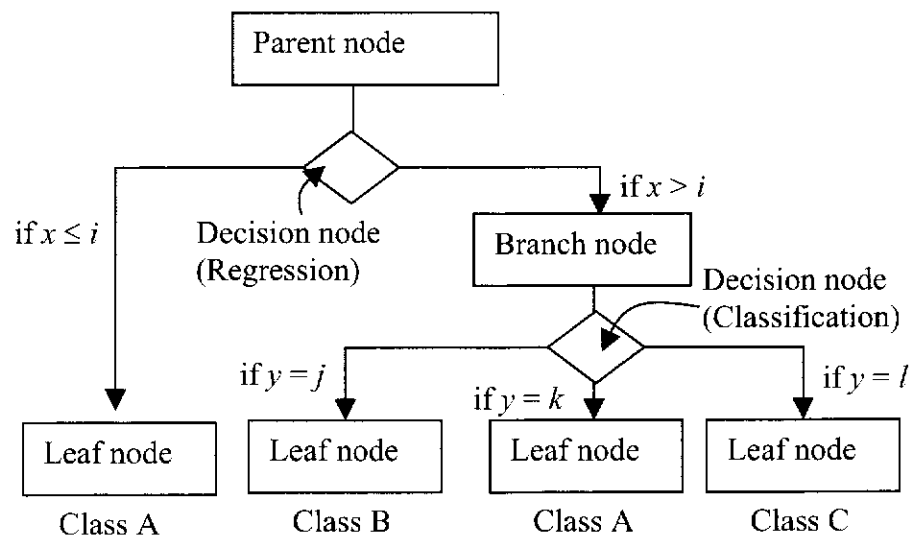


Figure 6.2 Classification and Regression Tree. x is continuous, y is categorical.

Classification and Regression Trees are thus called because input data can be both categorical (classification) and continuous (regression); the output however is always categorical. When CART consists of only categorical data, it is sometimes referred to as a Decision Tree, even though in reality classification is being carried out rather than decision analysis.

When a decision node relates to a continuous variable, regression analysis is used to define a rule that splits the dataset into two or more partitions. The rule is chosen, based on the data, to provide the greatest homogeneity within groups and the greatest heterogeneity between groups. When a decision node relates to a categorical variable, the dataset is partitioned into single categories or groups of categories.

Splitting of data into branches continues until some predefined limit is reached, such as number of data points in each leaf, total number of nodes or number of levels in

the tree. Various algorithms exist to determine the optimal size of a tree and to decide when to branch further and when to prune back.

Although decision node rules and the shape and size of the tree can be specified based on expert knowledge, in practice CART are usually data-driven.

CART is a technique that has only recently been applied to ecological habitat modelling. Debeljak *et al.* (2001) modelled red deer habitat suitability using machine-learned regression tree modelling, with nine independent variables. The output was presence/absence (inside home range / outside home range). They then combined models for four animals into a generic model based on expert opinion. Huettmann and Diamond (2001) predicted seabird distribution using both GLM and CART. Their model used 18 independent variables with nine seabird species. Kobler and Adamic (2000) identified brown bear habitat using automated machine learning to create a decision tree knowledge base, using 37 independent variables. They found that their decision tree mostly agreed with existing domain knowledge. De'Ath and Fabricius (2000) evaluated CART to analyse abundance of soft coral. Iverson *et al.* (1999) applied regression tree analysis to examine potential distribution of tree species under future climate change. Andersen *et al.* (2000) modelled desert tortoise habitat using CART.

CART has also been used in determining physical properties. Summerell *et al.* (2000) modelled clay distribution by building a decision tree using recursive partitioning. Lawrence and Wright (2001) applied CART to image classification using remotely sensed data. Evans and Caccetta (2000) used a decision tree classifier based on satellite images and landform data to identify areas at risk from dryland salinity.

Other data-driven rule-based methods include Genetic Algorithm for Rule-Set Prediction (GARP), an expert-system, machine learning approach to predictive modelling. It iteratively uses rule selection, evaluation and testing to produce a rule-set describing the species' habitat. The rules include environmental envelope definitions, logistic regression and categorical rules (Stockwell and Peterson, 2002). According to Anderson *et al.* (2003), GARP has proven successful in predicting

species' distribution in a wide variety of situations. It is especially useful with presence-only occurrence data.

6.4.2 Strengths and Weaknesses

Researchers have found that CART can deal with non-linear relationships, high-order interactions and missing values, whilst being simple to understand and yielding interpretable results (De'Ath and Fabricius, 2000).

Decision trees require a large amount of data to yield robust results, and can be sensitive to unbalanced data (i.e., discrepancy in the number of observations in each class) (Lawrence and Wright, 2001). However, where there is sufficient training data, CART generally performs well.

McKenny and Pedlar (2003) point out that the level of precision suggested by the splits at each node is probably not realised in the natural world. Because CART works by deciding a classification locally at each node, uncertainty surrounding the classification decision is not easily dealt with.

Although decision trees are usually built using machine-based learning, expert opinion can be incorporated in the decision rules (Debeljak *et al.*, 2001). Decision trees can also be useful for discovering patterns that can then be used to define a knowledge base. Decision trees are usually designed to predict presence/absence, although with enough data they can be used to classify multiple discrete states.

De'Ath and Fabricius (2000) found CART to be superior to linear models in their application. Andersen *et al.* (2000) found that CART facilitated clear and interpretable analysis.

Pontius *et al.* (2001) demonstrated that data-driven rule-based systems can be designed to work with data which varies in completeness, precision, currency and accuracy. Stockwell and Peterson (2002) found that genetic rule-based approaches such as GARP can develop accurate models based on relatively few data points.

As with other presence/absence models, decision trees are usually validated by counting true and false presences and absences. As the output of decision trees is always categorical, multi-state models can be validated by comparing correctly and incorrectly classified training data.

6.4.3 Applicability to Tropical Forage Selection

The amount of data required to specify robust trees is a drawback in the case of predicting species success for smallholder farmers in the tropics. However, the fact that expert opinion can be incorporated relatively easily is beneficial. In addition, trees can be easily interpreted for biological meaning. It is not clear, however, how to deal effectively with uncertainty in decision trees. CART could be a useful tool for organising data and incorporating expert knowledge. However, unless large amounts of data are available, other more knowledge-driven approaches are probably more appropriate for tropical forage selection.

6.5 Environmental Envelopes

6.5.1 Description

Environmental envelopes, or habitat envelopes, define an envelope in multi-dimensional attribute space within which the species is expected to be found. A number of different algorithms have been used to define these envelopes. A rectilinear envelope is equivalent to a very simple classification tree, and this concept of simple classification rules has long been used to derive classifications of vegetation and ecosystems (see for example Holdridge lifezones [Holdridge, 1967]).

With environmental envelopes, the response variable tends to be presence/absence, with responses classified as ‘present’ (or ‘suitable’) if they fall within a given percentile for all variables. An additional class of ‘marginal’ is often added for responses that fall within that percentile for some factors but outside for others (Figure 6.3).

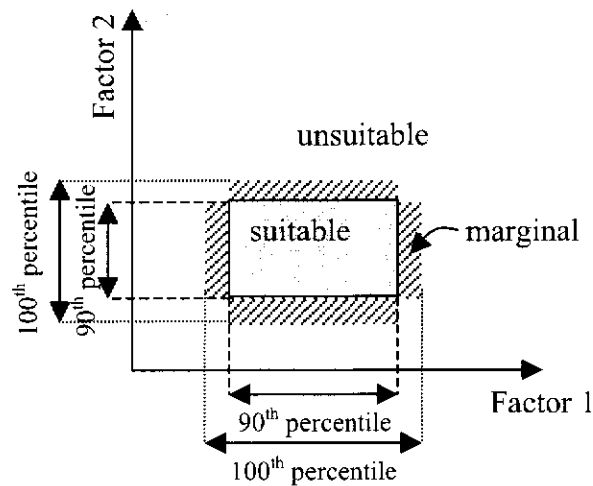


Figure 6.3 Environmental envelope for two factors, after Skidmore *et al.* (1996).

The most basic environmental envelopes are rectilinear. However, more sophisticated algorithms have been developed for defining environmental envelopes, including minimum bounding polygons and fuzzy clouds. In some cases, a habitat may be comprised of multiple, non-adjacent envelopes (e.g., FloraMap [Jones and Gladkov, 1999]).

Environmental envelopes are often formalised in specialised software packages with built-in GIS functionality. BIOCLIM (Busby, 1991) uses climate variables to form a multi-dimensional rectilinear environmental envelope for specific species. HABITAT (Walker and Cocks, 1991) extends the concepts of BIOCLIM, using convex polytope (multi-dimensional polygon) envelopes rather than rectilinear envelopes. BIOCLIM and HABITAT create continuous envelopes in multi-dimensional attribute space, including all data points (or all points within a given percentile). Therefore, any new location will be classified as 'suitable' if all its attributes' values fall within the given range. DOMAIN (Carpenter *et al.*, 1993) creates envelopes based on a point-to-point similarity metric known as the Gower metric, which indicates the degree of similarity between a new location and the most similar location in the database.

Booth and Jones (1998) used climatic mapping to define environmental envelopes for trees in Latin America. Booth (1995; 1999) combined climatic mapping with

growth simulation models to predict where trees will grow and how well. The main difference between BIOCLIM-like models and climatic mapping is that BIOCLIM was designed to study natural distributions, whereas climatic mapping programs are designed to assist plant introductions (Booth, 1999).

FloraMap (Jones and Gladkov, 1999) computes a climate probability model based on a set of species' collection points (known presence). PCA is used to produce a probability distribution in multiple dimensions. FloraMap also allows disjoint habitats to be defined for a single species (for example, native and naturalised populations of a species may have different distributions). In essence, FloraMap produces sophisticated fuzzy environmental envelopes, based only on species' occurrence locations.

Biomapper (Hirzel *et al.*, 2001) is software implementing Ecological Niche Factor Analysis (ENFA) to compute species' habitat suitability based on presence only data. ENFA is similar to PCA in that it reduces predictor variables to a few uncorrelated factors. Hirzel *et al.* (2001) explain that with ENFA, however, these factors retain ecological meaning.

6.5.2 Strengths and Weaknesses

A major advantage of environmental envelopes is the fact they can readily be interpreted in biological terms. This also means that where data is lacking, expert opinion can be used to delineate envelopes.

Environmental envelopes are usually validated using data not employed in the specification process. However, as with other methods reviewed, problems can arise when the data is only presence data, although some methods (e.g., Biomapper) claim to overcome these.

6.5.3 Applicability to Tropical Forage Selection

Environmental envelopes have already been used extensively for crop selection in tropical agriculture. Most databases on crop adaptation store environmental data in

formats that lend themselves to mapping using envelopes, even if this approach was not initially intended. For example, EcoCrop (FAO, 2000) lists minimum and maximum environmental criteria for various crops, sometimes including marginal boundaries. Therefore, environmental envelopes are a highly suitable method for use in tropical forage selection.

6.6 Fuzzy Rule-Based Methods

6.6.1 Description

Fuzzy logic (Zadeh, 1965) seeks to relax the crisp and deterministic classifications imposed by Boolean logic. Fuzzy membership generalises Boolean logic by assigning the value 1 to the state ‘true’, 0 to the state ‘false’ and allowing values between these two numbers. An example of a fuzzy membership function is given by:

$$\mu(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \quad (6.6)$$

where a is the limit below which x does not belong to the set and b is the limit above which x does belong to the set. All values of x between a and b are said to have fuzzy membership of the set and $\mu(x)$ is known as the fuzzy membership function. The shape of the function in Equation 6.6 is constant when x is less than a or greater than b (0 and 1 respectively) and linear when x lies between these values. Other function shapes are possible, including, but not limited to, triangular, trapezoidal, power, exponential and Gaussian functions (McBratney and Odeh, 1997). The only requirement is that $x \in [0, 1]$.

Methods are needed to combine multiple predictor variables X_i . In Boolean logic, two operators exist, namely, AND and OR. Fuzzy logic extends these concepts to fuzzy AND and fuzzy OR and adds the operators fuzzy algebraic product, fuzzy algebraic sum and fuzzy gamma operator (Bonham-Carter, 1994). Fuzzy AND and fuzzy OR are defined in Equations 6.7 and 6.8 respectively:

$$\mu(X_1, X_2, \dots, X_n) = \text{MIN}(\mu(X_1), \mu(X_2), \dots, \mu(X_n)) \quad (6.7)$$

$$\mu(X_1, X_2, \dots, X_n) = \text{MAX}(\mu(X_1), \mu(X_2), \dots, \mu(X_n)) \quad (6.8)$$

Fuzzy AND is defined by the smallest fuzzy membership of all variables and fuzzy OR is defined by the largest fuzzy membership of all variables. Fuzzy algebraic product is calculated from the product of all fuzzy memberships:

$$\mu(X_1, X_2, \dots, X_n) = \prod_{i=1}^n \mu(X_i) \quad (6.9)$$

Fuzzy algebraic sum is defined by:

$$\mu(X_1, X_2, \dots, X_n) = 1 - \prod_{i=1}^n (1 - \mu(X_i)) \quad (6.10)$$

The two operators above both take information from all variables, but fuzzy algebraic product is always ‘decreaseive’ (i.e., the result is always less than or equal to the smallest fuzzy membership) and fuzzy algebraic sum is always ‘increaseive’ (i.e., the result is always greater than or equal to largest fuzzy membership). Gamma operation (Equation 6.11) is a combination of these two operators:

$$\mu(X_1, X_2, \dots, X_n) = \left(\prod_{i=1}^n \mu(X_i) \right)^\gamma * \left(1 - \prod_{i=1}^n (1 - \mu(X_i)) \right)^{1-\gamma} \quad (6.11)$$

where γ is parameter between 0 and 1.

Numerous applications of fuzzy logic to soil science are found in the literature, including soil mapping (Assimakopoulos *et al.*, 2003), salinity changes (Metternicht, 2001) and soil classification and fuzzy measures of imprecisely defined soil phenomena (McBratney and Odeh, 1997). Sui (1992) applied fuzzy logic to land evaluation in China. Sasikala and Petrou (2001) assessed the risk of desertification after a forest fire using fuzzy logic. Mackinson (2000) describes CLUPEX, a fuzzy

logic expert system. MacMillan *et al.* (2000) applied a fuzzy logic rule-based system for automated landform classification.

6.6.2 Strengths and Weaknesses

GIS modellers have long argued that traditional Boolean logic is too deterministic to satisfactorily represent imprecisions and uncertainties in spatial data (see for example Sui, 1992; Zhang and Goodchild, 2002). Fuzzy logic allows both spatial uncertainty and attribute uncertainty to be explicitly modelled. Fuzzy logic also allows the inclusion of imprecise or vague expert knowledge.

However, one drawback is that the choice of which fuzzy operator to use is subjective. In addition, if inadequate expert knowledge is available, then rule definition may become problematic. Nevertheless, Mackinson (2000) successfully combined knowledge from a variety of sources in the knowledge base for CLUPEX. He points out that when knowledge is incomplete, rules can still be applied.

MacMillan *et al.* (2000) highlight the advantage of knowledge-based classifications over data-driven classifications for fuzzy rule bases, namely, that with a data-driven approach, the classification rules will be optimised for a particular site and knowledge-driven models will apply more generally. This observation applies for all data-driven and knowledge-driven models.

6.6.3 Applicability to Tropical Forage Selection

Fuzzy theory could be a valuable tool for forage selection. The many uncertainties, both in geographical and attribute space, could be addressed using fuzzy classification. Fuzzy logic provides a many-valued alternative to the binary nature of traditional Boolean logic, where values are true/false or presence/absence. This allows for classifications of 'marginally suitable', in addition to classifications of 'not suitable' and 'suitable'.

6.7 Bayesian Probability Models

6.7.1 Description

Bayesian methods provide a “formalism for reasoning under conditions of uncertainty, with degrees of belief coded as numerical parameters, which are then combined according to rules of probability theory” (Pearl, 1990). The term ‘Bayesian’ derives from Thomas Bayes, whose essay *Towards Solving a Problem in the Doctrine of Chances* was published in 1763 (reprinted in 1958). A Bayesian network consists of a Directed Acyclic Graph (DAG), linking nodes and rules for propagating probabilities from a parent node to its child nodes (Figure 6.4). DAG means that all links between nodes are directed (usually a cause-effect link) and that no loops exist within the network (acyclic). Each node represents a predictor or response variable.

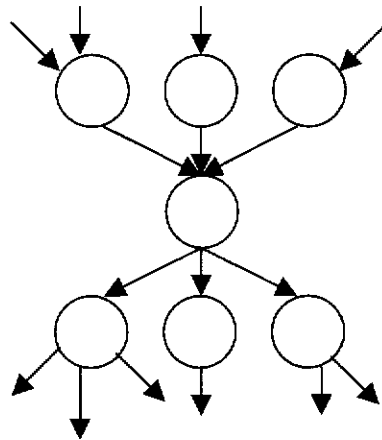


Figure 6.4 Typical node in a Bayesian schema, after Stassopoulou *et al.* (1998).

A simple Bayesian model defines prior and conditional probability distributions for each node and then uses combination rules to propagate conditional probability distributions through the network. The probability distributions may be derived from data, set by experts or defined from a combination of data and expert opinion. This process of combining probabilities produces conditional probabilities for each possible outcome. Where a node has multiple parents, a conditional probability table (CPT) is defined, defining probability distributions for each possible combination of parent node states.

The three basic axioms of probability theory for an event X are defined as follows:

$$\begin{aligned} 0 &\leq P(X) \leq 1 \\ P(X) &= 1 \text{ if } X \text{ is certain} \\ P(X_1 \text{ or } X_2) &= P(X_1) + P(X_2) \text{ if } X_1 \text{ and } X_2 \text{ are mutually exclusive} \end{aligned} \quad (6.12)$$

where $P(X)$ is the probability of event X occurring.

A ‘prior probability’ is an initial estimate that may be modified once more information becomes available. If Y is a response variable, then the prior probability of Y is denoted $P(Y)$. ‘Joint probability’ refers to the probability of two events occurring together, such as a trial site with low rainfall having a trial with a species that thrives. This is denoted by $P(X, Y)$, where X is a predictor variable (e.g., rainfall) and Y is a response variable. ‘Conditional probability’ is the probability of a response variable being in a given state, given that a predictor variable is in a particular state, and is denoted $P(Y | X)$.

Conditional probability can be calculated from prior probability and joint probability:

$$P(Y | X) = \frac{P(Y, X)}{P(X)} \quad (6.13)$$

From Equation 6.13, Bayes’ inversion formula can be derived:

$$P(Y | X) = \frac{P(Y)P(X | Y)}{P(X)} \quad (6.14)$$

Probability decomposition is given by:

$$P(Y) = \sum_i P(Y | X_i)P(X_i) \quad (6.15)$$

These equations form the basis for combining and updating probabilities in a Bayesian network. Further discussion on Bayesian networks can be found in Jensen (1996) and Pearl (1988), amongst others.

Bayesian modelling does not seek to predict exact outcomes, but rather the probabilities of various outcomes, given the effects of the input factors on each outcome. Therefore, if two variables strongly support a given outcome, then the combined effect of these two variables will produce a high probability for that outcome. Conversely, if one variable strongly supports an outcome but another variable does not, then the probability of that outcome occurring will be lessened. Uncertain data can be accounted for by reducing the probability that the data will support a given outcome.

If enough data is available, a Bayesian network can be learned from a database. This learning consists of both constructing an appropriate structure and calibrating parameters. Heckerman *et al.* (1995) suggest approaches to learning networks, finding high-scoring networks and evaluating learning algorithms.

‘Weights of evidence’ is a special case of Bayesian modelling, using natural logarithms of odds, or *logits* (Bonham-Carter, 1994). From Equation 6.2, the *logit* of Y given X is defined by:

$$\text{logit}(Y | X) = \ln \left(\frac{P(Y | X)}{P(\bar{Y} | X)} \right) \quad (6.16)$$

where $P(Y | X)$ is the conditional probability of Y given X , as defined above. \bar{Y} denotes *NOT* Y , and $P(\bar{Y}) = 1 - P(Y)$.

Based on the odds formulation, the concepts ‘sufficiency ratio’ and ‘necessity ratio’ are developed, also known as ‘likelihood ratios’. Taking the natural logarithms of these ratios yields weights of evidence (Bonham-Carter, 1994). The advantage of this formulation is that the likelihood ratios, and hence weights of evidence, can be interpreted to determine the relative importance of evidence, which may be easier for experts to estimate than probabilities. In addition, the *logit* formulation means that weights of evidence are additive, which simplifies computation.

Davis and Hall (2003) discuss the concepts of necessity, sufficiency and relevance, which are useful concepts in analysing the importance of predictor variables. These are defined as follows:

$$\begin{aligned}
 P(Y | X) \leq 1, P(Y | \bar{X}) &= 0 && \text{[necessary condition]} \\
 P(Y | X) = 1, P(Y | \bar{X}) &\leq 1 && \text{[sufficient condition]} \\
 P(Y | X) = 1, P(Y | \bar{X}) &= 0 && \text{[necessary and sufficient condition]} \\
 0 < P(Y | X) \leq 1, 0 \leq P(Y | \bar{X}) &\leq 1 && \text{[relevant or partially sufficient condition]}
 \end{aligned} \tag{6.17}$$

Dempster-Shafer belief models are an extension of Bayesian probability models, rejecting the Bayesian rule of additivity, in favour of a less restrictive formulation (Ducey, 2001). In Bayesian methods, ignorance in prior probability distributions is denoted with equal probability assignments. The Dempster-Shafer technique introduces the concepts of ‘belief’ and ‘plausibility’. The idea is that the true probability value will lie somewhere between these two values, and the level of uncertainty is embodied in these values. ‘Belief’ and ‘plausibility’ adhere to the following relation:

$$Bel(Y | X) \leq P(Y | X) \leq Pl(Y | X) \tag{6.18}$$

where *Bel* denotes belief and *Pl* denotes plausibility.

Bayesian probability modelling and related methods have been used in a number of habitat problems. Aspinall (1992) and Aspinall and Veitch (1993) applied Bayesian methods to modelling red deer habitats and curlew distributions in Scotland. Skidmore and Gauld (1996) compared a Bayesian technique with CART and environmental envelopes. Gu *et al.* (1996) used a belief network approach to examine the impact of climate change on faba bean production in Scotland. Asadi and Hale (2001) applied weights of evidence to map potential gold deposits in Iran.

Ducey (2001) applied Dempster-Shafer theory to forest management decisions. He found Dempster-Shafer a promising alternative for decision-making in environments with limited data and considerable uncertainty, where the goal is to select the best decision from a number of management alternatives. The main strength of the

Dempster-Shafer approach is in the way uncertainty is represented. Belief and plausibility can be interpreted as lower and upper limits on the possible values of a probability in any given situation.

Bayesian approaches have also been implemented in conjunction with other modelling techniques. For example, Mac Nally *et al.* (2003) used Poisson regression together with Bayesian modelling to predict butterfly species richness. Hooten *et al.* (2003) modelled plant species distribution using a generalised linear mixed model in a hierarchical Bayesian framework.

6.7.2 Strengths and Weaknesses

A major strength of probabilistic modelling is the ability to deal with uncertainty. Because Bayesian methods work with probabilities rather than absolute values, uncertainties can be explicitly included in the probability distributions and propagated through the model. Another strength is the ability to easily incorporate expert knowledge. Most people find Bayesian networks easy to construct and interpret (Heckerman, 1995).

Aspinall (1992) found a number of advantages in using Bayesian methods over alternatives, including statistical robustness of the method and the ability to quantify error and variation.

Seidel *et al.* (2003) compared rule-based heuristic decision support with Bayesian-based decision support. They claim that multiplying probabilities better represents the methods human experts apply than adding values in a rule-based system.

A Bayesian network can also be completely data-driven, with structure and probability distributions learned from the data. Because the representation has formal probabilistic semantics, it is suitable for statistical manipulation (Heckerman, 1995).

Proponents of Dempster-Shafer theory criticise Bayesian approaches because of difficulties expressing incomplete information or partial belief (Ducey, 2001). Dempster-Shafer models overcome this problem.

6.7.3 Applicability to Tropical Forage Selection

The decision problem is characterised by uncertainty, and probability-based methods are well equipped to incorporate uncertainty. In addition, the ability to incorporate both data and expert knowledge makes probability-based modelling appealing. Bayesian models can have varying degrees of complexity, but even complex models have clear biological meaning. Probability-based modelling is a promising approach to the problem of selecting tropical forages.

6.8 Other Methods

6.8.1 Description

Other methods that have been applied to ecological modelling include Boolean overlay, Canonical Correspondence Analysis (CCA) and cellular automata.

Boolean overlay is a very simple rule-based method, which is very easily implemented in GIS packages (see for example Bonham-Carter, 1994). CCA is an ordination technique based on reciprocal averaging of species and site scores (Guisan and Zimmermann, 2000).

In systems using cellular automata (for example PANTHER [Cramer and Portier, 2001]), the evaluation of a cell may have a direct impact on the evaluation of a neighbouring cell. This type of system is usually used to define the habitats of animals that are highly mobile over time, such as panthers (Cramer and Portier, 2001) and brown bears (Kobler and Adamic, 2000).

In the current decision problem, these methods appear less useful than those already discussed, and hence are not analysed here in further detail.

6.9 Conclusions

The main models considered were logistic regression, GLM, GAM, ANN, CART, environmental envelopes, fuzzy rule-based methods and Bayesian probability models. Some of these methods are empirical, data-driven methods and others are causal, knowledge-driven methods. Some techniques can be either, or a combination of both. The criteria of working with small datasets and expert knowledge simultaneously dictate that models that can be at least partly knowledge-based are preferable. Models depending mostly on expert knowledge will generally display low structural complexity and ease of communication.

For the decision problem of selecting forages, environmental envelopes are promising, and have already been spatially implemented in many habitat distribution problems. Fuzzy rule-based methods deal well with uncertainty and expert knowledge. Finally, Bayesian probability methods allow the combination of both data and knowledge and handle uncertainty well. It is proposed that a simple Bayesian probability model is well suited to the decision problem. It is envisaged that environmental envelope concepts and fuzzy rule-based methods display some overlap with a spatial implementation of a simple Bayesian model. Therefore, in the following chapter, fuzzy environmental envelopes are discussed, followed by a discussion on Bayesian modelling, before proceeding to the specification of a functional model to address the decision problem.

6.10 Summary

The purpose of this chapter was to review functional models which could be applied to the decision problem, that is, deciding which species are suitable in which locations. A number of criteria for selecting an appropriate model were defined, namely, ability to work with small datasets, ability to work with expert knowledge, ability to predict the full range of species' responses, low structural complexity, ease of communication and ease of spatial implementation.

This review contributes greatly to the process of selecting suitable models for problems similar to the decision problem discussed in this research. Often, models are selected and applied with little consideration to alternative models. A number of

review papers do exist which compare models (see for example Austin, 2002; Kriticos and Randall, 2001; Guisan and Zimmermann, 2000; Guisan *et al.*, 1999; Hill *et al.*, 1997; Manel *et al.*, 1997; Skidmore and Gauld, 1996). The majority of these only compare two or three techniques, although admittedly this is usually empirical comparison, and comparing more techniques is likely to be intractable. Only a handful of papers, however, present a theoretical comparison of a larger number of methods (e.g., Kriticos and Randall, 2001; Guisan and Zimmermann, 2000).

The following chapter proposes a probabilistic GIS model to address the decision problem.

CHAPTER 7. PROBABILISTIC GIS MODEL

In the previous chapter, a number of models were described which could potentially be applied to the decision problem. A number of criteria were suggested, and the most promising models for further investigation were environmental envelopes, fuzzy rule-based models and Bayesian probability models. Therefore, these approaches will now be discussed in more depth, and a probabilistic GIS method will be proposed as a model to adequately address the decision problem.

7.1 Fuzzy Envelopes

Environmental envelopes and rule-base methods were introduced in the previous chapter. Envelopes are defined in multi-dimensional variable space within which a species is expected to be found or expected to succeed. Fuzzy rule-based methods use fuzzy logic to produce a rule-base for classification. Here, a combination of these two methods is considered, which can be termed ‘fuzzy envelopes’.

Consider the forage species *Stylosanthes guianensis*, for which Ecocrop (FAO, 2000) publishes an optimal annual rainfall range of 900-2000mm, an absolute annual rainfall range of 500-4000mm, optimal soil pH of 4.5 – 6, and absolute soil pH of 4 – 7.7. Recalling Figure 6.3, this information can be graphically displayed (Figure 7.1).

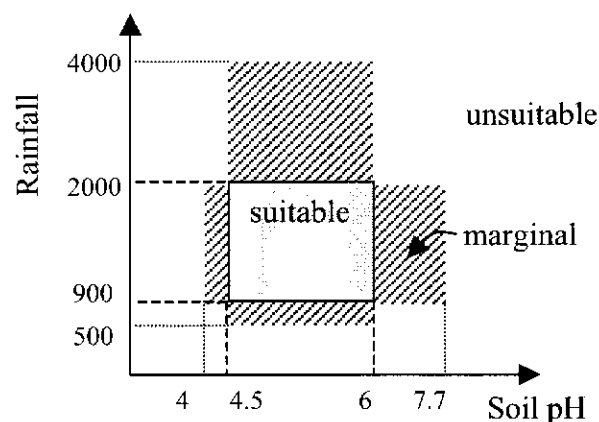


Figure 7.1 Environmental envelope for *S. guianensis* related to soil pH and rainfall.

The information on where the species is suitable can be written as a series of rules, namely:

$$\begin{aligned} &\text{IF soil pH} \in [4.5, 6] \text{ AND rainfall} \in [900, 2000] \\ &\quad \text{THEN } \textit{SUITABLE}, \\ &\quad \text{ELSE } \textit{NOT SUITABLE} \end{aligned} \tag{7.1}$$

where the square brackets denote a closed interval, that is, $4.5 \leq \text{soil pH} \leq 6$ and $900 \leq \text{rainfall} \leq 2000$.

In order to generalise the notation, let X^1 be the first predictor variable and X^2 the second predictor variable (in the example above, soil pH and rainfall respectively), and let x_a^1 and x_b^1 be the lower and upper limits of X^1 , and let x_a^2 and x_b^2 be the lower and upper limits of X^2 . Let S denote suitability of a species, where $S = 1$ means the species is suitable and $S = 0$ means the species is not suitable. Then the statement above can be rewritten as:

$$\begin{aligned} &\text{IF } X^1 \in [x_a^1, x_b^1] \text{ AND } X^2 \in [x_a^2, x_b^2] \\ &\quad \text{THEN } S = 1 \\ &\quad \text{ELSE } S = 0 \end{aligned} \tag{7.2}$$

This can be expanded for n variables to:

$$\begin{aligned} &\text{IF } X^1 \in [x_a^1, x_b^1] \text{ AND } X^2 \in [x_a^2, x_b^2] \text{ AND } \dots \text{ AND } X^n \in [x_a^n, x_b^n] \\ &\quad \text{THEN } S = 1 \\ &\quad \text{ELSE } S = 0 \end{aligned} \tag{7.3}$$

However, as also discussed in the previous chapter, the addition of marginal classes, where the species is neither wholly suitable nor wholly unsuitable, is often desirable. This is displayed graphically in Figure 7.1, where *S. guianensis* will be marginally suitable if pH is between 4 and 4.5 or between 6 and 7.7, and if rainfall is between 500 and 900mm or between 2000 and 4000mm, with the other variable remaining between the 'suitable' bounds. Equation 7.2 then becomes:

$$\begin{aligned}
& \text{IF } X^1 \in [x_a^1, x_b^1] \text{ AND } X^2 \in [x_a^2, x_b^2] \\
& \quad \text{THEN } S = 1 \\
& \text{ELSE IF } (X^1 \in [x_\alpha^1, x_a^1] \text{ OR } X^1 \in [x_b^1, x_\beta^1]) \text{ AND } X^2 \in [x_a^2, x_b^2] \\
& \quad \text{THEN } S = s \\
& \text{ELSE IF } X^1 \in [x_a^1, x_b^1] \text{ AND } (X^2 \in [x_\alpha^2, x_a^2] \text{ OR } X^2 \in [x_b^2, x_\beta^2]) \\
& \quad \text{THEN } S = s \\
& \text{ELSE } S = 0
\end{aligned} \tag{7.4}$$

where subscript α and β denote the lower and upper limit of absolute values, and s is some value between 0 and 1 denoting marginal suitability. Expanding this to the case of n variables, 7.4 can now be written:

$$\begin{aligned}
& \text{IF } X^i \in [x_a^i, x_b^i] \forall i \\
& \quad \text{THEN } S = 1 \\
& \text{ELSE IF } (X^i \in [x_\alpha^i, x_a^i] \text{ OR } X^i \in [x_b^i, x_\beta^i]) \text{ AND } X^j \in [x_a^j, x_b^j] \forall j, j \neq i \\
& \quad \text{THEN } S = s \\
& \text{ELSE } S = 0
\end{aligned} \tag{7.5}$$

In essence, Equation 7.5 states that if all variables are suitable, then the species is suitable. If all variables bar one are suitable, and one variable is marginal, then suitability is marginal. In all other cases the species is not suitable. It may better reflect reality to state that suitability is also marginal if more than one variable is marginal. This requires only a slight variation to Equation 7.5, namely:

$$\begin{aligned}
& \text{IF } X^i \in [x_a^i, x_b^i] \forall i \\
& \quad \text{THEN } S = 1 \\
& \text{ELSE IF } (X^i \in [x_\alpha^i, x_a^i] \text{ OR } X^i \in [x_b^i, x_\beta^i]) \text{ AND } X^j \in [x_a^j, x_b^j] \forall X^i \in \{\mathbf{X}^1\}, X^j \in \{\mathbf{X}^1\} \\
& \quad \text{THEN } S = s \\
& \text{ELSE } S = 0
\end{aligned} \tag{7.6}$$

where $\{\mathbf{X}^1\}$ and $\{\mathbf{X}^j\}$ are two exhaustive and mutually exclusive sets of all predictor variables. The suitability value S is therefore set to s when all predictor variables are

either between optimal and absolute limits (X^d) or within optimal limits (X^j), but none are outside of optimal limits.

If the three categories ‘suitable’, ‘marginal’ and ‘not suitable’ are adequate, then Equation 7.6 is satisfactory to model the suitability of a species. However, in reality, suitability is likely to be more of a gradient. For example, in Equation 7.6, the modeller may wish to differentiate between cases where most variables are suitable and cases where most variables are marginal (but all variables are either marginal or suitable, that is, $S = s$). In this case, S could take two values: s^1 meaning marginal but tending towards unsuitable and s^2 meaning marginal but tending towards suitable. Equation 7.6 would then have the added clauses:

IF $\{X^j\} > \{X^d\}$
 THEN $S=s^1$
 ELSE IF $\{X^j\} \leq \{X^d\}$
 THEN $S=s^2$

It is easy to imagine how this could be extended to multiple values of S . It is also clear that the meanings of the values of S become linguistically more difficult to define the more partitions there are of S . Even if it is only an estimate, there is merit in assigning numerical values to S . Rather than assigning discrete values to S , it can also be assigned a linear function, where the value of S varies between 0 and 1. This differs from standard linear regression in that above a certain threshold, S is set to a constant of 1 and, similarly, below a certain threshold, S is set to a constant of 0. The function where values vary between 0 and 1 only applies to values within marginal thresholds. Therefore, the model is constrained to producing values within the desired range, namely $[0,1]$.

This concept is derived from the theory of fuzzy sets (Zadeh, 1965). Fuzzy logic was discussed in Section 6.6 in the previous chapter.

Fuzzy logic focuses on ambiguities in describing events. In the previous equations, the value of s is ambiguous. The only information known is that it lies somewhere between 0 ('unsuitable') and 1 ('suitable'). This information is displayed graphically in Figure 7.2, with S as a linear function when conditions are marginal. If soil pH is 4.4, say, then suitability is 0.8.

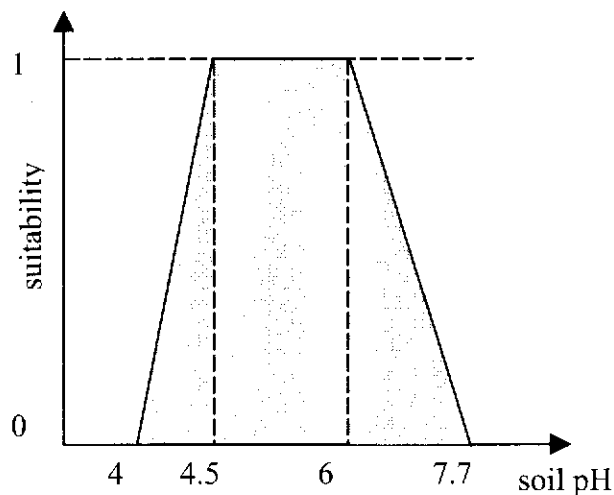


Figure 7.2 Fuzzy suitability of soil pH.

For n variables, each with fuzzy marginal suitability, the problem then arises of how to combine these assignments of suitability. Five fuzzy operators were described in Equations 6.7 – 6.11. Each of these combines fuzzy values in slightly different ways, and the choice of which fuzzy operator to use is subjective.

Returning, however, to the idea of discrete values of S , Figure 7.2 can be simplified into a stepped function where S is marginal (Figure 7.3). The purpose of this simplification is to facilitate the calculation of suitability when multiple variables are present. Rather than calculating suitability along multiple continuous gradients, it need only be calculated for a finite number of cases.

The question still remains of how exactly the values of S should be interpreted. From Figure 7.3, if soil pH is 6.5, then $S=0.75$. This can be interpreted as:

- Conditions are 75% suitable (fairly suitable) for the species
- Conditions are 75% likely to be suitable for the species

- Conditions will be suitable for the species 75% of the time

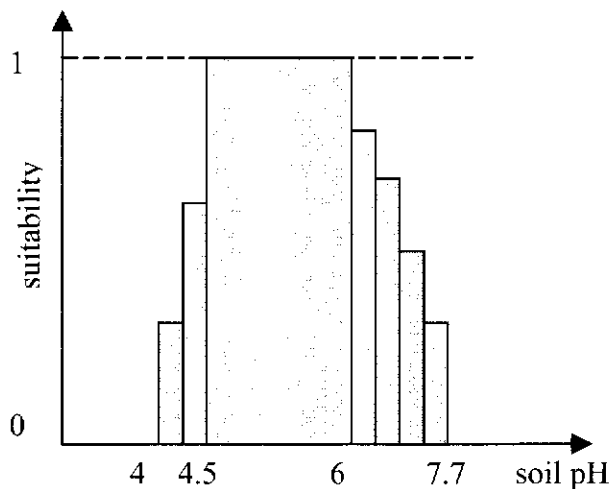


Figure 7.3 Stepped suitability of soil pH.

The first statement creates a new classification ('75% suitable'). The second statement maintains there are only two classifications ('suitable' and 'not suitable'), and it assigns a probability value of 75 percent to the class of 'suitable' (implying the conditions are 25 percent likely to be 'unsuitable'). The final statement differs from the second only in that it is now cast in frequentist terms, namely, if 100 instances of the species are observed with soil pH 6.5, then 75 of them will thrive and 25 of them will not.

It may be desirable to extend the model by creating both new classifications and assigning probabilities to these classifications. Hence, the question may be: "If soil pH is 6.5, what is the probability that the species will thrive (i.e., conditions are 'suitable'), what is the probability that the species will survive (i.e., conditions are 'marginally suitable'), and what is the probability that the species will not survive (i.e., conditions are 'unsuitable')?" Rather than assigning a single probability value to each possible value of pH, a distribution of probabilities needs to be specified, specifying the probability that a species will be 'suitable', 'marginally suitable' or 'unsuitable' for each value of pH.

The discussion will now turn to Bayesian models. At the end of the chapter, a method will be proposed based on the discussion of fuzzy envelopes and the discussion on Bayesian modelling.

7.2 Bayesian Models

Bayesian models were introduced in the previous chapter, and identified as a good candidate for modelling the decision problem. The specification of the model structure is discussed here. The discussion and formulation is drawn from a number of sources, including Bonham-Carter (1994), Corner *et al.* (2002) and Pearl (1988).

7.2.1 Formulation

Let X be an independent (predictor) variable and Y be a dependent (response) variable, and let y_i be a possible state of Y and x_j be a possible state of X . If Y has n possible states and X has m possible states, then by the total probability rule:

$$\begin{aligned} \sum_{i=1}^n P(Y = y_i) &= 1 \\ \sum_{j=1}^m P(X = x_j) &= 1 \end{aligned} \tag{7.7}$$

where $P(Y = y_i)$ denotes the probability that Y is in state y_i .

$P(Y)$ is the probability distribution over all possible states y_i . Say Y can take three different states: ‘not suitable’, ‘marginally suitable’ and ‘suitable’. Then the probability distribution could be written $P(Y) = (0.2, 0.3, 0.5)$, meaning that there is a 20 percent probability that the species is not suitable, a 30 percent probability that the species is marginally suitable and a 50 percent probability that the species is suitable. Note that the three probabilities sum to 1. This distribution can be displayed graphically (Figure 7.4).

Note that this has a different meaning from the suitability distribution in Figure 7.3. Fuzzy sets describe a degree of membership; in the example in Figure 7.3 this is the degree of membership in a set called ‘suitability’. In Figure 7.4 the interpretation is

probability of occurrence. It is also possible to define the response variable (and the predictor variables) as a continuous function, in which case the probability distribution is characterised by an integral. However, the discussion here will only consider distributions over discrete states.

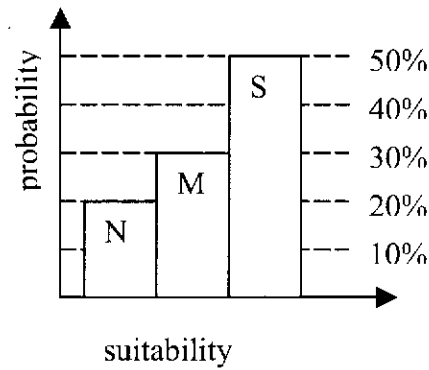


Figure 7.4 Probability distribution with three states. N='not suitable', M='marginally suitable', S='suitable'.

The conditional probability $P(Y = y_i | X = x_j)$ denotes the probability of y_i being the state of Y given that x_j is the state of X . For simplicity this can be written as $P(y_i | x_j)$. From Equation 6.13, this can be written:

$$P(y_i | x_j) = \frac{P(y_i, x_j)}{P(x_j)} \quad (7.8)$$

Similarly, the probability of x_j given y_i can be written:

$$P(x_j | y_i) = \frac{P(x_j, y_i)}{P(y_i)} \quad (7.9)$$

By definition, the probability of both x_j and y_i occurring is equal to the probability of both y_i and x_j occurring:

$$P(x_j, y_i) = P(y_i, x_j) \quad (7.10)$$

From 7.8, 7.9 and 7.10, the following equality can be derived:

$$P(x_j | y_i) = \frac{P(x_j)P(y_i | x_j)}{P(y_i)} \quad (7.11)$$

and similarly:

$$P(y_i | x_j) = \frac{P(y_i)P(x_j | y_i)}{P(x_j)} \quad (7.12)$$

Now consider a variable X which has m mutually exclusive and exhaustive states, $\{x_1, x_2, \dots, x_m\}$, then Equation 7.13 holds.

$$P(y_i) = P(y_i, x_1) + P(y_i, x_2) + \dots + P(y_i, x_m) = \sum_{j=1}^m P(y_i, x_j) \quad (7.13)$$

Substituting Equation 7.8 into the right-hand side of 7.13 gives:

$$P(y_i) = \sum_{j=1}^m P(x_j)P(y_i | x_j) \quad (7.14)$$

Substituting Equation 7.14 into Equation 7.11 gives:

$$P(x_j | y_i) = \frac{P(x_j)P(y_i | x_j)}{\sum_{j=1}^m P(x_j)P(y_i | x_j)} \quad (7.15)$$

Assuming Y has n mutually exclusive and exhaustive states, $\{y_1, y_2, \dots, y_n\}$, then Equation 7.16 can be derived in a similar fashion.

$$P(y_i | x_j) = \frac{P(y_i)P(x_j | y_i)}{\sum_{i=1}^n P(y_i)P(x_j | y_i)} \quad (7.16)$$

Equations 7.15 and 7.16 provide mechanisms for inverting causality. If the probability that y_i causes x_j is known (i.e., the probability of x_j given y_i), then the

probability can be calculated that if x_j is observed, then y_i is the cause, and vice versa. Framing this in the context of the decision problem, $P(y_i | x_j)$ is the probability that a species will succeed under certain conditions (say, rainfall > 800mm). Conversely, $P(x_j | y_i)$ is the probability that if a species succeeds, it is because it receives more than 800mm of rain in that location.

7.2.2 Calculating Posterior Probabilities Under Conditional Independence

The discussion will now consider the problem of calculating the posterior probability of a response based on multiple predictor variables. Suppose there are l independent variables denoted X^1, X^2, \dots, X^l . Say each variable can take on m different states, then let these states be denoted by x_j^k where $j = 1$ to m and $k = 1$ to l . Note, however, that each variable X^k need not have the same number of states m . For example, it might be convenient to split soil pH into three classes (say ‘acid’, ‘neutral’ and ‘alkaline’) and rainfall into five classes. Therefore, the number of states that variable X^k can take is denoted by m_k . Then their possible states can be denoted x_{jk}^k , for $j_k = 1$ to m_k and $k = 1$ to l . Then substituting a combination of states x_{jk}^k into Equation 7.16 yields:

$$P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) = \frac{P(y_i)P(x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l | y_i)}{\sum_{i=1}^n P(y_i)P(x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l | y_i)} \quad (7.17)$$

In order to proceed, an assumption of conditional independence (CI) of the independent variables X^k is made. Independence of variables is defined by:

$$P(x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) = P(x_{j1}^1)P(x_{j2}^2) \dots P(x_{jl}^l) = \prod_k P(x_{jk}^k) \quad (7.18)$$

and conditional independence can be thus be written:

$$P(x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l | y_i) = \prod_{k=1}^l P(x_{jk}^k | y_i) \quad (7.19)$$

Substituting 7.19 into the numerator and substituting 7.11 into the denominator of 7.17 gives:

$$P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) = \frac{P(y_i) \prod_{k=1}^l P(x_{jk}^k | y_i)}{\sum_{i=1}^n P(y_i) \frac{P(x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l)}{P(y_i)}} \quad (7.20)$$

Simplifying the denominator and substituting with 7.18 gives:

$$P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) = \frac{P(y_i) \prod_{k=1}^l P(x_{jk}^k | y_i)}{\sum_{i=1}^n \left(\prod_{k=1}^l P(x_{jk}^k) P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) \right)} \quad (7.21)$$

The product term in the denominator may be moved outside the summation, as it is independent of i , and rewriting this term in the numerator gives:

$$P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) = \frac{P(y_i) \prod_{k=1}^l \left(\frac{P(x_{jk}^k | y_i)}{P(x_{jk}^k)} \right)}{\sum_{i=1}^n P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l)} \quad (7.22)$$

It can be seen that the denominator is a normalising factor and is the same for all i . Therefore, it is sufficient to compute:

$$P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) \propto P(y_i) \prod_{k=1}^l \left(\frac{P(x_{jk}^k | y_i)}{P(x_{jk}^k)} \right) \quad (7.23)$$

for all states y_i of Y and then normalise across all values of i .

If $P(x_{jk}^k | y_i)$ is known for all i, j, k then Equation 7.23 can be used to calculate posterior probabilities under the assumption of conditional independence. However, if, instead, $P(y_i | x_{jk}^k)$ is known, then 7.23 can be rewritten using 7.11 to yield:

$$P(y_i | x_{j1}^1, x_{j2}^2, \dots, x_{jl}^l) \propto P(y_i) \prod_{k=1}^l \left(\frac{P(y_i | x_{jk}^k)}{P(y_i)} \right) \quad (7.24)$$

We now have methods for calculating posterior probabilities depending on which conditional probabilities are known. Values of y_i denote probability distribution over variable Y . Given all possible combinations of $P(y_i | x_{jk}^k)$, for all i, j, k , a full conditional probability table (CPT) can be created (Figure 7.5).

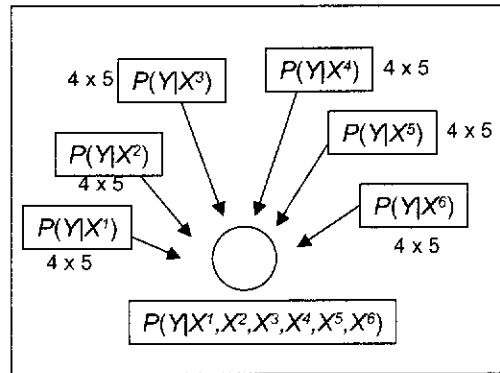


Figure 7.5 Populating a CPT for 6 conditionally independent variables X^k with 5 states each and a dependent variable Y with 4 states.

As an illustrative example, consider a case with two predictor variables, each of which has two possible states ('high' and 'low'), and a response variable, also with two possible states ('suitable' and 'not suitable'). A full CPT, derived using Equation 7.24, will provide a value for each conditional probability, namely:

$$\begin{array}{ll} P(y = \text{suitable} | x^1 = \text{high}, x^2 = \text{high}) & P(y = \text{notsuitable} | x^1 = \text{high}, x^2 = \text{high}) \\ P(y = \text{suitable} | x^1 = \text{high}, x^2 = \text{low}) & P(y = \text{notsuitable} | x^1 = \text{high}, x^2 = \text{low}) \\ P(y = \text{suitable} | x^1 = \text{low}, x^2 = \text{high}) & P(y = \text{notsuitable} | x^1 = \text{low}, x^2 = \text{high}) \\ P(y = \text{suitable} | x^1 = \text{low}, x^2 = \text{low}) & P(y = \text{notsuitable} | x^1 = \text{low}, x^2 = \text{low}) \end{array}$$

The dimension of the full CPT will be $l + 1$ (in the example above, the CPT is three-dimensional), where l is the number of predictor of variables. If the response variable has n possible states and each predictor variable x^j has k_j possible states, then the number of entries in the CPT is $n \times k_1 \times k_2 \times \dots \times k_l$.

7.2.3 Testing for Conditional Independence

The main assumption in the above approach is that of conditional independence (CI). If two variables are conditionally independent, then knowing the state of one variable has no bearing on the probability distribution of another variable once the state of the response variable is known. As Aspinall (1992) points out, the requirement of CI is often not met when dealing with environmental data.

The probability of datasets being conditionally dependent can be lessened by using fewer datasets. In addition, some authors (Corner *et al.*, 2002; Aspinall, 1992) question whether CI is operationally important. Corner *et al.* (2002) opine that functional independence is more critical. Even if two datasets are statistically dependent (for example, because they are derived from a common source), if they have different meanings in the model then it may still be valid to include them both. Pearl (1988) points out that in human reasoning, dependencies can usually be easily detected, even if this is difficult to ascertain numerically.

It may still be important however to consider CI. Various tests can be employed to check whether data is conditionally independent, including χ^2 , entropy and Cohen's kappa (Bonham-Carter, 1994). These tests can also be used to simply compare two datasets, for example, to check correlation between two maps. Low values mean that the maps are not correlated and high values imply correlation. This is a different, but related, concept to conditional independence. With CI, the aim is to examine whether the calculated (expected) values for joint probabilities are close to what would be observed if the full CPT for two variables X^1 and X^2 were known.

The chi-squared statistic compares expected values with observed values. In testing for conditional independence between two variables, the observed value is given by:

$$Obs = P(y_i | x_{j1}^1, x_{j2}^2) \quad (7.25)$$

and the expected value is given by:

$$Exp = P(y_i | x_{j1}^1)P(y_i | x_{j2}^2) \quad (7.26)$$

if X^1 and X^2 are conditionally independent. Chi-square related to y_i is then given by:

$$\chi^2 = \sum_{j1} \sum_{j2} \frac{(P(y_i | x_{j1}^1, x_{j2}^2) - P(y_i | x_{j1}^1)P(y_i | x_{j2}^2))^2}{P(y_i | x_{j1}^1)P(y_i | x_{j2}^2)} \quad (7.27)$$

Because the values being compared here are probabilities, the value of χ^2 is independent of unit of measurement and can be compared with tabled values to test for conditional independence (Bonham-Carter, 1994). Degrees of freedom in this test are determined by the number of categories for each variable. A low χ^2 (less than tabled χ^2 at a given level of confidence) means that observed and expected values are close, and the assumption of CI holds. A value of χ^2 above tabled χ^2 means that the assumption of CI may be violated.

Testing for conditional independence requires the calculation of observed and expected probabilities for each pair of variables, using Equations 7.26 and 7.27. In a sparse database this calculation may be problematic. In addition, the χ^2 statistic becomes unstable with low expected counts, as in a sparse database. Where $P(y_i | x_{j1}^1, x_{j2}^2)$ is observable, then the most meaningful indication of CI may be to simply calculate expected value as a percentage of observed value for all combinations of Y , X^1 and X^2 . Where $P(y_i | x_{j1}^1, x_{j2}^2)$ is not observable due to sparseness of the database or where no data exists, then CI cannot be assessed numerically. In this case, judgement is necessary to assess whether variables are conditionally independent or not. As mentioned above, using judgment to assess functional independence is considered a valid alternative (Corner *et al.*, 2002).

Although it has been stated that correlation is not the same as conditional independence, when correlation between two datasets is low, then the chances of conditional independence being violated are also low. Therefore, when CI cannot be empirically tested, testing for correlation may give an indication of whether CI could be violated. Correlation can be tested using a slightly different formulation of χ^2 . Here, the two entities being compared are not observed and predicted values for Y , but simply values of X^1 and X^2 independent of Y . The formula for χ^2 then becomes:

$$\chi^2 = \sum_j \frac{(P(x_j^1) - P(x_j^2))^2}{P(x_j^2)} \quad (7.28)$$

In this case, χ^2 does depend on unit of measure and therefore is not the best measure of correlation. Another possible measure of correlation is the joint information uncertainty measure (Press *et al.*, 1986). This is derived from the entropy of the two variables, defined as:

$$\begin{aligned} Entropy(X^1) &= -\sum_i P(x_i^1) \ln P(x_i^1) \\ Entropy(X^2) &= -\sum_j P(x_j^2) \ln P(x_j^2) \end{aligned} \quad (7.29)$$

where $P(x_i^1)$ can be interpreted in frequentist terms, namely, the proportion of entries in a database, or cells in a map, with value x_i^1 for variable X^1 .

Joint entropy of the two variables is then given by:

$$Entropy(X^1, X^2) = -\sum_i \sum_j P(x_i^1, x_j^2) \ln P(x_i^1, x_j^2) \quad (7.30)$$

Joint information uncertainty U is then given by:

$$U = 2 \left(\frac{Entropy(X^1) + Entropy(X^2) - Entropy(X^1, X^2)}{Entropy(X^1) + Entropy(X^2)} \right) \quad (7.31)$$

It can be shown that the value of U lies between 0 and 1. The value of $Entropy(X^1, X^2)$ is equal to $Entropy(X^1) + Entropy(X^2)$ when the proportions are perfectly balanced. Conversely, $Entropy(X^1, X^2)$ tends towards zero when at least one joint probability $P(x_i^1, x_j^2)$ tends towards zero or tends towards one, that is, the proportions are unbalanced. Therefore, when U is close to 1, X^1 and X^2 are correlated and the assumption of CI is likely to be violated. When U is close to 0, X^1 and X^2 are uncorrelated and the assumption of CI is likely to hold.

A third possibility for measuring correlation between two variables is Cohen's kappa (Cohen, 1960), which measures the amount of agreement between two sets of categorised data. Cohen's kappa κ is given by:

$$\begin{aligned} P_o &= \sum_i P(x_i^1, x_i^2) \\ P_e &= \sum_i P(x_i^1)P(x_i^2) \\ \kappa &= \frac{P_o - P_e}{1 - P_e} \end{aligned} \quad (7.32)$$

where P_o is the observed agreement between the two sets and P_e is the agreement expected by chance. The weighted kappa coefficient (Fleiss, 1981) is a generalisation of the simple kappa coefficient and takes account of classifications that do not agree through weighting coefficients. Weighted kappa κ_w is given by:

$$\begin{aligned} w_{ij} &= 1 - (C_i - C_j)^2 / (C_c - C_1)^2 \\ P_{ow} &= \sum_i \sum_j w_{ij} P(x_i^1, x_j^2) \\ P_{ew} &= \sum_i \sum_j w_{ij} P(x_i^1) P(x_j^2) \\ \kappa_w &= \frac{P_{ow} - P_{ew}}{1 - P_{ew}} \end{aligned} \quad (7.33)$$

where C_i is the score for column i (usually equal to i) and c is the number of columns. P_{ow} is the weighted observed agreement between the two sets and P_{ew} is the weighted agreement expected by chance. The weights are constructed following Fleiss and Cohen (1973) so that $0 \leq w_{ij} < 1$ where $i \neq j$ and $w_{ii} = 1$ for all i . The weights are also symmetrical.

However, kappa is only meaningful if the two variables are categorised into the same number of classes and if each class for one variable corresponds directly to a class for another variable. Cohen's kappa was developed for measuring agreement of the same classification from two different sources, and hence is not so useful in this situation.

Chi-squared, joint information uncertainty and kappa are all potential measures where data is available in discrete categories. When the two datasets being compared are continuous, then the correlation coefficient R can be used to test for correlation.

7.2.4 Dealing with Violations of Conditional Independence

If in the set of variables there are two variables X^j and X^l where conditional independence does not hold, then Equation 7.18 does not hold and it is not true that:

$$P(x_{j1}^1, x_{j2}^2 | y_i) = P(x_{j1}^1 | y_i)P(x_{j2}^2 | y_i) \quad (7.34)$$

which is the two variable case of Equation 7.19.

The product rule of probability states that:

$$P(x_{j1}^1, x_{j2}^2 | y_i) = P(x_{j2}^2 | x_{j1}^1, y_i)P(x_{j1}^1 | y_i) \quad (7.35)$$

or equivalently:

$$P(x_{j1}^1, x_{j2}^2 | y_i) = P(x_{j1}^1 | x_{j2}^2, y_i)P(x_{j2}^2 | y_i) \quad (7.36)$$

It is clear that these are equivalent to Equation 7.34 if and only if:

$$P(x_{j2}^2 | x_{j1}^1, y_i) = P(x_{j2}^2 | y_i) \text{ and } P(x_{j1}^1 | x_{j2}^2, y_i) = P(x_{j1}^1 | y_i) \quad (7.37)$$

That is, X^2 is independent of X^j conditioned on Y , and X^j is independent of X^2 conditioned on Y . It should be pointed out that even when X^j and X^2 are dependent, they may still be independent when conditioned on Y , and conversely X^j and X^2 could be independent, yet conditionally dependent with regards to Y .

From the above, it follows that if CI is violated, then Equations 7.20 – 7.24 do not hold either, and other methods are required to populate the CPT. One approach

would be to identify where the assumption of CI is violated, and, in these cases only, specify $P(y_i | x_{j1}^l, x_{j2}^2)$ instead. With l variables of m classes each (i.e., where each variable has the same number of classes) and a response variable with n possible states, if the entire CPT is to be calculated, the number of values required is $n \times m^l$. If the assumption of CI is satisfied, then the number of values is reduced to $n \times m \times l$. If CI is violated for one pair of variables, then the number of values is $n \times (m \times (l - 2) + m^2)$. To illustrate numerically, if there are six variables of five classes each and Y has four possible states, then if CI is satisfied, the number of values to be specified is 120. If CI is violated for one pair of variables, then the number rises to 180. If CI is violated for all variables, then the number of values to be specified is 62,500 (Table 7.1).

Situation	Values required	Numerical example
All variables satisfy CI	$n \times m \times l$	$4 \times 5 \times 6 = 120$
One pair violates CI	$n \times (m \times (l - 2) + m^2)$	$4 \times (5 \times 4 + 5^2) = 180$
Two distinct pairs violate CI	$n \times (m \times (l - 4) + 2 \times m^2)$	$4 \times (5 \times 2 + 2 \times 5^2) = 240$
One triplet violates CI	$n \times (m \times (l - 3) + m^3)$	$4 \times (5 \times 3 + 5^3) = 560$
Two triplets violate CI	$n \times (m \times (l - 6) + 2 \times m^3)$	$4 \times (2 \times 5^3) = 1,000$
All variables violate CI	$n \times m^l$	$4 \times 5^6 = 62,500$

Table 7.1 Number of values required to populate the full CPT.

The situations described in Table 7.1 are displayed graphically in Figure 7.6.

Another approach is to combine the two variables in some other way. Bonham-Carter (1994) suggests Boolean operators, principal components and multiple regression. These methods will generally be more tractable than populating the CPT for conditionally dependent variables. If a biologically meaningful relationship is known to exist between two variables, then often equations already exist to allow them to be combined. For example, it is known that organic matter and phosphorus both have bearing on soil fertility. Rather than treat them as separate variables, they could be combined using this known relationship.

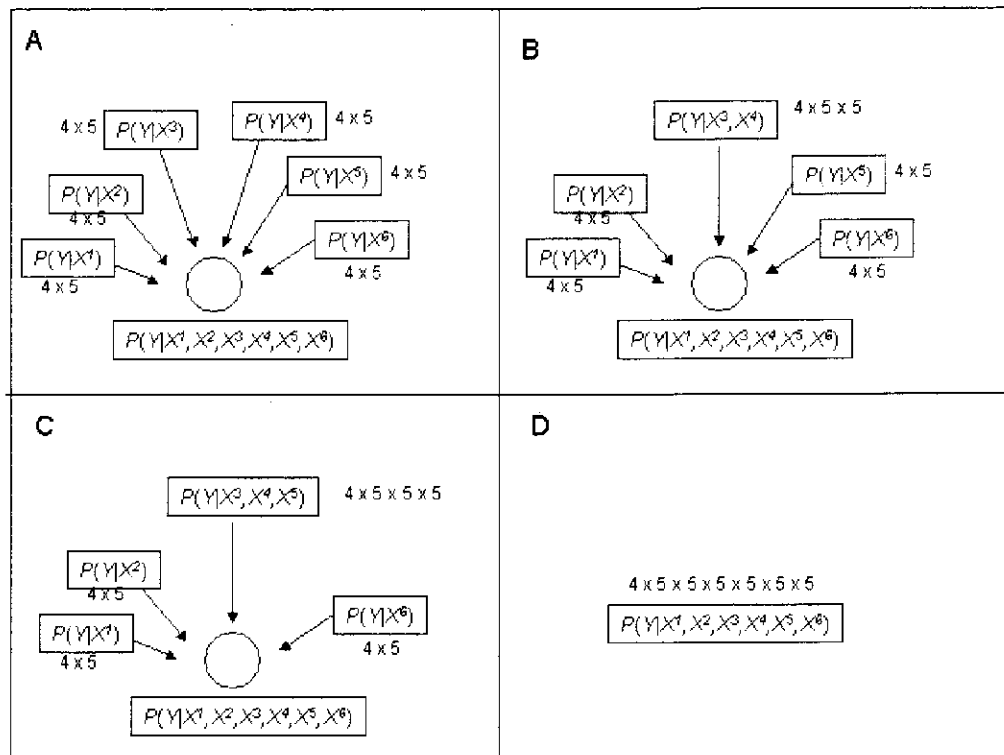


Figure 7.6 Populating a CPT for 6 independent variables X^k with 5 states each and a dependent variable Y with 4 states.

- A: CI is satisfied for all variables
- B: One pair (X^3, X^4) violates CI
- C: One triplet (X^3, X^4, X^5) violates CI
- D: All variables violate CI and the CPT must be directly populated

Although PCA effectively removes correlation, resulting variables are difficult to interpret. PCA is therefore only an option if a relatively large amount of data exists and probabilities will never need to be estimated using expert opinion.

7.2.5 Causality

Causality refers to the direction of flow in a directed acyclic graph (see Section 6.7). The relationship between two nodes in a graph can be interpreted as 'causal', from cause to effect, or 'diagnostic', from effect to cause. Pearl (1988) points out that rules expressed in causal form are usually assessed more reliably, although many expert systems, especially medical ones, use diagnostic reasoning. Essentially, $P(Y|X)$, the probability of Y given X , is only meaningful if either X causes Y or X is caused by Y . Causality is usually temporal, that is, if X causes Y , then X usually

occurs before Y . If there is no causal relationship, then the model cannot be ecologically meaningful, even though relationships may exist in the data. In the case described above, success of a species is in part caused by rainfall being above 800mm annually. Expert knowledge can be very powerful in defining causal relationships in a Bayesian network.

A database conveys no knowledge of causality, and probabilities derived from a database are simply frequency counts. However, an expert does think in terms of causality. In formulating the structure of the model, it is important to consider how the expert interprets these relationships. This is linked to the concept of belief, in that an expert's belief in an event happening is related to the probability of the event happening. Pearl (1988) stresses that, since beliefs are formed from experiences, it is valid to treat beliefs as frequentist probabilities.

A less intuitive approach in the decision problem is the diagnostic formulation $P(X | Y)$, that is, the probability that rainfall is above 800mm given that the species is successful. Obviously, the fact that the species is successful does not cause rainfall to be above 800mm. However, when an expert is asked to define where a species will thrive, they will usually give ranges for various variables. Say an expert states that a species will thrive when rainfall is between 800 and 1200mm. This is equivalent to stating that if the species is successful, then rainfall is most probably between 800 and 1200mm. If the species is not successful, then rainfall is most probably outside this range. This is a diagnostic relationship, therefore, the expert's beliefs can be represented by $P(X | Y)$. Seidel *et al.* (2003), in their analysis of using experts to specify a Bayesian system, found little difference between asking questions in the causal direction and asking questions in the diagnostic direction.

7.2.6 Uncertainty Measures

Defining the model in terms of probabilities gives some measure of certainty. For example, the probability distribution in Figure 7.4 states that the species is most likely to be suitable, but relays considerable uncertainty about this statement (only 50 percent probability). However, it would also be useful to know something about the certainty of the probability distribution itself. In the case where probability

distributions are derived from frequency counts in a trials database, if there are many different trials for a species under the same conditions (but at different locations), then a high certainty can be associated with the probability distribution. Conversely, if only a few trials have been recorded, then the probability distribution is very uncertain.

When probability distributions are elicited using expert knowledge, these measures can also be applied, depending on how much certainty the expert associates with their judgement.

This is a fairly simplistic approach to assessing certainty of probability distributions. For example, Dempster and Shafer's concepts of plausibility and belief (see Section 6.7) could be applied to define intervals around each probability value. However, the approach described above provides a sufficient measure of uncertainty, whilst at the same time not requiring any considerable amount of extra information from the database or the expert.

7.2.7 Sensitivity Analysis

There are various methods available to determine the sensitivity of the model to changes in states of the predictor variables. One method suggested by Dittmer and Jensen (1997) is 'what-if' analysis, which analyses the change in outcomes if the state of one variable is changed. A species may be suitable under a given set of circumstances, but may become less suitable if rainfall decreases or if soil is more acidic. Having information on how quickly suitability might change, especially when climatic changes occur, is important in the context of the decision problem.

In addition, sensitivity analysis indicates which variables are important in a particular case. If the response is highly sensitive to one predictor variable, and that particular predictor variable is very uncertain, then this could affect the decision.

7.3 Modelling in a GIS Context

An important facet of the model being developed is its ability to be implemented spatially. GIS modelling provides a means of scaling from local to regional predictions.

Environmental envelopes are generally straightforward to implement using GIS. In many cases, no other information is needed apart from presence data and climate data (for example, BIOCLIM [Busby, 1991] and FloraMap [Jones and Gladkov, 1999]). It is a straightforward process to include other variables, such as soil characteristics, as long as they can be mapped.

Fuzzy envelopes take even more advantage of the capabilities of GIS. Rather than simply displaying envelopes as simple presence/absence values on a map, fuzziness can be displayed using colours and shades to denote level of fuzzy membership. Probabilities calculated using Bayesian modelling can similarly be displayed.

Another advantage of spatial implementation is the increased capability to deal with and visualise spatial uncertainty. Almost all types of uncertainty associated with decision problem are spatially heterogeneous and, therefore, simply mapping uncertainty increases the decision-maker's ability to manage the impacts of uncertainty.

GIS is not only useful for visualisation of outputs, but also for processing and analysing spatial inputs. The variables used as predictor variables in the decision problem are all spatial in nature. GIS techniques such as map algebra allow equations, such as Bayesian joint probability calculation, to be implemented across all locations in space simultaneously.

Some of the uncertainty in the decision problem is introduced because of metrical and temporal uncertainty in the predictor variables. This uncertainty will vary for each variable and for each state of the variable. For example, DEMs are generally fairly accurate, depending on their resolution, because elevation values are directly calculated from satellite capture data. Conversely, soil maps are likely to contain a

great deal of uncertainty due to uncertainty in measurement techniques, classification techniques and due to high spatial heterogeneity of soil characteristics.

Corner *et al.* (2002) term this uncertainty in input data ‘map purity’. Map purity can be defined as the probability that a classification is the true classification on the ground and can be derived directly from experts, or, if such information exists, from metadata. In a Bayesian modelling context, map purity can be propagated through the network using Equation 7.38:

$$P(Y | x_j) = \sum_i P(Y | x_i) P(x_i | x_j) \quad (7.38)$$

where x_i are possible states of variable X . In the above equation, $P(x_i | x_j)$ is the probability that the true classification is x_i given that the mapped classification is x_j .

7.4 Proposed Modelling Approach

In Chapter 4, a number of information sources were discussed, including forage databases, expert knowledge and spatial data. The purpose of the functional model is to determine where a given forage species will succeed, or which forage species will succeed in a given location. A measure of success needs to be derived from data in forage databases and/or from expert knowledge. Although multiple measures of success could be defined (such as how quickly a species establishes and the size of the species’ yield is after a defined period of time), in this model it is proposed to use just one measure of success as the dependent variable Y in the model. This variable may be a function of other variables. Because the concern is how well a species will perform, the response variable needs to be discrete over a number of classes, rather than binary.

The independent variables X^* will be drawn from the sources of information previously discussed. The requirement is that there is some known functional relationship between the predictor variables and the dependent variable. This relationship should also be evident in the database – if it is not, then the validity of either the data or the relationship comes into question.

The proposed method is essentially to combine these predictor variables to define probabilistic environmental envelopes based on Bayesian modelling techniques. The response variable Y will be characterised by a probability distribution function, that is, the likelihood of occurrence for each possible outcome of the response variable. In addition, the level of certainty associated with the probability distribution is analogous to fuzzy membership.

The Bayesian model employed will be the most simple Bayesian network possible, with all input variables at one level, feeding simultaneously into the output variable (Figure 7.7).

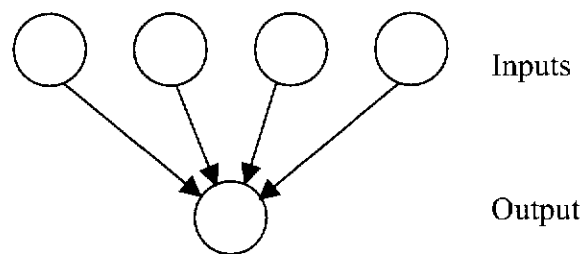


Figure 7.7 Simple Bayesian network.

It is not necessarily the case that all variables act simultaneously – for example, in reality, elevation may affect both temperature and soil characteristics, and temperature itself may affect soil characteristics, so that a more complicated network could be more valid. However, sparse datasets make defining such a structure more problematic. In addition, a simple structure makes both implementation and explanation of the model much more straightforward.

If information is available relating to map purity, then each of the input nodes receives this information (Figure 7.8). While the model being developed will allow for this eventuality, it is unlikely that such information will be available across large areas, especially when the input data may already be derived from a combination of sources.

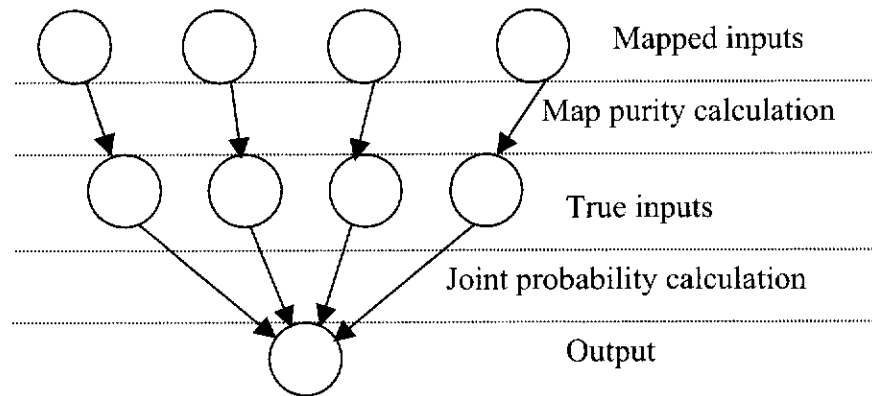


Figure 7.8 Simple Bayesian network with map purity.

The structure of the simple Bayesian network (Figure 7.7) allows one single calculation to be made for each possible combination of variables. This is analogous to defining environmental envelopes and to applying rules defined in a knowledge base. Depending on whether CI holds for all variables, joint probability is calculated according to the situations displayed in Figure 7.6. Where CI does not hold, experts are required to define the CPT for two or more variables simultaneously. Because, as shown in Table 7.1, the size of the CPT grows exponentially with number of variables and number of classes in each variable, this option would only be made available for pairs of variables.

For some variables, their relationship to success is not probabilistic in nature. Some forage species have very low tolerance to certain variables, such as drought, salinity or waterlogging. If one of these conditions is present, then it is valid to assign zero probability to the success of such a species, which is equivalent to excluding the species from consideration.

In the case of selecting forage species for a specific niche, the problem is not just to predict the success of one species, but of many species simultaneously at one location, in order to identify the most promising species. Therefore, the model needs to apply some sort of ranking or filter to identify a subset of species for consideration at any one location. Filters can be applied by removing from consideration any species which do not meet tolerance or use requirements, if these are defined.

Ranking can be applied once posterior probabilities have been derived for each variable for the conditions at the location under consideration.

From posterior probabilities, a ranked 'basket of options' can be selected for the location in question. At the same time, uncertainty information is retained, based on the database, expert knowledge and probability combinations. Sensitivity analysis can also be performed by inspecting how much probability distributions change, depending on changes in variable states. Posterior probability distributions, uncertainty information and sensitivity information can all be communicated using maps and graphs. This implementation is outlined in Figure 7.9.

7.5 Summary

The approach outlined above is based on a combination of fuzzy envelopes and Bayesian probability modelling. Rather than defining any state in a multi-state envelope as wholly 'suitable' or 'unsuitable', each state is assigned a probability distribution.

The model allows information from diverse sources on success of forages to be combined to predict success distributions for any combination of variables. The model incorporates uncertainty, retaining uncertainty information throughout the model and allowing this information to be displayed and interrogated in a GIS environment. The next step is to formalize this model as a spatial decision support system. This is described in the following chapters.

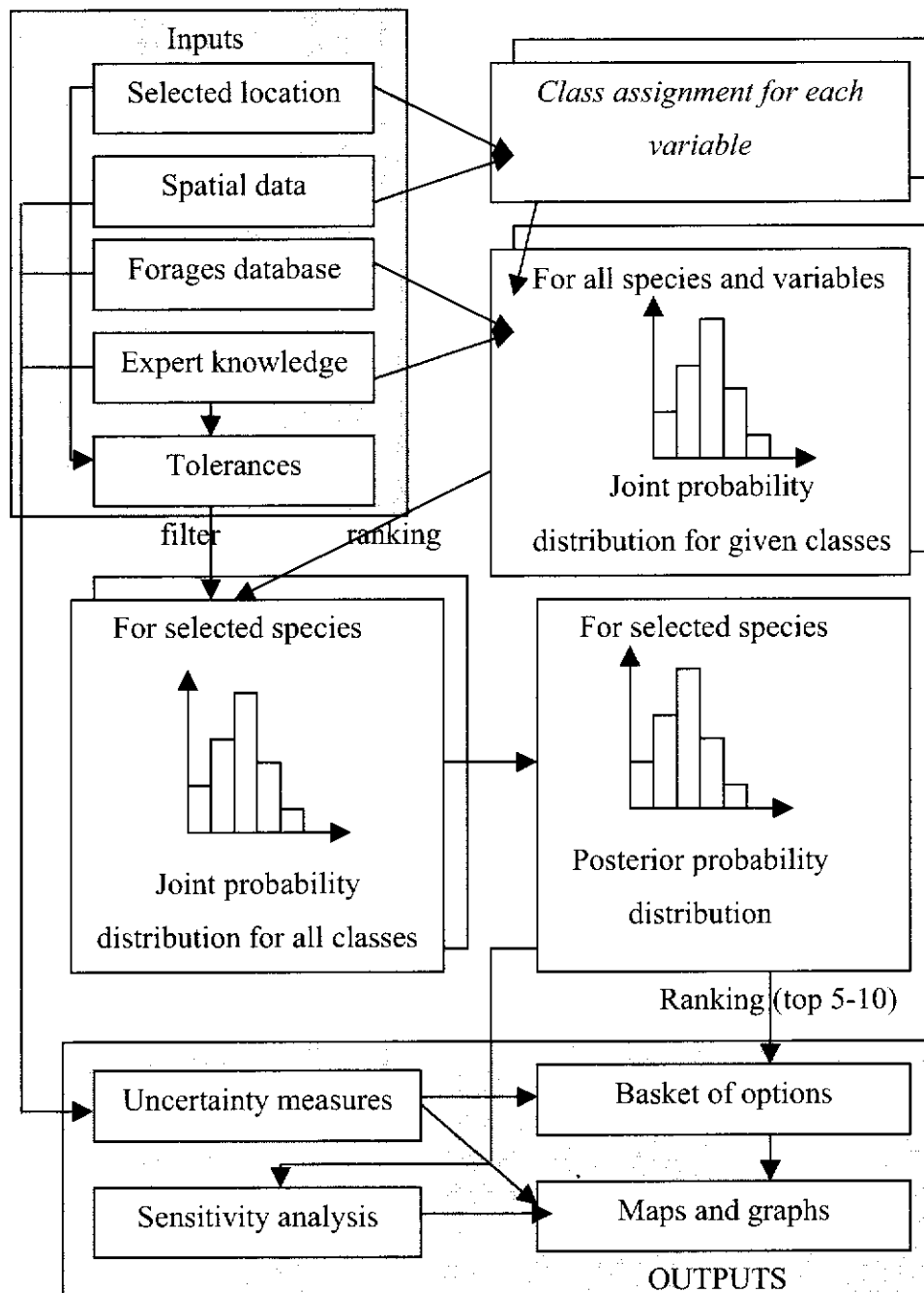


Figure 7.9 Model implementation diagram.

CHAPTER 8. DATA AND KNOWLEDGE SELECTION

In Chapter 4, Decision Support Systems (DSS) were introduced, with particular focus on DSS in agriculture and spatial DSS (SDSS). Although a number of issues were identified which need to be overcome for successful DSS development, an SDSS for forage selection has been identified as a potential method to provide information to agricultural decision-makers.

In the previous chapters, potential models were reviewed and a probabilistic GIS model was selected to functionally model forage species' success spatially. In this chapter, the sources of data and knowledge discussed in Chapter 4 will be analysed. Bearing in mind the probabilistic GIS model developed in Chapter 7, suitable predictor and response variables will be identified and processed to produce input data required for the SDSS discussed in the following chapter.

8.1 Predictor Variables

Predictor and response variables are used to specify a model, and predictor variables are also used in the implementation of the model in order to predict the values of response variables in situations where the response variable value is not already known. In the model specification stage, predictor and response variables need to be calculated from databases and knowledge bases containing information that relates predictor variables to response variables. In the implementation stage, predictor variables are drawn from spatial data, allowing the response variables to be spatially interpolated.

In the probabilistic model presented in the previous chapter, it was shown that all predictor variables feed simultaneously into the model. This does not mean, however, that all variables have equal weighting. If for a particular species a predictor variable has no impact, then the associated conditional probability distribution $P(Y|X)$ will simply be the same as the prior probability $P(Y)$. Therefore, the task here is to identify a suitable set of predictor variables which may impact on the suitability of any forage species.

8.1.1 RIEPT database

A major source of information, both for predictor and response variables for model specification, is the RIEPT database. In this section, the RIEPT database is analysed and issues are highlighted. The following chapter will discuss how these issues are dealt with in this research. This database contains adaptation, production and establishment data for forage trials mainly throughout Latin America (Central and South America), spanning 1979 – 1992 (and unofficially continuing at present). The purpose of the RIEPT database is to record information on forage trials under various environmental and management conditions. Initially, a large number of species were evaluated at a relatively small number of locations to assess adaptation of grasses and legumes in locations representative of major tropical ecosystems ('Adaptation' database). Further trials were then undertaken to evaluate productivity under cutting of certain species, selected from the adaptation trials but in different ecosystems ('Establishment' and 'Production' databases) (Barco *et al.*, 2002). The Establishment and Production databases contain records for countries in Africa, Asia (China) and the Caribbean, as well as the Latin American countries included in the Adaptation database.

Throughout the planning and development of the RIEPT database, a number of meetings were held and manuals of methodology produced, defining standardised methodologies for germplasm evaluation, analytical methods and procedures for soil and plant material evaluation (CIAT, 1980). Methodologies were produced, amongst others, for dealing with agronomic evaluation to determine germplasm adaptation to edaphic, climatic and biotic factors, and to determine seasonal dry matter yields (Lascano and Spain, 1992). Therefore, it is expected that there is a reasonable level of consistency across sites and years in the RIEPT database.

In total, the database records data on 11,211 trials of 1,798 accessions in 314 locations in the tropics. Of these, 2,539 trials of 929 accessions are in 28 locations in Central America (Figure 8.1). In Central America, there are 1,694 trials in the Adaptation database, 436 in the Establishment database and 409 in the Production database.

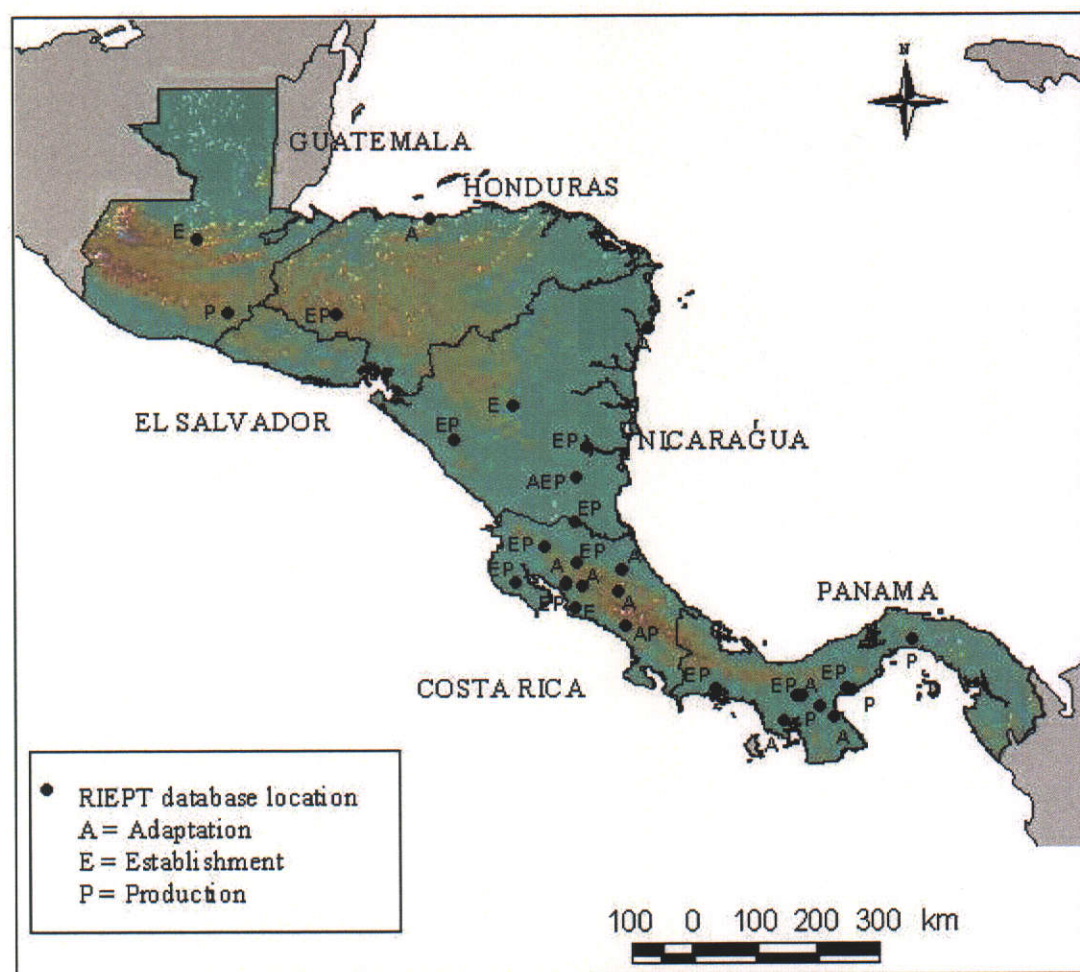


Figure 8.1 RIEPT trial locations in Central America.

Plants can be identified by genus, species and accession, and each species may have a number of accessions with unique characteristics. In RIEPT, accessions are identified by a unique CIAT number. 'Passport' data is also available for 824 of the accessions, namely, those which are held in CIAT's germplasm bank. Passport data gives details of the location where the accession was originally collected.

The Adaptation database covers trials in Latin America only. Examination of the locations of these trials and how many trials are carried out at each location shows bias in spatial distribution as well as in species distribution. Figure 8.2 shows the number of adaptation trials for all species in RIEPT.

Figure 8.3 gives an example of species bias, showing number of trials at each location for five selected species. This illustrates that locations with a large number

of trials may in fact be dominated by trials of a single species. While this may increase confidence in predicting the performance of the species under the conditions at the trial site, it does not add substantial information on the species' performance, as each trial at the site has exactly the same environmental characteristics. Although climate varies at a single site between seasons and between years, information on climate variability is not included in the RIEPT database.

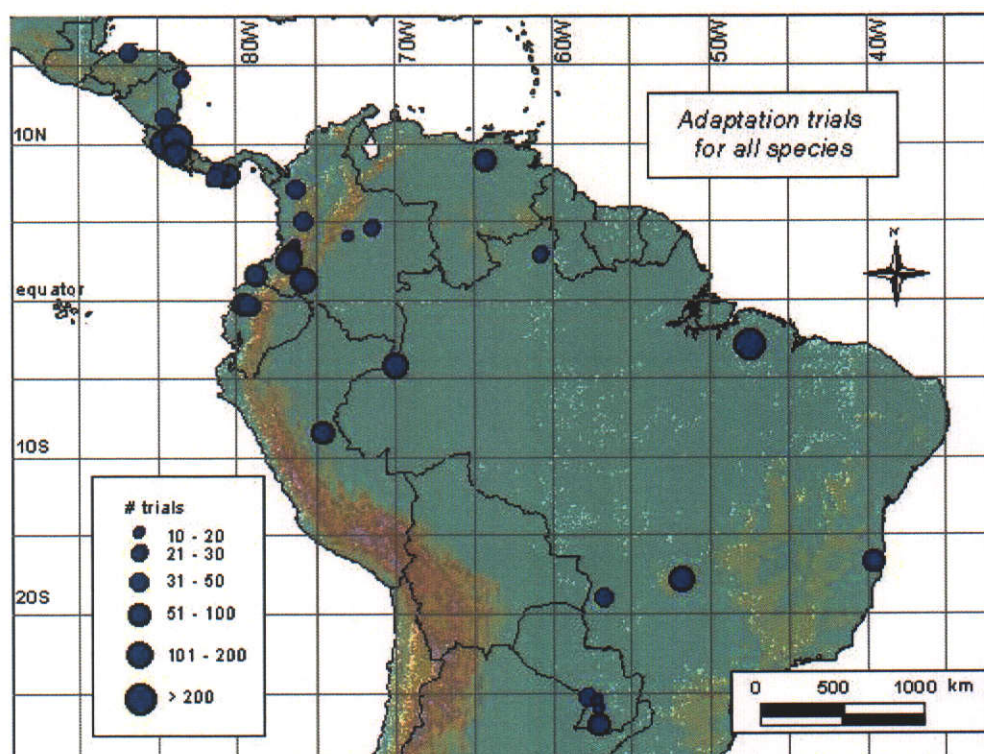


Figure 8.2 Adaptation trials for all species in RIEPT.

In the previous chapter it was shown that in order to specify the probabilistic GIS model the full CPT must be derived. One of the steps is to calculate the prior and joint probability distribution of the response variable, which will be a measure of success derived from the RIEPT database. It is therefore necessary to count the frequency of trial results under different conditions. The number of trials for each species and the spatial distribution of trials can, however, bias the probability distributions.

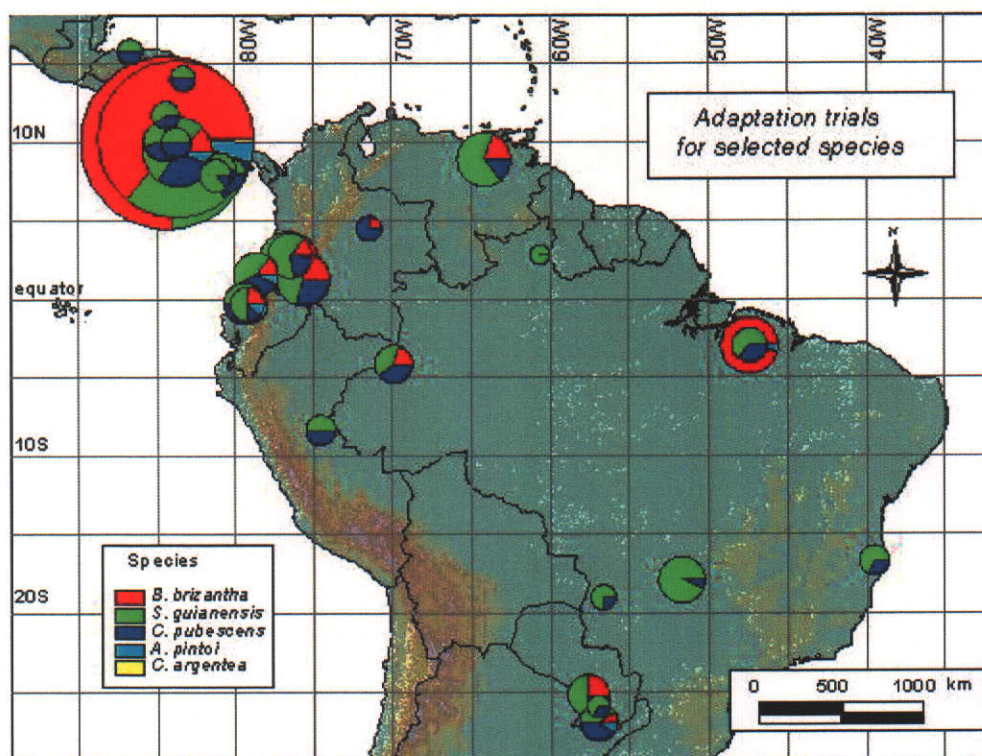


Figure 8.3 Adaptation trials for selected species in RIEPT. Size of circles denotes number of trials for the five species combined.

In the RIEPT database, trial results are given for multiple trials for multiple species at multiple locations. At some locations, dozens of trials have been carried out on a single species, while other locations may only have a single trial for that species. This has the potential to strongly bias the database towards conditions at locations with a larger number of trials. It is important that information from all trials be taken into account, while attempting to mitigate this bias. The approach chosen in this research is to treat each trial site as a single observation and record the results of each trial as a proportion of one observation. If a trial is carried out at a different location with the same conditions, then this is treated as a new observation. This approach removes the bias introduced by multiple trials of a single species at a location. However, it also reduces the number of observations in the database, for any one species, to the number of trial locations rather than the number of trials.

Climatic bias will still exist in the database. Sites with certain combinations of environmental characteristics will be absent from the database, but these missing values are not missing at random. Values are more likely to be missing because researchers had already decided not to trial under those conditions. The reasons

could be because the likely outcome of the trial is already sufficiently known (either known to be very good or very poor), because of inaccessibility (e.g., high elevations or remote locations) or because the locations of trial sites are determined in part by other non-random factors (e.g., locations of collaborators). If the trial sites are interpreted as samples, then it is clear that the sampling strategy is far from random.

Data recorded includes level of adaptation, percent cover, insect resistance, disease resistance, height and dry matter weight. Values are recorded at different times during a trial, but the timing is not the same for all trials, and not all values are recorded (i.e., missing attribute values). Trial locations have elevation, climate and soil data recorded, but there are gaps and inconsistencies in the data.

8.1.2 Representativeness of RIEPT Database

Location variables recorded in RIEPT are listed in Table 8.1. As part of the current research, statistical analyses were carried out on these variables in order to test the representativeness of the database. The analysis was intended to determine patterns in the data and correlations between variables. Firstly, the distribution of recorded elevation, rainfall, soil pH, soil texture and soil fertility in the database were compared with distributions across Central America, derived from GIS databases. For spatial variables represented as rasters, each 1km² raster cell is counted as a location, so there are around 600,000 data points for Central America. The *Indicadores Atlas* (Winograd *et al.*, 2000) estimates agricultural land use in Central America (33 percent of total land area), allowing distributions to be calculated for agricultural land only. This is assumed to be broadly representative of forage-growing land. Figure 8.4 compares cumulative frequency of elevation at locations for all of Central America, agricultural land in Central America, the entire RIEPT database and the RIEPT Adaptation database.

Latitude
Longitude
Elevation
Ecosystem
Mean monthly temperature
Minimum monthly temperature
Maximum monthly temperature
Solar radiation
Relative humidity
Hours of sun
Wind speed
Mean monthly precipitation
Number of dry months
Percentage sand 0-20cm and 20-40cm
Percentage silt 0-20cm and 20-40cm
Percentage clay 0-20cm and 20-40cm
Apparent density 0-20cm and 20-40cm
Field capacity 0-20cm and 20-40cm
Soil pH 0-20cm and 20-40cm
Organic matter 0-20cm and 20-40cm
Phosphorus 0-20cm and 20-40cm
Calcium 0-20cm and 20-40cm
Magnesium 0-20cm and 20-40cm
Potassium 0-20cm and 20-40cm
Sodium 0-20cm and 20-40cm
Aluminium 0-20cm and 20-40cm
Aluminium saturation 0-20cm and 20-40cm

Table 8.1 Variables relating to location included in the RIEPT database.

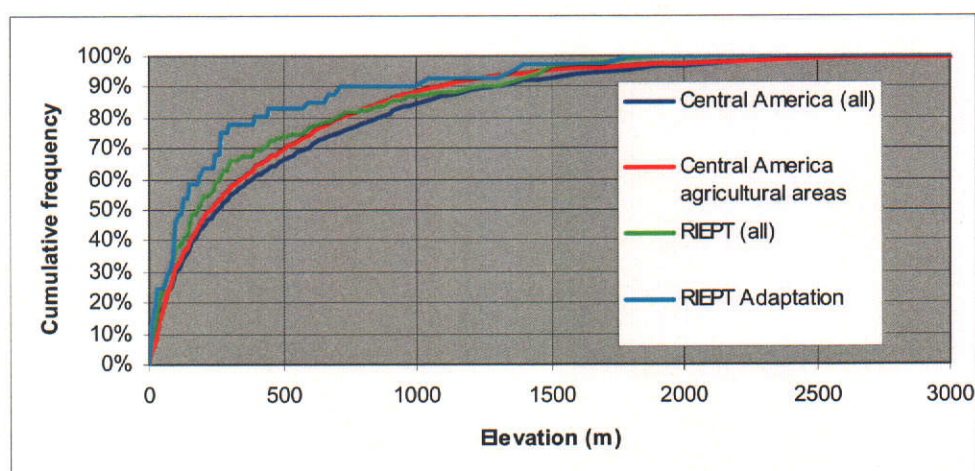


Figure 8.4 Comparison of cumulative frequency of location elevations.

Locations with lower elevation are over-represented in the RIEPT database, particularly in the Adaptation database. CIAT's Forage project mandate areas are the humid and subhumid tropics and subtropics, and at higher elevations (above approximately 2000 masl) conditions become temperate. In addition, most smallholder farmers are located in the lowlands in Central America. Also, agricultural activity generally diminishes at higher elevations, and this is reflected in the difference between the curves for all Central America and for agricultural land only. In the RIEPT Adaptation database, 78 percent of locations lie below 300m, whereas in Central America only 54 percent of all land lies below 300m (57 percent of agricultural land).

Cumulative frequency of annual rainfall is shown in Figure 8.5, comparing the same sources of data. Agricultural land in Central America does not appear to favour any particular ranges of rainfall. The RIEPT database slightly over-represents drier locations. The RIEPT Adaptation database under-represents some classes of rainfall and over-represents others.

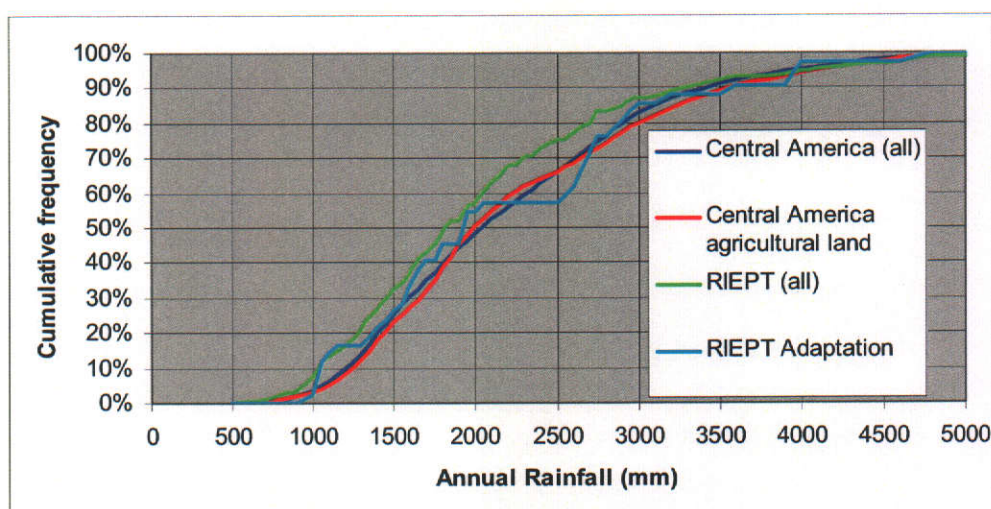


Figure 8.5 Comparison of cumulative frequency of rainfall.

Soil characteristics were derived for Central America and for agricultural land in Central America based on FAO soil maps and classifications used in the *Indicadores Atlas* (Winograd *et al.*, 2000). Because of the nature of soil maps and the processes used to derive soil properties (discussed in Chapter 4), the classifications for GIS data may not be accurate. Figure 8.6 shows soil pH for the sources of data discussed

above. The majority of RIEPT trials are in locations with acidic to moderately acidic soils (pH between 4.5 and 6).

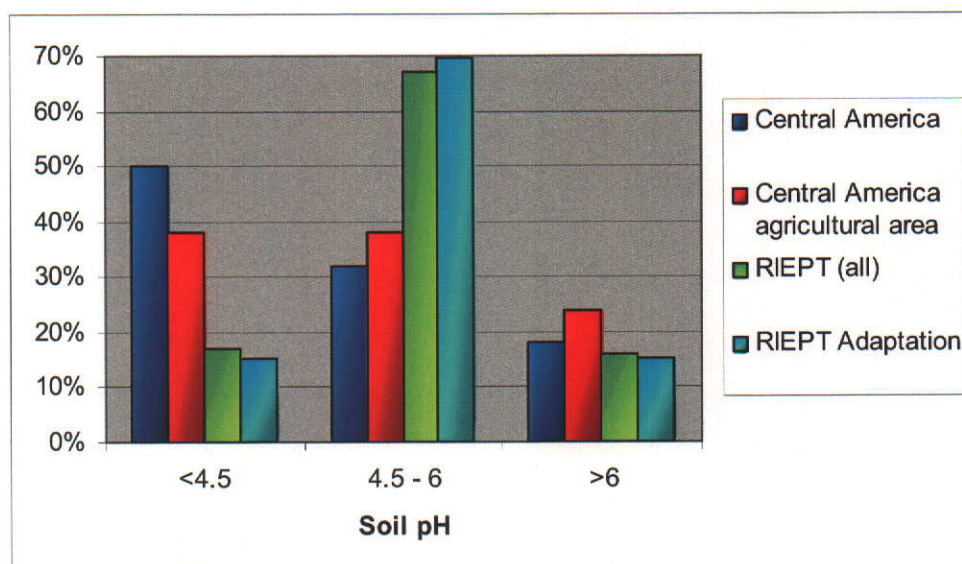


Figure 8.6 Comparison of percentage of area or locations with soil pH in classes shown.

Figure 8.7 shows percentage of area and percentage of RIEPT locations with soil texture classed as clay, loam or sand. For the RIEPT data, these classifications were derived from percentage sand, clay and silt at 0-20cm depth and converted to these three classes based on an FAO soil scheme (Verheye and Ameryckx, 1984) (Figure 8.8).

The majority of land (both agricultural and non-agricultural) in Central America is clay (fine texture). However, the RIEPT database has more locations in areas with loamy soil (medium texture).

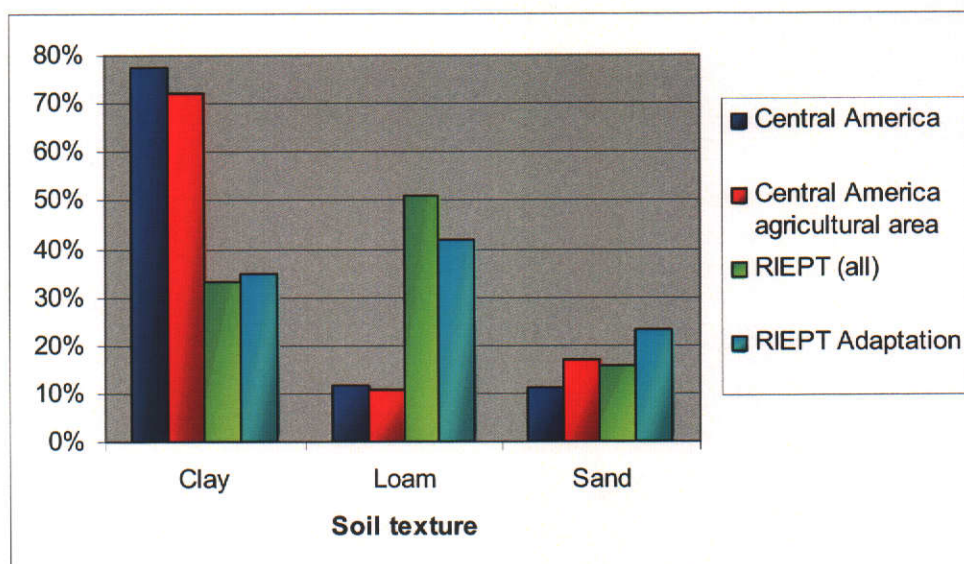


Figure 8.7 Comparison of percentage of area or locations with soil texture in classes shown.

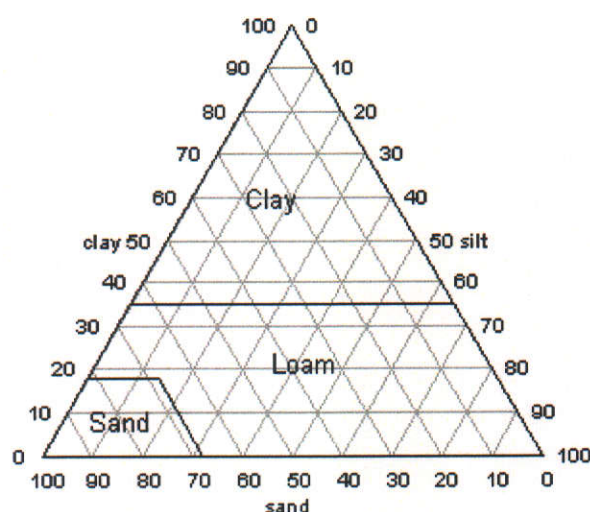


Figure 8.8 FAO soil classification. Source: Verheye and Ameryckx (1984).

Classification of soil fertility is shown in Figure 8.9. Soil fertility can be calculated from values for organic matter and phosphorus in the RIEPT database. This process is described later (see Table 8.8). Central America has a fairly even distribution of soil fertility, although agricultural land favours soils with higher fertility. However, the RIEPT database markedly over-represents low-fertility soils because most smallholder farmers are located on marginal land, which is usually less fertile. When

farmers do have more fertile land available, it is often reserved for crops, and forages are mostly planted on less fertile soils.

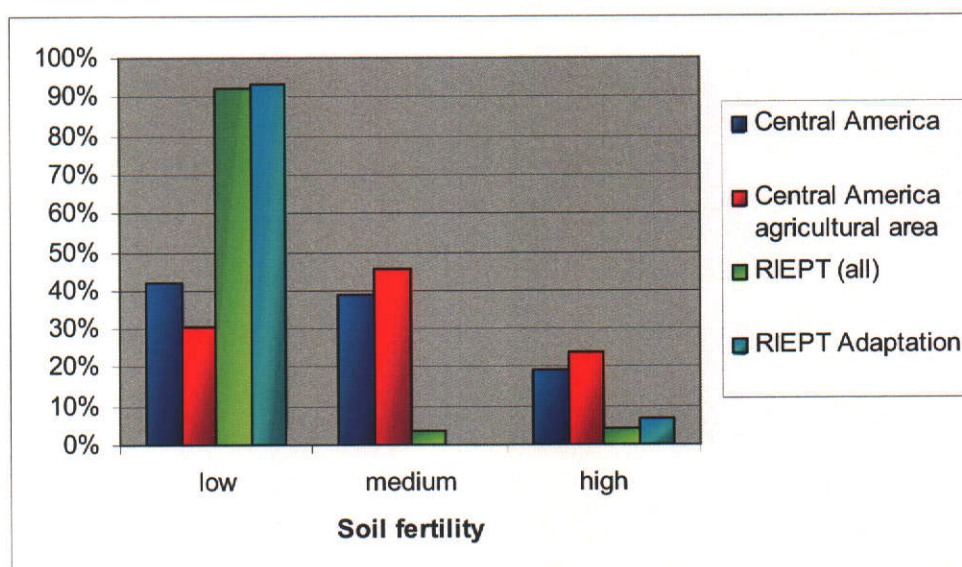


Figure 8.9 Comparison of percentage of area or location with soil fertility in classes shown.

The analysis presented thus far illustrates both the spatial and attribute biases present in the RIEPT database. The database is moderately biased spatially and also in relation to number of forage species. However, climatic and edaphic attributes are relatively well represented, compared with the distribution of these attributes across Central America as a whole and agricultural land in Central America. Where the distribution varies, it is generally by design in the RIEPT database. This needs to be taken into account when calculating prior and joint probabilities.

8.1.3 Accuracy of RIEPT Attribute Data

In order to check the accuracy of RIEPT attribute data, these values were compared with corresponding mapped values at the same locations, and the coefficient of determination R^2 was calculated (Table 8.2). The coefficient of determination is the square of the correlation coefficient R , and can be interpreted as the proportion of variance explained by the regression. Correlation was not as high as would be expected for attributes representing the same information. This could be for a number of reasons, including errors in the spatial data, uncertainties in the spatial

data (introduced at various stages during spatial processing), positional errors in the RIEPT database (incorrect latitude and longitude recorded), measurement errors in the RIEPT database and data entry errors.

Elevation	Mean temp	Rainfall	Dry months
0.56	0.38	0.55	0.34

Table 8.2 R^2 between RIEPT data and GIS data.

The low correlation values mean that care must be taken when using values derived from RIEPT and from spatial data in the same calculation. As discussed earlier, uncertainty can be spatial (location uncertainty) or aspatial (attribute uncertainty). Inspection of the RIEPT database suggested that errors are most likely to be present because latitude and longitude were inaccurately recorded – most trials were recorded before widespread use of global positioning systems (GPS). It is possible that some sites were identified by name, and that latitude and longitude values were added later. Also trials may have been recorded with an identical location description, although in reality the trials have differing edaphic and possibly climatic characteristics. Therefore, the values for the other variables may well be correct, but will appear incorrect when compared with mapped values. For this reason, it was decided to use values from the RIEPT database (and not the corresponding mapped values) in order to specify the CPT.

Soil characteristics were not directly compared between the RIEPT database and spatial data. Uncertainties in the spatial soils data mean that characteristics identified for point locations are unlikely to be a true representation, and therefore comparisons are not expected to be useful in assessing the quality of the RIEPT soils data.

Despite these biases and inaccuracies in the RIEPT database, it is still a valuable source of data as input to the probabilistic model, in defining conditional probability distributions. Also, data from RIEPT is only one source of information used as input to the model, and will be supplemented with expert knowledge, particularly where RIEPT data is uncertain. Once prior probabilities have been defined, these discrepancies will not adversely bias the calculations.

8.1.4 Correlation Analysis of Potential Predictor Variables

As a first step in selecting suitable variables as predictor variables, correlation analysis was carried out between all pairs of variables in the RIEPT database across all locations (Appendix A). The climatic variables sourced from GIS data for Central America were also analysed for correlation. This gives an indication of which variables are correlated in general and, also, which variables in RIEPT display more correlation than would be expected from a random sample of locations in Central America. Even though data (both spatial and RIEPT) exists for mean temperature, minimum temperature and maximum temperature, as it is already known that these are highly correlated, for simplicity, only mean temperature is considered here. Results of correlation analysis for four climatic variables are shown in Table 8.3.

R^2	Elevation	Mean temperature	Annual rainfall	Dry months
Elevation	1 / 1			
Mean temperature	0.86 / 0.50	1 / 1		
Annual rainfall	0.04 / 0.00	0.02 / 0.00	1 / 1	
Dry months	0.05 / 0.04	0.02 / 0.02	0.56 / 0.44	1 / 1

Table 8.3 R^2 values for all Central America / all RIEPT locations.

As expected in the tropics, elevation and mean temperature are strongly correlated. Mean annual rainfall and number of dry months are also correlated. However, the RIEPT database tends to show less correlation than Central America as a whole. This suggests that RIEPT locations may not be climatically representative of Central America.

Although number of dry months is a variable included in the RIEPT database, this was a late addition to the database (i.e., not recorded at the time of the trial) and the source of the data is often unknown (Franco, *pers. comm.*, 2002). However, because monthly rainfall is recorded for most locations, the dry season length can be derived in an alternative manner. A dry month can be defined in various ways, and a 'dry month' in the tropics may be different to a 'dry month' in a temperate zone. Davis (2000) defines a dry month as the month in which total rainfall (in mm) is less than twice the mean temperature (in degrees C). Bonan (2002) defines a dry month as a

month with less than 60mm of rain. In the tropics, these definitions are roughly equivalent. Therefore, in the current research, the latter definition is used. The largest number of consecutive dry months in a 12 month period is then counted. As a check, the correlation coefficient has been calculated between number of dry months recorded in the database and length of dry season derived using this method. The result is $R^2 = 0.62$, indicating high correlation, but not as high as would be expected for data supposedly conveying the same information. From these considerations, it was decided that length of dry season derived from rainfall data is likely to be more reliable. An additional benefit is that the same method can be used to derive dry season length from GIS data. Comparing the length of dry season derived in the same way from both GIS data and RIEPT locations, R^2 is calculated at 0.44 (an increase from 0.34 in Table 8.2).

Other variables in the RIEPT database that show strong cross-correlation are all soil properties at 0-20cm depth with their corresponding properties at 20-40cm (R^2 between 0.55 and 0.94). Percentages of sand, silt and clay are also strongly correlated. Other notable correlations are between: field capacity, organic matter and apparent density; magnesium and calcium; potassium and sodium; pH, calcium and aluminium saturation; and aluminium and aluminium saturation (see Appendix A).

Because ecosystem is a categorical variable and cannot be transformed to a linear scale, standard correlation analysis could not be carried out for comparison with other variables without first transforming these variables.

Some values in the RIEPT database are more reliable and complete than others. There are 314 unique locations in the database of which only four have entries for all fields. The number of complete entries in each field ranges from 30 (field capacity 20-40cm depth) to 304 (elevation and dry months). Variables with limited data (less than half the records complete) are solar radiation, hours of sun, wind speed, aluminium saturation and most soil variables at 20-40cm depth (excluding potassium and pH).

8.1.5 Specifying Predictor Variables

It is clear that there are too many variables in the RIEPT database, with too many correlations and gaps to incorporate them all directly in the functional model. Valid values of predictor variables must be available at all locations, not just at those in the RIEPT database. Therefore, variables in the model must either be available as spatial data for Central America or they must be variables that a farmer can be expected to know the value of at their specific location. However, because uncertainty can be explicitly modelled, it is permissible for some variables to be undefined in some cases. The discussion will now explore which variables to use as predictor variables in the model and how to define categories for the selected variables.

Techniques exist to combine a large number of variables into a smaller number of variables while retaining most of the relevant information from all variables. PCA is one such technique. However, PCA can also obscure the biological meaning of the data when variables are combined. Another approach is to use biologically or ecologically meaningful formulae to transform the data.

It was therefore decided to use well-defined process-based methods to combine variables where necessary and to split each resulting variable into five classes. This was an iterative process involving expert consultation, and some variables were initially split into fewer classes. Five classes emerged as being the optimal number of classes for experts to consider in this case study. Four classes do not always capture as much variation as experts would like, and more than five classes complicate the knowledge elicitation task unnecessarily. It was also found that constraining all variables to the same number of classes simplified the model, both conceptually and from the point of view of software implementation.

Elevation data is well represented in RIEPT, with values recorded for almost all locations. In addition, spatial Digital Elevation Models (DEMs) are generally of high accuracy. At 1km² resolution, each gridcell represents an average of elevation within the area covered by the gridcell. In many locations, elevation varies at a smaller resolution, and a better representation of reality could be obtained by using 90x90m elevation and climate surfaces, where available. Temperature obviously has

a more direct relationship with forage species' success than elevation, but is less well represented in RIEPT and likely to be slightly less accurate than elevation data. Because temperature and elevation are highly correlated in the tropics, elevation can be seen as a proxy for temperature. Therefore, elevation has been chosen as the first variable for inclusion in the model.

Other important climatic drivers of forage success are rainfall patterns. Although monthly rainfall data is available both in the RIEPT database and as spatial data, including 12 separate variables on rainfall (one for each month) is unfeasible in a knowledge-driven model. Therefore, two variables were derived from these, namely, annual rainfall and length of dry season (discussed above). Although there is some correlation between the two, they provide different information in the modelling context and therefore can be classed as functionally independent (see Section 7.2.3).

Ecosystem is also considered an important climatic driver of forage success. Various methods exist to calculate ecosystem classifications from climatic variables. The ecosystem classification used in RIEPT consists of five categories, defined as in Table 8.4 below.

Ecosystem	WSPE	WS	WSMT
Isohyperthermic savannahs	901-1060 mm	6-8 months	> 23.5 C
Isothermic savannahs	901-1060 mm	6-8 months	< 23.5 C
Seasonal semi-evergreen forests	1061-1300 mm	8-9 months	> 23.5 C
Humid tropical forests	> 1300 mm	> 9 months	> 23.5 C
Poorly drained tropical savannah	Not defined in terms of WSPE, WS and WSMT		

Table 8.4 Ecosystem classification used in RIEPT. WSPE = wet season potential evaporation. WS = wet season. WSMT = wet season monthly temperature.

In these calculations, wet season is defined as the months where the monthly availability index (MAI) is greater than 0.33. MAI is the ratio of dependable precipitation to potential evapotranspiration. Dependable precipitation is precipitation that is equalled or exceeded in three out of four years. Monthly

potential evaporation can be calculated by the reduced Penman method after Linacre (1977). This calculation is based on temperature, diurnal temperature range, latitude and elevation. The above is the method used in RIEPT to define ecosystems.

In the current research, problems were encountered in trying to recreate these ecosystems for Central America, mostly because of the classification not covering all possible eventualities (the definitions only account for 51 percent of the area of Central America) and the lack of definition for poorly drained tropical savannahs. In addition, this ecosystem classification was shown by Schmidt (2001) to have little relationship to variation in forage success.

An alternative ecosystem classification is Holdridge lifezones. Holdridge (1967) classified 38 zones based on mean annual temperature and annual precipitation. Of these, 17 are represented in Central America (Table 8.5).

	Mean Temperature (degrees C)						
Annual Rainfall (mm)	<1.5	1.5-3	3-6	6-12	12-18	18-24	>24
<125	Polar	Dry tundra	Boreal desert	Cool temperate desert	Warm temperate desert	Subtropical desert	Tropical desert
125-250		Moist tundra	Dry scrub	Cool temperate desert scrub	Warm temperate desert scrub	Subtropical desert scrub	Tropical desert scrub
250-500		Wet tundra	Moist forest (puno)	Steppe	Thorn steppe	Subtropical thorn woodland	Tropical thorn woodland
500-1000		Rain tundra	Wet forest (paramo)	Cool temperate moist forest	Warm temperate dry forest	Subtropical dry forest	Very dry forest
1000-2000			Rain forest (rain paramo)	Cool temperate wet forest	Warm temperate moist forest	Subtropical moist forest	Tropical dry forest
2000-4000				Cool temperate rainforest	Warm temperate wet forest	Subtropical wet forest	Tropical moist forest
4000-8000					Warm temperate rainforest	Subtropical rainforest	Tropical wet forest
> 8000							Tropical rainforest

Table 8.5 Holdridge lifezones. Shaded cells are Holdridge lifezones represented in Central America.

In consultation with forage experts, it was decided to discard the RIEPT ecosystem classification and instead create a classification based on Holdridge lifezones. This classification is summarised in Table 8.6.

Name	Definition
Tropical and subtropical wet and rain forest	T > 24 C and R > 4000 mm; or T 18-24 C and R > 2000mm
Tropical and subtropical moist forest	T > 24 C and R 2000-4000 mm; or T 18-24 C and R 1000-2000 mm
Tropical and subtropical dry forest	T > 24 C and R 1000-2000 mm; or T 18-24 C and R 500-1000 mm
Tropical and subtropical very dry forest and thorn woodland	T > 24 C and R < 1000 mm; or T 18-24 C and R < 500 mm
Temperate	T < 18 C

Table 8.6 Ecosystem classification based on Holdridge lifezones. T = mean annual temperature, R = annual rainfall.

Holdridge lifezones were also calculated from the RIEPT data using the same calculations. Analysis was then carried out to measure the level of agreement between RIEPT classifications and classifications at the same locations according to GIS data. It was found that 51 percent of classifications agreed and 41 percent were one class different over two-dimensional space. Cohen's kappa (κ) measures the amount of agreement between two datasets. Applying this to the above data, $\kappa = 0.33$, suggesting weak agreement. This is analogous to the low correlations in Table 8.2 and again suggests poor agreement between the RIEPT database and mapped data.

The remaining climate variables listed in Table 8.1 are solar radiation, relative humidity, hours of sun and wind speed. Of these, relative humidity is the only variable with more than half the records complete, and none were considered to add sufficient information to be included as predictor variables in the model.

Sand, silt and clay percentages can be combined to form texture categories. Soil texture classes in RIEPT are defined in agreement with USDA definitions, i.e., clay < 0.002mm particle size, silt 0.002 – 0.05mm and sand 0.05 – 2.0mm (USDA, 1993). The 12 classes found in the USDA texture triangle can be reclassified into five classes shown in Table 8.7. Figure 8.10 shows the relationship between percentage clay, silt and sand, and the categories.

Category	USDA Definition
Clay (Very fine)	Clay, silty clay, sandy clay
Clay loam (Fine)	Clay loam, silty clay loam, sandy clay loam
Loam (Medium)	Loam, silt loam, silt
Sandy loam (Coarse)	Sandy loam
Sand (Very coarse)	Sand, loamy sand

Table 8.7 Soil texture classification based on USDA texture triangle.

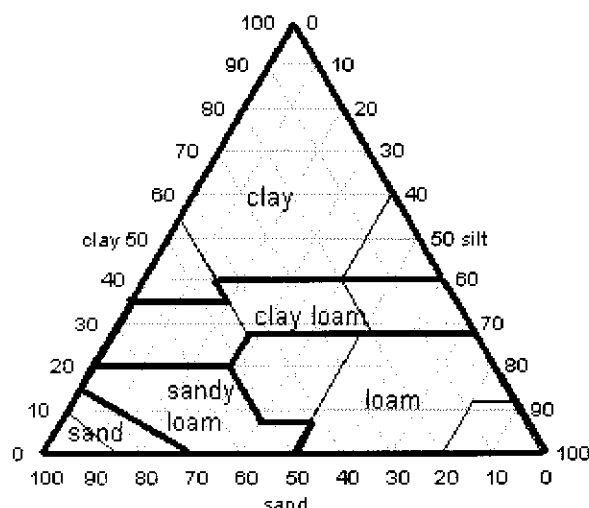


Figure 8.10 Soil texture classification based on USDA triangle. Thin black lines denote original USDA boundaries, bold black lines denote classification boundaries as in Table 8.7. Source: USDA, 1993.

Soil fertility depends on acidity (pH), organic matter (OM), phosphorus (P) and potassium (K), as well as on other minerals such as calcium (Ca), magnesium (Mg) and aluminium (Al). In the RIEPT database, Ca, Mg and Al are all correlated with pH. However, pH, OM, P and K show no correlations. The majority of sites have high potassium ($K > 0.15$ meq/100g). Therefore it was decided that pH should be retained as a separate variable, and that OM and P combined give a reasonable measure of soil fertility (Amezquita, *pers. comm.*, 2003) (Table 8.8).

Fertility	Organic matter				
Phosphorus	< 2%	2-3%	3-5%	5-6%	> 6%
< 2%	Very low	Very low	Very low	Very low	Very low
2-15%	Low	Low	Low	Low	Low
15-25%	Medium	Medium	Medium	High	High
25-30%	Medium	High	High	High	Very high
> 30%	High	High	High	Very high	Very high

Table 8.8 Soil fertility classification based on organic matter and phosphorus.

Additional edaphic variables listed in Table 8.1 are apparent density, field capacity, sodium and aluminium saturation. Apparent density and field capacity are both somewhat correlated with organic matter, and sodium is highly correlated with potassium. In addition these variables all have less than half their records complete, therefore it was considered irrelevant to include them as variables in the model.

The variables identified thus far for inclusion in the model are elevation, annual rainfall, length of dry season, Holdridge lifezones, soil pH, soil texture and soil fertility. These variables are incorporated into the SDSS which is described in the following chapters. A part of this process is eliciting expert knowledge relating these variables to the success of various forage species. During an iteration of this expert knowledge elicitation, issues arose regarding Holdridge lifezones. Although experts had helped identify the classification presented in Table 8.6, they did not intuitively relate these classifications to forage success when defining conditional probabilities. Instead, they consistently referred back to the definition of the ecosystem in terms of temperature and rainfall. As both of these are already accounted for (temperature with the proxy elevation), it was decided that Holdridge lifezones are probably not functionally independent from elevation and rainfall. Therefore, including Holdridge lifezones as a variable in the model would not add more information and, moreover, could cause the assumption of conditional independence to be violated. Instead, the underlying variables of elevation and annual rainfall are used.

In the discussion to date, only biophysical variables have been considered. This is primarily because the RIEPT database only includes biophysical variables. However, socio-economic data is available as GIS data and can therefore be associated with locations in the RIEPT database. Socio-economic variables that could influence forage selection were discussed in Chapter 4, namely, population

density, access to market and livestock density, as well as other variables derived from population censuses. However, whilst socio-economic variables have direct bearing on the strategic decision of whether to adopt forages, they have less bearing on the decision of which forages to trial, once it has already been decided to adopt forages.

Another issue with socio-economic data is that, by its nature, it is almost always aggregated. Census data, for example, is usually published at a level comprising a large number of households (such as district level). The data can be used to characterise the population of the entire district, but it is impossible to say anything about the socio-economic characteristics of a household at a single location.

For these reasons, socio-economic data has not been included directly as a variable in the functional model. However, it is recognised that socio-economic data may be useful at a regional level, when deciding for example where to target certain forages. In addition, as discussed in Chapter 4, socio-economic data may be of more value when applying the model to cash crops (e.g., distance to market becomes critical). Therefore, the possibility remains for socio-economic data to be included as variables in the model implementation in the future.

In consultation with forage experts, the selected variables were partitioned into five categories each (Table 8.9). This was an iterative process throughout the development of the SDSS and the resulting categories are believed by experts to best represent the expected variation in forage species' responses.

Variable	Class 1	Class 2	Class 3	Class 4	Class 5
Elevation	0-500m	500-1000m	1000-1500m	1500-2000m	>2000m
Rainfall	0-500mm	500-800mm	800-1200mm	1200-1800mm	>1800mm
Dry months	0-2	3-4	5-6	7-8	>8
Soil pH	Very acid (<4.5)	Acid (4.5-5.5)	Moderately acid (5.5-6.5)	Neutral (6.5-7.5)	Alkaline (>7.5)
Soil texture	Clay	Clay loam	Loam	Sandy loam	Sand
Soil fertility	Very low	Low	Medium	High	Very high

Table 8.9 Variable categories selected for SDSS development.

Now that the predictor variables have been categorised into discrete classes, the joint information uncertainty measure U defined in Equation 7.31 can be calculated to check for correlation between variables and hence potential violation of the assumption of CI. The calculated values for GIS data (elevation, rainfall and dry months) and RIEPT data (all variables) are shown in Table 8.10.

U	Elevation	Rainfall	Dry months	Soil pH	Soil texture	Soil fertility
Elevation	1 / 1					
Rainfall	0.06 / 0.03	1 / 1				
Dry months	0.03 / 0.05	0.26 / 0.17	1 / 1			
Soil pH	0.02	0.05	0.04	1		
Soil texture	0.03	0.04	0.02	0.04	1	
Soil fertility	0.03	0.04	0.03	0.06	0.04	1

Table 8.10 Joint information uncertainty for GIS data / RIEPT data (RIEPT data only for soil factors).

All values of U are low, indicating independence, except for rainfall compared with number of dry months. Therefore, these two variables are somewhat correlated. However, as previously discussed, these two variables are functionally independent in this decision problem. Even though it is possible that conditional independence is violated for this variable pair, they are both retained in the model as independent variables.

In addition, the agreement between RIEPT data and categorised mapped data can be assessed using Cohen's kappa on the variable categories. The results are shown in Table 8.11.

Elevation	Rainfall	Dry months
0.72	0.69	0.81

Table 8.11 Cohen's kappa comparing RIEPT and GIS data for three variables.

The kappa values suggest that there is very good agreement between classifications for RIEPT and GIS data. Therefore, despite the low correlations shown in Table 8.3, once categorised, the variables agree more closely between the two data sources.

8.1.6 Expert Knowledge

Expert knowledge has been identified as an important source of information for the model. Expert knowledge can be incorporated into the model in two ways. The first is to use published expert knowledge, such as the SoFT database, and derive ways of translating this to the variable classes described above. The second way is as an ad-hoc process, allowing experts to directly introduce knowledge for each variable for selected species.

The SoFT database is still in development and will not be available to the public until mid 2005. However, some preliminary data has been made available for use in this project. This data is discussed here with the caveat that classifications and data are preliminary and subject to change.

The expert knowledge in SoFT is formalised as a series of rules describing each forage species and its optimal conditions (Table 8.12). SoFT currently contains data on 155 forage species, and this number is increasing as further information is collected from forage experts around the world. Apart from the data listed in Table 8.12, fact sheets also exist (or are in development) for each species, with more detailed information and recommendations.

Both descriptions and optimal conditions are important in selecting forage species. Intended forage use is clearly an important management variable and should be included in the model as an initial filter. Other descriptive variables are not selected for inclusion in the model at this stage.

A number of the variables describing optimal conditions can be related to the variables identified as predictor variables, namely, latitude x altitude, rainfall, months without rain, soil texture, soil fertility and soil pH. In addition, a number of variables have been selected as filtering variables, namely, soil salinity, drought tolerance, seasonal waterlogging, frost tolerance and shade tolerance. These variables will be referred to as 'filter' variables in the discussion on model implementation, as they are used to filter out unsuitable species (i.e., species which do not match required level of tolerance to stresses).

SoFT variable	Type
Family	Plant description
Life cycle	
Intended forage use	
Growth form	
Habit	
Production potential	
Time to flowering	
Latitude x Altitude	Growing conditions
Rainfall	
Months without rain	
Soil texture	
Soil fertility	
Soil pH	
Level of available soil Al/Mn	
Soil salinity	
Soil drainage	
Ability to grow in cool growing season temperatures	
Photoperiod (day length)	
Drought tolerance	
Seasonal waterlogging	
Frost tolerance (foliage damage)	
Light (shade tolerance)	
Grazing pressure	
Tolerance to cutting	

Table 8.12 Variables in SoFT database.

In the SoFT database, the rules are logical IF...THEN rules, classifying each species as either suitable or unsuitable. In the case of rainfall, marginal limits are also sometimes given, meaning that rainfall in these ranges is marginally, but not optimally, suitable. From these rules environmental envelopes can be constructed, as in Figure 7.1.

The classifications in the SoFT database need to be translated to the variable categories in Table 8.9. This process is not always straightforward and causes some loss of information. The processes developed to achieve the translation are described in the following chapter.

Expert knowledge also exists apart from that formalised in the SoFT database. Forage experts at CIAT provided knowledge for certain species on an ad-hoc basis.

The benefit of this is that knowledge can be directly elicited to relate to the categories defined for each variable, whereas with SoFT data the process is somewhat convoluted and inaccurate. The concept is that RIEPT and SoFT data are automatically incorporated in the model, and additional expert knowledge can be included where necessary (where there is insufficient data from the databases or where considerable uncertainty is present).

8.2 Response Variables

As previously reported, the trial data in RIEPT is in three main databases, namely, Adaptation, Establishment and Production. For each species in the database, location is recorded, along with a number of measurements relating to how the species performed in the trial (Table 8.13).

Adaptation database	Classification
Adaptation	Poor, adequate, good, excellent
Cover	Percentage
Dry matter	Weight
Number of plants	Numeric
Resistance to insects	Yes / No
Resistance to diseases	Yes / No

Establishment database	Classification
Height at 9 weeks	Height
Cover at 9 weeks	Percentage
Insects at 9 weeks	Yes / No
Diseases at 9 weeks	Yes / No

Production database	Classification
Dry weight at 12 weeks	Weight
Cover at 12 weeks	Percentage
Height at 12 weeks	Height

Table 8.13 Trial variables in the RIEPT database.

From these, response variables must be selected. Although it is possible to model multiple response variables, it was decided to initially select just one. Some of the measurements are correlated and some are problematic for other reasons.

In the Adaptation database, correlation was found between adaptation and percent cover ($R^2 = 0.46$). Multiple evaluations are held during each trial, over a period of up to 24 weeks. Initial evaluation is defined as the evaluation at 4 weeks, or the closest record, and final evaluation is the one at 10 weeks, or closest record. Analysis showed strong correlation for adaptation, cover, number of plants and dry matter weight, between initial evaluation and final evaluation (see Appendix A).

In the Establishment and Production databases, no significant correlations were found. Correlation analysis between Adaptation, Establishment and Production databases cannot be directly performed because they relate to separate trials.

The measures relating to dry matter weight, number of plants and plant height are all problematic because of variation between different forage species. The measurements could conceivably be transformed to a common scale for all species, but there could be problems in defining the maximum for each individual species. In addition, plant height, for example, might not have the same significance for different species.

Percent cover has the potential to be a valid measurement which has the same meaning across all species. As percent cover is included in all three databases, it could be a way of utilising the larger number of species present in the Adaptation database, along with the larger number of locations present in the Establishment and Production databases. However, because percent cover is measured at different times for trials in the different databases, the measures are not directly comparable.

As percent cover and level of adaptation are positively correlated, albeit weakly, it is reasonable then to select the latter as the response variable. In addition, as this is a variable in the Adaptation database, it applies to a larger number of species than the other two databases. However, as discussed above, fewer trial sites are included in the Adaptation database.

Resistance to insects and diseases is also very important in forage selection. Although this information is included in RIEPT, it is considered as secondary to

level of adaptation. As just one response variable is being selected, resistance to diseases and insects is not considered at this point.

8.3 Summary of Predictor and Response Variables

Six predictor variables were identified from an analysis of the RIEPT database, GIS data and in consultation with forage experts. In addition, the SoFT database provides information on potential uses and tolerances, which are included as filter variables. These are summarised in Table 8.14.

Data	RIEPT	GIS	SoFT
Elevation	Raw value	1km grid (Jones, 2001)	Latitude x Altitude
Annual rainfall	Derived from monthly rainfall	Derived from monthly 1km grid (Jones, 2001)	Rainfall
Length of dry season	Derived from monthly rainfall	Derived from monthly 1km grid (Jones, 2001)	Months without rain
Soil pH	Raw value	FAO-derived coverage 55km grid (Batjes, 1996)	Soil pH
Soil texture	Derived from % clay and sand at 0-20cm	FAO-derived coverage 55km grid (Batjes, 1996)	Soil texture
Soil fertility	Derived from OM and P at 0-20cm	FAO-derived coverage 55km grid (Batjes, 1996)	Soil fertility
Filter variables			Intended forage use Soil salinity Drought tolerance Seasonal waterlogging Shade tolerance Frost tolerance

Table 8.14 Predictor variables and filter variables for SDSS implementation.

Figure 8.11 shows how data from the RIEPT database was combined to derive the six predictor variables identified for use in the model.

One single response variable has been identified, namely, level of adaptation from the Adaptation database. In the database, the level is identified as one of four values;

in Spanish these are '*Malo*', '*Regular*', '*Bueno*' and '*Excelente*'. These can be translated as 'Poor', 'Adequate', 'Good' and 'Excellent'.

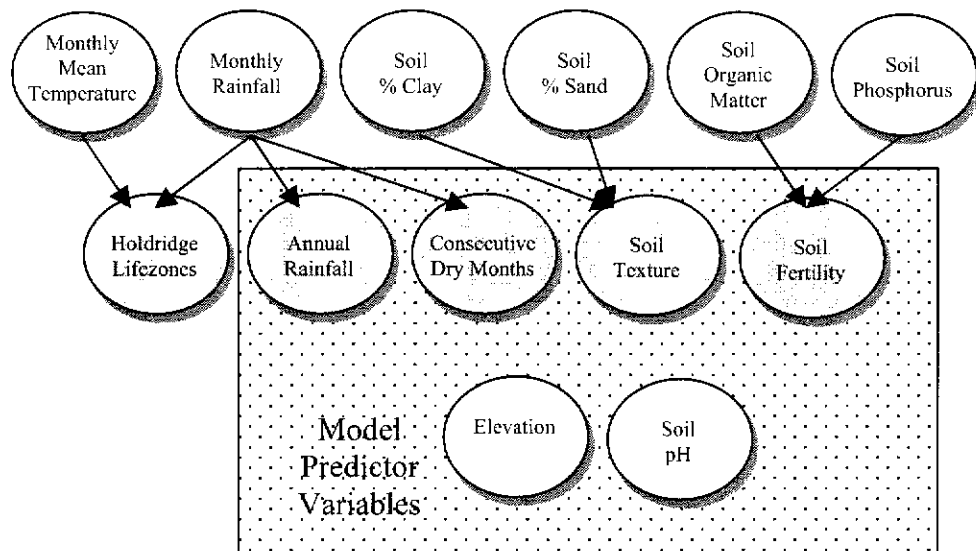


Figure 8.11 RIEPT data in model. Modified data is shaded, unmodified data is white.

8.4 Summary

In this chapter, the derivation of six predictor variables, six filter variables and one response variable was described. A large amount of data and information is available, and so a process was needed in order to select the most useful data and reduce it to a manageable and valid subset of variables. Variables and their categories were defined based on statistical analysis, functional equations and expert opinion.

The resulting variables will be incorporated into the probabilistic GIS model which in turn is implemented as an SDSS. The next chapter turns to the specification of the SDSS, the implementation of which is described in the subsequent chapter.

CHAPTER 9. SPATIAL DECISION SUPPORT SYSTEM

In Chapter 4, the necessary steps in DSS development were identified as needs analysis, design, implementation, capacity-building, fostering uptake and evaluation. The need for an SDSS for selecting forages was identified in Chapter 2, and the requirements for an SDSS to address the decision problem were analysed in subsequent chapters. The functional model design was developed in Chapters 5, 6 and 7. In this chapter, the discussion turns to design of the SDSS itself, based on a probabilistic GIS modelling approach and using the variables identified in Chapter 8.

The steps in implementing the SDSS are presented and, in the subsequent chapter, the SDSS that was consequently implemented is described and discussed. The SDSS is called ‘CaNaSTA’ (Crop Niche Selection in Tropical Agriculture) (*canasta* is Spanish for basket, and the tool aims to offer a basket of options to farmers, particularly in Spanish-speaking Central America).

The design and implementation of the SDSS are discussed with reference to the case study of selecting forage species in Central America.

9.1 Overcoming Potential Barriers to Uptake

A number of barriers to the successful uptake of DSS in agriculture, particularly the developing world, were discussed in Chapter 4. Those relating to design considerations include complex design and presentation of DSS, unrealistic requirements for monitoring data, the irrelevance and inflexibility of many DSS, lack of user confidence and poor data availability and quality.

The problem of poor data availability and quality is overcome partly by allowing uncertainty to be explicit in the structure of the model itself. Also, the incorporation of expert knowledge supplements unavailable data.

There are no requirements for the farmer to provide monitoring data as input to the DSS. The SDSS will be implemented as a stand-alone piece of software, meaning

there is no requirement for other software to be present on the user's computer (such as proprietary GIS or database software).

The design and presentation of the SDSS is intended to be simple and functional. At the most basic level, users need only be able to interpret maps in order to select their location. Displaying information in a number of ways (maps, graphs and numerically) should assist the user with understanding the outputs of the model in order to make better-informed decisions.

Lack of user confidence can be addressed by reducing and describing all sources of uncertainty in the model and by allowing the user's own knowledge to be incorporated.

9.2 Conceptual Model

9.2.1 Nomenclature

When the response variable is discussed generally, it is denoted Y and its possible states are denoted y_i , meaning that variable y is in state i . Specifically, the response variable 'adaptation level' is denoted by A and its possible states 'poor', 'adequate', 'good' and 'excellent' are denoted by a_p , a_a , a_g , and a_e respectively.

The prior probability distribution of adaptation for a species is written $P(A)$ and the single prior probability value that adaptation is excellent, for example, is written $P(a_e)$. The values of the prior distribution may be written (P_p, P_a, P_g, P_e) , which is shorthand for writing $P(a_p) = P_p$, $P(a_a) = P_a$, $P(a_g) = P_g$, and $P(a_e) = P_e$.

In the previous chapter, six predictor variables were identified. Where predictor variables are discussed in general, they are written X^j ($j = 1$ to 6), with potential classes that X^j can belong to denoted as x_k^j ($k = 1$ to 5). Where they are discussed specifically, they will be denoted as shown in Table 9.1, where the classes are those defined in Table 8.9.

Variable	Notation	Classes
Elevation	E	e_1, e_2, e_3, e_4, e_5
Rainfall	R	r_1, r_2, r_3, r_4, r_5
Dry months	D	d_1, d_2, d_3, d_4, d_5
Soil pH	H	h_1, h_2, h_3, h_4, h_5
Soil texture	T	t_1, t_2, t_3, t_4, t_5
Soil fertility	F	f_1, f_2, f_3, f_4, f_5

Table 9.1 Notations for predictor variables.

The full matrix of conditional probability distributions (the CPT) of adaptation, given the above variables, is written:

$$P(A | E, R, D, H, T, F) \quad (9.1)$$

The conditional probability distribution of adaptation, given that elevation is between 500 and 1000m (class 2), rainfall is between 800 and 1200mm (class 3), the length of the dry season is 5-6 months (class 3), soil pH is acid (class 2), soil texture is clay (class 1) and soil fertility is medium (class 3), is written:

$$P(A | e_2, r_3, d_3, h_2, t_1, f_3) \quad (9.2)$$

The single conditional probability value that adaptation is excellent, based on the same information, is given by:

$$P(a_e | e_2, r_3, d_3, h_2, t_1, f_3) \quad (9.3)$$

9.2.2 SDSS Design

The SDSS was designed as two related stand-alone components. The main component carries out posterior distribution calculations to produce species recommendations and maps as required. This component assumes prior and joint probabilities have already been defined, either from data or expert knowledge or a combination of both. In addition, suitable uses and level of tolerance to stresses are also assumed to have been defined. The secondary component allows for data entry and manipulation, and its main purpose is in building the knowledge base.

The modelling processes, i.e., selecting suitable species for a location and selecting suitable locations for a species, are carried out entirely within the main component. Posterior probability calculations are performed as required, based on prior and joint probability distributions stored in the knowledge base, along with spatial data and user inputs. The main component is the only component that most users will need to access. Spatial data for elevation, annual rainfall and number of dry months is stored in a format accessible to the SDSS, as a 1km grid for the extent of Central America. The knowledge base consists of prior and joint probabilities for each species related to each response variable, as well as filter variable thresholds for each species. User input is also required, to define locations and/or species of interest and to specify information such as local soil characteristics and desired forage use.

The following sections describe the steps in the modelling process and consider how they may be implemented within an SDSS.

9.3 Building the Knowledge Base

9.3.1 Joint and Conditional Probability Distributions

The primary model inputs are the prior and conditional adaptation distributions for each species. Joint probabilities are derived from the RIEPT database, from the SoFT knowledge base and from additional expert knowledge. These three sources complement each other, filling in gaps and often providing higher certainty regarding probability distributions. From these sources, conditional and prior probability distributions can be calculated. Probabilities are specified for all species in RIEPT and in SoFT. Expert knowledge can either update these probabilities or define probabilities for a completely new species.

In the case of RIEPT, the values in the database can be translated directly from counts to joint probabilities. The relationship between counts, joint probabilities and prior probabilities are represented in a numerical example in Table 9.2 below.

	Adaptation				
Rainfall	Poor	Adequate	Good	Excellent	Total / Prior
r_1	0 / 0.00	1 / 0.02	0 / 0.00	0 / 0.00	1 / 0.02
r_2	0 / 0.00	2 / 0.04	1 / 0.02	0 / 0.00	3 / 0.06
r_3	2 / 0.04	5 / 0.10	2 / 0.04	1 / 0.02	10 / 0.20
r_4	1 / 0.02	2 / 0.04	3 / 0.06	10 / 0.20	16 / 0.32
r_5	0 / 0.00	0 / 0.00	6 / 0.12	14 / 0.28	20 / 0.40
Total / Prior	3 / 0.06	10 / 0.20	12 / 0.24	25 / 0.50	50 / 1.00

Table 9.2 Frequency counts / joint probability values for rainfall class against adaptation class.

In the case of SoFT, a number of manipulations need to be carried out to translate the information in the SoFT knowledge base to adaptation probability distributions. This process is described in detail in Section 9.3.5 below.

Defining probability distributions from expert knowledge involves directly eliciting conditional probability distributions for each predictor variable for each species. When the conditional probabilities are defined using expert knowledge, the expert also assesses how certain they are in their probability definitions.

9.3.2 Potential Uses and Filter Variable Thresholds

Potential use and tolerances to environmental stresses (such as drought or frost) are Boolean in nature, that is, a species either meets the requirements or it does not. These are not used in the probability distribution calculations but are used to filter out unsuitable species.

Potential forage uses were initially identified in conjunction with forage experts at CIAT. The SoFT database also lists potential forage uses, and these are categorised slightly differently to those identified at CIAT. Because SoFT is the most comprehensive source of information relating forage uses to each individual forage species, it is reasonable to use the definitions from SoFT. The original list compiled at CIAT is compared below with the classifications defined in SoFT (Table 9.3).

Defined at CIAT	Defined in SoFT
Pasture	Long term pasture (>4 years)
Improved fallow	Short term pasture (phase / ley / improved fallow) (<3 years)
Cut and carry	Cut and carry
Hay	Conservation (hay/silage/leaf meal)
Silage	Intercropping
Live barriers	Green manure / mulch
Living fences	Ground cover (erosion control)
Green manure	Agroforestry
Cover or erosion control	Hedgerows
Dry season supplement	Living fences
Short term rapid use	Ponded pasture
Long term rapid use	Irrigated pasture
Concentrate	
Agroforestry	
Soil recuperation	
Intercropping	
Ponded pasture	
Feed for monogastrics	
Feed for fish	

Table 9.3 Potential forage uses defined at CIAT (left) and in SoFT (right).

There is a many-to-many relationship between forages and potential uses – that is, a selected forage can be multi-use, and for a selected use multiple forages may be suitable. The definition of use is binary – either a species is suitable or not suitable. If a species is not suitable for the intended use accorded it in SoFT's knowledge base, it is filtered out of the selection set. Using the list of potential uses defined at CIAT would require experts to individually identify potential uses for each forage species. Therefore, it is more straightforward to use the potential uses defined in SoFT. For species which are in RIEPT, but not in SoFT, potential uses need to be defined on a case-by-case basis from expert knowledge or literature.

The other filter variables (tolerances to stresses) are not binary, however, but multi-state. These variables are also defined in SoFT, and their possible states are listed in Table 9.4 below.

Variable	States
Soil salinity	Low Medium High
Drought tolerance	Persists Stays green Continues to grow
Seasonal waterlogging	< 1 week < 1 month > 1 month
Frost tolerance (foliage damage)	Persists, but burnt off by frost Stays green Continues to grow
Light (shade tolerance)	Full sun Light shade Heavy shade

Table 9.4 Filter variables in SoFT.

These states are interpreted as follows. The user of the SDSS can identify four different levels of tolerance, namely ‘none’, ‘low’, ‘moderate’ and ‘high’. Taking the example of drought tolerance, if the user indicates the level of tolerance required as ‘none’, then the filter variable is not activated and no species are filtered out. This means that species may be selected that would die if there is a drought, but if drought tolerance is identified as ‘none’, this means drought conditions are not expected to be an issue. If the user selects ‘low’ tolerance to drought required, then all species which persist, stay green or continue to grow are included. If the user indicates that ‘moderate’ tolerance is required, then only species which stay green or continue to grow are included. Finally, if tolerance required is ‘high’, then only species which continue to grow are included and all others are filtered out. The process is the same for the other filter variables.

9.3.3 Joint and Conditional Probabilities from RIEPT

Joint probabilities can be read directly from the RIEPT database, as shown in Table 9.2. Calculating conditional probabilities is then a matter of applying Equation 7.8 or 7.9. For example, the conditional probability that adaptation is good, given that rainfall is in class 5, is given by:

$$P(a_g | r_5) = \frac{P(a_g, r_5)}{P(r_5)} = \frac{0.12}{0.40} = 0.30 \quad (9.4)$$

Similarly, the conditional probability that rainfall is in class 5, given that adaptation is good, is given by:

$$P(r_5 | a_g) = \frac{P(r_5, a_g)}{P(a_g)} = \frac{0.12}{0.24} = 0.50 \quad (9.5)$$

The conditional probability distribution of adaptation, given a certain rainfall class, can therefore be calculated by normalising across a row, and the conditional probability distribution of rainfall, given a certain adaptation class, can be calculated by normalising down a column in Table 9.2.

In addition, certainty values need to be assigned to the conditional probability distribution for each class of each variable for each species. When the conditional probabilities are calculated from RIEPT, the level of certainty depends on how many trials exist for the species for that variable class. If less than two trials exist, certainty is 'low'. If up to five trials exist, certainty is 'medium' and otherwise certainty is 'high'.

Recall from Equation 7.12 that causality for conditional probabilities can be inverted. In the example above, the conditional probability in Equation 9.4 can be derived from the conditional probability in Equation 9.5 as follows:

$$P(a_g | r_5) = \frac{P(a_g)P(r_5 | a_g)}{P(r_5)} = \frac{0.24 \cdot 0.50}{0.40} = 0.30 \quad (9.6)$$

9.3.4 Prior Probability Distributions

Prior probability distributions are required both for predictor and response variables. For predictor variables, prior probability is denoted $P(X)$ and for the response

variable (adaptation), prior probability is denoted $P(A)$. Recall that the prior probability is equal to the sum of joint probabilities (see Equation 7.13), that is:

$$P(a_i) = \sum_j P(a_i, x_j) \quad (9.7)$$

and

$$P(x_j) = \sum_i P(a_i, x_j) \quad (9.8)$$

Therefore, in most cases, once joint probabilities are known, prior probabilities can be calculated. For example, from Table 9.2, the prior probability distribution for rainfall is given by:

$$P(R) = (0.02, 0.06, 0.20, 0.32, 0.40)$$

and the prior probability distribution for adaptation is given by:

$$P(A) = (0.06, 0.20, 0.24, 0.50)$$

Note that $\sum_j P(r_j) = 1$ and $\sum_i P(a_i) = 1$.

However, prior probability distributions for the predictor variables must be the same across all species, and prior probability distributions for the response variable (adaptation) must be the same across all variables, for a given species. The approach for ensuring these equalities is discussed in Section 9.3.8 below.

9.3.5 Joint and Conditional Probabilities from SoFT

The SoFT database was described in the previous chapter. The variables in SoFT that relate to the predictor variables are listed below, along with the categories defined in SoFT (Table 9.5).

Predictor variable	SoFT variable	SoFT Classification
Elevation	Latitude x Altitude	Tropics, 0-1000m Tropics, 1000-2000m Tropics, 2000-2500m Tropics, 2500m+ Subtropics, 0-1000m Subtropics, 1000-2000m Subtropics, 2000-2500m Subtropics, 2500m+
Rainfall	Rainfall	Range (defined by absolute lower limit, optimal lower limit, optimal upper limit and absolute upper limit)
Dry months	Length of dry period	< 1 month up to 3 months up to 6 months up to 9 months > 9 months
Soil pH	Soil pH	Strongly acidic (< 5.0) Mildly acidic (5.0 – 6.5) Neutral Alkaline (> 7.5)
Soil texture	Soil texture	Sand / sandy loam (light) Loam / clay loam (medium) Heavy clay (heavy)
Soil fertility	Soil fertility	Low Medium High

Table 9.5 SoFT variables related to predictor variables.

The first task, therefore, is to relate the categories in SoFT to the categories already defined for predictor variables. These relationships are a mixture of one-to-one, one-to-many, many-to-one and many-to-many, and are shown in Table 9.6 below.

Recall that SoFT uses binary rules to classify species as either suitable for the given environment or not suitable, except in the case of rainfall where both optimal and absolute limits are given. In order to convert these judgements to adaptation probability distributions, the following assumptions were made, in consultation with CIAT forage experts (who are also involved in the SoFT project).

Variable	SoFT classes	Defined classes
Elevation	Tropics 0-1000m	0-500m
	Subtropics 0-1000m	500-1000m
	Tropics 1000-2000m	1000-1500m
	Subtropics 1000-2000m	1500-2000m
	Tropics 2000-2500m	>2000m
	Tropics 2500m+	
	Subtropics 2000-2500m	
	Subtropics 2500m+	
Rainfall	Ranges	Linear transformation
Dry months	< 1 month	0-2 months
	up to 3 months	
	up to 6 months	3-4 months
		5-6 months
	up to 9 months	7-8 months
	> 9 months	> 8 months
Soil pH	Strongly acidic (< 5.0)	Very acid (<4.5)
		Acid (4.5-5.5)
	Mildly acidic (5.0 – 6.5)	Moderately acid (5.5-6.5)
	Neutral	Neutral (6.5-7.5)
	Alkaline (> 7.5)	Alkaline (>7.5)
Soil texture	Heavy clay (heavy)	Clay
	Loam / clay loam (medium)	Clay loam
		Loam
	Sand / sandy loam (light)	Sandy loam
		Sand
Soil fertility	Low	Very low
		Low
	Medium	Medium
	High	High
		Very high

Table 9.6 Transforming SoFT categories to predictor variable categories.

9.3.6 Transforming SoFT Categories to CaNaSTA Categories

The process to transform SoFT categories to CaNaSTA categories consists of two steps. The first step transforms a SoFT evaluation of 'suitable' or 'not suitable' to probability distributions. The second step then translates these distributions to the categories defined for the predictor variables. In the first step, the probabilities are set for 'poor', 'adequate', 'good' and 'excellent' adaptation. If SoFT data indicates suitability for a given species, then the probability distribution is set to (0, 0, 0.5, 0.5), that is, adaptation is either 'good' or 'excellent', with equal probability. Certainty is flagged as 'high'. If SoFT data does not indicate suitability,

then this does not necessarily mean that conditions are unsuitable. Therefore, while the distribution is set to (0.5, 0.5, 0, 0), certainty is flagged as 'low'.

In the next step, SoFT ranges need to be transformed to CaNaSTA ranges. The simplest case is a one-to-one relationship, such as 'Heavy clay (heavy)' → 'Clay', where the probability distribution and certainty value are simply copied. With a one-to-many relationship, such as 'Loam / clay loam (medium)' → 'Clay loam' and 'Loam', the probability distribution and certainty value are simply copied for both categories.

In the case of a many-to-one relationship, such as '< 1 month' and 'up to 3 months' → '0-2 months', the distributions and certainty values are averaged. Therefore, if the probability distribution for '< 1 month' is (0, 0, 0.5, 0.5) with certainty 'high' and the probability distribution for 'up to 3 months' is (0.5, 0.5, 0, 0) with certainty 'low', then the probability distribution for '0-2 months' is set to (0.25, 0.25, 0.25, 0.25) with certainty 'medium'.

The case of many-to-many (as occurs for elevation) is simply an extension of these methods. The distribution values are averaged for 'Tropics 0-1000m' and 'Subtropics 0-1000m' and the averaged distribution is applied to both '0-500m' and '500-1000m'.

Rainfall is treated differently because, rather than being categorised, optimal and absolute limits are given (although in some cases only optimal limits are given, or only one of the absolute limits are given). The probability distribution for the range within the optimal limits is set to (0, 0, 0.5, 0.5) with certainty 'high'. The probability distribution for the range within the absolute limits (but outside the optimal limits) is set to (0, 0.1, 0.8, 0.1) with certainty 'medium' and the probability distribution for the range outside of the absolute limits is set to (0.5, 0.5, 0, 0) with certainty 'low'. These distributions are then combined proportionately to calculate probability distributions for the predefined classes. Certainty values are similarly combined proportionately, with the dominating certainty value being carried over.

To illustrate, consider a rainfall range in SoFT given as (600-)1200-2500(-4500), that is, the optimal rainfall range is 1200-2500mm and the absolute rainfall range is 600-4500mm. Probability distributions are then set as follows (Table 9.7).

Range	Suitability	Adaptation Distribution	Certainty
0-600mm	Unsuitable	(0.5, 0.5, 0, 0)	Low
600-1200mm	Marginal	(0, 0.1, 0.8, 0.1)	Medium
1200-2500mm	Suitable	(0, 0, 0.5, 0.5)	High
2500-4500mm	Marginal	(0, 0.1, 0.8, 0.1)	Medium
>4500mm	Unsuitable	(0.5, 0.5, 0, 0)	Low

Table 9.7 Probability distributions and certainty values for rainfall ranges in SoFT.

The next step is to distribute these proportionally across the predefined rainfall classes, as shown in Table 9.8 below. In order to define proportions relating to the uppermost range (>2500mm), a practical upper limit of 5000mm rainfall is defined.

Class	Ranges in SoFT	Calculation	Adaptation Distribution	Certainty
< 500mm	0-600mm	$1 \times (0.5, 0.5, 0, 0)$	(0.5, 0.5, 0, 0)	Low
500-800mm	0-600mm 600-1200mm	$0.33 \times (0.5, 0.5, 0, 0) + 0.67 \times (0, 0.1, 0.8, 0.1)$	(0.17, 0.23, 0.53, 0.07)	Medium
800-1200mm	600-1200mm	$1 \times (0, 0.1, 0.8, 0.1)$	(0, 0.1, 0.8, 0.1)	Medium
1200-1800mm	1200-2500mm	$1 \times (0, 0, 0.5, 0.5)$	(0, 0, 0.5, 0.5)	High
>1800mm	1200-2500mm 2500-4500mm 4500-5000mm	$0.22 \times (0, 0, 0.5, 0.5) + 0.62 \times (0, 0.1, 0.8, 0.1) + 0.16 \times (0.5, 0.5, 0, 0)$	(0.08, 0.14, 0.61, 0.17)	Medium

Table 9.8 Transforming SoFT rainfall probability distributions to predefined classes.

9.3.7 Calculating Prior Probabilities for Adaptation

As discussed above, prior probabilities can be calculated directly from joint probabilities. However, problems arise where zero counts are encountered for row or column totals, that is, prior probabilities are zero. This situation arises when there are no trials in the database for either a particular variable class or for a particular adaptation class.

The probability distribution that ultimately needs to be calculated (from Equation 7.24) is:

$$P(A|E, R, D, H, T, F) = P(A) \frac{P(A|E)}{P(A)} \frac{P(A|R)}{P(A)} \frac{P(A|D)}{P(A)} \frac{P(A|H)}{P(A)} \frac{P(A|T)}{P(A)} \frac{P(A|F)}{P(A)} \quad (9.9)$$

The above equation assumes that causal conditional probabilities are known; when frequency counts are being directly read from a database, then the equation can be written using diagnostic conditional probabilities.

When joint probabilities are known rather than conditional probabilities (as is in fact the case when reading directly from a database), then from Equation 7.8, Equation 9.9 is equivalent to:

$$P(A|E, R, D, H, T, F) = P(A) \frac{P(A, E)}{P(A)P(E)} \frac{P(A, R)}{P(A)P(R)} \frac{P(A, D)}{P(A)P(D)} \frac{P(A, H)}{P(A)P(H)} \frac{P(A, T)}{P(A)P(T)} \frac{P(A, F)}{P(A)P(F)} \quad (9.10)$$

Hence it should be clear that if any prior probability value is zero, then problems will ensue through attempting to divide by zero. The case where adaptation priors are zero is treated slightly differently to the case where predictor variable priors are zero, but in both cases the approach in this research is to replace the priors with a reasonable non-zero estimate.

In the case where adaptation priors are zero, two distinct situations can occur. The first is where $P(a_i) = 0$ across all predictor variables. This situation can occur when there are no entries in the database of trials of a species where adaptation is, say, 'poor'. In this case, the approach is to simply set $P(a_i) = \alpha$, where α is a very small positive number. Joint probabilities $P(a_i, x_j)$ are set to $\alpha/5$. Setting joint probabilities to an equal distribution implies, correctly, that no information is known about the distribution (complete ignorance).

The second situation can occur when there are fewer entries in the database for one variable than for another. For a given species, say $\sum_j P(a_p, e_j) = 0.01$, which might

equate to just one frequency count for poor adaptation. If this one record in the database is missing information for soil fertility, then $\sum_j P(a_p, f_j) = 0$.

However, the prior probability distribution must be equal for all variables, that is, the following equality must hold:

$$\begin{aligned} P(a_i) &= \sum_j P(a_i, e_j) = \sum_j P(a_i, r_j) = \sum_j P(a_i, d_j) = \\ &= \sum_j P(a_i, h_j) = \sum_j P(a_i, t_j) = \sum_j P(a_i, f_j) \end{aligned} \quad (9.11)$$

Therefore, the above situation can be remedied by setting $P(a_p) = 0.01$ and setting each joint probability $P(a_p, f_j) = 0.002$, again under conditions of complete ignorance. Incidentally, these adjustments need to be made whenever prior adaptation probabilities are not equal across variables. In non-zero cases, adjustments are made so that priors agree, and the joint probability distributions stay as close to their original distributions as possible.

For species which are only present in SoFT, prior probability distributions for adaptation could be defined under complete ignorance, that is, set to (0.25, 0.25, 0.25, 0.25). However, for each predictor variable, the following equality must be satisfied (see Equation 6.15):

$$P(A) = \sum_j P(A | x_j) P(x_j) \quad (9.12)$$

To illustrate, consider the information in SoFT for *Stylosanthes guianensis*. Transforming the information given using the methods described above, the following conditional probability distributions are defined for elevation, rainfall and dry months, shown in Table 9.9 below.

Variable	Class 1	Class 2	Class 3	Class 4	Class 5
Elevation	(0, 0, 0.5, 0.5)	(0, 0, 0.5, 0.5)	(0, 0, 0.5, 0.5)	(0, 0, 0.5, 0.5)	(0.5, 0.5, 0, 0)
Rainfall	(0.5, 0.5, 0, 0)	(0.17, 0.23, 0.53, 0.07)	(0, 0.1, 0.8, 0.1)	(0, 0, 0.5, 0.5)	(0.08, 0.14, 0.61, 0.17)
Dry months	(0.25, 0.25, 0.25, 0.25)	(0, 0, 0.5, 0.5)	(0, 0, 0.5, 0.5)	(0, 0, 0.5, 0.5)	(0.5, 0.5, 0, 0)

Table 9.9 Conditional probabilities $P(A | x_i)$ for *S. guianensis*.

Calculating $P(A)$ based on conditional probabilities and prior probabilities for each variable separately, from Equation 9.10 above, gives:

$$\begin{aligned}
 P(A) &= \sum_j P(A | e_j) P(e_j) = (0,0,0.5,0.5) \cdot 0.69 + (0,0,0.5,0.5) \cdot 0.19 + \\
 &\quad (0,0,0.5,0.5) \cdot 0.07 + (0,0,0.5,0.5) \cdot 0.02 + (0.5,0.5,0,0) \cdot 0.03 \\
 &= (0.015, 0.015, 0.485, 0.485)
 \end{aligned}$$

$$\begin{aligned}
 P(A) &= \sum_j P(A | r_j) P(r_j) = (0.5,0.5,0,0) \cdot 0.02 + (0.17,0.23,0.53,0.07) \cdot 0.06 + \\
 &\quad (0,0.1,0.8,0.1) \cdot 0.19 + (0,0,0.5,0.5) \cdot 0.32 + (0.08,0.14,0.61,0.17) \cdot 0.41 \\
 &= (0.053, 0.100, 0.594, 0.253)
 \end{aligned}$$

$$\begin{aligned}
 P(A) &= \sum_j P(A | d_j) P(d_j) = (0.25,0.25,0.25,0.25) \cdot 0.15 + (0,0,0.5,0.5) \cdot 0.13 + \\
 &\quad (0,0,0.5,0.5) \cdot 0.29 + (0,0,0.5,0.5) \cdot 0.42 + (0,0,0.5,0.5) \cdot 0.01 \\
 &= (0.038, 0.038, 0.463, 0.463)
 \end{aligned}$$

The method employed in this research is to average the values of $P(A)$ over all variables. In the case with three variables above, $P(A)$ becomes (0.033, 0.051, 0.513, 0.400). The example above considers just three variables, however, in practice all six predictor variables are considered. Conditional probability values are then adjusted so that Equation 9.10 holds for each variable. This is done in an iterative process. In the first iteration, the joint probabilities are multiplied by a constant so that they sum to the calculated $P(A)$ (down columns in Table 9.2). However, now the joint probabilities will no longer sum to the desired priors for the variables, so they are again multiplied by a constant so they sum to the desired $P(X)$ (across rows in Table 9.2). These steps are repeated until the sum down columns is close to $P(A)$

and the sum across rows is close to $P(X)$. This process attempts to retain the relative proportions in the conditional distributions whilst satisfying equations 9.7 and 9.8.

9.3.8 Calculating Prior Probabilities for Predictor Variables

When predictor variable prior probabilities are zero, this is because no records are found in the database in a particular class of the variable. For example, if $P(r_1) = 0$ for a particular species, this means that there are no trials for the species where rainfall is less than 500mm. However, there could still be a location under consideration where rainfall is in fact in this class. In this case, the best assumption to make is to set prior probability distributions to the same distribution that is found in the area under consideration, in this case Central America. Distributions for each predictor variable have been calculated from spatial data for agricultural land only, and are displayed in Table 9.10.

Note that for soil variables, these percentages are derived from the classifications applied to FAO soil maps, where soil pH, soil texture and soil fertility are split into three classes each. In order to distribute these over five classes, the top and bottom classes are each divided by two. This is equivalent to saying that where soil fertility is classed as ‘low’ from the FAO classification, there is equal probability that it could, in fact, be ‘very low’ or ‘low’ in the classification in Table 9.10.

Variable	Class				
	1	2	3	4	5
Elevation	0.69	0.19	0.07	0.02	0.03
Rainfall	0.02	0.06	0.19	0.32	0.41
Dry months	0.15	0.13	0.29	0.42	0.01
Soil pH	0.19	0.19	0.38	0.12	0.12
Soil texture	0.36	0.36	0.10	0.09	0.09
Soil fertility	0.15	0.16	0.45	0.12	0.12

Table 9.10 Prior probabilities derived from spatial data for predictor variables across all agricultural land in Central America.

9.3.9 Combining Data Sources

RIEPT and SoFT contain information on different forage species, but there is substantial overlap. In addition, experts may update probability distributions for a species already in the database or may define distributions for new species. Ultimately, however, one unique 'master' set of prior and conditional probability distributions is required for each species.

A number of update rules are possible. The rule set implemented is as follows:

1. If only one source of data exists, use it
2. If expert knowledge is one of the sources, use it
3. If sources of data have different certainty levels, use the source of data with the highest certainty
4. If sources of data have the same certainty levels, use data averages.

The process for defining conditional probability distributions, using the update rules, is conceptualised in Figure 9.1 below

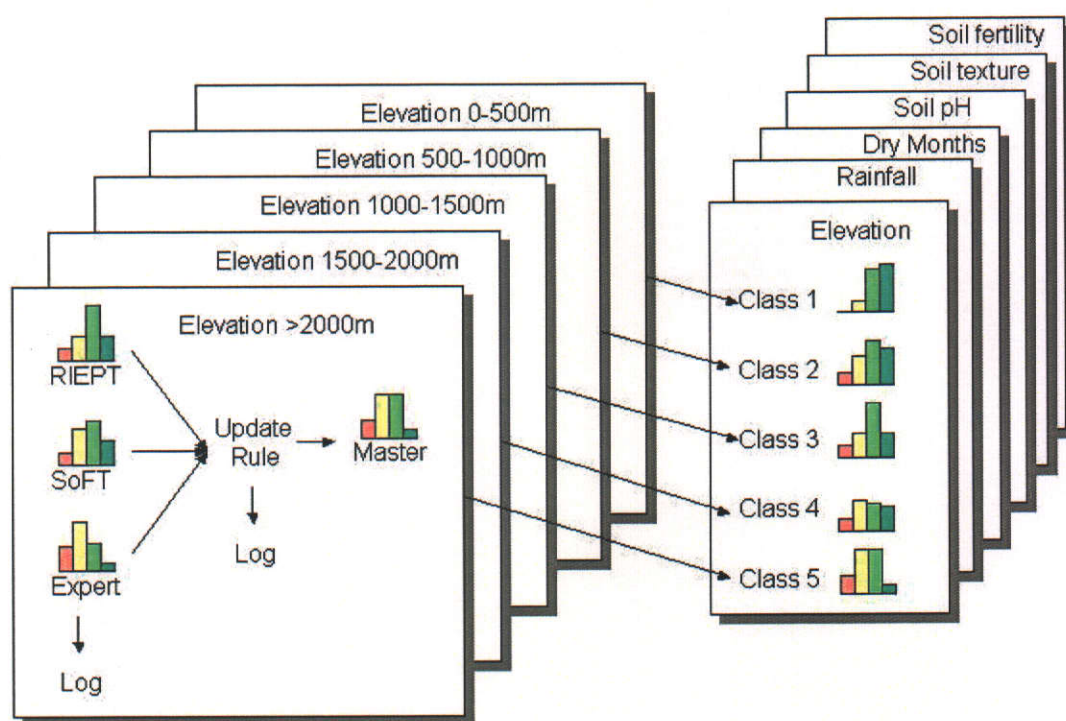


Figure 9.1 Defining conditional probability distributions.

9.4 Model Calculations

9.4.1 Calculating Posterior Probability Distributions

Once conditional and prior probability distributions have been defined, the full CPT can be calculated using Equation 9.9. The suitability value is calculated directly from the posterior probability distribution $P(A | e, r, d, h, t, f)$ using Equation 9.4. The certainty value is calculated from the individual certainty values associated with each $P(A | x)$. Representing 'High' certainty with the value '2', 'Medium' with '1' and 'Low' with '0', these values are simply averaged.

The conditional and prior probability distributions in the 'master' file are then combined using Equation 7.24, to define a full CPT for each species. In addition, potential uses and thresholds for filter variables are stored for each species (Figure 9.2).

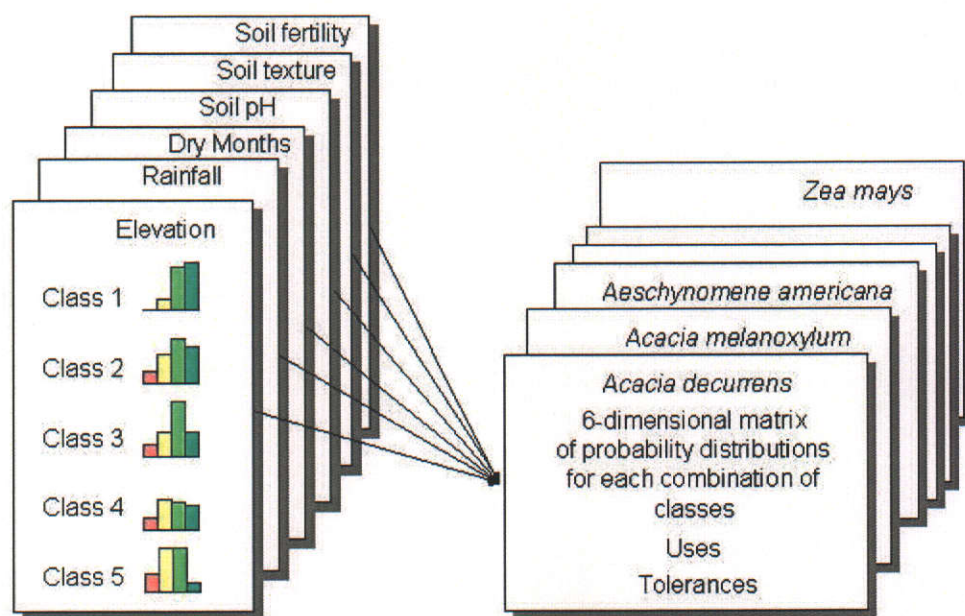


Figure 9.2 Defining CPT for each species.

9.4.2 Suitability and Sensitivity

In order to rank species, one value is required rather than the full adaptation probability distribution – note that this is only for ranking purposes, and full distribution information is retained. The ranking value is the suitability score and is determined using the following equation.

$$Score = \left((P(a_e) + P(a_g)) + \frac{P(a_e)}{P(a_e) + P(a_g)} \right) / 2 \quad (9.13)$$

This is the average of the probability of adaptation being ‘excellent’ or ‘good’, and the proportion of that probability value that is ‘excellent’. In this way, the score reflects the nature of the adaptation distribution more comprehensively than if only probability ‘excellent’, or probability ‘good’ or ‘excellent’ were used. Table 9.11 shows an example of ranking calculation for five species based on the score calculated in Equation 9.13.

Species	$P(a_e)$	$P(a_g)$	$Score$	Rank
<i>A</i>	0.86	0.07	0.93	1
<i>B</i>	0.61	0.28	0.86	2
<i>C</i>	0.00	0.92	0.46	3
<i>D</i>	0.11	0.10	0.37	4
<i>E</i>	0.01	0.45	0.23	5

Table 9.11 Ranking calculation example for five species based on the score value.

Although species *C* has an overall probability of 0.92 that adaptation will be ‘good’ or ‘excellent’, the fact that probability of ‘excellent’ adaptation is 0.00 reduces the overall score. Conversely, species *D*’s score is increased from 0.11 (probability ‘excellent’) and 0.21 (probability ‘good’ or ‘excellent’) to 0.37, because of the relatively large proportion of probability ‘excellent’. As the score ignores the probability values for ‘poor’ and ‘adequate’, it is not a complete summary of the probability distribution. However, for the purposes of ranking it reflects the essential characteristics of the probability distribution.

Another value associated with the posterior probability distribution is sensitivity, as described in Section 7.2.7. Sensitivity is evaluated from changes in the adaptation

distribution as the states of predictor variables change. Some variables may be more sensitive than others, for a given species. Sensitivity is computed by calculating in how many instances the adaptation distribution changes significantly if only one variable changes into an adjacent state.

As with the suitability ranking, a single value of sensitivity is needed, even though full sensitivity information is retained and can be queried by the user. Firstly, the difference in suitability score is calculated when one variable changes to an adjacent class. This yields 12 values (six variables in two directions each). Two measures can be combined to determine a measure of sensitivity, namely, the maximum difference in suitability score and the average of all difference scores:

$$\text{Sensitivity} = |\max(d_i) - \text{ave}(d_i)| \quad (9.14)$$

where d_i is the difference in suitability scores between the i^{th} value and the original value.

9.5 Model Inputs and Outputs

9.5.1 User Inputs

Recalling Figure 7.8, the first input required from the user is selection of a location. Elevation, annual rainfall and number of dry months can then be determined from spatial data, both for the selected location and for the surrounding area. The size and resolution of the surrounding area depends on the scale and extent chosen by the user, as discussed further in Chapter 10. The process of selecting a location can make full use of the spatial capabilities of an SDSS. Firstly, roads, rivers, towns and administrative boundaries can be displayed in order to help the user find the location of interest. Secondly, standard GIS procedures can be used to navigate around the map, including zooming, panning and identifying location attributes. Lastly, as all spatial data is geographically referenced, locations can be identified by latitude and longitude.

Because soil information is not included as spatial data, information on soil pH, texture and fertility need to be provided by the user. However, if any of these values are unknown, this simply means that this variable has no influence on the posterior probability calculations.

Other information required from the user is the intended use(s) of the selected forage, any tolerances required and the level of risk-aversion of the farmer in question. Based on this information, it is a straightforward process to filter out any forage species that do not match intended uses or required tolerances.

Based on these inputs, posterior probability distributions are calculated for each species not already filtered out. Probability distributions are calculated for the combination of variables at the location of interest.

9.5.2 Model Outputs

There are a number of outputs possible, depending on which information the user requires. The first is a ranked list of suitable species for a location, with various associated measures. These are posterior adaptation distribution, suitability score, level of certainty, sensitivity of each variable for a selected species and a combined sensitivity score. The process for calculating suitability and sensitivity is described in Section 9.4.2.

At the same time, the certainty value associated with the posterior probability is calculated based on the certainty values of each joint probability distribution. If risk-aversion is identified as low, then species are ranked purely by the ranking value. If risk-aversion is identified as high, then species where the probability distribution is associated with low certainty are excluded, and those with medium certainty are ranked lower. If risk-aversion is medium, then those with low certainty are not excluded, but ranked lower. In this way, if the farmer is identified as risk-averse, then species with lower certainty will be penalised, even if the probability score is high. Level of risk-aversion is used to rank all potentially suitable species. If risk-aversion is low, then the species with the highest probability of adaptation is ranked first. If risk-aversion is high, then probability of adaptation is traded off against

certainty, and certainty must be high before a species is recommended. If risk-aversion is medium, then species with low certainty for the given combination of factors are excluded.

In order to display a map of species' adaptation, posterior probability distributions are required, not just for the location identified but also for the surrounding area. To produce these maps the full CPT must be specified, rather than just for the states of the variables at the selected location. In the first instance, the full CPT is calculated only for the top five ranked species. However, the user may select a species further down the list, at which point the necessary calculations are performed. The full CPT initially need only be calculated for the three variables that have associated spatial data, namely, elevation, rainfall and dry months. This requires $4 \times 5 \times 5 \times 5 = 500$ values to be specified for each species, so in total 2,500 values (top five species).

Each cell on the map is then assigned a posterior probability distribution and the associated certainty value. This information can be displayed spatially in a number of ways, including most likely adaptation class, probability of adaptation being 'good' or 'excellent' and by the suitability score.

The spatial variation of soil variables is undefined in this implementation of the SDSS because soil maps are not being used. Therefore, all probability distributions across the map can be updated using a constant probability distribution for each soil variable. If the user wishes to examine the effect of a soil variable changing, then the entire map is simply updated using the new probability distribution. This allows the user to examine the effect of certain soil characteristics whilst retaining the spatial variability of elevation and climate data.

A user may also wish to know where a certain forage species will be suitable. In this case, the process is much more straightforward than the one described above. Once the user has selected the species, the full CPT is calculated for that species. The results can then be displayed in the same manner as described above. If the user selects a small enough scale, then it is valid to include spatial soil data, in which case this is also included in the display. In addition, population density data and access to

market can be displayed, although in this implementation they are not incorporated in the probability calculations.

9.6 Summary

A decision support system can aid in the decision making process of which forage species to adopt. Decision support systems in agriculture have been available for decades, but uptake is poor. Ways to overcome potential issues were discussed in the first section of this chapter.

Sources of information to develop the SDSS were discussed in the previous chapter. In this chapter, it was shown how prior and conditional probability distributions can be derived from these sources. When data is sparse or missing, it can be difficult to specify some probability distributions. Methods have been devised to overcome this difficulty, including defining prior probability distributions from spatial data. In addition, information from SoFT is extracted, including potential forage use and a number of filter variables.

By allowing multiple data sources to be used, the model is strengthened, as it is not reliant on a single database or a single expert. Where certainty in a probability distribution is low, this is flagged, and the information is retained throughout the modelling process.

This implementation therefore allows for sparse and uncertain data, works with expert knowledge and deals with uncertainty. In addition, the concept of the modelling process is fairly intuitive to follow, without the need for the user to understand the equations. Therefore, a user can decide how much faith to put in the information, because the entire process is transparent. In supporting decision-making, it has been suggested earlier in this research that reduction of structural and translational uncertainty may be more important than reduction of metrical uncertainty. This implementation attempts to present accurate results (low metrical uncertainty), but at the same time concentrates on delivering a structurally uncomplicated model and providing results that are straightforward in their interpretation.

This process therefore addresses most of the problems encountered with other agricultural DSS and SDSS.

The following chapter describes the implementation of CaNaSTA, software designed for Crop Niche Selection in Tropical Agriculture, based on the methods described in this chapter and in the research as a whole.

CHAPTER 10. IMPLEMENTATION

The previous chapters described the model selected to support decision-making for tropical forage adoption. In order to make the model accessible to its intended users, it has been implemented as stand-alone software, called CaNaSTA (Crop Niche Selection in Tropical Agriculture). The approach and software design are described in this chapter.

The current implementation of CaNaSTA is for the case study of tropical forage species in Central America.

10.1 Approach and Objectives

The design of the tool is based on the research described in the previous chapters. The guiding principles for design of the software include ease of use, flexibility and transparency. The tool must be simple to navigate and intuitive to use, and the results must be easy to interpret. Minimal training should be required for the tool to be used effectively. Transparency means that the user should be able to clearly see where the results come from and how much confidence can be placed in them.

As previously identified, intended users are extension workers, NGOs, development projects, and scientists from national research and international research institutions involved in tropical agriculture. To a limited extent, the product may also be useful for educational purposes.

10.2 Software Design

Software has been developed based on the conceptual model presented in the previous chapter. The software is called CaNaSTA and has been developed using Borland Delphi 6 (Borland Software Corporation, 2002) and ESRI MapObjects LT (ESRI, 2000).

As discussed in the previous chapter, the SDSS consists of the decision support component itself, a number of databases and a spatial GUI which accepts user input

and provides output. In addition to the CaNaSTA SDSS, a program called ‘CaNaSTA Manager’ has been developed, which allows direct interaction with the data in the databases. An overview of the SDSS is shown in Figure 10.1.

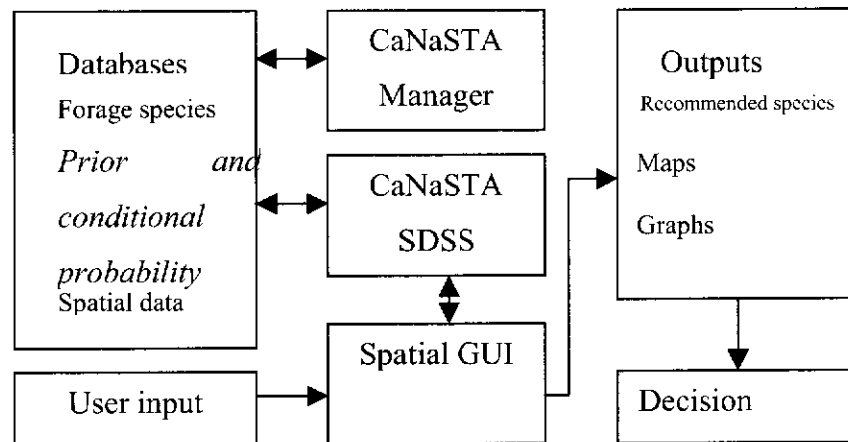


Figure 10.1 Overview of CaNaSTA.

CaNaSTA consists of three main modules. In the first module, the user selects a location and defines some additional characteristics. The primary output is a ranked list of suitable forage species. The second module allows the user to select a species, and the primary output is a map showing where the species is suitable. The third module allows for interactive updating of data (Figure 10.2).

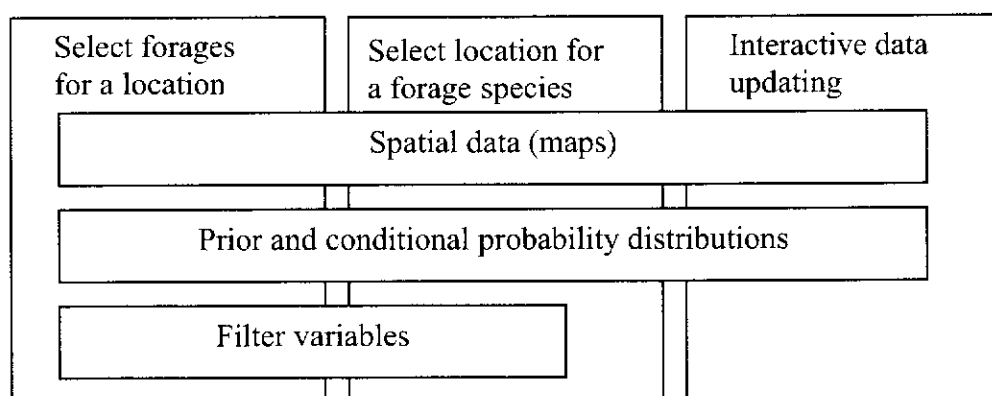


Figure 10.2 Three main modules in CaNaSTA.

Each module employs common routines, and at relevant stages output is available in map form, numerically and graphically. The routines are grouped into libraries depending on their function. The main libraries are the map routines library, the grid

routines library and the probability routines library. The main screens are the location selection screen, the location characteristics screen, the data updating screen and various results screens. These main libraries and screens are briefly described below. These screens are presented here as an overview and are discussed in more detail in the context of using CaNaSTA in Section 10.3.

Additional software has been developed for the uploading of new data and for manual alteration and deletion. This software (CaNaSTA Manager) is intended for use by those who maintain the data for CaNaSTA.

10.3 Libraries

10.3.1 Map Routine Library

The map routine library contains procedures for loading, displaying and navigating spatial data. Spatial data format is the ESRI shapefile format for vector data (administrative boundaries, roads, rivers and towns) and DIVA GRD format (see Grid Routine Library) for grid data (climate data and adaptation probability output maps). This format was chosen because MapObjects LT does not work with ESRI grid formats, and because the DIVA format had already been implemented using the combination of Borland Delphi and MapObjects LT (see Hijmans *et al.*, 2004b). Routines are also stored here for converting continuous grid data into classes and, in conjunction with the probability library, converting multiple class grids into combined adaptation grids.

10.3.2 Grid Routine Library

DIVA GRD format was developed by Hijmans *et al.* (2004b) for use with the DIVA-GIS software. DIVA-GIS is free software for mapping and analysing natural species' distribution and biological diversity. There are four files associated with each grid layer: a bitmap file for display, a 'world' file associated with the bitmap, an 'info' file containing the parameters and filenames associated with the grid and a sequential file containing the actual data (for more details see Hijmans *et al.*, 2001).

The grid routine library contains procedures for reading, writing and manipulating the grid files.

10.3.3 Probability Routine Library

The probability routine library defines structures for storing prior and joint probability distributions for each variable and provides routines for calculating and storing posterior probability distributions, following the methods described in the previous chapters.

10. 4 Screens

10.4.1 Location Selection Screen

The location selection screen shows a map of Central America with various options for navigating the map. An area of interest can be chosen by zooming, panning or selecting a predefined scale. Six scales have been defined (Table 10.1).

Scale	Extent	Output cell size
1	1920 × 1920 km (full extent)	32 × 32 km
2	960 × 960 km	16 × 16 km
3	480 × 480 km	8 × 8 km
4	240 × 240 km	4 × 4 km
5	120 × 120 km	2 × 2 km
6	60 × 60 km	1 × 1 km

Table 10.1 Predefined scales in location selection screen in CaNaSTA.

Views of Central America at these six scales are shown in Figure 10.3 below.

When zooming in is selected with a user-defined rectangle (by dragging and clicking the mouse), scale is snapped to next largest predefined scale. Upon processing, cells are amalgamated to the required output cell size. This ensures that the output grid is always a standard 60 x 60 cells. Subsequent map processing is greatly simplified and resulting data files associated with the grids are kept small.

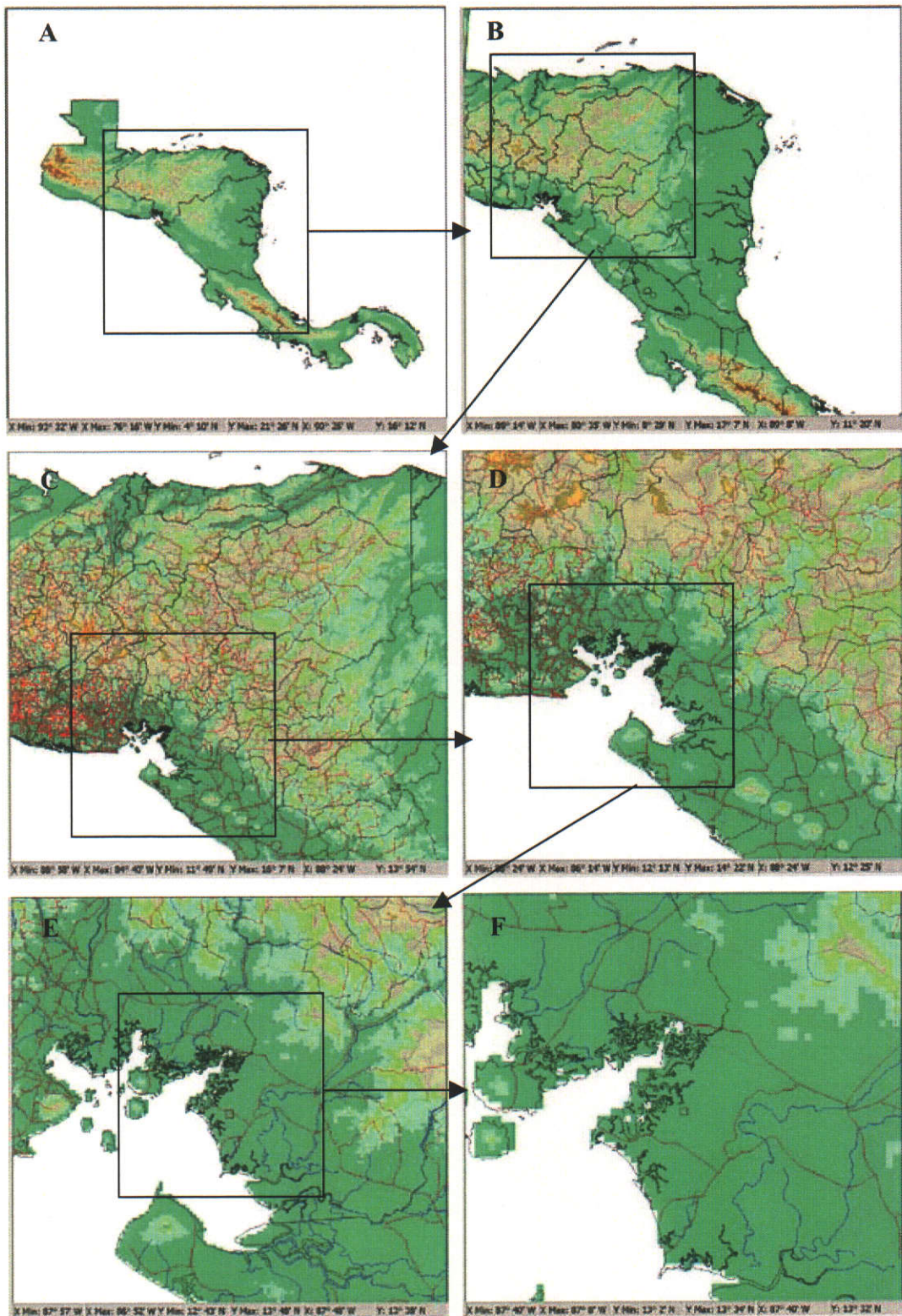


Figure 10.3 Views at six scales implemented in CaNaSTA. A: 1920x1920 km (full extent). B: 960x960km. C: 480x480km. D: 240x240km. E: 120x120km. F: 60x60km.

The area of interest can also be defined by entering the desired latitude and longitude or by repositioning the focus box. The focus box is a square which is always at the

centre of the screen and highlights the cell for which data is reported in some screens. In addition, identifying features can be displayed and queried, including roads, rivers, department names and municipality names. Elevation is automatically displayed, as are administrative boundaries. Although databases exist with populated places, these are too numerous to effectively display without a substantial amount of processing to determine which populated places to display at which scale. In addition, too many labels can clutter the display unnecessarily. Therefore, although the intention is to show towns as identifying features, at this stage this has not yet been implemented.

10.4.2 Location Characteristics Screen

This screen shows characteristics for a selected location (Figure 10.4), namely, the location highlighted by the focus box in the location selection screen. Elevation, annual rainfall and length of dry season are displayed as classes and are derived from GIS data. The user has the option to select soil pH, soil texture and soil fertility, where known. The user is also provided with tools to calculate soil texture and soil fertility, from percent sand and clay, and levels of organic matter and phosphorus, respectively.

The user also has the option to identify whether any tolerances are required in the selected species. The tolerance options are shade, waterlogging, drought, frost and salinity.

Intended use of the forage species must be identified by the user. This can be a single use or multiple uses. Finally, the user is prompted for the farmer's level of risk aversion, being 'low', 'medium' or 'high'.

Characteristics at X: 87° 10' W Y: 15° 2' N

Climate ☐

Elevation (masl) 1500-2000

Annual rainfall (mm) 800-1200

Length of dry season (months) 5-6

Soil ☐

Soil pH Neutral (6.5-7.5)

Soil texture Clay loam Calculate

Soil fertility Low Calculate

Tolerances required ☐

Shade Waterlogging

Drought Frost

Salinity

none high none high

Intended use ☐ (check all that apply)

☒ Any use

☐ long term pasture

☐ short term pasture

☐ cut and carry

☐ conservation

☐ intercropping

☐ green manure/mulch

☐ ground cover (erosion control)

☐ agroforestry

☐ hedgerows

☐ living fences

☐ ponded pasture

☐ irrigated pasture

Risk aversion ☐

low medium high

OK

Figure 10.4 Locations characteristics screen.

10.4.3 Data Updating Screen

This is the mechanism for updating data using expert knowledge. When this screen (Figure 10.5) is shown, a species and a variable have already been selected. The probability distribution is shown broken down by class, both numerically and graphically. Numerically, a table is displayed where the column headers are the adaptation categories ('poor', 'adequate', 'good' and 'excellent') and the row headers are the variable classes (e.g., in the case of elevation '0-500m', '500-1000m', '1000-1500m', '1500-2000m' and '> 2000m'). These same figures are represented in a graph where each set of bars represents one variable class. Clicking on one set of bars displays a graph where the individual bars can be manipulated by dragging them up or down. This, in turn, updates all probabilities in the table.

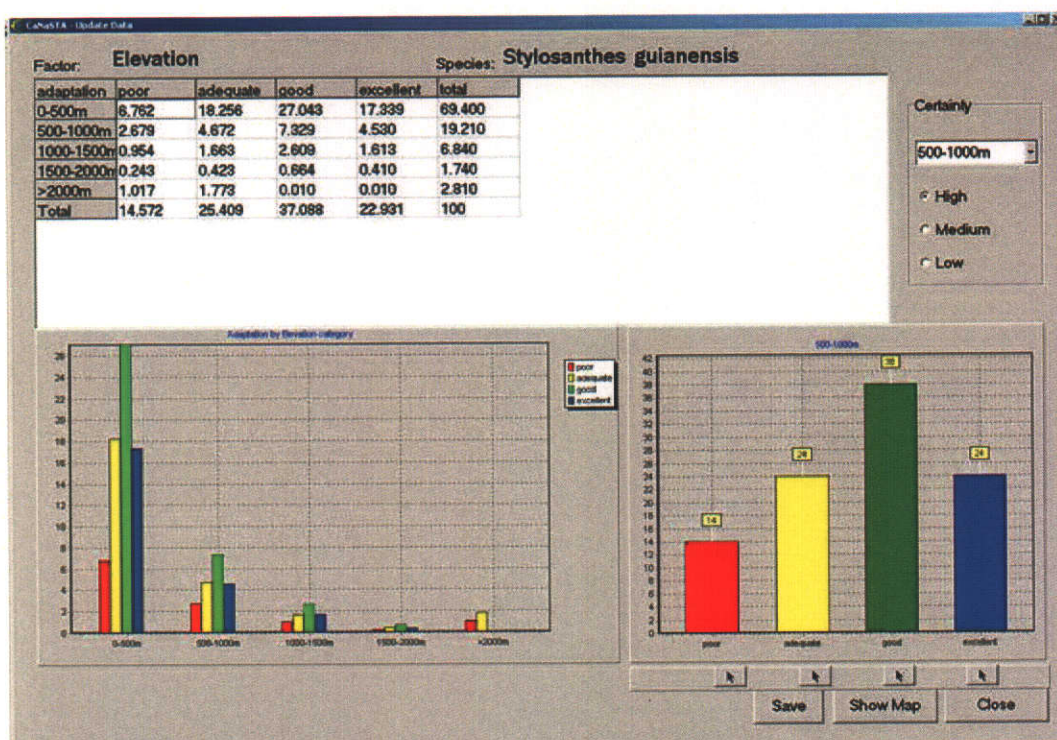


Figure 10.5 Data updating screen.

In addition, the level of certainty for each probability distribution is displayed as either 'high', 'medium' or 'low', and this level can also be updated by the expert.

10.4.4 Results Screens

Results screens differ slightly, depending on which module is being used. However they all contain a map, a legend and some identifying data (Figure 10.6). The map is static in extent but dynamic in content. That is, the spatial extent and resolution have already been defined by the user in the location selection screen.

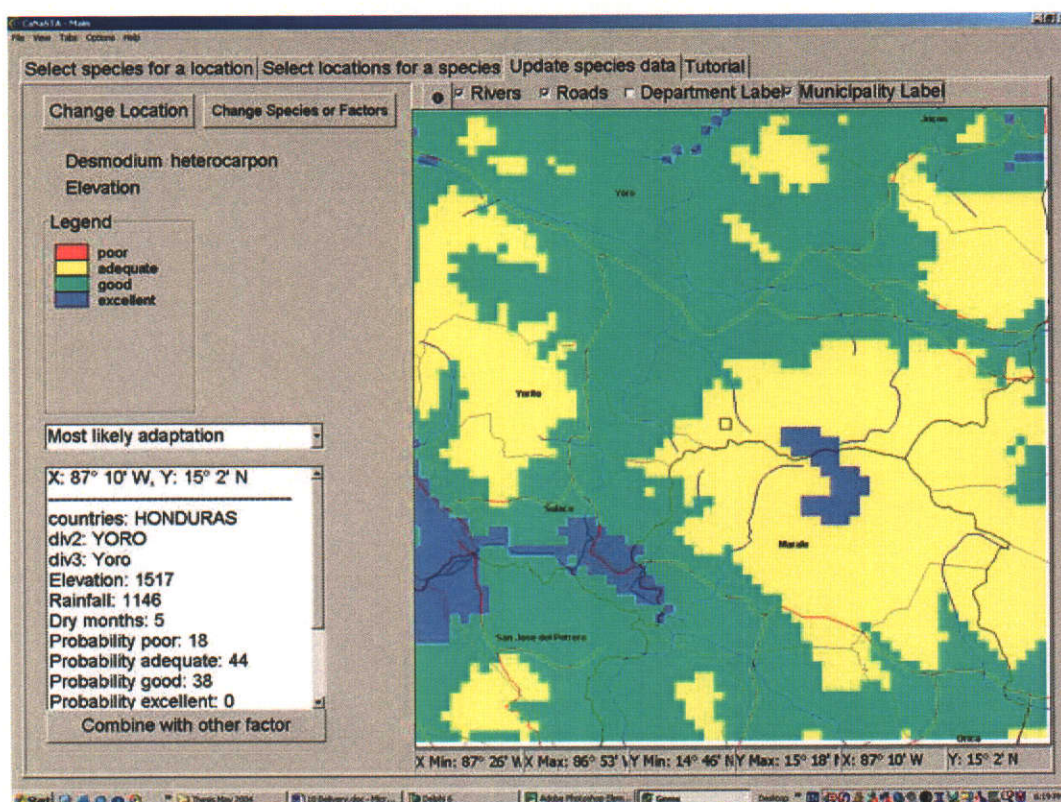


Figure 10.6 Results screen.

Data is reported for the cell in the focus box. In Figure 10.6, this is the text in the white box in the lower left corner of the screen. This data includes latitude, longitude, country, department, municipality, elevation, rainfall and dry months. If adaptation has been calculated, then it will also display the posterior probability of adaptation being 'poor', 'adequate', 'good' or 'excellent', the most likely adaptation class, the score calculated from the adaptation distribution and the certainty level associated with the adaptation calculation in each location. The map in Figure 10.6 shows most likely adaptation class for the species *Desmodium heterocarpon* based on elevation alone.

10.4.5 CaNaSTA Manager Tool

This tool is a separate application, allowing an administrator to manage users, variable data, species data and probabilities. The tool shares libraries with CaNaSTA, namely, read, write and storage routines. It does not access probability calculation routines, map display routines or grid handling routines. In each of the screens, data can be added, changed or deleted.

10.5 Using CaNaSTA

10.5.1 Selecting Species for a Location

When CaNaSTA is first launched, a screen is shown with an empty map and a button to select a location. At this point the 'Select Species for Location' tab is selected, but the other tabs ('Select Location for a Species', 'Update Species Data' and 'Tutorial') are available (Figure 10.7).

Clicking on 'Select Location' displays the location selection screen, displaying a map of Central America with elevation and country boundaries (Figure 10.8). Using the map navigation tools, an area of interest can then be selected.

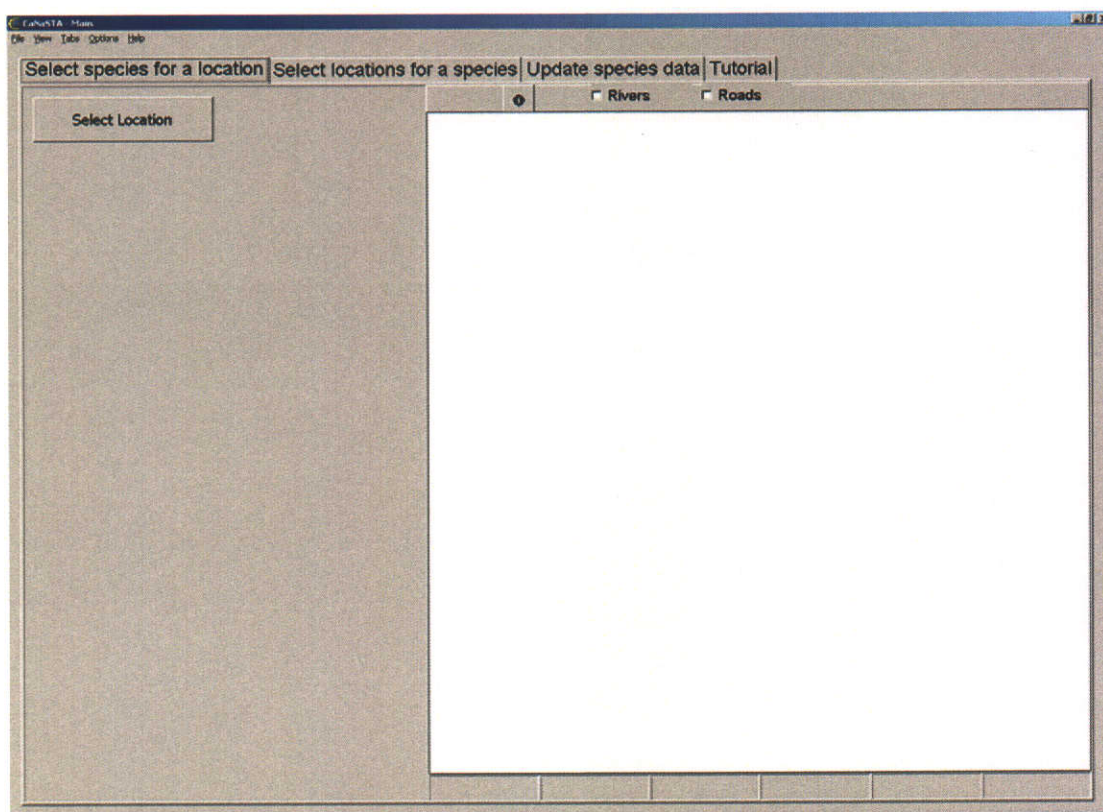


Figure 10.7 Empty map when CaNaSTA is first launched.

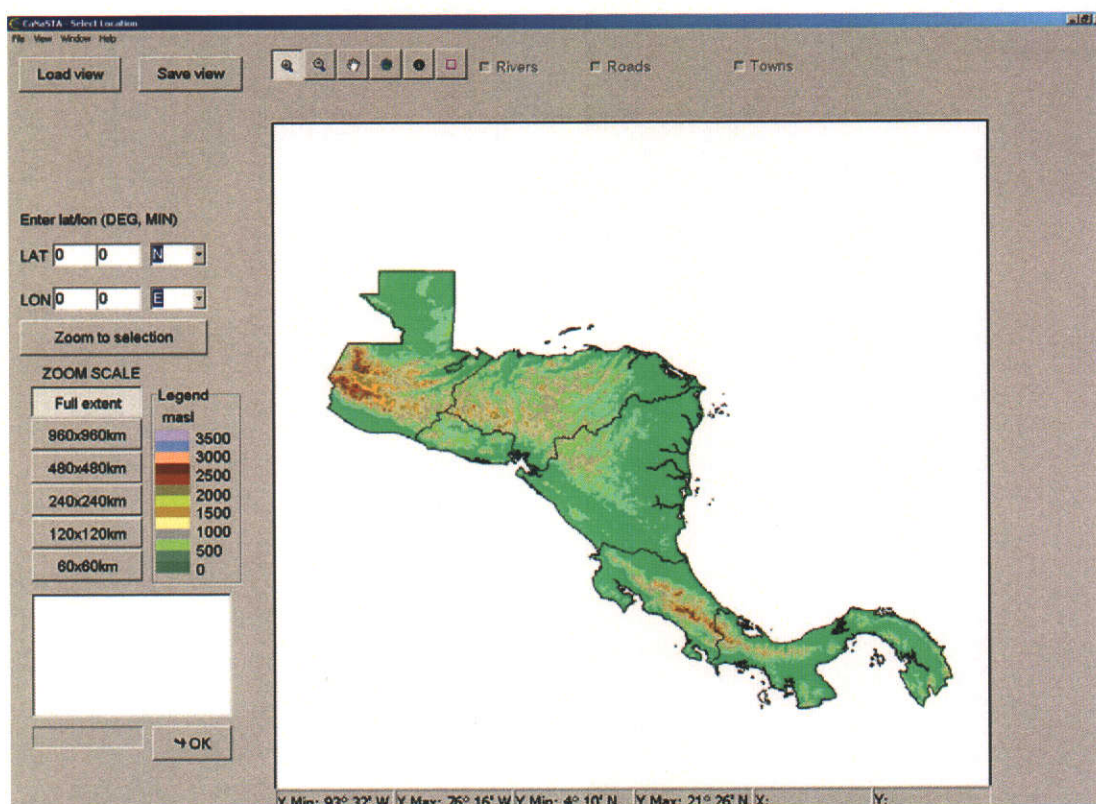


Figure 10.8 Initial 'Select Location' screen.

In Figure 10.9, a location near Luquigüe in Honduras (see Figure 2.8) has been selected where the elevation is 1517m, annual rainfall is 1146mm and the length of the dry season is five months. The labels shown in this figure are not town names but municipality labels.

Selecting the 'OK' button returns the user to the 'Select Species for a Location' screen, with the map now displaying data for the extent chosen in the previous screen (Figure 10.10). The user can choose to display maps for elevation, rainfall or dry months, each of which is categorised according to the predefined classification. Although in the display the variables are categorised into classes, clicking anywhere on the map will show the actual figures for elevation, rainfall and dry months in a separate popup box.

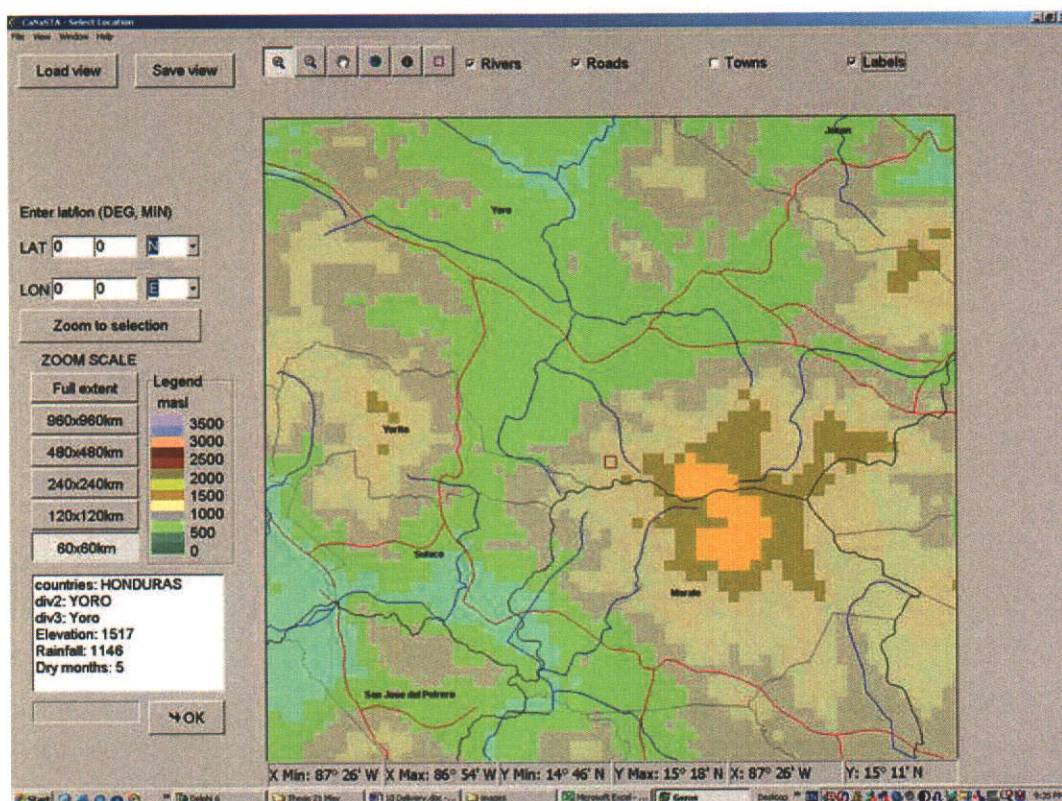


Figure 10.9 'Select Location' screen with a location near Luquigüe selected.

Selecting 'Define Characteristics' brings up the location characteristics screen (Figure 10.11). Elevation, rainfall and dry months classes are shown. The user can select drop down lists to choose soil pH, soil texture and soil fertility classes. In this example, soil pH has been chosen as 'Acid (4.5 – 5.5)' and soil texture is being calculated based on percentage sand and percentage clay. Any of the soil characteristics can be left as 'unknown'. No tolerance levels have been selected. 'Cut and carry' has been selected as the intended use of the forage; risk aversion has been identified as 'low'. Clicking on OK returns the user to the Select Species for a Location screen.

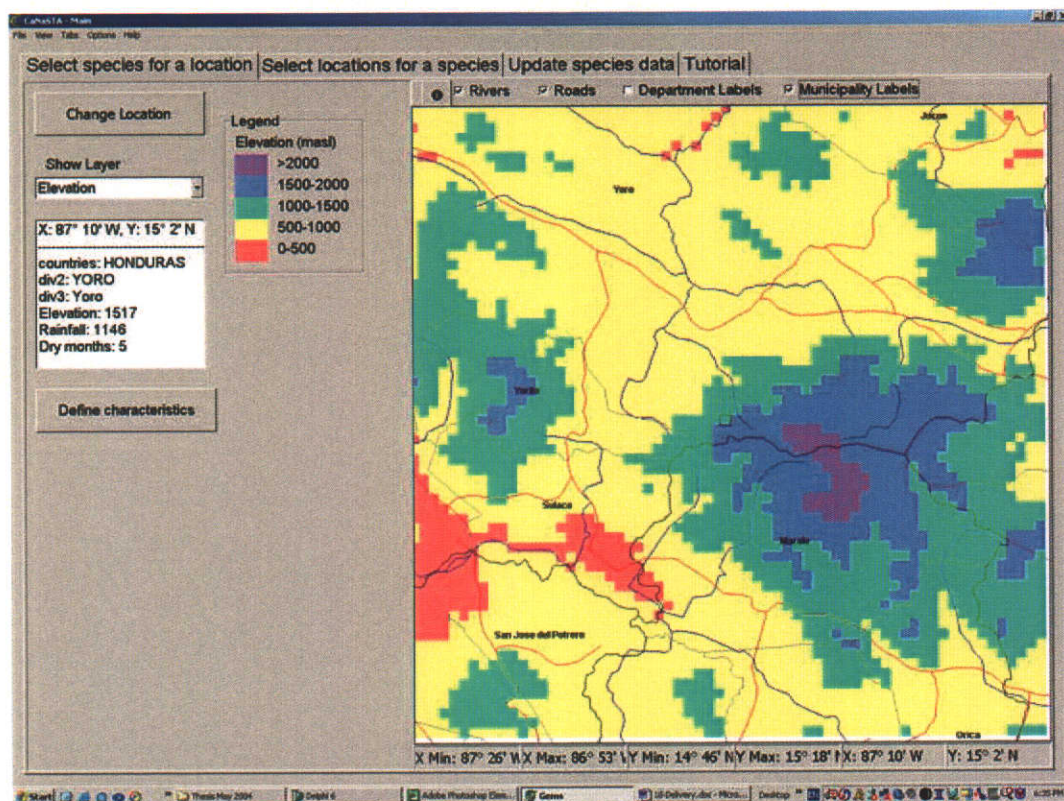


Figure 10.10 Elevation classes for Luquigüe.

Figure 10.11 Location characteristics.

A new button now appears, namely, 'Calculate Suitable Species'. Selecting this button performs the necessary calculations and returns a ranked list of suitable species for the location (Figure 10.12). The ranking is determined by calculating the score value defined in Equation 9.13.

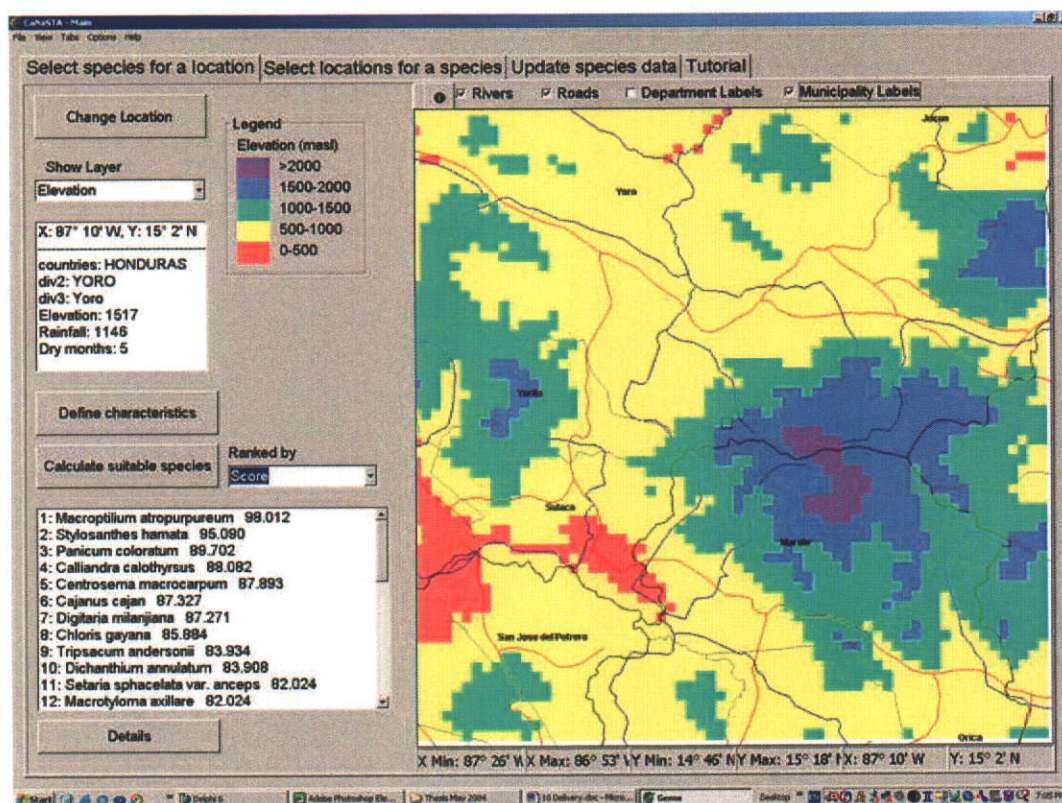


Figure 10.12 Ranked list of suitable species.

Selecting the 'Details' button returns suitability, certainty and stability details for the top five species (Figure 10.13). 'Stability' represents the value calculated in the sensitivity Equation 9.14.

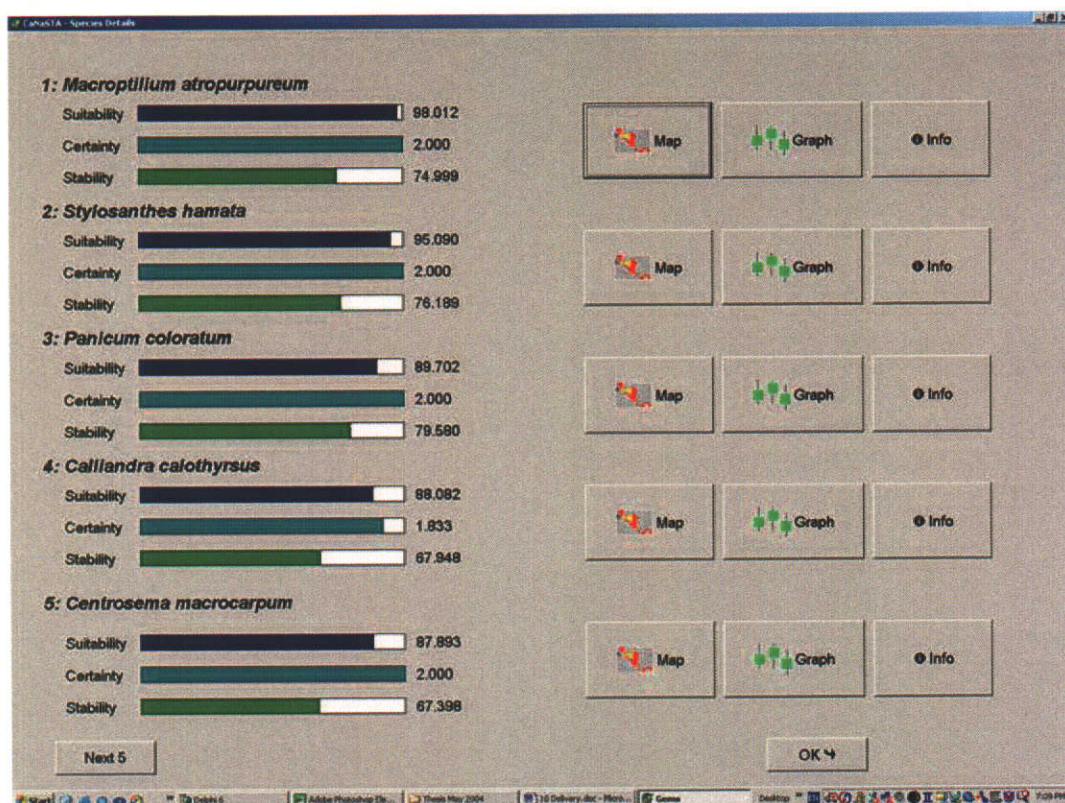


Figure 10.13 Suitability, certainty and stability values for top five species.

Selecting the 'Map' button for a species displays the suitability map for the chosen species for the previously selected location and its surrounding region (Figure 10.14).

Selecting the graph button displays additional information for the selected species, including adaptation distribution and stability rating broken down by variable (Figure 10.15). This screen summarises the available information for one species, showing the selected variable classes and the posterior adaptation distribution for the selected location. The stability display on the right hand side shows the amount by which the adaptation score would change, if each variable changes one class down or one class up. The size of the change is reflected in the size of the bars, with the colour of the bars showing whether the change in score is negative (red) or positive (green).

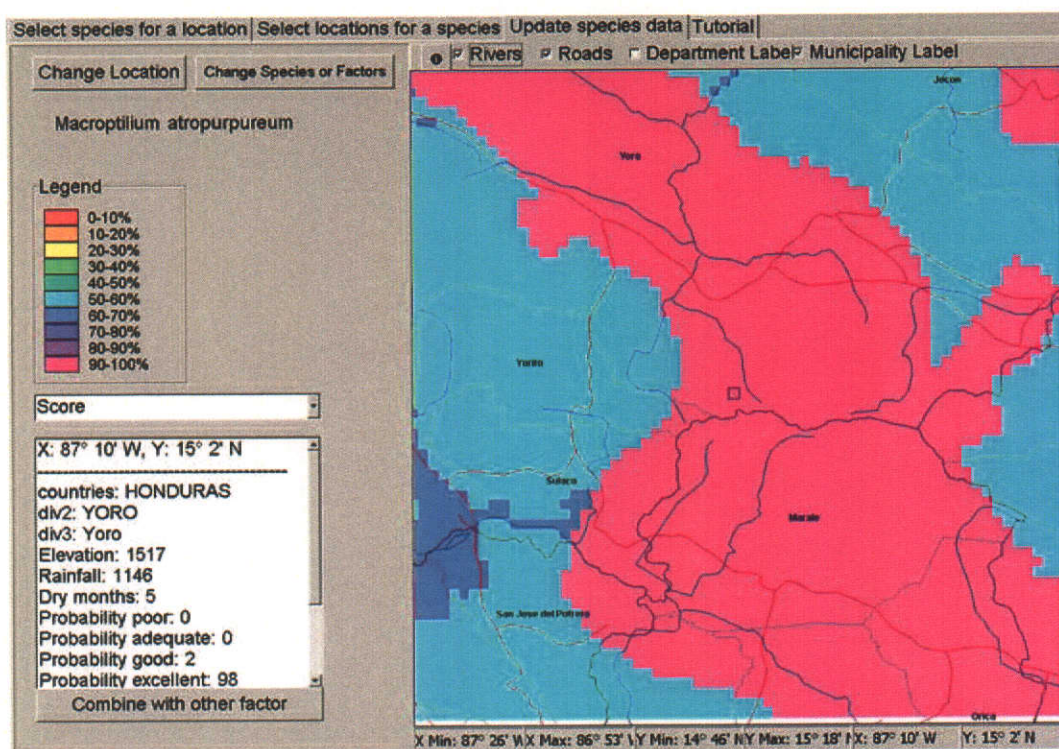


Figure 10.14 Combined suitability score for top selected species.

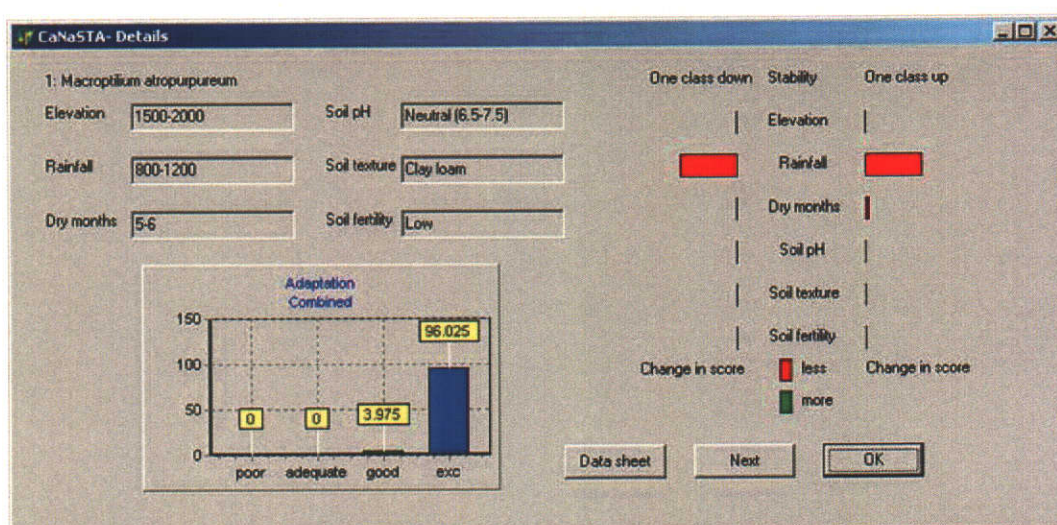


Figure 10.15 Suitability and stability details.

10.5.2 Selecting Locations for a Species

The 'Select Location for a Species' tab allows the user to select or change the location of interest in the same way as in the 'Select Species for a Location' tab. Once an extent has been chosen, elevation classes are displayed. The user selects a

species from the drop down list and then selects which variables to use in the posterior probability calculations (Figure 10.16).

Selecting the 'Calculate probabilities' button calculates the posterior adaptation distribution for the selected species in the selected location (Figure 10.17).

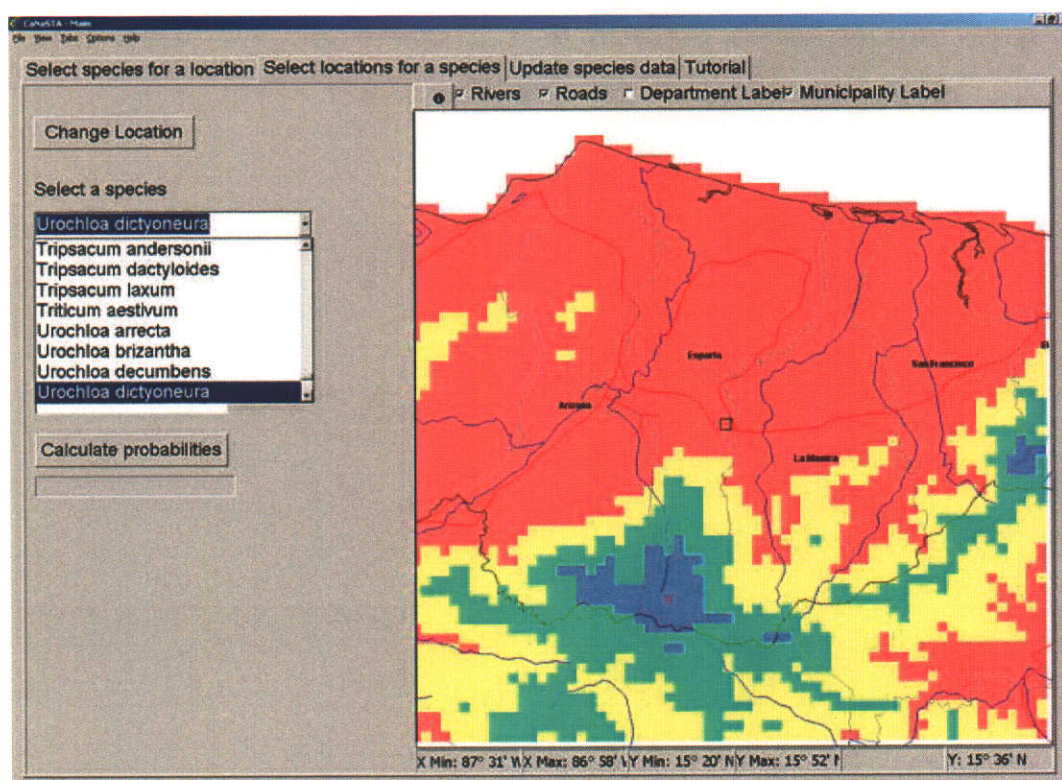


Figure 10.16 Selecting locations for a species.

In addition, access to market and population density can be displayed (Figure 10.18). These maps are currently not incorporated in the analysis but may be useful in the decision of where a species is suitable.

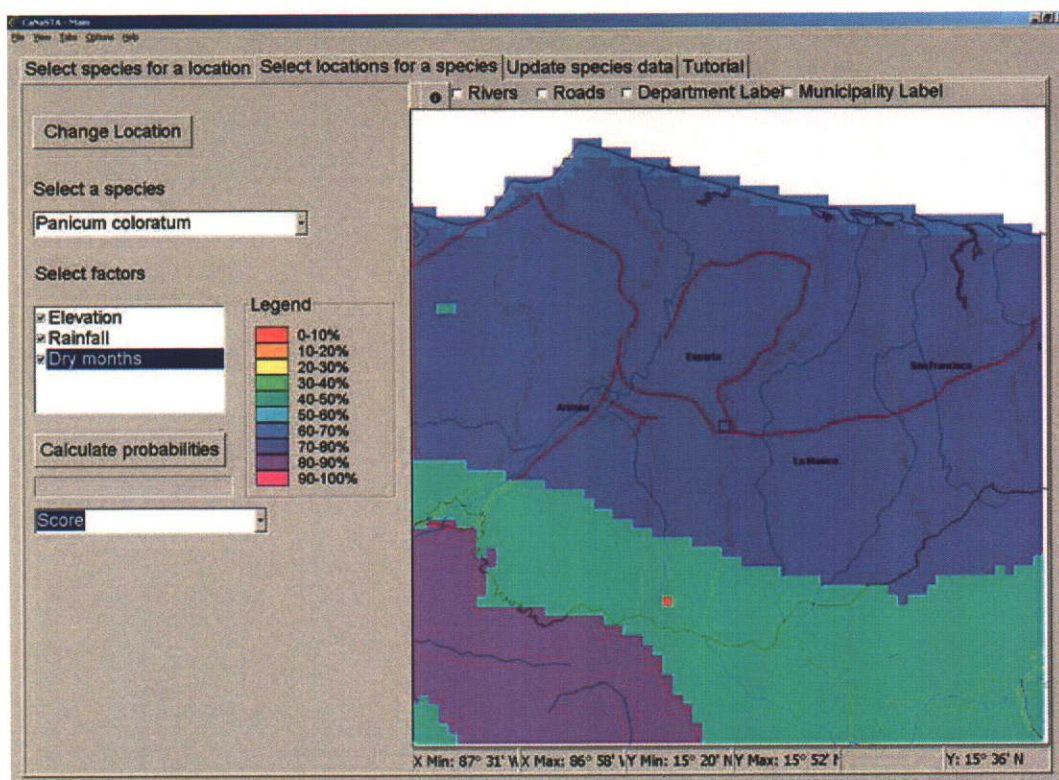


Figure 10.17 Score for selected species in selected location.

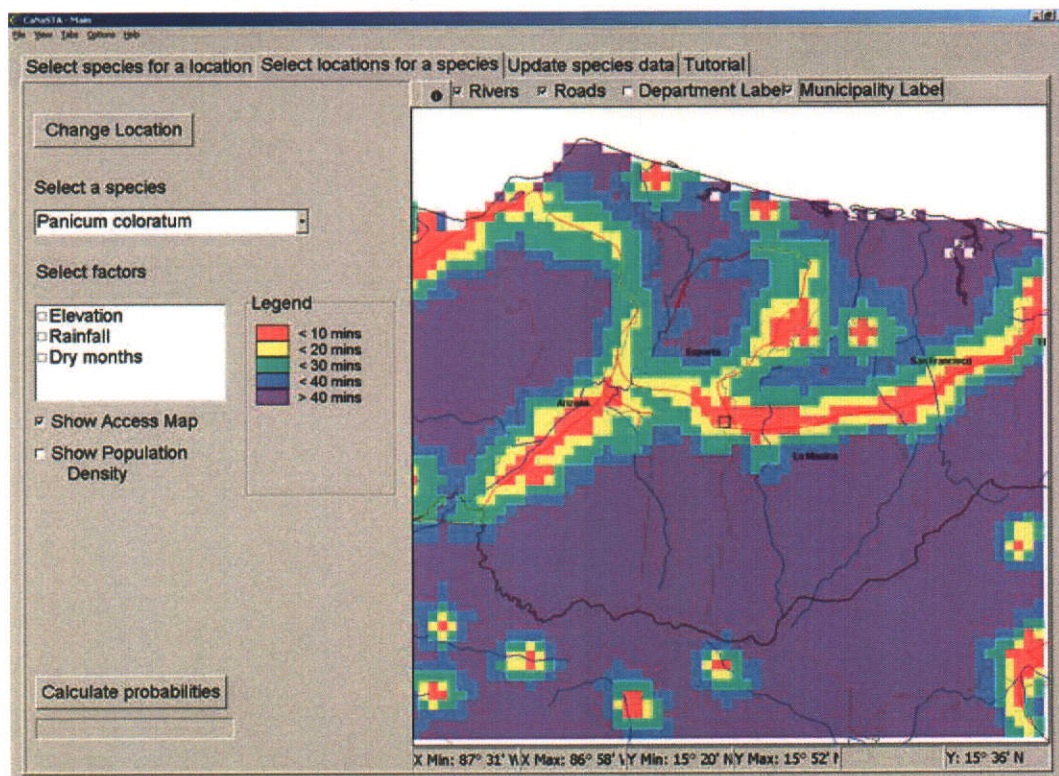


Figure 10.18 Access to market.

10.5.3 Updating Species Data

The 'Update Species Data' (Figure 10.19) screen allows conditional probability values to be adjusted for individual species. These updates are reflected in the maps that are displayed when the user selects a location, and therefore the impact of the updates can be immediately assessed. In this screen a password is required. The reason for this is to be able to log changes made to probability data. Passwords are stored with weak encryption and new user accounts are easily created.

Once a user has logged in, they are prompted to select genus, species and predictor variable from drop down boxes (Figure 10.20). Clicking on 'Update' takes the user to the 'Update Data' screen (Figure 10.21). Here, the user can adjust individual probability values by clicking and dragging on the bar chart displayed in the lower right corner. They can also update the certainty value for each class.

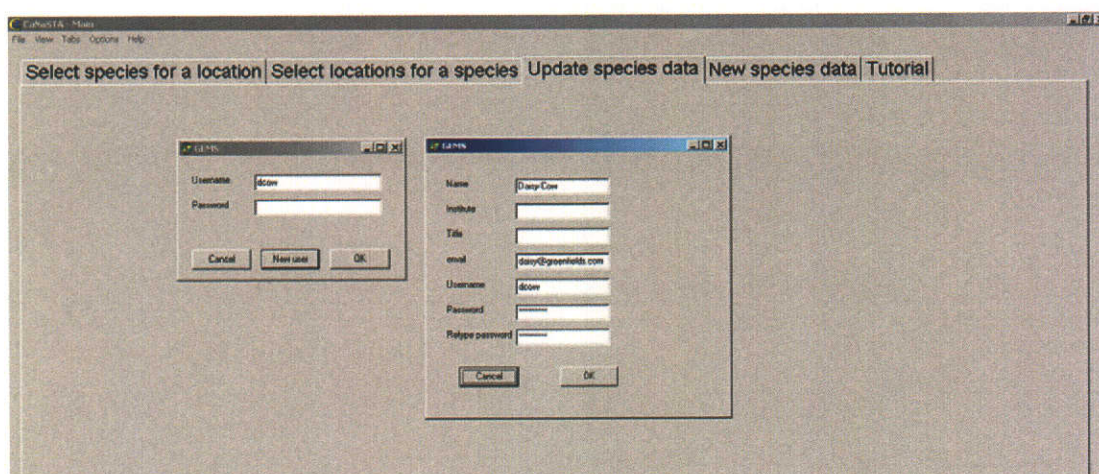


Figure 10.19 New user prompt.

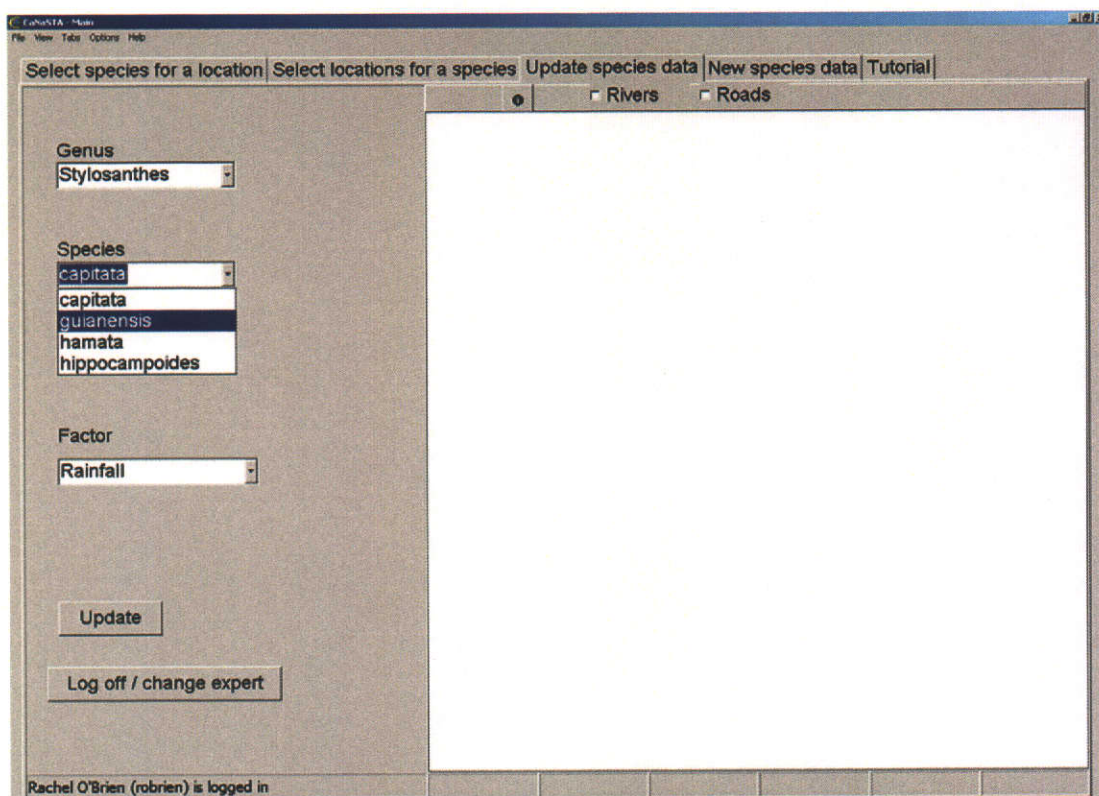


Figure 10.20 Species and variable selection.



Figure 10.21 Updating probabilities.

Selecting 'Show Map' returns the user to the 'Update Species Data' tab, but this time with the map displaying adaptation values for the selected location (Figure 10.22).

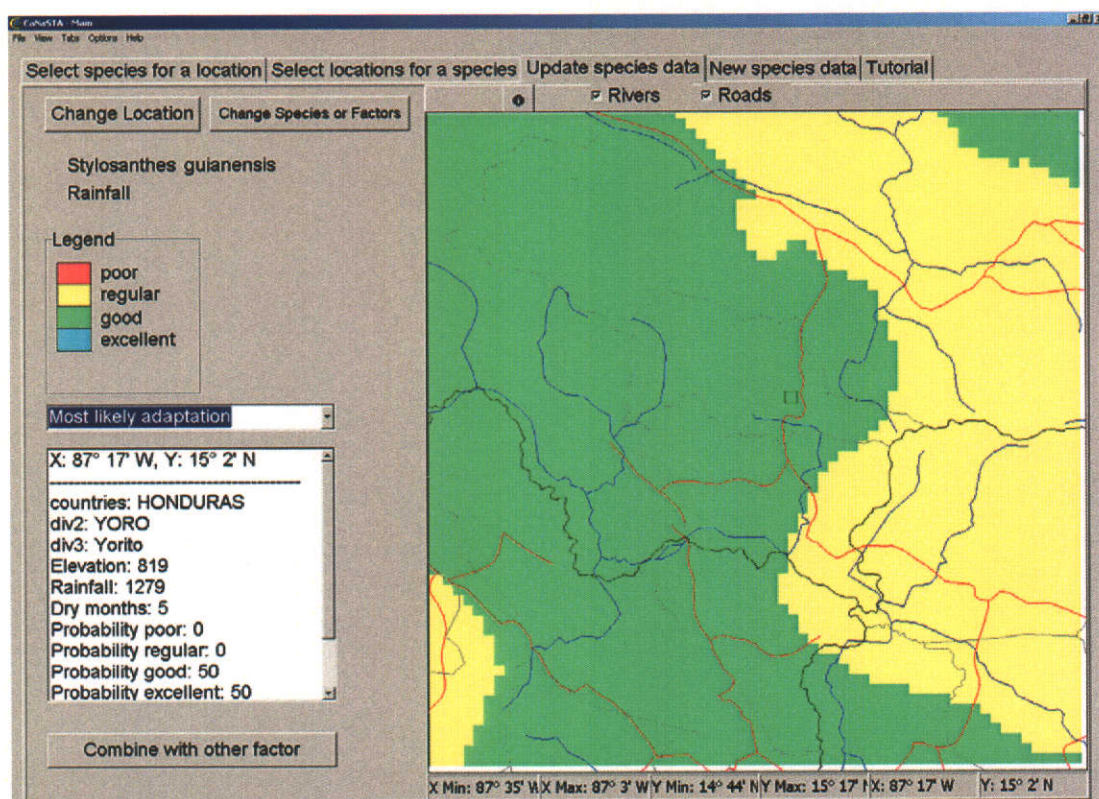


Figure 10.22 Most likely adaptation for *S. guianensis* based on rainfall.

Clicking on 'Combine with other Factor' brings up a box which allows the user to select multiple variables from which joint probabilities will be calculated (Figure 10.23). Although in most cases a user is expected to want to use all available variables, it is possible to select any combination of any number of variables. The posterior probability distribution is then calculated based only on data for the variables selected.

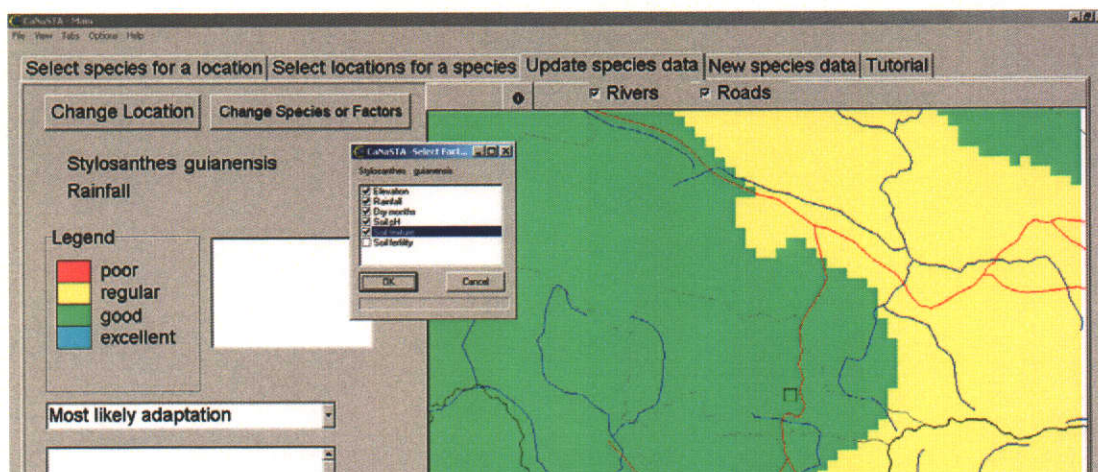


Figure 10.23 Combining multiple variables.

If soil variables are chosen, then initially they are set to unknown and the map is displayed for the combination of climate variables selected. Clicking on 'Change Value' allows a value to be chosen for the selected soil variable (Figure 10.24).

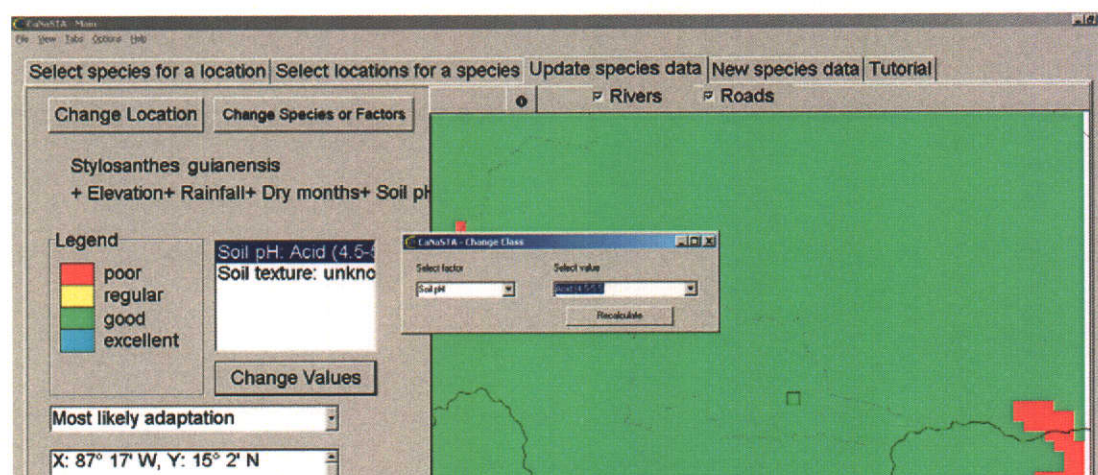


Figure 10.24 Changing value for soil pH.

The adaptation map is then redisplayed with the assumption that the selected soil property holds for the entire displayed area. The effects of changing any variable can be examined by the user.

With all results screens, the user can choose to display different views, namely, 'probability poor', 'probability adequate', 'probability good', 'probability excellent', 'probability good or excellent', 'most likely adaptation', 'score' and 'certainty' (Figure 10.25).

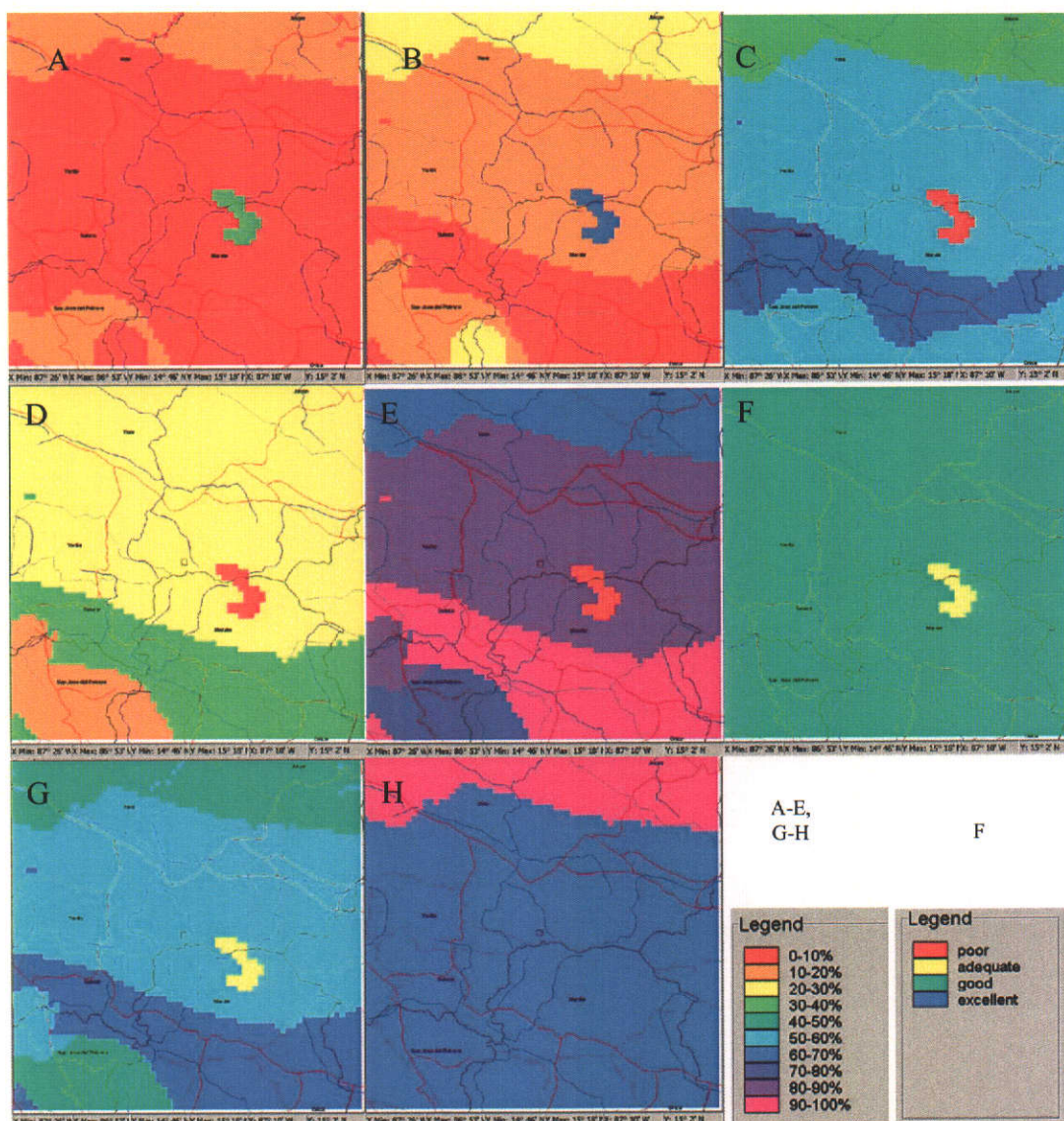


Figure 10.25 Different views of adaptation distributions. A. Probability of poor adaptation. B. Probability of adequate adaptation. C. Probability of good adaptation. D. Probability of excellent adaptation. E. Probability of good or excellent adaptation. F. Most likely adaptation class. G. Score calculated from adaptation distribution. H. Level of certainty.

10.6 Using CaNaSTA Manager

CaNaSTA Manager allows the edition, updating and deletion of data in each of CaNaSTA's databases. User accounts can be modified and passwords can be updated.

Screens for managing variable and species data allow these to be added, changed or deleted. A variable can only be deleted if there is no associated joint probability

distribution data for any species. A species can only be deleted if there is no associated joint probability distribution data for that species.

Joint probability distribution data can also be added, updated or deleted. Utilities are available to load both RIEPT and SoFT data into the appropriate formats and to manipulate the data to fit the requirements discussed in the previous chapter. Probability values and certainty values can be updated manually. This functionality is especially useful when small errors are known to exist in RIEPT or SoFT, such as typographical errors.

Forage uses and tolerances to stresses can also be added, updated or deleted. Where available, they are read directly from the SoFT knowledge base. For species present in CaNaSTA but not in SoFT, this information needs to be added manually, and this can be done here.

10. 7 Summary

CaNaSTA software has been developed as an implementation of the modelling approach described in this research. CaNaSTA recommends species for a given location and situation and recommends locations for a given species. In addition, users can update data interactively and examine results through maps, tables and graphs.

CaNaSTA is implemented as a standalone software package using Borland Delphi and ESRI MapObjects LT. Because it is standalone, the user does not need to have any GIS or database programs installed in order to run CaNaSTA.

Incorporating spatial capabilities into an agricultural DSS, as in CaNaSTA, facilitates data input, allows more informative output of results and allows spatial variability to be made explicit, both of results and of uncertainties related to the results.

The next chapter will compare some of the outputs of CaNaSTA with other methods for recommending forage species for niche locations. This will assess the accuracy of the methods used.

CHAPTER 11. RESULTS AND DISCUSSION

In the previous chapter, the SDSS 'CaNaSTA' was described. CaNaSTA is an implementation of the concepts and methods developed throughout this thesis. In this chapter, output from CaNaSTA is presented and discussed. The discussion will then turn to the stated objectives in developing CaNaSTA and to what extent these were achieved.

The main objective of this research was not to develop the software tool itself, but rather to investigate ways of providing decision support in uncertain and risky environments. A supporting objective was to develop an appropriate model for providing this decision support. The development of CaNaSTA serves to illustrate the implementation of the model. However, the fully fledged production of CaNaSTA is outside of the scope of this research. CaNaSTA nevertheless provides a viable test bed for the method. At the time of writing, CaNaSTA has been developed to the stage of functionality for most intended tasks, and all RIEPT data and preliminary SoFT data available at time of writing are included.

The model can be assessed in a number of ways. The first is to check the accuracy of the functional model by comparing results from the SDSS with results from other sources. A second assessment is the comparison of the process of decision-making using the SDSS with other methods of addressing the decision problem. Finally, an assessment is needed of how well the SDSS meets the stated objectives.

11.1 Accuracy of Model

To check the accuracy of the functional model, results are compared with results from a number of sources (Table 11.1). Each of these sources can be used for different approaches to validation. SoFT, EcoCrop, Lexsys and direct elicitation can all supply a list of suitable species, given a number of requirements, but their knowledge bases have been defined independently. There may, however, be some overlap in the experts used to define the knowledge bases. FloraMap can produce maps for a single species at a time, based on location data for wild accessions. The

output from the other sources can also be used to produce single-species maps, based on environmental characteristics.

Output from FloraMap based on wild accessions
SoFT output
EcoCrop output
Lexsys output
Direct elicitation from forage experts

Table 11.1 Data and knowledge for functional model assessment.

11.1.1 Selecting Species for a Location

Five locations in Central America have been chosen for validation, with different predictor variable values. In addition, an intended forage use was chosen for a hypothetical farmer at each location and at a particular level of risk-aversion. The locations are summarised in Table 11.2.

	Luquigüe	San Dionisio-Wibuse	El Corozo	Near Flores	Esparza
Country	Honduras	Nicaragua	Nicaragua	Honduras	Costa Rica
Department	Yoro	Matagalpa	Matagalpa	Atlantida	Puntarenas
Municipality	Yorito	San Dionisio	San Dionisio	Esparta	Esparza
Lat	15°02'N	12°45'N	12°47'N	15°36'N	9°59'N
Lon	87°10'W	85°49'W	85°54'W	87°15'W	84°40'W
Elevation	1514m	430m	650m	70m	145m
Rainfall	1146mm	900mm	800mm	2606mm	2277mm
Dry months	5	5	6	0	5
Soil pH	Neutral	Moderately acid	Moderately acid	Acid	Moderately acid
Soil texture	Clay loam	Clay	Sandy loam	Loam	Sandy loam
Soil fertility	Low	Medium	High	High	Very high
Intended use	Cut and carry	Pasture	Cut and carry	Pasture	Cut and carry, living barriers
Risk level	Risk neutral	Risk neutral	Risk averse	Risk averse	Risk neutral

Table 11.2 Summary of selected locations.

In order to elicit expert knowledge, the situations above were described in a questionnaire sent to a number of forage experts (Appendix C). They were asked to suggest appropriate forage species for different intended uses, given their experience. The results from this questionnaire can be compared to results from CaNaSTA, indicating how well CaNaSTA imitates expert opinion. The questionnaire was sent to ten forage experts, five of whom replied. Four of these work, or have worked, in Central America. In the case of tropical forages, it is of note that the number of experts worldwide is small (probably less than 200 with international experience [Peters, *pers. comm.*, 2004]). In addition, SoFT, EcoCrop and Lexsys were queried, based on the climatic and edaphic variables listed in Table 11.3, as well as intended use.

CaNaSTA	SoFT	EcoCrop	Lexsys
Elevation	Altitude	Altitude	Altitude
Rainfall	Rainfall	Rainfall range	Precipitation
Dry months	Dry season		
Soil pH	Soil pH	pH range	pH range
Soil texture	Soil texture	Soil texture	Soil type
Soil fertility	Soil fertility	Soil fertility	Fertility requirement
Intended use	Intended use	Main use	

Table 11.3 Comparison of query variables for different sources.

Each source suggests a number of species for each set of criteria. The top ten species recommended by CaNaSTA for each situation are listed in Tables 11.4 – 11.8 below. Because the knowledge base of CaNaSTA is partly based on the SoFT knowledge base, a high level of agreement is expected between CaNaSTA and SoFT. EcoCrop and Lexsys give an independent comparison, although it is likely that some of the information in the different databases are sourced from the same forage agronomists in some cases. However, not all species included in CaNaSTA are present in EcoCrop and Lexsys, and vice versa. This is particularly the case for Lexsys, as it contains only legumes. When SoFT, EcoCrop or Lexsys are queried, they return a list of all species matching the query criteria. In the case of EcoCrop, the user can choose to return a certain number of species, but these are simply the first species the query finds in the database, and not necessarily the best.

The fact that none of the experts listed a particular species as suitable does not mean it is not suitable in their opinion. It could be that the expert is unfamiliar with the species or simply did not recall it at the time of completing the questionnaire. However, it is useful to compare the species recommended by the experts with the ranking given to the species by CaNaSTA.

Table 11.4 shows the species recommended for the farmer in Luquigüe under the conditions listed in Table 11.2.

A	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
1	0.98	<i>Macroptilium atropurpureum</i>	Yes	Yes	Yes	No
2	0.95	<i>Stylosanthes hamata</i>	Yes	Yes	No	No
3	0.90	<i>Panicum coloratum</i>	Yes	Yes	N/A	No
4	0.88	<i>Calliandra calothyrsus</i>	No	Yes	N/A	No
5	0.88	<i>Centrosema macrocarpum</i>	Yes	Yes	N/A	Yes
6	0.87	<i>Cajanus cajan</i>	Yes	Yes	Yes	No
7	0.87	<i>Digitaria milanjiara</i>	Yes	No	N/A	Yes
8	0.86	<i>Chloris gayana</i>	Yes	No	N/A	No
9	0.84	<i>Tripsacum andersonii</i>	Yes	No	N/A	No
10	0.84	<i>Dichanthium annulatum</i>	Yes	No	N/A	No

B	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
12	0.82	<i>Macrotyloma axillare</i>	Yes	No	No	Yes
15	0.78	<i>Lablab purpureus</i>	Yes	Yes	Yes	Yes
17	0.75	<i>Cratylia argentea</i>	Yes	N/A	N/A	Yes
19	0.72	<i>Leucaena leucocephala</i>	No	Yes	Yes	Yes
21	0.69	<i>Andropogon gayanus</i>	No	No	N/A	Yes
22	0.62	<i>Panicum maximum</i>	No	Yes	N/A	Yes
23	0.58	<i>Pennisetum purpureum</i>	No	Yes	N/A	Yes
	0.90	<i>Bothriochloa insculpta</i>	No	Yes	N/A	Yes
	0.81	<i>Vigna unguiculata</i>	No	No	No	Yes
	0.77	<i>Clitoria ternatea</i>	No	Yes	Yes	Yes
	0.50	<i>Desmodium velutinum</i>	No	N/A	N/A	Yes
	0.48	<i>Brachiaria brizantha</i>	No	No	N/A	Yes
	0.03	<i>Axonopus scoparius</i>	No	No	N/A	Yes

Table 11.4 Species suggested for farmer in Luquigüe. A: Top ten. B: Ranking (within top 50) of additional species suggested by experts. N/A means the species is not present in the database.

Of the top ten species recommended for Luquigüe by CaNaSTA, only one (*C. calothyrsus*) is not listed as suitable by SoFT. Examination shows that, according to SoFT, the amount of rain is too low at this location for this species to be suitable.

However, as it is only just below the threshold, it is appropriate that the species is included by CaNaSTA.

Of the species recommended by the experts, nine were selected by CaNaSTA and six were not. *B. insculpta* is excluded because it is listed as not suitable for ‘cut and carry’ in SoFT, but would otherwise have been ranked highly. *A. scoparius* is excluded because, according to SoFT, the number of dry months is too high and the soil fertility too low for the species to be suitable in this location. The others are all excluded because for one of the variables the selected category has certainty ‘Low’, but, of these, *V. unguicalata* and *C. ternaeta* score otherwise well. Species excluded by CaNaSTA are reflections of exclusions in the SoFT knowledge base.

Table 11.5 lists the species recommended by CaNaSTA for San Dionisio-Wibuse.

A	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
1	0.99	<i>Macroptilium atropurpureum</i>	Yes	Yes	Yes	No
2	0.98	<i>Guasuma ulmifolia</i>	Yes	Yes	N/A	No
3	0.97	<i>Dichanthium aristatum</i>	Yes	Yes	N/A	Yes
4	0.96	<i>Bothriochloa pertusa</i>	Yes	No	N/A	No
5	0.95	<i>Chloris gayana</i>	Yes	Yes	N/A	No
6	0.95	<i>Canavalia brasiliensis</i>	Yes	N/A	N/A	No
7	0.92	<i>Stylosanthes hamata</i>	Yes	No	No	Yes
8	0.92	<i>Centrosema macrocarpum</i>	Yes	No	N/A	No
9	0.91	<i>Macroptilium bracteatum</i>	Yes	Yes	N/A	No
10	0.91	<i>Desmanthus virgatus</i>	Yes	Yes	N/A	No

B	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
11	0.90	<i>Urochloa mosambicensis</i>	Yes	No	N/A	Yes
18	0.86	<i>Panicum maximum</i>	Yes	Yes	N/A	Yes
24	0.82	<i>Clitoria terneata</i>	Yes	Yes	No	Yes
40	0.76	<i>Leucaena leucocephala</i>	Yes	Yes	Yes	Yes
47	0.71	<i>Stylosanthes seabrana</i>	Yes	N/A	N/A	Yes
49	0.65	<i>Cenchrus ciliaris</i>	Yes	Yes	N/A	Yes
	0.66	<i>Cratylia argentea</i>	No	N/A	N/A	Yes
	0.56	<i>Brachiaria decumbens</i>	No	No	N/A	Yes
	0.45	<i>Brachiaria brizantha</i>	No	Yes	N/A	Yes
	0.28	<i>Arachis pintoii</i>	No	No	N/A	Yes

Table 11.5 Species suggested for farmer in San Dionisio – Wibuse. A: Top ten. B: Ranking (within top 50) of additional species suggested by experts. N/A means the species is not present in the database.

Of the top ten species, all are recommended by SoFT. Of the species recommended by the experts, eight are in the top 50 recommended by CaNaSTA and four are not. *C. argentea* is excluded because of insufficient certainty, relative to the farmer's risk profile, when soil texture is clay, but otherwise would have been included with a score of 0.66. *B. decumbens*' score is reduced because, according to SoFT, the species is not adapted to clay soils. *B. brizantha* scores 0.45, which is still feasible, but the lower rainfall is detrimental for this species. *A. pintoii* scores 0.28, the low rainfall again reducing the overall score.

Table 11.6 shows species recommended for the farmer in El Corozo.

A	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
1	0.98	<i>Macroptilium atropurpureum</i>	Yes	Yes	Yes	No
2	0.95	<i>Stylosanthes hamata</i>	Yes	No	No	No
3	0.92	<i>Centrosema macrocarpum</i>	Yes	No	N/A	Yes
4	0.91	<i>Macroptilium bracteatum</i>	Yes	Yes	N/A	No
5	0.90	<i>Panicum coloratum</i>	Yes	No	N/A	No
6	0.88	<i>Medicago sativa</i>	Yes	Yes	No	No
7	0.87	<i>Cajanus cajan</i>	Yes	No	Yes	No
8	0.87	<i>Digitaria milaniana</i>	Yes	No	N/A	Yes
9	0.86	<i>Chloris gayana</i>	Yes	Yes	N/A	No
10	0.85	<i>Panicum maximum</i>	No	Yes	N/A	Yes

B	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
17	0.78	<i>Lablab purpureus</i>	Yes	Yes	Yes	Yes
20	0.77	<i>Pennisetum purpureum</i>	Yes	No	N/A	Yes
23	0.75	<i>Cratylia argentea</i>	No	N/A	N/A	Yes
26	0.73	<i>Leucaena leucocephala</i>	No	Yes	N/A	Yes
27	0.72	<i>Saccharum officinarum</i>	Yes	No	N/A	Yes
28	0.51	<i>Brachiaria brizantha</i>	Yes	Yes	N/A	Yes
	0.81	<i>Vigna unguiculata</i>	No	Yes	Yes	Yes
	0.72	<i>Macrotyloma daltonii</i>	Yes	N/A	N/A	Yes
	0.65	<i>Cenchrus ciliaris</i>	Yes	Yes	N/A	Yes

Table 11.6 Species suggested for farmer in El Corozo. A: Top ten. B: Ranking (within top 50) of additional species suggested by experts. N/A means the species is not present in the database.

Of the top ten species recommended by CaNaSTA, nine are selected by SoFT and eight are selected by at least one other source. Of the species recommended by experts, nine are selected by CaNaSTA and three are not. *C. ciliaris* and *M. daltonii*

are excluded because they are not listed as suitable for 'cut and carry'. Otherwise they would score well, with 0.65 and 0.72 respectively. *V. unguicalata* scores 0.81 but certainty is low for the variable 'dry months'. Had the farmer been classed as risk-taking (low risk-aversion) then this species would have been included in the recommendations. Of the top ten species, all are recommended by SoFT except for *P. maximum*, which is not suitable for sandy loam soils. However, as all other variables are favourable, CaNaSTA assigns a high score. It is notable that experts also suggested this species.

Table 11.7 lists species recommended by CaNaSTA for the farmer near Flores.

A	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
1	0.87	<i>Brachiaria brizantha</i>	Yes	Yes	N/A	Yes
2	0.85	<i>Panicum maximum</i>	Yes	Yes	N/A	Yes
3	0.85	<i>Axonopus scoparius</i>	Yes	Yes	N/A	No
4	0.83	<i>Cratylia argentea</i>	Yes	N/A	N/A	No
5	0.83	<i>Lotus spp.</i>	No	N/A	N/A	No
6	0.83	<i>Cenchrus ciliaris</i>	No	No	N/A	No
7	0.83	<i>Secale cereale</i>	No	No	N/A	No
8	0.82	<i>Clitoria ternatea</i>	No	Yes	No	No
9	0.80	<i>Codariocalyx gyroides</i>	Yes	N/A	N/A	No
10	0.80	<i>Macroptilium atropurpureum</i>	No	Yes	Yes	No

B	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
14	0.78	<i>Centrosema macrocarpum</i>	Yes	No	N/A	Yes
17	0.76	<i>Centrosema pubescens</i>	Yes	Yes	No	Yes
19	0.76	<i>Paspalum atratum</i>	Yes	N/A	N/A	Yes
	0.64	<i>Desmodium ovalifolium</i>	No	No	N/A	Yes
	0.52	<i>Arachis pintoii</i>	Yes	Yes	N/A	Yes
	0.50	<i>Brachiaria decumbens</i>	No	No	N/A	Yes

Table 11.7 Species suggested for farmer near Flores. A: Top ten. B: Ranking (within top 50) of additional species suggested by experts. N/A means the species is not present in the database.

Only five of the top ten species are also recommended by SoFT. The remainder of the top ten species are not selected in SoFT, although they agree on all variables except for rainfall, which at over 2500mm is too high for most species. CaNaSTA ranks these species lower, but still includes them for consideration. Of the species suggested by the experts, five are recommended by CaNaSTA and three are not. *A. pintoii* scores 0.52 and only just misses out on the top 50. *B. decumbens* scores 0.50

and *D. ovalifolium* scores 0.64, and these are excluded because of low certainty when soil fertility is high.

Table 11.8 lists the species recommended for the farmer in Esparza, where a species is required for both ‘cut and carry’ and ‘living fences’.

A	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
1	0.89	<i>Brachiaria brizantha</i>	Yes	Yes	N/A	Yes
2	0.83	<i>Cratylia argentea</i> (*)	Yes	N/A	N/A	Yes
3	0.82	<i>Clitoria ternatea</i>	No	Yes	No	No
4	0.80	<i>Macroptilium atropurpureum</i>	No	Yes	Yes	No
5	0.79	<i>Panicum maximum</i>	Yes	Yes	N/A	Yes
6	0.77	<i>Sesbania grandiflora</i>	Yes	Yes	N/A	No
7	0.77	<i>Centrosema macrocarpum</i>	Yes	No	N/A	Yes
8	0.76	<i>Paspalum atratum</i>	Yes	N/A	N/A	No
9	0.76	<i>Paspalum guenoarum</i>	Yes	N/A	N/A	No
10	0.76	<i>Andropogon gayanus</i>	Yes	No	N/A	No

B	Score	Species ranking in CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
15	0.74	<i>Gliricidia sepium</i>	Yes	Yes	No	Yes
20	0.73	<i>Saccharum officinarum</i>	Yes	Yes	N/A	Yes
28	0.60	<i>Pennisetum purpureum</i> (*)	Yes	Yes	N/A	Yes
29	0.60	<i>Codariocalyx gyroides</i>	No	N/A	N/A	Yes
38	0.42	<i>Leucaena leucocephala</i> (*)	Yes	Yes	Yes	Yes
39	0.36	<i>Cajanus cajan</i> (*)	Yes	Yes	Yes	Yes
	0.52	<i>Desmodium velutinum</i>	No	N/A	N/A	Yes
	0.24	<i>Desmodium cinereum</i>	No	N/A	N/A	Yes

Table 11.8 Species suggested for farmer in Esparza. A: Top ten. B: Ranking (within top 50) of additional species suggested by experts. Species marked with (*) are also suitable as living fences. N/A means the species is not present in the database.

All species in the top ten are recommended by at least one other source. Of the species recommended by the experts, ten are listed by CaNaSTA and two are not. *D. velutinum* scores 0.52 and *D. cinereum* scores 0.24. In both cases, certainty is low when soil fertility is very high. Two of the species in the top ten, *C. ternatea* and *M. atropurpureum*, are not recommended by SoFT because rainfall is too high in this location for these species. However, because rainfall is only marginally too high (both species have absolute rainfall limits of 2000mm) and all other variables are favourable, CaNaSTA ranks them highly.

M. atropurpureum is consistently selected across all situations. This is mostly because SoFT shows the species to be widely adapted, except for locations with very low or very high rainfall or moderately to very acidic soils. Despite the fact that the location near Flores has both acidic soils and high rainfall, the species is still in the top ten with a high score. This is due to the fact that the species has a prior adaptation distribution (derived from RIEPT) biased towards 'good' and 'excellent'. This highlights the impact that prior probability distributions have on the final distribution. In this case, it means that the species overall has good adaptation, and conditions need to be quite poor to reduce the level of adaptation.

Similarly, other species which appear suitable in multiple situations are widely adapted, particularly where edaphic characteristics are concerned. They also tend to have prior adaptation distributions biased towards 'good' and 'excellent'. This is valid, because it reflects the fact that these species adapted well in most trials. However, this also highlights an area of further research, as these distributions may also be artefacts of biases in the database.

Because the knowledge base for CaNaSTA is partly reliant on information from SoFT, it would be expected that the two sources would show a high level of agreement. Where CaNaSTA recommends a species and SoFT does not, it is usually because the situation is favourable for all variables bar one. This highlights a strength of CaNaSTA. Rather than excluding a species, if all other variables are highly suitable then the species will still be included for consideration.

SoFT, EcoCrop and Lexsys recommend many more species than have been listed in the tables above. Each of these simply recommends all species which fit the criteria, and in some cases hundreds of species may be listed. This again highlights a benefit of CaNaSTA, namely the score and ranking system, which allows the most suitable species to be considered first.

Experts recommended a number of species which they considered suitable in each situation. They were not asked to comment on the species recommended by CaNaSTA, therefore no conclusions can be drawn when CaNaSTA lists a species and the experts do not. However, of the species recommended by experts, 69 percent

were ranked highly by CaNaSTA (Table 11.9). Of the remaining species, 56 percent were excluded from CaNaSTA because of low certainty and a further 17 percent because the species was designated unsuitable for the intended use.

Recommended	Low certainty	Unsuitable for intended use	Not recommended
41 (69%)	10 (17%)	3 (5%)	5 (8%)

Table 11.9 Recommendations by CaNaSTA for species recommended by experts.

11.1.2 Selecting Locations for a Species

Five species are considered here, namely, *Arachis pintoii*, *Brachiaria brizantha*, *Cratylia argentea*, *Centrosema pubescens*¹ and *Stylosanthes guianensis*. The purpose of the comparison is to evaluate how closely CaNaSTA agrees with experts and other sources when concentrating on a selected species. Each species has between five and 263 records in the RIEPT Adaptation database, spread over a number of trial sites (Table 11.10). CaNaSTA scores are based on a combination of this data and the SoFT knowledge base.

Species	Records	Trial sites
<i>A. pintoii</i>	10	4
<i>B. brizantha</i>	263	5
<i>C. pubescens</i>	111	32
<i>C. argentea</i>	5	5
<i>S. guianensis</i>	121	14

Table 11.10 Records in RIEPT Adaptation for selected species.

Experts were asked to select whether each species is 'suitable', 'marginal' or 'not suitable' under the conditions described. The expert opinion did not necessarily agree. Variations in expert opinion were expected, as the experts have different levels of experience with the species and have worked in different locations.

¹ *Centrosema pubescens* was renamed at the end of 2003 to *Centrosema molle*, therefore there may be some inconsistencies between information in the databases, knowledge bases and expert knowledge (Peters, pers. comm., 2004).

There are methods available, such as the Delphi method (Linstone and Turoff, 1975), to facilitate the convergence of expert opinion from a number of sources. In this case, however, no effort has been made to seek agreement between experts. Five species were considered at five locations each, giving 25 situations in total. Cohen's weighted kappa was calculated to measure the level of agreement between each pair of experts for the 25 situations. The results are given in Table 11.11 below, showing that overall agreement was relatively low between experts.

Experts	1	2	3	4	5
1	1				
2	0.28	1			
3	0.28	0.21	1		
4	0.21	0.19	0.21	1	
5	0.33	0.09	0.19	0.24	1

Table 11.11 Weighted kappa (κ_w) for each pair of experts.

In order to compare expert knowledge with other sources, the expert knowledge assessments are amalgamated by recording the most common classification given by the five experts. The suitability of each of these species is also calculated using CaNaSTA and the score is recorded. The suitability is also read directly from the knowledge bases in SoFT, EcoCrop and Lexsys, where information is available. Suitability is calculated based on the same variables mentioned in Table 11.3 above. This information is summarised in Tables 11.12 – 11.16 below.

Table 11.12 shows the suitability of these five species for Luquigüe, ignoring intended use.

Species	CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
<i>C. argentea</i>	0.75	Yes	N/A	N/A	Yes
<i>B. brizantha</i>	0.48	No	Marginal	N/A	Yes
<i>S. guianensis</i>	0.46	Marginal	Marginal	Marginal	Marginal
<i>A. pintoii</i>	0.25	No	No	N/A	Marginal
<i>C. pubescens</i>	0.00	No	No	No	Yes

Table 11.12 Suitability of selected species for Luquigüe. For CaNaSTA, the score value is given.

The scores calculated by CaNaSTA agree with expert assessment, except in the case of *C. pubescens*. The reason this species is excluded by SoFT is because of the high elevation of Luquigüe. This translates to the very low score in CaNaSTA, since the probability distribution and hence score are partly derived from the SoFT knowledge base.

Table 11.13 shows the suitability of these five species for San Dionisio-Wibuse, again ignoring intended use. None of the scores calculated by CaNaSTA are very high, but the higher scores correspond to species selected by experts as suitable.

Species	CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
<i>C. argentea</i>	0.66	No	N/A	N/A	Yes
<i>S. guianensis</i>	0.55	Marginal	Marginal	No	Yes
<i>C. pubescens</i>	0.49	No	Marginal	Yes	Yes
<i>B. brizantha</i>	0.45	No	Yes	N/A	Yes
<i>A. pinto</i>	0.28	No	No	N/A	No

Table 11.13 Suitability of selected species for San Dionisio-Wibuse. For CaNaSTA, the score value is given.

Table 11.14 shows the suitability of the five species for El Corozo. The highest two scores given by CaNaSTA correspond to the species selected as suitable by the experts.

Species	CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
<i>C. argentea</i>	0.75	No	N/A	N/A	Yes
<i>C. pubescens</i>	0.58	No	Marginal	Marginal	Yes
<i>B. brizantha</i>	0.51	Marginal	Yes	N/A	Marginal
<i>A. pinto</i>	0.26	No	No	N/A	No
<i>S. guianensis</i>	0.26	No	Marginal	No	Marginal

Table 11.14 Suitability of selected species for El Corozo. For CaNaSTA, the score value is given.

Table 11.15 shows suitability of the five species for the farmer near Flores. All five species score relatively high in CaNaSTA. *A. pinto* scores relatively low, even though the species is recommended as suitable by all other sources. Although this score does not mean adaptation is poor (adaptation distribution shows 21 percent

‘good’ and 27 percent ‘excellent’ adaptation), the final distribution is affected by the prior adaptation distribution, which favours ‘adequate’ adaptation.

Species	CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
<i>B. brizantha</i>	0.87	Marginal	Marginal	N/A	Yes
<i>C. argentea</i>	0.83	Marginal	N/A	N/A	Marginal
<i>C. pubescens</i>	0.76	Yes	Marginal	No	Marginal
<i>S. guianensis</i>	0.59	No	Marginal	No	Marginal
<i>A. pintoii</i>	0.52	Yes	Marginal	N/A	Yes

Table 11.15 Suitability of selected species for near Flores. For CaNaSTA, the score value is given.

Table 11.16 shows the scores calculated by CaNaSTA for the five species in Esparza. The two species chosen as suitable by the experts score highly in CaNaSTA.

Species	CaNaSTA	SoFT	EcoCrop	Lexsys	Experts
<i>B. brizantha</i>	0.89	Yes	Yes	N/A	Yes
<i>C. argentea</i>	0.83	No	N/A	N/A	Yes
<i>C. pubescens</i>	0.76	No	Marginal	No	Marginal
<i>A. pintoii</i>	0.52	Marginal	Marginal	N/A	No
<i>S. guianensis</i>	0.33	No	Marginal	No	Marginal

Table 11.16 Suitability of selected species for Esparza. For CaNaSTA, the score value is given.

The CaNaSTA scores can be interpreted as ‘suitable’ (score > 0.66), ‘marginal’ (0.33 < score < 0.66) and ‘not suitable’ (score < 0.33). Analysis can then be carried out on the agreement between CaNaSTA and expert opinion. Table 11.17 shows the confusion matrix for expert opinion versus suitability as assessed by CaNaSTA.

	Experts		
CaNaSTA	Not suitable	Marginal	Suitable
Not suitable	2	2	0
Marginal	1	4	6
Suitable	0	3	6

Table 11.17 Confusion matrix for expert assessment vs. CaNaSTA.

From this, Cohen’s weighted kappa can be calculated, giving $\kappa_w = 0.41$, which indicates moderate agreement. In addition, κ_w can be calculated for each expert separately compared to the CaNaSTA results (Table 11.18). In some instances, these

agreements are substantially higher than between individual expert pairs (Table 11.11 above).

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
CaNaSTA	0.18	0.09	0.34	0.49	0.42

Table 11.18 Weighted kappa for CaNaSTA recommendations against individual experts.

In order to compare all sources of information, weighted kappa is calculated, again for the three classes of 'suitable', 'marginal' and 'not suitable'. Lexsys is not included in this analysis. The results are shown in Table 11.19 below. CaNaSTA shows reasonable levels of agreement with all three sources, and agrees highest with EcoCrop.

	CaNaSTA	SoFT	EcoCrop	Experts
CaNaSTA	1			
SoFT	0.34	1		
EcoCrop	0.58	0.34	1	
Experts	0.41	0.06	0.34	1

Table 11.19 Weighted kappa (κ_w) for recommendations from all sources.

One of the strengths of CaNaSTA is the ability to display maps showing species' suitability over a region. The output of knowledge-based systems can also be used to produce maps on an ad-hoc basis. Maps for these five species have been created for Central America and San Dionisio-Wibuse / El Corozo region using CaNaSTA and, as a comparison, the rule base of EcoCrop (except for *C. argentea*, which is not present in EcoCrop). The CaNaSTA maps were created based on elevation, rainfall and number of dry months, and are displayed showing the score value. The EcoCrop maps are created based on elevation and both optimal and absolute limits for rainfall and temperature. In addition, FloraMap, coupled with accession data, was used to produce suitability maps for *S. guianensis*, *A. pintoii* and *C. argentea*. These maps are based on 422 accessions in Latin America (21 in Central America) for *S. guianensis*, 132 accessions in Brazil for *A. pintoii* and 52 accessions in Brazil and Bolivia for *C. argentea*.

Figure 11.1 shows elevation, annual rainfall and length of dry season maps for Central America and for the San Dionisio-Wibuse / El Corozo region, for comparison purposes.

Figure 11.2 compares suitability of *S. guianensis* throughout Central America and in the San Dionisio region based on information in FloraMap, EcoCrop and CaNaSTA. At the regional level, all three sources show overlap, with the species predicted to do well in northern Guatemala, most of Honduras and the western part of Nicaragua. Costa Rica is marginally suitable. CaNaSTA appears to agree quite well with EcoCrop, except in El Salvador where both EcoCrop and FloraMap predict better suitability than CaNaSTA. At the local level, CaNaSTA and EcoCrop roughly agree, but FloraMap gives markedly different results.

Figure 11.3 compares suitability of *A. pintoii* throughout Central America and in the San Dionisio region based on information in FloraMap, EcoCrop and CaNaSTA. Both FloraMap and EcoCrop show low suitability throughout most of Central America, but do not show agreement on where suitability is higher. CaNaSTA shows agreement with FloraMap in most of Honduras, with low suitability in the central region and higher towards the border with Nicaragua. At the local scale, both FloraMap and EcoCrop show only slight suitability around Muy Muy. CaNaSTA shows high suitability in a much larger region, with suitability dropping markedly with the lower rainfall in the west of the region.

Figure 11.4 compares suitability of *C. argentea* throughout Central America and in the San Dionisio region based on information in FloraMap and CaNaSTA. EcoCrop does not hold any information for this species. At the regional level, CaNaSTA shows much higher suitability than FloraMap. At the local level, both sources agree with higher suitability in the western portion of the region, but CaNaSTA assigns higher suitability to the entire region than FloraMap.

Figure 11.5 compares suitability of *C. pubescens* throughout Central America and in the San Dionisio region based on information in EcoCrop and CaNaSTA. FloraMap outputs have not been produced for this species. At the regional level, both sources agree with low suitability in central Guatemala, central Honduras and central Costa

Rica and higher suitability in northern Guatemala. At the local level, both sources agree with higher suitability in the eastern portion of the region and lower suitability in the west.

Figure 11.6 compares suitability of *B. brizantha* throughout Central America and in the San Dionisio region based on information in EcoCrop and CaNaSTA. FloraMap outputs have not been produced for this species. Both sources agree that the species is widely suitable throughout Central America and particularly in the north of Guatemala, the west of Nicaragua and most of Honduras. At the local level, both sources again agree that the species is suitable throughout the region, although CaNaSTA once again divides along the boundary between two rainfall classes.

In order to carry out statistical comparison, 904 evenly spaced points (0.2 degrees spacing) were selected in Central America, and the values at these points recorded for CaNaSTA, FloraMap and EcoCrop suitability maps. The values reflected in each map do not necessarily reflect the same levels suitability, as both source of data and method of assigning probabilities differ. However, for the purpose of analysis, the values in each map were divided into three classes, shown in Table 11.20 below.

	Class 1	Class 2	Class 3
CaNaSTA	Score < 0.33	Score 0.33 - 0.67	Score >= 0.67
FloraMap	Prob < 0.33	Prob 0.33 – 0.67	Prob >= 0.67
EcoCrop	Not suitable	Marginal	Suitable

Table 11.20 Classes for maps from different sources.

Based on these classifications, the joint information uncertainty statistic (Equation 7.31) was calculated for each map pair for each species (Table 11.21). The only moderate agreement is between CaNaSTA and EcoCrop for *C. pubescens*. All other comparisons show very low levels of agreement, although the agreement between CaNaSTA and EcoCrop tends to be slightly higher than agreements with FloraMap.

		CaNaSTA	FloraMap	EcoCrop
CaNaSTA	<i>S. guianensis</i>	1	0.05	0.09
	<i>A. pintoii</i>	1	0.02	0.07
	<i>C. argentea</i>	1	0.01	N/A
	<i>C. pubescens</i>	1	N/A	0.17
	<i>B. brizantha</i>	1	N/A	0.08
FloraMap	<i>S. guianensis</i>		1	0.04
	<i>A. pintoii</i>		1	0.02

Table 11.21 Joint information uncertainty for map comparisons.

CaNaSTA shows low to moderate agreement spatially with FloraMap and EcoCrop for the species selected, but agreement between FloraMap and EcoCrop is also poor. The maps for CaNaSTA and EcoCrop are based on different spatial variables, with CaNaSTA considering elevation, annual rainfall and length of dry season and EcoCrop considering elevation, rainfall and mean annual temperature. FloraMap maps are not based on a knowledge base, but purely on locations of known accessions in the wild. Therefore, FloraMap may misrepresent the *adaptation* of a species spatially, although it gives a good indication of where species may be found in the wild.

The comparison of maps is a useful exercise, but as no ‘ground truth’ maps exist of where species adapt well, it is difficult to draw strong conclusions from the comparison. This comparison draws attention to the fact that between different sources, there is low agreement as to where a species adapts well spatially. Although various publications, knowledge bases and databases exist, as outlined in Chapter 4, very few maps exist showing which locations are suitable for which species. Even though validation of these maps is inconclusive, it provides a starting point in the assessment of which species are suitable where.

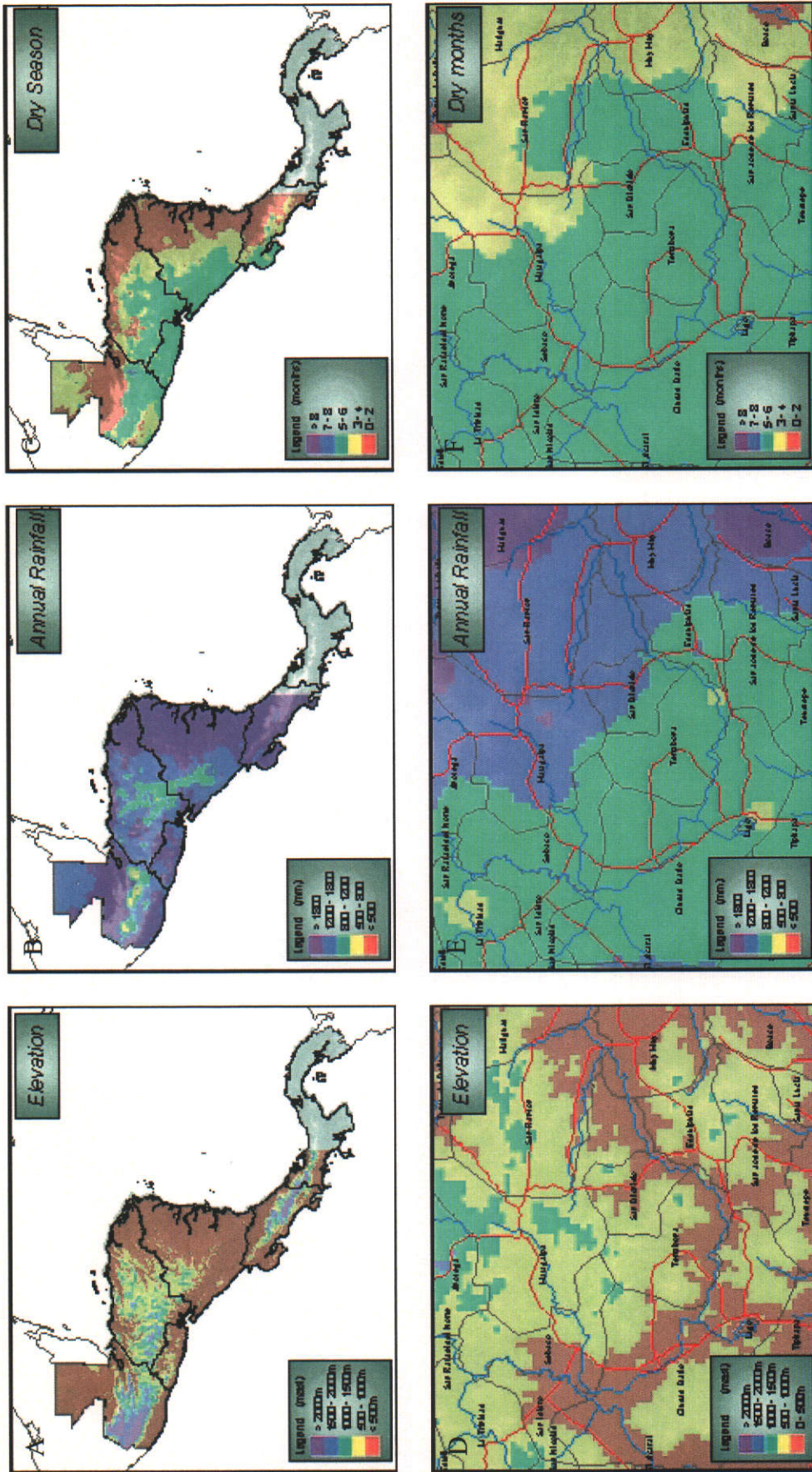


Figure 11.1 A. Elevation, Central America. B. Annual rainfall, Central America. C. Dry months, Central America. D. Elevation, San Dionisio region. E. Annual rainfall, San Dionisio region. F. Dry months, San Dionisio region.

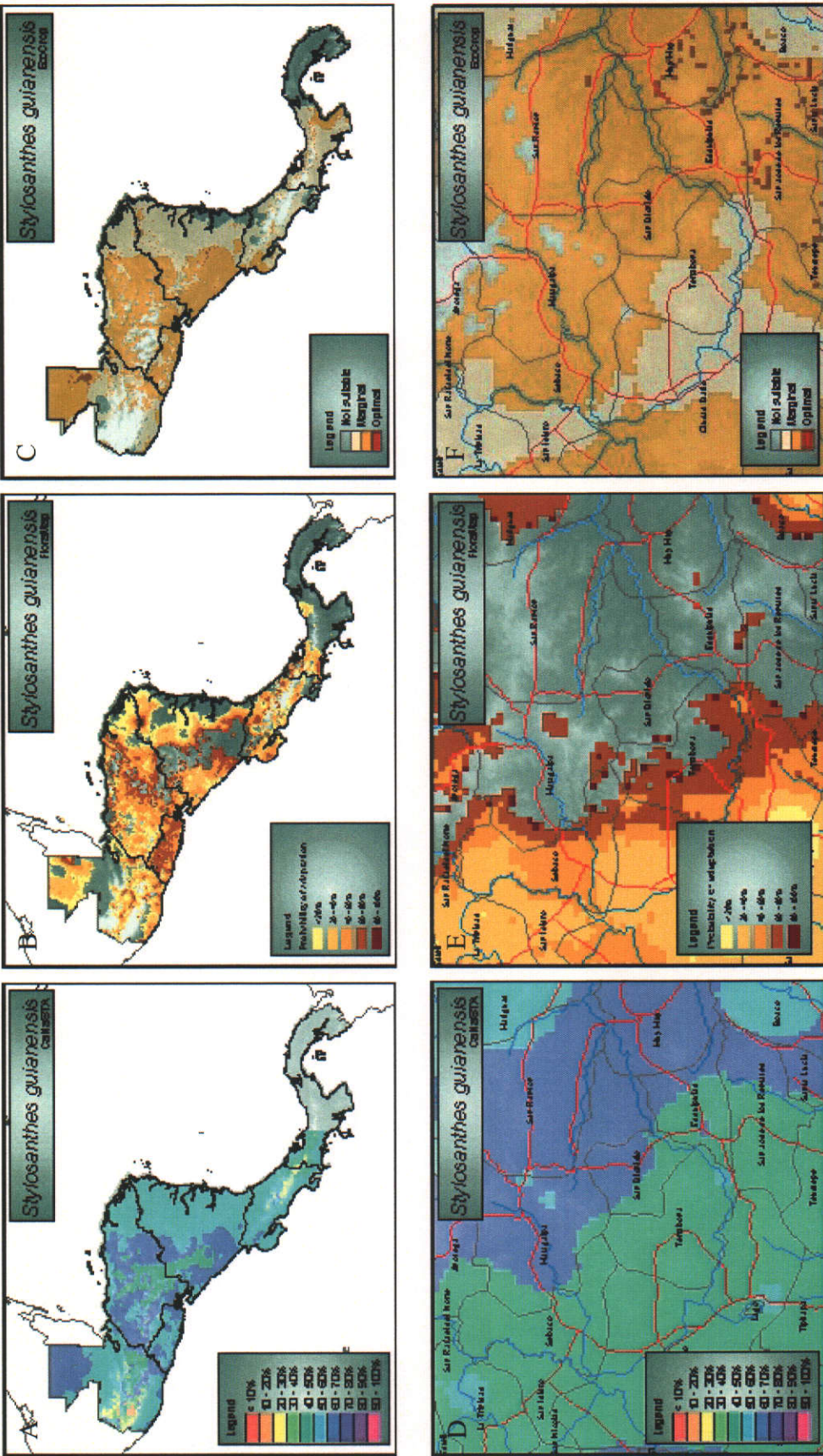


Figure 11.2 Probability of adaptation of *Stylosanthes guianensis*. A. CaNaSTA, Central America. B. FloraMap, Central America. C. EcoCrop, Central America. D. CaNaSTA, San Dionisio. E. FloraMap, San Dionisio. F. EcoCrop, San Dionisio.

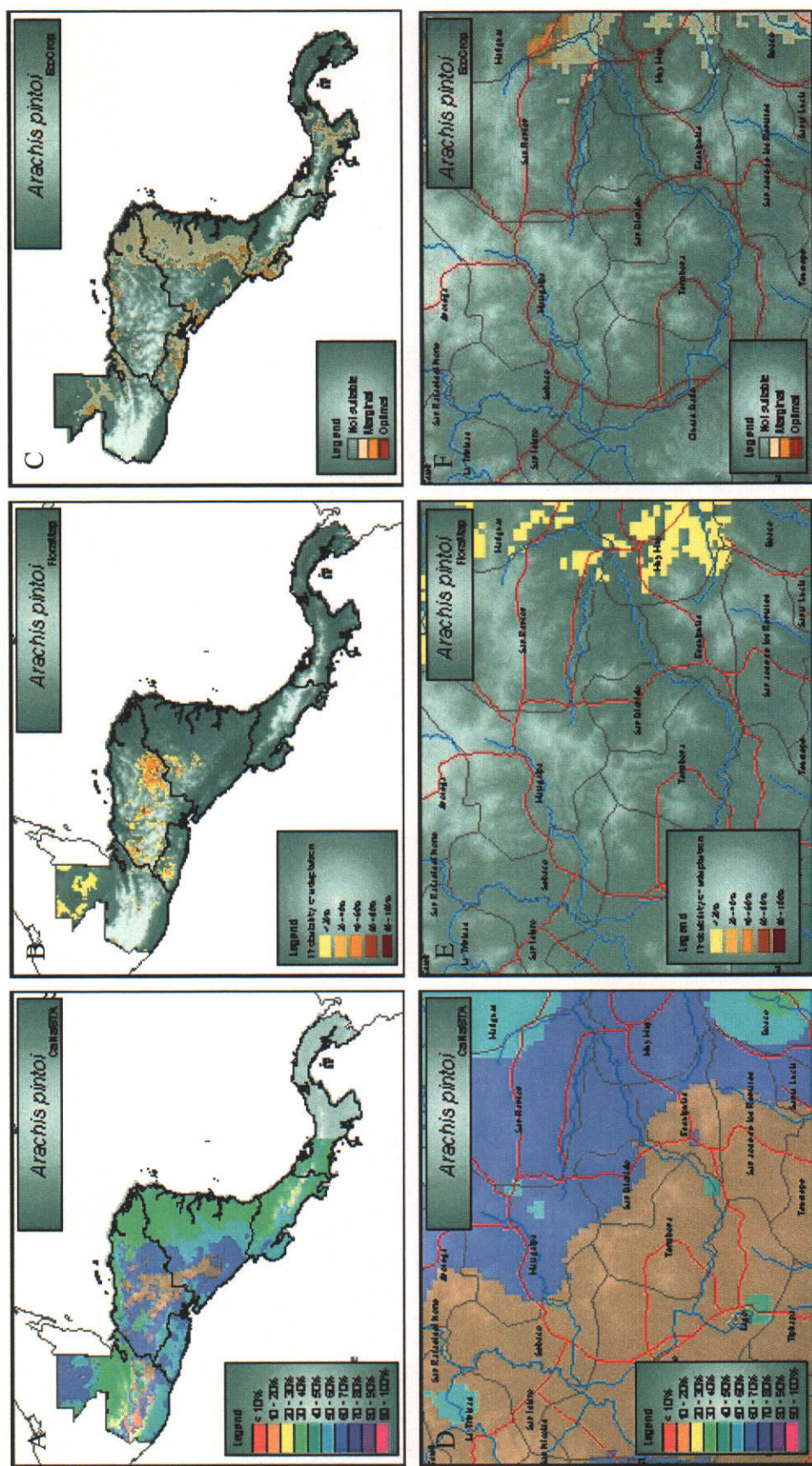


Figure 11.3 Probability of adaptation of *Arachis pintoi*. A. CaNaSTA, Central America. B. FloraMap, Central America. C. EcoCrop, Central America. D. CaNaSTA, San Dionisio. E. FloraMap, San Dionisio. F. EcoCrop, San Dionisio.

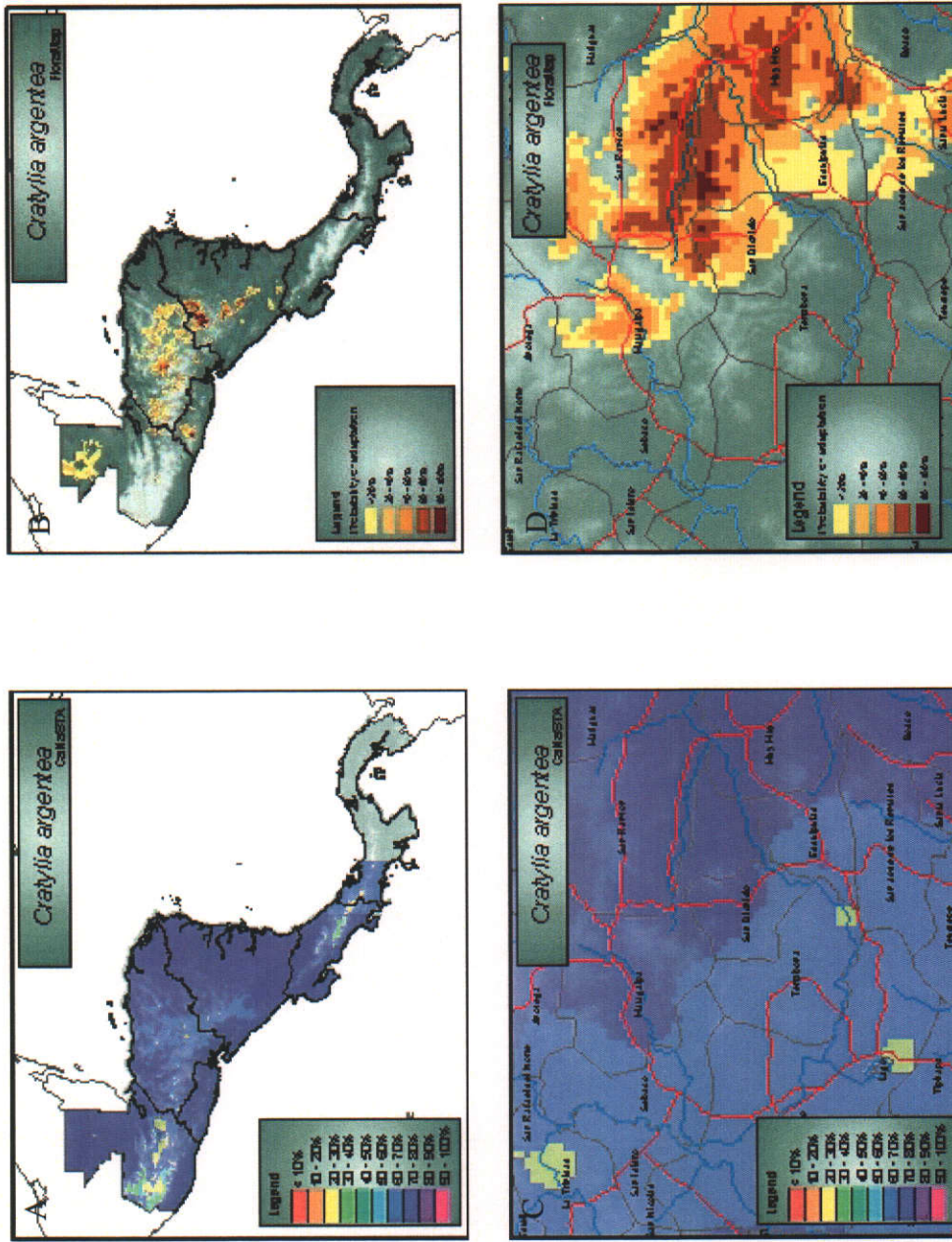


Figure 11.4 Probability of adaptation of *Cratylia argentea*. A. CaNaSTA, Central America. B. FloraMap, Central America. C. CaNaSTA, San Dionisio. D. FloraMap, San Dionisio.

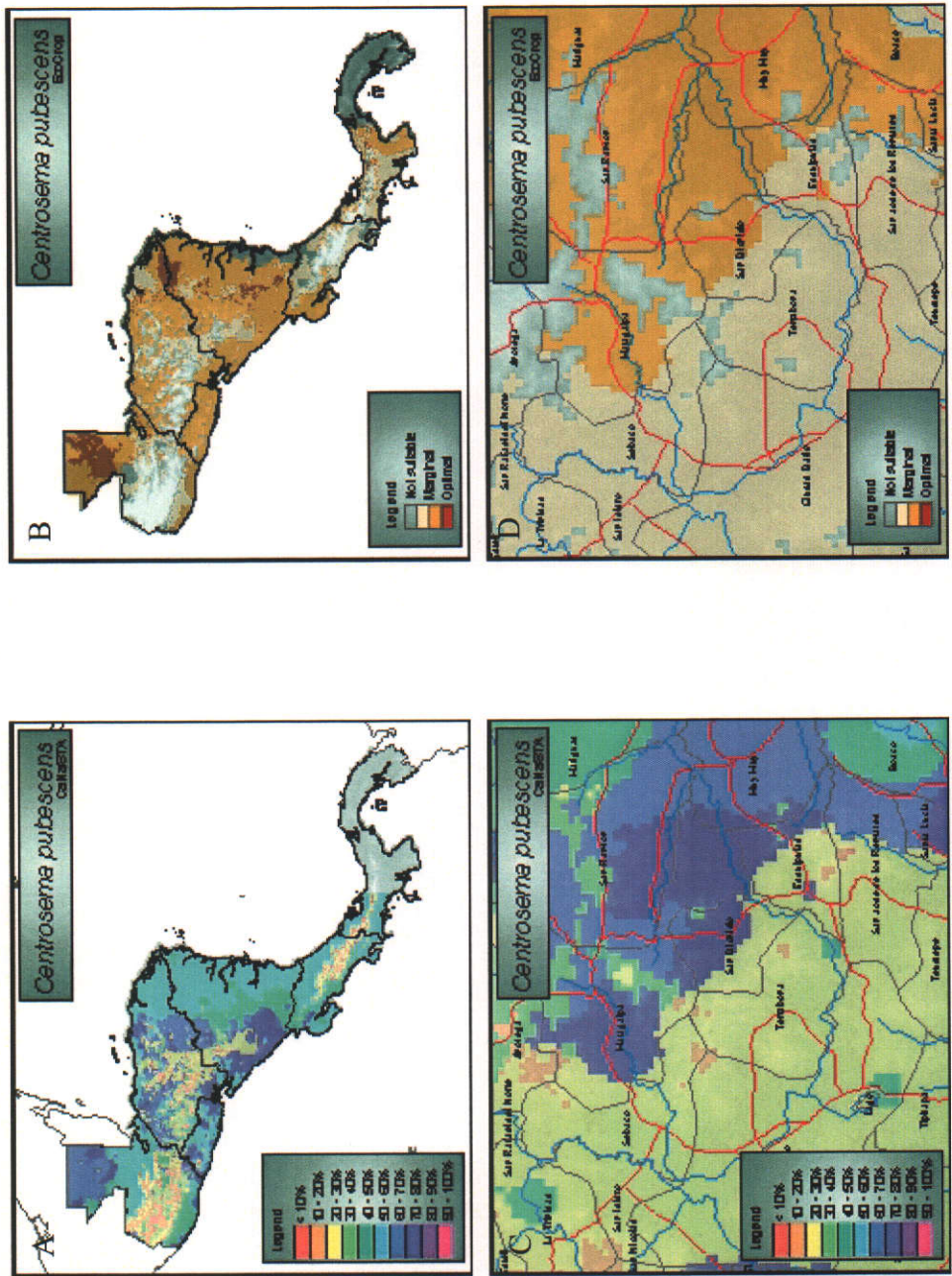


Figure 11.5 Probability of adaptation for *Centrosema pubescens*. A. CaNaSTA, Central America. B. EcoCrop, Central America. C. CaNaSTA, San Dionisio. D. EcoCrop, San Dionisio.

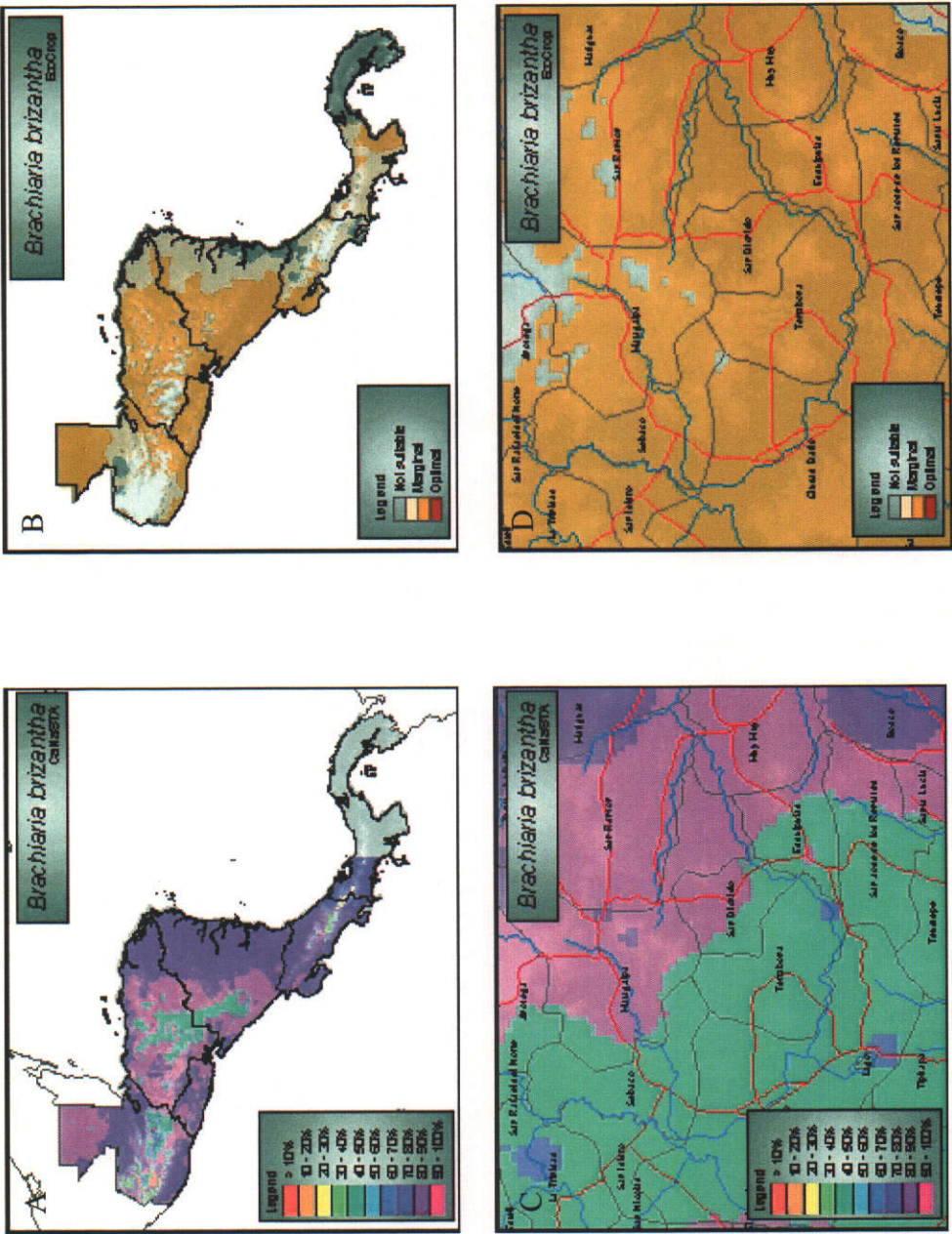


Figure 11.6 Probability of adaptation for *Brachiaria brizantha*. A. CaNaSTA, Central America. B. EcoCrop, Central America. C. CaNaSTA, San Dionisio. D. EcoCrop, San Dionisio.

11.2 Appropriateness of Method

11.2.1 Comparison with Other Models

Alternative software systems exist that were designed to assist with the decision of what forages to plant where. Most of these systems are not spatially implemented, and are simply in the form of a Boolean rule-base. If species fit all criteria, they are included in the list of suitable species. EcoCrop provides both optimal and absolute limits for temperature, rainfall, latitude, soil pH, light intensity, soil depth, soil texture, soil fertility, soil aluminium toxicity, soil salinity and soil drainage. In this way, species could be classified along a gradient of 'not suitable' to 'suitable', depending on the number of variables within optimal and absolute limits. As described in Chapter 9, SoFT defines a binary rule base, except in the case of rainfall where both optimal and absolute limits are given. SoFT is still in development at time of writing, and is subject to change. However, at present, SoFT simply returns a list of suitable species, without any kind of ranking. EcoCrop similarly returns all suitable species. EcoCrop also has an option to return only a set number of species, but these are simply the first species encountered in the database, and not necessarily the highest ranked.

The information in EcoCrop and SoFT could be used to create fuzzy envelopes of species suitability. This is the approach taken in the figures above to create the EcoCrop maps of species' suitability. However, as these systems do not have inherent spatial functionality, maps can only be produced on an ad-hoc basis.

FloraMap, on the other hand, is a spatially enabled tool, but does not work with a rule base or attribute data of any kind that directly relates species success to environmental variables. Instead these linkages are made through spatial data, that is, known climate at locations where a species is known to exist. FloraMap is designed precisely to function well in situations where location data is known ('presence'), but little else.

CaNaSTA lists suitable species, as with SoFT, EcoCrop and Lexsys, but, in addition, it ranks these species and provides dynamic maps of where the species is likely to adapt. Other expert systems often contain more variables for querying and provide

more information on the species' management and use. Although CaNaSTA is currently implemented to take account of six predictor variables, more variables can easily be added if information is available. It is hoped that rather than duplicating information on species' management and use, CaNaSTA and SoFT can be linked in the future, providing CaNaSTA with this information. In turn, SoFT can benefit from the maps, probability distributions and ranking system developed in CaNaSTA.

In an optimal situation, CaNaSTA would be tested by using it to recommend species for hundreds of farmers in varying conditions. Each farmer would adopt a few of the species suggested, and after two or three seasons adaptation would be assessed and compared with the recommendation of the tool. This approach is unrealistic for a number of reasons. Firstly, the sheer scale of such an assessment is impracticable. It is also unrealistic to wait for years to validate the output. In addition, the need to validate 'poor' adaptation would require species to be planted in environments where they are not expected to thrive. Finally, inherent uncertainties in the environment mean that multiple trials would need to be carried out for each species under each set of conditions in order to test the predicted adaptation distribution.

11.2.2 Feedback from Users

As CaNaSTA is still in development at the time of writing, it has not yet been released to users. However, a number of forage experts have either been involved in the development or have been exposed to the project at various stages. A questionnaire was sent to these experts soliciting their opinion on CaNaSTA as a tool for forage selection.

Four experts responded to this questionnaire and all thought that CaNaSTA would be useful in their research. They agreed that CaNaSTA would also be useful for NGOs, extension agents and scientists at international and national research centres. In addition, it was suggested that the tool may be useful for national government, better-off farmers, universities and consultants. They agreed that there is a need for this type of software and that an SDSS is an appropriate way to deliver information to forage professionals, but only to farmers through an intermediary.

CD-ROM and Internet were identified as appropriate delivery mechanisms, and possibly hard copy print outs of maps and information. Finally, it was agreed that CaNaSTA should work well with crops other than forages as long as input data is available.

Issues raised in the questionnaire responses include questions about species management, access to irrigation and fertilizer and presence of pests, and how these impact on which species should be recommended. In addition, the importance of subspecies and varieties was raised, i.e., in some cases it is important to distinguish between varieties when making a recommendation. The problem of seed availability and its impact on forage adoption was also raised. All these factors impact on forage adoption and need to be considered when making decisions about species selection.

11.3 Achievement of Stated Aims

The objectives of this research were to investigate ways of providing decision support in risky and uncertain environments and to develop an appropriate model to facilitate the decision process. The model criteria, as developed in Chapter 5, are the ability to work with small datasets and expert knowledge, the ability to predict a range of species' responses, low structural complexity, ease of communication and ability to implement spatially.

In Chapter 9, some of the attributes required for a successful SDSS were presented. These include the requirement of the SDSS to be stand-alone, with simple and functional design and presentation. Incorporating users' knowledge is important, as is explicitly dealing with sources of uncertainty.

11.3.1 Ability to Work with Small Datasets

CaNaSTA attempts to use all available data, even when it is sparse or uncertain. In the RIEPT database, in some cases very few trials exist for a given species. Even where no information exists from other sources (i.e., SoFT or expert knowledge), this data is still included in the model but certainty in the results is flagged as 'low'. This in turn determines, in part, whether the species is recommended in a given situation.

11.3.2 Ability to Work with Expert Knowledge

Expert and farmer knowledge are incorporated alongside data from a trials database. Combining information from various sources increases the strength of the model and allows for the inclusion of a larger amount of information. The SDSS is implemented so that information can be added continuously, as new data or expert knowledge becomes available.

Currently, farmer knowledge is only added as information on local soil characteristics. However, it is also possible for users to update species' probability distributions, if they consider they have the expertise to do so. In the future, the SDSS will have the ability to accept farmers' descriptions of forage suitability in their situation and to add this to the database as an additional data point.

11.3.3 Ability to Predict a Range of Species Responses

The model chosen is probabilistic, and it predicts species responses as the probability of the response being in certain states. This modelling approach addresses metrical and temporal uncertainty by utilising probability distributions rather than single values. The number of states has been set at four (adaptation classes), but could easily be changed to a different number if required. Bayesian calculus can also be applied to continuous gradients rather than discrete classes, and it is therefore theoretically possible to produce continuous probability surfaces. However, in the current implementation of CaNaSTA, the response is limited to discrete classes.

11.3.4 Low Structural Complexity

The model chosen is relatively straightforward and can be seen as a summary of available data and knowledge. This approach reduces structural uncertainty, thereby reducing overall uncertainty and allowing the user to have greater confidence in the model.

11.3.5 Ease of Communication

By implementing the model as an SDSS, the information is expected to reach its intended audience more readily. The low complexity model, the design of the GUI, the spatial aspect of CaNaSTA and the overall design of the SDSS are all intended to allow for more efficient communication of the results. In addition, the mere existence of the tool should assist with communication of trials results from the scientific community to farmers' advisors.

11.3.6 Ability to Implement Spatially

All the models considered in Chapter 6 can be implemented spatially. The challenges were to select appropriate spatial data and to implement a standalone SDSS. The spatial implementation of the model helps to reduce and describe all sources of uncertainty.

11.3.7 Appropriateness of Agricultural SDSS

The SDSS developed addresses the obstacles encountered by some agricultural DSS. CaNaSTA is standalone, and it has simple and functional design and presentation. Users' knowledge is incorporated to a certain extent, and uncertainty is dealt with explicitly. CaNaSTA has the ability to aid decision-making in tropical agriculture effectively.

11.4 Summary

In this chapter, the results of the research were presented and discussed. Accuracy of the model was checked by comparing results from CaNaSTA with results from a number of other sources. CaNaSTA showed reasonable agreement with these sources when recommending species for a given location. For five selected species, CaNaSTA showed moderate agreement with experts regarding suitability in selected locations. Spatial comparisons were also made for selected species by visually inspecting maps produced from different sources, and CaNaSTA showed moderate agreement spatially with other sources.

CaNaSTA provides some benefits over existing systems for determining which forage species are suitable where. CaNaSTA provides adaptation distributions, ranks suitable species and produces dynamic maps of species' suitability.

The development of the probabilistic GIS model and its development as an SDSS meet the objective of providing decision support in risky and uncertain environments. The appropriate reduction and description of different types of uncertainty allow farmers to better manage the risks associated with decision-making in uncertain environments. This is achieved in the implementation of CaNaSTA.

The final chapter of this thesis will summarise the conclusions of the research.

CHAPTER 12. CONCLUSIONS

This research has explored the nature of spatial decision problems and put forward an approach to supporting decision-making in tropical agriculture. It has been shown that farmers' decision problems can be supported by providing information and that delivery of information is improved through the use of computer tools and GIS.

The combination of data and expert knowledge in a spatial environment allows spatial and aspatial uncertainty to be explicitly modelled. This is an original approach to the problem of helping farmers decide what to plant where.

12.1 The Decision Problem

Improved forages are often a suitable option for smallholder farmers seeking to sustainably improve livelihoods. However, for a number of reasons, forage adoption is low, particularly in the case of legumes. One of these reasons is uncertainty on the part of the farmer about how particular forage species will perform in specific environments. Providing information on forages and their suitability to particular biophysical and socio-economic niches can equip farmers with the ability to make better-informed decisions.

Risk and uncertainty are factors in most decision-making, especially when the decision-making process has spatial aspects. In the case of supporting farmers' decisions about forage selection, there are a number of sources of risk and uncertainty. Decision makers in tropical agriculture include farmers, extension workers, NGOs, development agencies and national and international agricultural research institutions involved in tropical agriculture.

12.2 Addressing the Decision Problem

Functional models are needed to support farmers' tactical decisions, but the results of the model also need to be made available to the decision maker. A DSS, and in particular an SDSS, was identified as a well developed method for achieving this

aim. Because information sources include not only data but also knowledge, it is appropriate to develop an SDSS incorporating expert system concepts.

In selecting a model to address the decision problem, a number of criteria should be considered. The first criteria are the ability to work with small datasets and expert knowledge and the ability to predict a range of species' responses. In addition, the model must display low structural complexity and must be easy to communicate and to implement spatially.

A probabilistic GIS model was identified as a suitable model and was developed based on data and knowledge available for Central America and forage crops. The model allows information from diverse sources on success of forages to be combined to predict success distributions for any combination of variables. The model incorporates uncertainty, retaining uncertainty information throughout the model and allowing this information to be displayed and interrogated in a GIS environment.

12.3 SDSS Development and Implementation

This model was implemented as the stand-alone software CaNaSTA. The processes and methods used address many of the problems encountered with other agricultural DSS and SDSS. CaNaSTA recommends species for a given location and situation, and recommends locations for a given species. In addition, users can update data interactively and examine results through maps, tables and graphs. Incorporating spatial capabilities into an agricultural DSS, as in CaNaSTA, facilitates data input, allows more informative output of results and allows spatial variability to be made explicit, both of results and of uncertainties related to the results.

Many of the issues encountered with other agricultural DSS and SDSS were overcome in this implementation. The spatial implementation assists with interpretation of the information and explicitly shows spatial variation and uncertainty. The design and presentation of the SDSS is intended to be simple and functional. The user does not need to provide monitoring data or to have access to any additional proprietary software.

The software has not yet been released to users, but a number of forage experts have been involved in the development of CaNaSTA. Their input has guided the development and implementation of the tool. Limited testing has been carried out by comparing expert recommendations and recommendations from other knowledge bases with those made by CaNaSTA for specific farmers' situations. CaNaSTA shows considerable overlap with recommendations from other sources. In addition, CaNaSTA provides details on the likely adaptation distribution of each species at each location, as well as measures of sensitivity and certainty. Rather than simply classify each species as either suitable or unsuitable for each niche, more detailed information is given, allowing the user to make a better-informed decision.

12.4 Further Research and Development

Further research is needed in order to validate the model once the final SoFT knowledge base is available. Adaptation distributions and maps need to be assessed for accuracy by experts, and the knowledge base updated where necessary. This validation process also provides an alternative way for experts to express their knowledge regarding species' spatial adaptation. The validity of the outscaling to regions outside of Central America can also be tested in this way.

Further work is required to validate the use of the tool in the field. This research should follow two strands. The first is to further validate the accuracy of the software output in terms of species recommended for locations and locations suggested for species. This validation should be implemented on an ongoing basis as more trials are completed in different locations (including on-farm trials and adoption). The second strand of research should assess adoption and use of the completed software in order to validate the chosen form of delivery. Results from this research will inform future versions of the software and reveal whether other modes of delivery might be more appropriate. Participatory research methodology was briefly described in the introductory chapter. Further development of the tool should be along participatory guidelines, with shared ownership of research and a community focus. In addition, more effort is needed to incorporate farmers' own knowledge in the tool and in the wider research.

This development is needed on a continual basis in order to maintain CaNaSTA (and any successors and derivatives) as a useful and accurate tool. The fact that such updates can easily be accommodated is an advantage of the method.

Two steps in the DSS development process that have not yet been addressed are capacity building and fostering uptake. It is hoped that use of the software will be largely intuitive and that little training will be necessary. Both help files and a user manual are in production. However, this process will be monitored and adjusted to provide appropriate capacity building. Uptake can be primarily fostered by making potential users aware of the tool and its utility. The tasks of capacity building, fostering uptake and maintaining the software will ultimately be the responsibility of CIAT.

Further research is needed to successfully incorporate socio-economic data. Market access and market price information could be valuable, particularly where cash crops are concerned. The incorporation of market information adds a temporal dimension to the decision problem, with market access and market prices potentially impacting on decisions of not only what to plant, but also when to plant and how to manage the crop.

A number of issues have been raised in the course of this research which should be investigated or analysed further. With regard to predictor variables, additional variables such as aspect, slope and temperature could prove beneficial. Because temperature is highly correlated with elevation, the addition of this variable would need to be carefully handled. In addition, distinguishing subtropical and tropical environments could be helpful. The temporal dimension could also be addressed in regards to seasonality of rainfall and other climatic factors, which could be important for species with short growing periods. The potential impact of extreme climate events is also of importance when selecting species, and this should also be considered.

Further research is also warranted on the implementation of the probabilistic GIS model. Sensitivity to prior distributions should be analysed, as well as the impacts of biases in databases and in expert knowledge. The possibility of implementing

continuous probability distributions (rather than stepped) could improve the mapped outputs. Also, methods for ‘smoothing’ adaptation scores across categories (to avoid sudden ‘jumps’ when changing from one variable category to another) could be investigated.

12.5 Applicability to Other Fields

Within tropical agriculture, the research is clearly applicable to any new crops or forages. It is particularly relevant to niche crops that are not as widely adopted as could be expected, such as tropical fruits, nuts and coffee. For many of these crops only limited data is available. CaNaSTA can function with expert knowledge alone, where data is missing.

In a broader sense, the functionality of the model implemented in CaNaSTA could be applied to any situation where a mixture of sparse data and expert knowledge is available about the location of an entity given spatial characteristics of the location. This could include animal habitats, land use planning and site selection. This model is not appropriate to all spatial decision problems, however, where there is uncertainty, sparse data and expert knowledge, the approach could prove beneficial.

12.6 Lessons Learned

The development of a fully functioning computer tool, even just as a proof of concept, is a complex task. Ideally, it should be undertaken by a team of developers in conjunction with stakeholders. Although some stakeholders were involved in the development of CaNaSTA, the fact that the sole developer was also the researcher raised a number of issues. Aside from time constraints, the focus of the researcher on software development made objectivity in its assessment problematic. Ideally, a much larger team of people should have been involved in all stages from software development to GUI design to testing and debugging.

Tropical agriculture is a complex but well-researched field. Researchers in tropical developing agriculture include plant biologists, entomologists, bio-technicians, economists, anthropologists and geographers, to name just a few. In addition, farmers themselves are custodians of vast amounts of knowledge, which is only just

beginning to be tapped by researchers. Although many of these people were involved at various stages of the research, in hindsight, better use could have been made of their accumulated knowledge and experience.

12.7 Summary

The research reported in this thesis addresses the problem of making tactical decisions about land use under uncertainty. A model has been proposed and implemented as a software tool for use by those advising smallholder farmers in forage adoption.

Providing information to farmers to support their decision-making in uncertain environments is a meaningful goal. Farmers want more and better information and extension workers want to find ways to provide this information. Scientists often have this information and want to find meaningful and consistent ways of providing it to those who could benefit from it. This research has contributed to these goals.

The research shows that even with limited data and knowledge, results can be obtained that support the farmers' decision-making process. When uncertainties are made explicit, farmers can then make less risky decisions by taking these uncertainties into account.

Providing access to decision support through a Spatial Decision Support System, such as CaNaSTA, ensures that the information is delivered in a consistent and robust manner. Trial data and expert knowledge previously inaccessible to farmers are made available so that decisions taken are better informed.

These decisions will increase the adoption of appropriate forages, contributing towards sustainability, improving meat and milk quality, combating food problems and, ultimately, improving the livelihoods of smallholder farmers and their communities in the developing tropics.

REFERENCES

- Abadi Ghadim, A. (2000) Risk, Uncertainty and Learning in Farmer Adoption of a Crop Innovation, *PhD Thesis*, Faculty of Agriculture, University of Western Australia, Perth, 280 pp.
- Agumya, A. and Hunter, G.J. (2002) Responding to the Consequences of Uncertainty in Geographical Data, *International Journal of Geographical Information Science*, Vol. 16, No. 5, pp. 405-417.
- Amezquita, E. (2003). *Personal Communication*. CIAT, Soil Project, Cali, Colombia.
- Andersen, M.C., Watts, J.M., Freilich, J.E., Yool, S.R., Wakefield, G.I., McCauley, J.F. and Fahnestock, P.B. (2000) Regression-Tree Modeling of Desert Tortoise Habitat in the Central Mojave Desert, *Ecological Applications*, Vol. 10, No. 3, pp. 890-900.
- Anderson, R.P., Lew, D. and Peterson, A.T. (2003) Evaluating Predictive Models of Species' Distributions: Criteria for Selecting Optimal Models, *Ecological Modelling*, Vol. 162, pp. 211-232.
- Antle, J.M. (1987) Econometric Estimation of Producers' Risk Attitudes, *American Journal of Agricultural Economics*, Vol. 69, No. 3, pp. 509-522.
- Antonic, O., Pernar, N. and Jelaska, S.D. (2003) Spatial Distribution of Main Forest Soil Groups in Croatia as a Function of Basic Pedogenetic Factors, *Ecological Modelling*, Vol. 170, No. 2-3, pp. 363-371.
- Argel, P., Hidalgo, C., González, J., Lobo, M., Acuña, V. and Jiménez, C. (2001) Cultivar Veraniega (*Cratylia argentea* (Desv.) O. Kuntze) Una Leguminosa Arbustiva para la Ganadería de América Latina Tropical, Consorcio Tropileche, San José, Costa Rica, 22 pp.
- Argel, P., Hidalgo, C. and Lobo, M. (2000) Pasto Toledo (*Brachiaria brizantha* CIAT 26110) Gramínea de Crecimiento Vigoroso con Amplio Rango de Adaptación a Condiciones de Trópico Húmedo y Subhúmedo, Consorcio Tropileche, San José, Costa Rica, 15 pp.
- Argel, P. and Villarreal, M. (1998) Nuevo Maní Forrajero Perenne (*Arachis pintoi* Krapovickas y Gregory) Cultivar Porvenir - Leguminosa Herbácea para Alimentación Animal, el Mejoramiento y Conservación del Suelo y el Embellecimiento del Paisaje, *Boletín Técnico*, MAG/CIAT, San José, Costa Rica, 32 pp.
- Asadi, H.H. and Hale, M. (2001) A Predictive GIS Model for Mapping Potential Gold and Base Metal Mineralization in Takab Area, Iran, *Computers & Geosciences*, Vol. 27, No. 8, pp. 901-912.

- Aspinall, R. (1992) An Inductive Modelling Procedure Based on Bayes' Theorem for Analysis of Pattern in Spatial Data, *International Journal of Geographical Information Systems*, Vol. 6, No. 2, pp. 105-121.
- Aspinall, R. (2002) Use of Logistic Regression for Validation of Maps of the Spatial Distribution of Vegetation Species Derived from High Spatial Resolution Hyperspectral Remotely Sensed Data, *Ecological Modelling*, Vol. 157, pp. 301-312.
- Aspinall, R. and 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*, Vol. 59, No. 4, pp. 537-543.
- Assimakopoulos, J.H., Kalivas, D.P. and Kollias, V.J. (2003) A GIS-Based Fuzzy Classification for Mapping the Agricultural Soils for N-Fertilizers Use, *The Science of the Total Environment*, Vol. 309, pp. 19-33.
- Austin, G.E., Thomas, C.J., Houston, D.C. and 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*, Vol. 33, pp. 1541-1550.
- Austin, M.P. (2002) Spatial Prediction of Species Distribution: An Interface Between Ecological Theory and Statistical Modelling, *Ecological Modelling*, Vol. 157, pp. 101-118.
- Barbosa, A.M., Real, R., Olivero, J. and Vargas, J.M. (2003) Otter (*Lutra lutra*) Distribution Modeling at Two Resolution Scales Suited to Conservation Planning in the Iberian Peninsula, *Biological Conservation*, Vol. 114, pp. 377-387.
- Barco, F., Franco, M.A., Franco, L.H., Hincapie, B., Lascano, C., Ramírez, G. and Peters, M. (2002) *Forrajes Tropicales: Base de Datos de Recursos Genéticos Multipropósito, Versión 1.0 (Tropical Forages Database)*, <http://www.ciat.cgiar.org/catalogo/producto.jsp?codigo=P0219>.
- Bar-Hillel, M. (1983) The Base Rate Fallacy Controversy, in: *Decision Making Under Uncertainty*, Scholz, R. W. (ed.), Elsevier, Amsterdam, pp. 39-61.
- Batjes (1997) A World Data Set of Derived Soil Properties by FAO-UNESCO Soil Unit for Global Modelling, *Soil Use and Management*, Vol. 13, pp. 9-16.
- Bayes, T. (1763) An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society*, Vol. 53, pp. 370-418.
- Bayes, T. (1958) An Essay Towards Solving a Problem in the Doctrine of Chances, *Biometrika*, Vol. 45, pp. 296-315 (Reprint).

- Berger, T. (2001) Agent-based Spatial Models Applied to Agriculture: a Simulation Tool for Technology Diffusion, Resource Use Changes and Policy Analysis, *Agricultural Economics*, Vol. 25, pp. 245-260.
- Bonan, G. (2002) *Ecological Climatology*, Cambridge University Press, Cambridge. 678 pp.
- Bonham-Carter, G.F. (1994) *Geographic Information Systems for Geoscientists*, Pergamon, Ottawa. 398 pp.
- Booth, T.H. (1995) Predicting Plant Growth: Where Will it Grow? How Well Will it Grow?, *Proceedings of Third International Conference / Workshop on Integrating GIS and Environmental Modelling*, National Center for Geographic Information and Analysis, Santa Fe, New Mexico, 21-25 January 1995.
- Booth, T.H. (1999) Matching Germplasm to Geography: Environmental Analysis for Plant Introduction, in: *Linking Genetic Resources and Geograpy: Emerging Strategies for Conserving and Using Crop Biodiversity*, CSSA (ed.), CSSA, Madison.
- Booth, T.H. and Jones, P.G. (1998) Identifying Climatically Suitable Areas for Growing Particular Trees in Latin America, *Forest Ecology and Management*, Vol. 108, pp. 167-173.
- Bordogna, G. and Pasi, G. (2000) Modeling Linguistic Qualifiers of Uncertainty in a Fuzzy Database, *International Journal of Intelligent Systems*, Vol. 15, pp. 995-1014.
- Borland Software Corporation (2002) Borland Delphi 6.0 Software, <http://www.borland.com> (computer program).
- Bradshaw, C.J.A., Davis, L.S., Purvis, M., Zhou, Q. and Benwell, G.L. (2002) Using Artificial Neural Networks to Model the Suitability of Coastline for Breeding by New Zealand Fur Seals (*Arctocephalus forsteri*), *Ecological Modelling*, Vol. 148, pp. 111-131.
- Brooker, S., Hay, S.I. and Bundy, D.A.P. (2002) Tools from Ecology: Useful for Evaluating Infection Risk Models?, *Trends in Parasitology*, Vol. 18, No. 2, pp. 70-74.
- Burrough, P.A., van Gaans, P.F.M. and Hootmans, R. (1997) Continuous Classification in Soil Survey: Spatial Correlation, Confusion and Boundaries, *Geoderma*, Vol. 77, pp. 115-135.
- Busby, J.R. (1991) BIOCLIM - A Bioclimatic Analysis and Prediction System, in: *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, Margules, C. R. and Austin, M. P. (eds.), CSIRO, Melbourne, Australia, pp. 64-68.

- CABI (2003) CABI Compendia, <http://www.cabi.org/compendia/index.asp>.
- Carey, P.D. and Brown, N.J. (1994) The Use of GIS to Identify Sites That Will Become Suitable For a Rare Orchid, *Himantoglossum hircinum* L., in a Future Changed Climate, *Biodiversity Letters*, Vol. 2, pp. 117-123.
- Carpenter, G., Gillison, A.N. and Winter, J. (1993) DOMAIN: A Flexible Modelling Procedure for Mapping Potential Distributions of Plants and Animals, *Biodiversity and Conservation*, Vol. 2, pp. 667-680.
- Chalmers, A. (1999) *What is This Thing Called Science?*, University of Queensland Press, St Lucia, Queensland. 266 pp.
- Chambers, R. (1980) The Small Farmer is a Professional, *Ceres*, Vol. March-April, pp. 19-23.
- Chipeta, S., Hoydahl, E. and Krog, J. (2003) *Livestock Services and the Poor: A Global Perspective - Collecting, Coordinating and Disseminating Experience*, Danida, IFAD and The World Bank, 90 pp.
- CIAT (1980) *CIAT Report*, CIAT, Cali, Colombia, 101 pp.
- CIAT (1999) *Atlas de Honduras*, CIAT, Cali, Colombia (CD-ROM).
- CIAT (2003) *IntelAgro Version 1.0: Herramienta para Zonificación Agro Ecológica e Inteligencia de Mercados con Interfase Tipo GIS*, Rural Agroenterprise Development Project, Centro Internacional de Agricultura Tropical, Cali, Colombia (computer program).
- CIESEN, IFPRI and WRI (2000) *Gridded Population of the World, Version 2*, <http://sedac.ciesin.columbia.edu/plue/gpw>.
- Civco, D.L. (1993) Artificial Neural Networks for Land-Cover Classification and Mapping, *International Journal of Geographical Information Systems*, Vol. 7, No. 2, pp. 173-186.
- Clark, M.J. (2002) Dealing with Uncertainty: Adaptive Approaches to Sustainable River Management, *Aquatic Conservation: Marine and Freshwater Ecosystems*, Vol. 12, pp. 347-363.
- Coenen, F., Eaglestone, B. and Ridley, M. (2001) Verification, Validation and Integrity Issues in Expert and Database Systems: Two Perspectives, *International Journal of Intelligent Systems*, Vol. 16, pp. 425-447.
- Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, Vol. 20, pp. 37-46.
- Corner, R.J., Hickey, R.J. and Cook, S.E. (2002) Knowledge Based Soil Attribute Mapping in GIS: The Expector Method, *Transactions in GIS*, Vol. 6, No. 4, pp. 383-402.

- Cox, P.G. (1996) Some Issues in the Design of Agricultural Decision Support Systems, *Agricultural Systems*, Vol. 52, No. 2-3, pp. 355-381.
- Cramb, R.A. (2000) Processes Influencing the Successful Adoption of New Technologies by Smallholders, in: *Working with Farmers: the Key to Adoption of Forage Technologies*, Stür, W. W., Horne, P. M., Hacker, J. B. and Kerridge, P. C. (eds.), Australian Centre for International Agricultural Research, Canberra, pp. 11-22.
- Cramer, P. and Portier, K. (2001) Modeling Florida Panther Movements in Response to Human Attributes of the Landscape and Ecological Settings, *Ecological Modelling*, Vol. 140, pp. 51-80.
- Crosetto, M. and Tarantola, S. (2001) Uncertainty and Sensitivity Analysis: Tools for GIS-Based Model Implementation, *International Journal of Geographical Information Science*, Vol. 15, No. 5, pp. 415-437.
- Crossland, M.D., Wynne, B.E. and Perkins, W.C. (1995) Spatial Decision Support Systems: An Overview of Technology and a Test of Efficacy, *Decision Support Systems*, Vol. 14, pp. 219-235.
- Davis, J.P. and Hall, J.W. (2003) A Software-Supported Process for Assembling Evidence and Handling Uncertainty in Decision-Making, *Decision Support Systems*, Vol. 35, No. 3, pp. 415-433.
- Davis, R. (2000) *Global Mapping, Global Forest Resources Assessment 2000*, <http://www.fao.org/DOCREP/004/Y1997E/y1997e1g.htm>.
- Davis, T.J. and Keller, C.P. (1997) Modelling and Visualizing Multiple Spatial Uncertainties, *Computers and Geosciences*, Vol. 23, No. 4, pp. 397-408.
- de la Rosa, D., Mayol, F., Moreno, J.A., Bonsón, T. and Lozano, S. (1999) An Expert System Neural Network Model (ImpelERO) for Evaluating Agricultural Soil Erosion in Andalusia Region, Southern Spain, *Agriculture, Ecosystems and Environment*, Vol. 73, No. 3, pp. 211-226.
- De'ath, G. and Fabricius, K.E. (2000) Classification and Regression Trees: a Powerful Yet Simple Technique for Ecological Data Analysis, *Ecology*, Vol. 81, No. 11, pp. 3178-3192.
- Debeljak, M., Dzeroski, S., Jerina, K., Kobler, A. and Adamic, M. (2001) Habitat Suitability Modelling for Red Deer (*Cerphus elaphus* L.) in South-Central Slovenia with Classification Trees, *Ecological Modelling*, Vol. 138, pp. 321-330.
- Delgado, C., Rosegrant, M., Steinfeld, H., Ehui, S. and Courbois, C. (1999) *Livestock to 2020: The Next Food Revolution*, www.ifpri.org/2020/briefs/number61.htm.

- Denzin, N. and Lincoln, Y. (2000) The Discipline and Practice of Qualitative Research, in: *Handbook of Qualitative Research*, Denzin, N. and Lincoln, Y. eds.), Sage Publications, Thousand Oaks, CA, pp. 1-29.
- Desnoyers, D. (2001) *Arc-WofE - Weights of Evidence Extension for ArcView/Spatial Analyst*, <http://ntserv.gis.nrcan.gc.ca/wofe/>.
- Dittmer, S.L. and Jensen, F.V. (1997) Tools for Explanation in Bayesian Networks with Application to an Agricultural Problem, *Proceedings of Proceedings of the First European Conference for Information Technology in Agriculture*, Copenhagen, Denmark, June 15-18, 1997.
- Dixon, J. (2003) Characterising Livestock Production Systems and Their Linkages to Poverty, in: *Pro-Poor Livestock Policy Initiative Information Resources*, FAO (ed.)
- Donnelly, J.R., Freer, M., Salmon, L., Moore, A.D., Simpson, R.J., Dove, H. and Bolger, T.P. (2002) Evolution of the GRAZPLAN Decision Support Tools and Adoption by the Grazing Industry in Temperate Australia, *Agricultural Systems*, Vol. 74, pp. 115-139.
- Drumm, D., Purvis, M. and Zhou, Q. (1999) Spatial Ecology and Artificial Neural Networks: Modeling the Habitat Preference of the Sea Cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands, *Proceedings of SIRC 99 - The 11th Annual Colloquium of the Spatial Information Research Centre*, University of Otago, Dunedin, New Zealand, 13-15 December 1999.
- Ducey, M.J. (2001) Representing Uncertainty in Silvicultural Decisions: an Application of the Dempster-Shafer Theory of Evidence, *Forest Ecology and Management*, Vol. 150, No. 3, pp. 199-211.
- Eade, J., Farrow, A., Knapp, R., Leclerc, G., Nelson, A. and Winograd, M. (2000) *Accessibility Analyst*, <http://www.ciat.cgiar.org/catalogo/producto.jsp?codigo=P0173>.
- Edwards, W. and Fasolo, B. (2001) Decision Technology, *Annual Review of Psychology*, Vol. 52, pp. 581-606.
- Elith, J., Burgman, M.A. and Regan, H.M. (2002) Mapping Epistemic Uncertainties and Vague Concepts in Predictions of Species Distribution, *Ecological Modelling*, Vol. 157, pp. 313-329.
- ESRI (1992) *Digital Chart of the World [machine readable data file]*, ESRI, Redlands, CA.
- ESRI (1997) *World Bioclimatic Soils Regions*, www.geographynetwork.com.
- ESRI (2000) *MapObjects LT 1.0. Software Program*, <http://www.esri.com>.
- ESRI (2001) *ArcView 3.x*, <http://www.esri.com/software/arcview/index.html>.

- Evans, F. and Caccetta, P. (2000) Broad-Scale Spatial Prediction of Areas at Risk from Dryland Salinity, *Cartography*, Vol. 29, No. 2, pp. 33-40.
- FAO (2000) *Ecocrop*, <http://ecocrop.fao.org>.
- FAO (2003) *First Meeting of the Steering Committee for FAO's Pro-Poor Livestock Policy Initiative*,
http://www.fao.org/ag/againfo/projects/en/pplpi/docarc/sc_meeting_report_01.pdf.
- FAO (2004) *Problem Soils*,
<http://www.fao.org/waicent/faoinfo/agricult/agl/agll/prosoil/table.htm>.
- FAO, AGA and FRG (2004) *Animal Feed Resources Information System*,
<http://www.fao.org/ag/aga/agap/frg/afris/default1.htm>.
- FAO, UNEP and CGIAR (2004) *Poverty Mapping*, <http://www.povertymap.net/>.
- FAO Agriculture Department (2004) *Global Livestock Production and Health Atlas*,
<http://www.fao.org/ag/againfo/resources/en/glipha/default.html>.
- FAOSTAT (2004) *FAOSTAT Data*, <http://faostat.fao.org/faostat/>.
- Ferson, S. and Ginzburg, L.R. (1996) Different Methods are Needed to Propagate Ignorance and Variability, *Reliability Engineering and System Safety*, Vol. 54, pp. 133-144.
- Feyerabend, P.K. (1975) *Against Method: Outline of an Anarchistic Theory of Knowledge*, New Left Books, London. 279 pp.
- Fleiss, J. (1981) *Statistical Methods for Rates and Proportions*, Wiley, New York. 321 pp.
- Fleiss, J. and Cohen, J. (1973) The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability, *Educational and Psychological Measurement*, Vol. 33, pp. 613-619.
- Franco, L.H. (2002) *Personal Communication*, CIAT, Tropical Forages, Cali, Colombia.
- Fresco, L.O. (2001) *Agriculture After September 11*,
<http://www.fao.org/ag/magazine/0112sp2.htm>.
- Girard, N. and Hubert, B. (1999) Modelling Expert Knowledge with Knowledge-Based Systems to Design Decision Aids - The Example of a Knowledge-Based Model on Grazing Management, *Agricultural Systems*, Vol. 59, pp. 123-144.

- Glenz, C., Massolo, A., Kuonen, D. and Schlaepfer, R. (2001) A Wolf Habitat Suitability Prediction Study in Valais (Switzerland), *Landscape and Urban Planning*, Vol. 55, pp. 55-65.
- Gu, Y., Crawford, J.W., Peiris, D.R., Grashoff, C., McNicol, J.W. and Marshall, B. (1996) Modelling Faba Bean Production in an Uncertain Future Climate, *Agricultural and Forest Meteorology*, Vol. 79, pp. 289-300.
- Guisan, A., Edwards, T.C. and Hastie, T. (2002) Generalized Linear and Generalized Additive Models in Studies of Species Distributions: Setting the Scene, *Ecological Modelling*, Vol. 157, pp. 89-100.
- Guisan, A., Weiss, S.B. and Weiss, A.D. (1999) GLM Versus CCA Spatial Modeling of Plant Species Distribution, *Plant Ecology*, Vol. 143, pp. 107-122.
- Guisan, A. and Zimmerman, N. (2000) Predictive Habitat Distribution Models in Ecology, *Ecological Modelling*, Vol. 135, No. 2-3, pp. 147-186.
- Hall, G.B., Feick, R.D. and Bowerman, R.L. (1997) Problems and Prospects for GIS-Based Decision Support Applications in Developing Countries, *South African Journal of Geoinformation*, Vol. 17, No. 3, pp. 81-87.
- Hardaker, J.B., Huirne, R.B.M. and Anderson, J.R. (1997) *Coping With Risk in Agriculture*, CAB International, Wallingford, UK. 274 pp.
- Hart, P.E., Duda, R.O. and Einaudi, M.T. (1978) PROSPECTOR - A Computer-Based Consultation System for Mineral Exploration, *Mathematical Geology*, Vol. 10, No. 5, pp. 589-610.
- Heckerman, D. (1995) *A Tutorial on Learning Bayesian Networks*, MSR-TR-95-06, Microsoft Research, Redmond, 41 pp.
- Heckerman, D., Geiger, D. and Chickering, D. (1995) *Learning Bayesian Networks: The Combination of Knowledge and Statistical Data*, MSR-TR-94-09, Microsoft Corporation, Redmond, 54 pp.
- Heuvelink, G. (1998) *Error Propagation in Environmental Modelling with GIS*, Taylor & Francis, London. 144 pp.
- Hijmans, R.J., Cruz, M., Rojas, E. and Guarino, L. (2001) *DIVA-GIS: A Geographic Information System for the Management and Analysis of Genetic Resources Data - Manual*, http://gis.cip.cgiar.org/gis/Tools/diva/DIVA_manual.pdf
- Hijmans, R.J., Cameron, S., Parra, J., Jones, P.G., Jarvis, A. and Richardson, K. (2004a) *WORLDCLIM Version 1.2*, <http://biogeo.berkeley.edu/worldclim/worldclim.htm>.
- Hijmans, R.J., Guarino, L., Mathur, P. and Jarvis, A. (2004b) *DIVA-GIS: A Geographic Information System for the Management and Analysis of Genetic Resources Data*, <http://www.diva-gis.org/>.

- Hill, M.J. (2000) Applications for Spatial Data in Grassland Monitoring and Management, in: *Spatial Information for Land Use Management*, Hill, M. J. and Aspinall, R. (eds.), Gordon & Breach Science Publishers, New York, pp. 113-127.
- Hirzel, A. and Guisan, A. (2002) Which is the Optimal Sampling Strategy for Habitat Suitability Modelling, *Ecological Modelling*, Vol. 157, pp. 331-341.
- Hirzel, A., Hausser, J. and Perrin, N. (2001) *Biomapper 1.0*, <http://www.unil.ch/biomapper/>.
- Holdridge, L.R. (1967) *Life Zone Ecology*, Tropical Science Center, San Jose, Costa Rica, 206 pp.
- Holmann, F. and Lascano, C.E. (eds.) (2001) *Sistemas de Alimentación con Leguminosas para Intensificar Fincas Lecheras - Un Proyecto Ejecutado por el Consorcio Tropileche*, Tropileche, Cali, Colombia, 109 pp.
- Hooten, M.B., Larsen, D.R. and Wikle, C.K. (2003) Predicting the Spatial Distribution of Ground Flora on Large Domains Using a Hierarchical Bayesian Model, *Landscape Ecology*, Vol. 18, pp. 487-502.
- Horne, P. and Stür, W. (1999) *Developing Forage Technologies with Smallholder Farmers - How to Select the Best Varieties to Offer Farmers in Southeast Asia*, ACIAR Monograph No. 62, ACIAR, CIAT, 80 pp.
- Huettmann, F. and Diamond, A.W. (2001) Seabird Colony Locations and Environmental Determination of Seabird Distribution: a Spatially Explicit Breeding Seabird Model for the Northwest Atlantic, *Ecological Modelling*, Vol. 141, pp. 261-298.
- Humphreys, L.R. (1994) *Tropical Forages: Their Role in Sustainable Agriculture*, Longman Scientific and Technical, Essex, London. 414 pp.
- Iverson, L.R., Prasad, A. and Schwartz, M.W. (1999) Modeling Potential Future Individual Tree-species Distributions in the Eastern United States under a Climate Change Scenario: a Case Study with *Pinus virginiana*, *Ecological Modelling*, Vol. 115, No. 1, pp. 77-93.
- Jennings, D. and Wattam, S. (1994) *Decision Making: An Integrated Approach*, Pitman, London. 319 pp.
- Jensen, F.V. (1996) *An Introduction to Bayesian Networks*, UCL Press, London. 178 pp.
- Jones, P.G. (2001) *Interpolated Climate Grids for Central America at 30 Arc Seconds. [Machine Readable Dataset]*, CIAT, Cali, Colombia.
- Jones, P.G. and Gladkov, A. (1999) *FloraMap: A Computer Tool for Predicting the Distribution of Plants and Other Organisms in the Wild*, CIAT, Cali.

- Jones, P.G. and Thornton, P.K. (2003) The Potential Impacts of Climate Change on Maize Production in Africa and Latin America in 2055, *Global Environmental Change*, Vol. 13, pp. 51-59.
- Jongeneel, C.J.B. and Koppelaar, H. (1999) Gödel Pro and Contra AI: Dismissal of the Case, *Engineering Applications of Artificial Intelligence*, Vol. 12, pp. 655-659.
- Jungermann, H. (1983) The Two Camps on Rationality, in: *Decision Making Under Uncertainty*, Scholz, R. W. (ed.), Elsevier, Amsterdam, pp. 63-86.
- Katz, S.S. (1991) Emulating the Prospector Expert System with a Raster GIS, *Computers and Geoscience*, Vol. 17, No. 7, pp. 1033-1050.
- Kemmis, S. and McTaggart, R. (2000) Participatory Action Research, in: *Handbook of Qualitative Research*, Denzin, N. and Lincoln, Y. (eds.), Sage Publications, Thousand Oaks, CA, pp. 567-605.
- Kingwell, R.S. (1994) Risk Attitude and Dryland Farm Management, *Agricultural Systems*, Vol. 45, No. 2, pp. 191-202.
- Kobler, A. and Adamic, M. (2000) Identifying Brown Bear Habitat by a Combined GIS and Machine Learning Method, *Ecological Modelling*, Vol. 135, No. 2-3, pp. 291-300.
- Köppen, W. (1923) *Die Klimate der Erde*, de Gruyters, Berlin, Leipzig. 388 pp.
- Kriticos, D.J. and Randall, R.P. (2001) A Comparison of Systems to Analyze Potential Weed Distributions, in: *Weed Risk Assessment*, Groves, R. H., Panetta, F. D. and Virtue, J. G. (eds.), CSIRO Publishing, Melbourne.
- Lascano, C.E. and Spain, J.M. (1992) Methodological Challenges in Pasture Research, in: *Pastures for the Tropical Lowlands: CIAT's Contribution*, CIAT (ed.), CIAT, Cali, Colombia, pp.29-42.
- Laurance, W.F. (1997) A Distributional Survey and Habitat Model for the Endangered Northern Bettong *Bettongia tropica* in Tropical Queensland, *Biological Conservation*, Vol. 82, No. 1, pp. 47-60.
- Lawrence, R.L. and Wright, A. (2001) Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis, *Photogrammetric Engineering and Remote Sensing*, Vol. 67, No. 10, pp. 1137-1142.
- Leemans, R. (1990) *Global Data Sets Collected and Compiled by the Biosphere Project*, IIASA, Laxenburg, Austria.
- Lehmann, A. (1998) GIS Modeling of Submerged Macrophyte Distribution Using Generalized Additive Models, *Plant Ecology*, Vol. 139, pp. 113-124.

- Lehmann, A., Overton, J. and Leathwick, J.R. (2002) GRASP: Generalized Regression Analysis and Spatial Prediction, *Ecological Modelling*, Vol. 157, pp. 189-207.
- Li, W., Wang, Z., Ma, Z. and Tang, H. (1997) A Regression Model for the Spatial Distribution of Red-Crown Crane in Yancheng Biosphere Reserve, China, *Ecological Modelling*, Vol. 103, No. 2-3, pp. 115-121.
- Linacre, E.T. (1977) A Simple Formula for Estimating Evaporation Rates in Various Climates Using Temperature Data Alone, *Agricultural Meteorology*, Vol. 18, pp. 409-424.
- Lindner, R.K. (1987) Adoption and Diffusion of Technology: an Overview, in: *Technological Change in Postharvest Handling and Transportation of Grains in the Humid Tropics*, Champ, B. R., Highley, E. and Remenyi, J. V. (eds.), Australian Centre for International Agricultural Research, ACIAR Proceedings No. 19, Canberra, pp. 144-151.
- Linstone, H.A. and Turoff, M. (eds.) (1975) *The Delphi Method - Techniques and Applications*, Addison-Wesley Publishing, Reading, Massachusetts. 608 pp.
- Lipton, M. and Longhurst, R. (1989) *New Seeds and Poor People*, Unwin Hyman, London. 473 pp.
- Lobo, M. and Acuña, V. (2001) Productividad Forrajera de *Cratylia argentea* cv. veraniega en el Trópico sub Húmedo de Costa Rica, *Proceedings of XLVII Reunión - Programa Cooperativo Centroamericano para el Mejoramiento de Cultivos y Animales*, Araya, R. and Cháves, N. (eds.), pp. 83-84.
- Lobo, M.V. and Solano, J.A. (eds.) (1997) *Especies Forrajeras Liberadas en Costa Rica*, Progas-MAG-BID, San José. 69 pp.
- Mac Nally, R., Fleishman, E., Fay, J.P. and Murphy, D.D. (2003) Modelling Butterfly Species Richness using Mesoscale Environmental Variables: Model Construction and Validation for Mountain Ranges in the Great Basin of Western North America, *Biological Conservation*, Vol. 110, pp. 21-31.
- Mackinson, S. (2000) An Adaptive Fuzzy Expert System for Predicting Structure, Dynamics and Distribution of Herring Shoals, *Ecological Modelling*, Vol. 126, No. 2-3, pp. 155-178.
- Mackinson, S. (2001) Integrating Local and Scientific Knowledge: An Example in Fisheries Science, *Environmental Management*, Vol. 27, No. 4, pp. 533-545.
- MacMillan, R.A., Pettapiece, W.W., Nolan, S.C. and Goddard, T.W. (2000) A Generic Procedure for Automatically Segmenting Landforms into Landform Elements Using DEMs, Heuristic Rules and Fuzzy Logic, *Fuzzy Sets and Systems*, Vol. 113, No. 1, pp. 81-109.
- MAGFOR, INEC and CIAT (2001) *Atlas Rural de Nicaragua* (CD-ROM)

- Manel, S., Dias, J.M., Buckton, S.T. and Ormerod, S.J. (1999) Alternative Methods for Predicting Species Distribution: an Illustration with Himalayan River Birds, *Journal of Applied Ecology*, Vol. 36, pp. 734-747.
- Manel, S., Dias, J.-M. and Ormerod, S. (1999) Comparing Discriminant Analysis, Neural Networks and Logistic Regression for Predicting Species Distributions: a Case Study with a Himalayan River Bird, *Ecological Modelling*, Vol. 120, No. 2-3, pp. 337-347.
- Manel, S., Williams, H.C. and Ormerod, S.J. (2001) Evaluation Presence-Absence Models in Ecology: The Need to Account for Prevalence, *Journal of Applied Ecology*, Vol. 38, pp. 921-931.
- Manshard, W. (1968) *Tropical Agriculture - A Geographical Introduction and Appraisal*, Longman, London. 226 pp.
- Marra, M., Pannell, D.J. and Abadi Ghadim, A. (2003) The Economics of Risk, Uncertainty and Learning in the Adoption of New Agricultural Technologies: Where are we on the Learning Curve?, *Agricultural Systems*, Vol. 75, No. 2-3, pp. 215-234.
- Mas, J.F., Puig, H., Palacio, J.L. and Sosa-Lopez, A. (2003) Modelling Deforestation using GIS and Artificial Neural Networks, *Environmental Modelling & Software*, Vol. 19, No. 5, pp. 461-471.
- McBratney, A.B. and Odeh, I.O.A. (1997) Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions, *Geoderma*, Vol. 77, No. 2-4, pp. 85-113.
- McCown, R.L., Hochman, Z. and Carberry, P.S. (2002) Probing the Enigma of the Decision Support System for Farmers: Learning from Experience and from Theory, *Agricultural Systems*, Vol. 74, pp. 1-10.
- McKenney, D.W. and Pedlar, J.H. (2003) Spatial Models of Site Index Based on Climate and Soil Properties for Two Boreal Tree Species in Ontario, Canada, *Forest Ecology and Management*, Vol. 175, pp. 497-507.
- Messing, B. (1997) Combining Knowledge with Many-Valued Logics, *Data & Knowledge Engineering*, Vol. 23, pp. 297-315.
- Metternicht, G. (2001) Assessing Temporal and Spatial Changes of Salinity Using Fuzzy Logic, Remote Sensing and GIS. Foundations of an Expert System, *Ecological Modelling*, Vol. 144, No. 2-3, pp. 163-179.
- Mugglin, A.S., Carlin, B.P., Zhu, L. and Conlon, E. (1999) Bayesian Areal Interpolation, Estimation and Smoothing: an Inferential Approach for Geographic Information Systems, *Environment and Planning A*, Vol. 31, pp. 1337-1352.

- Mummery, D. and Battaglia, M. (2001) Applying PROMOD Spatially Across Tasmania with Sensitivity Analysis to Screen for Prospective Eucalyptus globulus Plantation Sites, *Forest Ecology and Management*, Vol. 140, pp. 51-63.
- Natsios, A.S. (2001) *Statement by USAID Administrator Andrew S. Natsios*, http://www.usaid.gov/press/spe_test/speeches/2001/sp010514.html.
- Neill, S.P. and Lee, D.R. (2001) Explaining the Adoption and Disadoption of Sustainable Agriculture: The Case of Cover Crops in Northern Honduras, *Economic Development and Cultural Change*, Vol. 49, No. 4, pp. 793-820.
- NGA (2004) *GEOnet Names Server (GNS)*, National Geospatial-Intelligence Agency.
- NOAA (1999) *GLOBE Global Land One Kilometer Base Elevation*, <http://www.ngdc.noaa.gov/seg/topo/globe.shtml>.
- Norman, A.L. and Shimer, D.W. (1994) Risk, Uncertainty, and Complexity, *Journal of Economic Dynamics and Control*, Vol. 18, No. 1, pp. 231-249.
- NSW Agriculture (2001) *Selecting Pastures for your District*, <http://www.agric.nsw.gov.au/reader/1484>.
- Openshaw, S. (1996) Fuzzy Logic as a New Scientific Paradigm for Doing Geography, *Environment and Planning A*, Vol. 28, pp. 761-768.
- Pannell, D.J., Malcolm, B. and Kingwell, R.S. (2000) Are we Risking Too Much? Perspectives on Risk in Farm Modelling, *Agricultural Economics*, Vol. 23, No. 1, pp. 69-78.
- Partridge, I. (2003) *Better Pastures for the Tropics and Subtropics*, <http://www.tropicalgrasslands.asn.au/pastures/default.htm>.
- Passioura, J.B. (1996) Simulation Models: Science, Snake Oil, Education, or Engineering?, *Agronomy Journal*, Vol. 88, pp. 690-694.
- Pearce, J. and Ferrier, S. (2001) The Practical Value of Modelling Relative Abundance of Species for Regional Conservation Planning: A Case Study, *Biological Conservation*, Vol. 98, No. 1, pp. 33-43.
- Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufman, San Mateo, CA. 552 pp.
- Pearl, J. (1990) Bayesian Decision Methods, in: *Readings in Uncertain Reasoning*, Shafer, G. and Pearl, J. (eds.), Morgan Kaufmann, San Mateo, CA, pp. 345-352.

- Pearson, R.G., Dawson, T.P., Berry, P.M. and Harrison, P.A. (2002) SPECIES: A Spatial Evaluation of Climate Impact on the Envelope of Species, *Ecological Modelling*, Vol. 154, No. 3, pp. 289-300.
- Peters, M. (2004) *Personal Communication*, CIAT, Forages Project, Cali, Colombia.
- Peters, M., Franco, L.H., Schmidt, A. and Hincapie, B. (2003a) *Especies Forrajeras Multipropósito: Opciones para Productores de Centroamérica*, CIAT, Cali, Colombia.
- Peters, M., Horne, P., Schmidt, A., Holmann, F., Kerridge, P.C., Tarawali, S.A., Schultze-Kraft, R., Lascano, C.E., Argel, P., Stür, W., Fujisaka, S., Müller-Sämann, K. and Wortmann, C. (2001) *The Role of Forages in Reducing Poverty and Degradation of Natural Resources in Tropical Production Systems*, Agricultural Research and Extension Network, Vol. Network Paper No. 117.
- Peters, M., Lascano, C.E., Roothaert, R. and de Haan, N.C. (2003b) Linking Research on Forage Germplasm to Farmers: the Pathway to Increased Adoption--a CIAT, ILRI and IITA Perspective, *Field Crops Research*, Vol. 84, No. 1-2, pp. 179-188.
- Pijanowski, B.C., Brown, D.G., Shellito, B.A. and Manik, G.A. (2002) Using Neural Networks and GIS to Forecast Land Use Changes: a Land Transformation Model, Computers, *Environment and Urban Systems*, Vol. 26, pp. 553-575.
- Pontius, R.G.J., Cornell, J.D. and Hall, C.A.S. (2001) Modeling the Spatial Pattern of Land-Use Change with GEOMOD2: Application and Validation for Costa Rica, Agriculture, *Ecosystems and Environment*, Vol. 85, No. 1-3, pp. 191-203.
- Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge. 1020 pp.
- PROSEA (2001) *Plant Resources of South East Asia*, <http://www.prosea.nl/>.
- Rossiter, D.G. (1996) *ALES*, <http://wwwscas.cit.cornell.edu/landeval/ales/alesback.htm>.
- Rowe, W.D. (1994) Understanding Uncertainty, *Risk Analysis*, Vol. 14, No. 5, pp. 743-750.
- Sandoval, B., Lobo, M. and Hidalgo, C. (2001) Evaluación del Efecto de la Asociación de *Brachiaria decumbens* cv pasto peludo con *Leucaena leucocephala* Cultivada en Franjas sobre la Producción de Leche en Vacas de Doble Propósito, *Proceedings of XLVII Reunión - Programa Cooperativo Centroamericano para el Mejoramiento de Cultivos y Animales*, Araya, R. and Chaves, A. (eds.), PCCMCA, San José, Costa Rica, p. 82.

- Sasikala, K.R. and Petrou, M. (2001) Generalised Fuzzy Aggregation in Estimating the Risks of Desertification of a Burned Forest, *Fuzzy Sets and Systems*, Vol. 118, No. 1, pp. 121-137.
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London. 430 pp.
- Schmidt, A. (2001) Genotype x Environment Interactions in *Desmodium ovalifolium* Wall., *PhD Thesis*, Institute for Plant Production and Agroecology in the Tropics and Subtropics, Hohenheim University, Hohenheim.
- Scholz, R.W. (1983) Introduction to Decision Making Under Uncertainty: Biases, Fallacies, and the Development of Decision Making, in: *Decision Making Under Uncertainty*, Scholz, R. W. (ed.), Elsevier, Amsterdam, pp. 3-18.
- Schultze-Kraft, R. and Peters, M. (1997) Tropical Legumes in Agricultural Production and Resource Management: An Overview. Presented at the Tropentag JLU Giessen 22-23 May 1997, *Giessener Beiträge zur Entwicklungsforschung*, Vol. 24, pp. 1-17.
- Seidel, M., Breslin, C., Christley, R.M., Gettinby, G., Reid, S.W.J. and Revie, C.W. (2003) Comparing Diagnoses from Expert Systems and Human Experts, *Agricultural Systems*, Vol. 76, pp. 527-538.
- Seoane, J., Bustamante, J. and Diaz-Delgado, R. (2003) Competing Roles for Landscape, Vegetation, Topography and Climate in Predictive Models of Bird Distribution, *Ecological Modelling*, Vol. 171, No. 3, pp. 209-222.
- Shafer, G. and Pearl, J. (eds.) (1990) *Readings in Uncertain Reasoning*, Morgan Kaufmann, San Mateo, CA. 768 pp.
- Shively, G.R. (1997) Consumption Risk, Farm Characteristics, and Soil Conservation Adoption Among Low-Income Farmers in the Philippines, *Agricultural Economics*, Vol. 17, pp. 165-177.
- Skidmore, A.K. and Gauld, A. (1996) Classification of Kangaroo Habitat Distribution Using Three GIS Models, *International Journal of Geographical Information Systems*, Vol. 10, No. 4, pp. 441-454.
- Skidmore, A.K., Turner, B.J., Brinkhof, W. and Knowles, E. (1997) Performance of a Neural Network: Mapping Forests Using GIS and Remotely Sensed Data, *Photogrammetric Engineering and Remote Sensing*, Vol. 63, No. 5, pp. 501-514.
- SOFI (2000) *Food Insecurity and Vulnerability Information and Mapping Systems*, www.fivims.net.
- Staal, S.J. (2003) *Market Opportunities for Milk Producer Organisations*, <http://www.cgiar.org/ilri/news/milkmarket.cfm>.

- Staal, S.J., Baltenweck, I., Waithaka, M.M., deWolff, T. and Njoroge, L. (2002) Location and Uptake: Integrated Household and GIS Analysis of Technology Adoption and Land Use, With Application to Smallholder Dairy Farms in Kenya, *Agricultural Economics*, Vol. 1682, pp. 1-21.
- Stassopoulou, A., Petrou, M. and Kittler, J. (1998) Application of a Bayesian Network in a GIS Based Decision Making System, *International Journal of Geographical Information Science*, Vol. 12, No. 1, pp. 23-45.
- Stephens, W. and Middleton, T. (2002) Why has the Uptake of Decision Support Systems been so Poor?, in: *Crop Soil Simulation Models: Applications in Developing Countries*, Matthews, R. and Stephens, W. (eds.), CAB International, Cranfield, pp.129-147.
- Stockwell, D.R.B. and Peterson, A.T. (2002) Effects of Sample Size on Accuracy of Species Distribution Models, *Ecological Modelling*, Vol. 148, pp. 1-13.
- Stür, W. and Horne, P. (2001) *Developing Forage Technologies with Smallholder Farmers - How to Grow, Manage and Use Forages*, ACIAR Monograph No. 88, ACIAR, CIAT, 96 pp.
- Stür, W., Horne, P., Jr, G.F.A., Phengsavanh, P. and Kerridge, P.C. (2002) Forage Options for Smallholder Crop-Animal Systems in Southeast Asia: Working with Farmers to Find Solutions, *Agricultural Systems*, Vol. 71, No. 1-2, pp. 75-98.
- Stür, W.W., Ibrahim, T., Tuhulele, M., Binh, L.H., Gabunada, F., Ibrahim, Nakamane, G., Phimphachanhvongsod, V., Guodao, L. and Horne, P.M. (1999) Adaptation of Forages to Climate, Soils and Use in Smallholder Farming Systems in Southeast Asia, in: *Working with Farmers: the Key to Adoption of Forage Technologies*, Stür, W. W., Horne, P. M., Hacker, J. B. and Kerridge, P. C. (eds.), Australian Centre for International Agricultural Research, Canberra, pp. 112-119.
- Sui, D.Z. (1992) A Fuzzy GIS Modeling Approach for Urban Land Evaluation, Computers, *Environment and Urban Systems*, Vol. 16, pp. 101-115.
- Summerell, G.K., Dowling, T.I., Richardson, D.P., Walker, J. and Lees., B. (2000) Modelling Current Parna Distribution in a Local Area, *Australian Journal of Soil Research*, Vol. 38, No. 4, pp. 867-873.
- Thomas, D. and Sumberg, J.E. (1995) A Review of the Evaluation and Use of Tropical Forage Legumes in Sub-Saharan Africa, *Agriculture, Ecosystems and Environment*, Vol. 54, pp. 151-163.
- Thomas, V.L., Aiken, R.M. and Waltman, W. (1999) *Agri-FACTS: Agricultural Farm Analysis and Comparison Tool*, Proceedings of 19th Annual ESRI International User Group Conference, ESRI, San Diego, California, July 26-30.

- Thornton, P.K., Kruska, R.L., Henniger, N., Kristjanson, P.M., Reid, R.S., Atieno, F., Odero, A.N. and Ndegwa, T. (2002) *Mapping Poverty and Livestock in the Developing World*, ILRI, Nairobi, Kenya. 124 pp.
- Tomlin, C.D. (1990) *Geographic Information Systems and Cartographic Modeling*, Prentice-Hall, New Jersey. 249 pp.
- Tversky, A. and Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases, *Science*, Vol. September 1974.
- United Nations (2004) *United Nations Statistical Division - Millennium Indicators*, http://unstats.un.org/unsd/mi/mi_series_results.asp?rowID=607&fID=r5&cglD=419.
- University of Wales (2003) *Lexsys KBS*, <ftp://ftp.bangor.ac.uk/pub/departments/af/LEXSYS/>.
- Upton, M. (1996) *The Economics of Tropical Farming Systems*, Cambridge University Press, Cambridge. 374 pp.
- Upton, M. (2004) *The Role of Livestock in Economic Development and Poverty Reduction*, PPLPI Working Paper No. 10, FAO Pro-Poor Livestock Policy Initiative, Rome, Italy, 57 pp.
- USDA (1993) *Soil Survey Manual*, USDA Handbook 18, United States Department of Agriculture, Washington, D.C..
- USDA (2002) *Plants Database*, <http://plants.usda.gov/>.
- USGS (2003) *SRTM (Shuttle Radar Topography Mission)*, <ftp://edcsgs9.cr.usgs.gov/pub/data/srtm/>.
- van Asselt, M.B.A. (2000) *Perspectives on Uncertainty and Risk - The PRIMA Approach to Decision Support*, Kluwer Academic Publishers, Dordrecht. 434 pp.
- Veitch, S.M. and Bowyer, J.K. (1996) ASSESS: A GIS-Based System for Selecting Suitable Sites, in: *Raster Imagery in Geographic Information Systems*, Morain, S. and Lopez Baros, S. (eds.), Onword Press, Santa Fe, pp. 182-191.
- Verheye, W. and Ameryckx, J. (1984) Mineral Fractions and Classification of Soil Texture, *Pedologie*, Vol. 2, pp. 215-225.
- Walker, D.H. (2002) Decision Support, Learning and Rural Resource Management, *Agricultural Systems*, Vol. 72, pp. 113-127.
- Walker, P.A. and Cocks, K.D. (1991) HABITAT: a Procedure for Modelling a Disjoint Environmental Envelope for a Plant or Animal Species, *Global Ecology and Biogeography Letters*, Vol. 1, pp. 108-118.

- Weischet, W. and Caviedes, C.N. (1993) *The Persisting Ecological Constraints of Tropical Agriculture*, Longman Group UK, Harlow. 319 pp.
- Welsh, A.H., Cunningham, R.B., Donnelly, C.F. and Lindenmayer, D.B. (1996) Modelling the Abundance of Rare Species: Statistical Models for Counts with Extra Zeros, *Ecological Modelling*, Vol. 88, pp. 297-308.
- White, J.W., Corbett, J.D. and Dobermann, A. (2002) Insufficient Geographic Characterization and Analysis in the Planning, Execution and Dissemination of Agronomic Research?, *Field Crops Research*, Vol. 4067, pp. 1-10.
- Wilbanks, T. (1986) Communication Between Hard and Soft Sciences, in: *Interdisciplinary Analysis and Research*, Chubin, D. E., Porter, A. L., Rossini, F. A. and Connolly, T. (eds.), Lomond Publications, Maryland, pp. 131-140.
- Winograd, M., Farrow, A., Aguilar, M. and Kok, K. (2000) *Indicadores de Sustentabilidad Rural: Una Visión para América Central*, World Bank, CIAT, PNUMA, Canada, (CD-ROM).
- World Bank, CCAD and WICE (2003) *Regional Ecosystems Map*, www.worldbank.org/ca-env.
- WRI (2001) *World Resources Institute*, <http://www.wri.org/wr2000>.
- Yamada, K., Elith, J., McCarthy, M. and Zerger, A. (2003) Eliciting and Integrating Expert Knowledge for Wildlife Habitat Modelling, *Ecological Modelling*, Vol. 165, pp. 251-264.
- Zadeh, L.H. (1965) Fuzzy Sets, *Information Control*, Vol. 8, pp. 338-353.
- Zhang, J. and Goodchild, M. (2002) *Uncertainty in Geographical Information*, Taylor and Francis, London. 266 pp.
- Zhu, L., Healey, R.G. and Aspinall, R. (1998) A Knowledge-Based Systems Approach to Design of Spatial Decision Support Systems for Environmental Management, *Environmental Management*, Vol. 22, No. 1, pp. 35-48.
- Zhu, X., Aspinall, R. and Healey, R.G. (1996) ILUDSS: A Knowledge-Based Spatial Decision Support System for Strategic Land-Use Planning, *Computers and Electronics in Agriculture*, Vol. 15, pp. 279-301.

APPENDIX A
FORAGE DATABASE ANALYSIS

LOCATION CHARACTERISTICS ANALYSIS

Coefficient of determination R^2 and number of observations n for location characteristics in the CIAT Forages database.

Key:

Elev:	Elevation (masl)
Tave:	Mean average temperature (averaged over 12 months) (°C)
Tmin:	Minimum annual temperature (min. monthly minimum) (°C)
Tmax:	Maximum annual temperature (max. monthly maximum) (°C)
SR:	Solar radiation (averaged over 12 months) (Langleys/day)
RH:	Relative humidity (averaged over 12 months) (%)
HS:	Hours of sunshine (summed over 12 months)
WS:	Wind speed (averaged over 12 months) (km/hour)
Rain:	Mean annual rainfall (summed over 12 months) (mm)
S%20 (S%40):	Percentage of sand 0-20cm (20-40cm)
L%20 (L%40):	Percentage of silt 0-20cm (20-40cm)
C%20 (C%40):	Percentage of clay 0-20cm (20-40cm)
AD20 (AD40):	Apparent density 0-20cm (20-40cm) (g/cc)
FC20 (FC40):	Field capacity 0-20cm (20-40cm) (% humidity)
pH20 (pH40):	pH in H ₂ O 1:1 0-20cm (20-40cm)
OM20 (OM40):	Organic matter 0-20cm (20-40cm) (%)
P20 (P40):	Phosphorus 0-20cm (20-40cm) (ppm)
Ca20 (Ca40):	Calcium 0-20cm (20-40cm)
Mg20 (Mg40):	Magnesium 0-20cm (20-40cm)
K20 (K40):	Potassium 0-20cm (20-40cm)
Na20 (Na40):	Sodium 0-20cm (20-40cm)
Al20 (Al40):	Aluminium 0-20cm (20-40cm)
Al%20 (Al%40):	Aluminium saturation 0-20cm (20-40cm) (%)

R^2 n	Elev	Tave	Tmin	Tmax	SR	RH	HS	WS	Rain	DryM
Elev	1									
Tave	0.50 236	1								
Tmin	0.09 165	0.43 165	1							
Tmax	0.49 184	0.55 184	0.02 165	1						
SR	0.01 120	0.05 120	0.04 83	0.04 84	1					
RH	0.00 175	0.00 175	0.08 131	0.04 132	0.01 105	1				
HS	0.00 101	0.00 101	0.02 89	0.01 90	0.02 66	0.21 98	1			
WS	0.01 93	0.01 93	0.01 77	0.04 77	0.00 57	0.00 90	0.04 73	1		
Rain	0.01 263	0.01 242	0.02 166	0.05 167	0.04 121	0.16 175	0.10 101	0.01 93	1	
DryM	0.04 299	0.02 261	0.01 181	0.08 182	0.03 136	0.07 192	0.00 114	0.01 107	0.43 283	1
S%20	0.00 223	0.00 200	0.02 140	0.00 140	0.01 100	0.01 149	0.01 92	0.00 88	0.03 216	0.00 233

R^2 n	Elev	Tave	Tmin	Tmax	SR	RH	HS	WS	Rain	DryM
L%20	0.00 223	0.00 200	0.01 140	0.00 140	0.00 100	0.05 149	0.00 92	0.00 88	0.02 216	0.00 233
C%20	0.00 223	0.00 200	0.00 140	0.00 140	0.01 100	0.00 149	0.00 92	0.00 88	0.01 216	0.00 233
AD20	0.23 55	0.00 55	0.02 46	0.23 46	0.11 36	0.06 46	0.06 37	0.08 35	0.02 55	0.10 59
FC20	0.20 38	0.12 37	0.00 32	0.07 32	0.18 24	0.04 31	0.10 29	0.00 24	0.14 37	0.12 41
pH20	0.01 257	0.01 230	0.00 161	0.01 162	0.00 117	0.01 171	0.00 98	0.02 94	0.05 249	0.06 267
OM20	0.14 225	0.07 205	0.00 146	0.08 146	0.01 101	0.04 152	0.12 90	0.00 84	0.12 220	0.05 234
P20	0.00 249	0.01 221	0.00 157	0.01 158	0.00 111	0.01 166	0.01 95	0.00 91	0.01 240	0.00 256
Ca20	0.02 231	0.01 206	0.01 148	0.02 148	0.00 103	0.00 154	0.01 85	0.00 85	0.04 225	0.06 238
Mg20	0.01 231	0.01 206	0.01 148	0.02 148	0.00 103	0.00 153	0.00 79	0.00 85	0.05 225	0.04 238
K20	0.00 246	0.00 220	0.00 156	0.00 156	0.00 113	0.01 165	0.02 96	0.00 92	0.01 239	0.00 255
Na20	0.01 74	0.00 72	0.01 50	0.02 50	0.01 35	0.01 58	0.02 31	0.05 34	0.01 74	0.01 76
Al20	0.00 196	0.01 172	0.02 129	0.04 130	0.01 100	0.08 135	0.04 81	0.01 76	0.09 189	0.05 195
Al%20	0.00 176	0.05 156	0.07 121	0.04 121	0.03 89	0.01 128	0.00 78	0.02 75	0.07 172	0.02 176
S%40	0.01 150	0.00 134	0.02 103	0.00 104	0.00 64	0.01 103	0.02 75	0.02 73	0.06 146	0.03 150
L%40	0.00 150	0.00 134	0.01 103	0.00 104	0.00 64	0.01 103	0.01 75	0.01 73	0.01 146	0.01 150
C%40	0.02 150	0.01 134	0.00 103	0.02 104	0.01 64	0.00 103	0.00 75	0.05 73	0.05 146	0.08 150
AD40	0.42 48	0.22 48	0.01 42	0.35 42	0.19 30	0.02 39	0.07 34	0.07 32	0.10 48	0.25 48
FC40	0.12 30	0.04 30	0.01 27	0.04 27	0.01 18	0.01 24	0.07 23	0.10 19	0.12 30	0.05 30
pH40	0.01 168	0.00 151	0.04 114	0.05 114	0.00 71	0.00 116	0.00 79	0.10 77	0.05 164	0.06 168
OM40	0.29 148	0.15 134	0.01 101	0.12 101	0.02 62	0.00 103	0.04 71	0.00 68	0.04 145	0.03 148
P40	0.00 161	0.01 144	0.00 108	0.00 108	0.00 70	0.01 111	0.01 74	0.05 72	0.00 157	0.00 161
Ca40	0.01 153	0.01 135	0.00 102	0.03 102	0.00 61	0.00 104	0.01 69	0.03 67	0.04 149	0.07 153
Mg40	0.02 149	0.02 131	0.00 99	0.03 99	0.00 59	0.02 101	0.04 67	0.02 65	0.04 145	0.02 149
K40	0.01 162	0.00 144	0.01 110	0.01 110	0.25 70	0.00 112	0.02 77	0.05 75	0.05 158	0.07 162
Na40	0.01 50	0.01 49	0.03 43	0.05 43	0.00 23	0.05 39	0.02 26	0.01 28	0.02 50	0.02 50
Al40	0.01 121	0.04 104	0.04 84	0.08 84	0.02 58	0.06 87	0.02 60	0.01 55	0.13 117	0.06 121
Al%40	0.00 111	0.07 97	0.17 79	0.02 79	0.00 50	0.01 83	0.00 56	0.07 52	0.11 109	0.04 111

R^2 n	S%20	L%20	C%20	AD20	FC20	pH20	OM20	P20	Ca20	Mg20
S%20	1									
L%20	0.29 223	1								
C%20	0.66 223	0.03 223	1							
AD20	0.03 54	0.00 54	0.03 54	1						
FC20	0.15 36	0.20 36	0.15 36	0.35 34	1					
pH20	0.00 222	0.00 222	0.00 222	0.00 55	0.03 37	1				
OM20	0.04 203	0.04 203	0.00 203	0.29 52	0.50 36	0.01 225	1			
P20	0.00 215	0.00 215	0.00 215	0.00 55	0.00 37	0.09 247	0.01 218	1		
Ca20	0.03 206	0.03 206	0.01 206	0.00 49	0.01 33	0.40 230	0.01 205	0.05 224	1	
Mg20	0.02 207	0.00 207	0.02 207	0.01 50	0.01 34	0.27 230	0.00 206	0.04 223	0.51 228	1
K20	0.00 219	0.00 219	0.00 219	0.00 53	0.01 37	0.01 245	0.00 217	0.02 236	0.00 230	0.03 230
Na20	0.06 68	0.00 68	0.11 68	0.01 29	0.01 18	0.16 73	0.01 70	0.01 67	0.04 72	0.29 74
Al20	0.10 173	0.02 173	0.07 173	0.10 40	0.00 30	0.17 195	0.01 168	0.00 191	0.01 181	0.04 182
Al%20	0.08 156	0.00 156	0.07 156	0.05 40	0.08 31	0.40 175	0.00 153	0.03 172	0.19 167	0.21 164
S%40	0.90 147	0.25 147	0.57 147	0.06 47	0.07 32	0.00 149	0.06 132	0.00 146	0.03 136	0.04 136
L%40	0.33 147	0.61 147	0.01 147	0.00 47	0.07 32	0.01 149	0.02 132	0.02 146	0.13 136	0.01 136
C%40	0.55 147	0.05 147	0.83 147	0.05 47	0.10 32	0.00 149	0.01 132	0.00 146	0.00 136	0.04 136
AD40	0.03 47	0.01 47	0.07 47	0.83 48	0.28 30	0.00 48	0.32 45	0.00 48	0.00 43	0.01 44
FC40	0.44 29	0.03 29	0.09 29	0.25 28	0.55 30	0.02 30	0.33 29	0.07 30	0.07 27	0.09 28
pH40	0.00 154	0.02 154	0.01 154	0.01 47	0.00 32	0.72 167	0.01 149	0.09 164	0.33 150	0.22 150
OM40	0.05 136	0.01 136	0.01 136	0.31 45	0.43 31	0.03 147	0.82 146	0.00 145	0.02 133	0.01 134
P40	0.01 147	0.00 147	0.01 147	0.02 45	0.12 31	0.09 160	0.00 143	0.77 160	0.02 145	0.01 145
Ca40	0.03 141	0.05 141	0.00 141	0.00 41	0.02 28	0.31 152	0.01 136	0.03 151	0.93 150	0.44 149
Mg40	0.02 138	0.00 138	0.04 138	0.00 40	0.00 27	0.10 148	0.00 133	0.01 147	0.27 145	0.65 148
K40	0.00 150	0.01 150	0.00 150	0.03 43	0.02 30	0.28 161	0.01 144	0.02 159	0.23 150	0.16 150
Na40	0.08 45	0.01 45	0.19 45	0.01 26	0.06 16	0.22 49	0.00 47	0.02 48	0.01 48	0.32 49
Al40	0.12 112	0.02 112	0.08 112	0.07 33	0.06 23	0.10 120	0.00 104	0.00 119	0.00 109	0.00 111
Al%40	0.12 101	0.00 101	0.08 101	0.07 31	0.21 22	0.43 110	0.01 94	0.06 110	0.21 103	0.18 100

R^2 n	K20	Na20	Al20	Al%20	S%40	L%40	C%40	AD40	FC40	pH40
K20	1									
Na20	0.63 74	1								
Al20	0.01 192	0.01 46	1							
Al%20	0.12 175	0.05 45	0.41 171	1						
S%40	0.02 146	0.08 44	0.19 108	0.12 99	1					
L%40	0.01 146	0.01 44	0.07 108	0.04 99	0.34 150	1				
C%40	0.01 146	0.17 44	0.11 108	0.07 99	0.64 150	0.00 150	1			
AD40	0.00 46	0.09 26	0.08 32	0.04 32	0.04 46	0.03 46	0.08 46	1		
FC40	0.02 30	0.04 16	0.00 21	0.05 22	0.40 29	0.10 29	0.18 29	0.22 28	1	
pH40	0.22 161	0.22 50	0.10 119	0.34 109	0.00 150	0.03 150	0.01 150	0.03 46	0.01 29	1
OM40	0.00 142	0.00 48	0.02 102	0.00 92	0.03 131	0.02 131	0.00 131	0.34 44	0.28 29	0.01 147
P40	0.00 154	0.01 47	0.01 116	0.05 107	0.01 143	0.01 143	0.00 143	0.05 44	0.26 28	0.08 158
Ca40	0.10 152	0.01 49	0.00 110	0.16 102	0.03 137	0.14 137	0.00 137	0.00 41	0.07 26	0.29 152
Mg40	0.01 148	0.01 49	0.02 109	0.13 98	0.04 134	0.00 134	0.04 134	0.01 40	0.08 25	0.08 148
K40	0.79 161	0.63 50	0.00 116	0.16 107	0.02 143	0.03 143	0.00 143	0.02 42	0.02 27	0.25 158
Na40	0.81 49	0.94 49	0.00 27	0.25 26	0.13 45	0.00 45	0.19 45	0.03 26	0.19 26	0.24 50
Al40	0.00 117	0.00 29	0.74 118	0.17 101	0.14 110	0.03 110	0.10 110	0.09 32	0.01 20	0.11 120
Al%40	0.18 108	0.12 28	0.40 100	0.82 106	0.11 99	0.01 99	0.08 99	0.07 30	0.05 20	0.44 109

R^2 n	OM40	P40	Ca40	Mg40	K40	Na40	Al40	Al%40	
OM40	1								
P40	0.02 141	1							
Ca40	0.03 135	0.02 146	1						
Mg40	0.00 132	0.01 142	0.32 148	1					
K40	0.01 139	0.01 151	0.22 150	0.02 147	1				
Na40	0.00 48	0.01 46	0.01 49	0.01 49	0.66 50	1			
Al40	0.00 103	0.01 116	0.01 112	0.00 111	0.01 115	0.11 29	1		
Al%40	0.01 92	0.08 107	0.20 103	0.12 99	0.19 106	0.04 28	0.30 102	1	

ADAPTATION CHARACTERISTICS ANALYSIS

Coefficient of determination R^2 and number of observations n for adaptation characteristics in the CIAT Forages Adaptation database.

Key:

Adap ini:	Adaptation (initial evaluation)
Cov ini:	Percent cover (initial evaluation)
NPl ini:	Number of plants (initial evaluation)
DM ini:	Dry matter weight (initial evaluation)
Adap fin:	Adaptation (final evaluation)
Cov fin:	Percent cover (final evaluation)
NPl fin:	Number of plants (final evaluation)
DM fin:	Dry matter weight (final evaluation)
Ins:	Resistance to insects
Dis:	Resistance to diseases

R^2 n	Adap ini	Cov ini	NPl ini	DM ini	Adap fin	Cov fin	NPl fin	DM fin	Ins	Dis
Adap ini	1									
Cov ini	0.44 2,045	1								
NPl ini	0.07 845	0.19 792	1							
DM ini	0.03 739	0.14 745	0.19 193	1						
Adap fin	0.30 2,014	0.28 1,523	0.00 405	0.03 739	1					
Cov fin	0.24 1,542	0.57 1,512	0.08 346	0.06 636	0.55 1,746	1				
NPl fin	0.00 300	0.14 239	0.33 277	0.13 143	0.06 305	0.45 260	1			
DM fin	0.02 732	0.07 758	0.08 170	0.59 573	0.06 918	0.08 863	0.22 136	1		
Ins	0.00 3,024	0.00 2,248	0.01 870	0.07 873	0.03 2,224	0.01 1,767	0.00 313	0.21 930	1	
Dis	0.01 3,024	0.03 2,248	0.00 870	0.10 873	0.13 2,224	0.12 1,767	0.01 313	0.23 930	0.15 3,467	1

APPENDIX B
FARMER SURVEYS

Nombre. Juan Geias Lopez

Comunidad. Esquipulas

Yo, Juan Geias Lopez, mayor de edad, afirmo que la información que brindo es de carácter voluntario y de colaboración a la persona que realiza la entrevista

Nº Cabañas

172 ~~grados~~ / 24 ordena 872 un ordeno

3 equinos

Area tierra

54 m² → potrero (31)

estrella - 14 m²

jaragua - 60 m²

brizante + maíz forajero - 10 m²

Verano y rotación

Invierno heno + sorgo + terciopelo regallero

Datos personales

★ → 2 estudiantes (en casa)

Mano de obra: 2 permanente + productor

temporales 525 promedio (en total equivalente

5 permanente)

Juan Geias Lopez

Nombre: Julio Rivas Pastrán

Comunidad: Susutí (2.5 km de Oroquieta N)

Yo, Julio Rivas Pastrán, mayor de edad, afirmo que la información que brindo es de carácter voluntario y de colaboración a la persona que realiza la entrevista.

Nº cabezas (vacuna, equinos, cerdos, gallinas...)

22 vacunas (4 ordeas) cerano + comecastro

2 bestia

Area tierra (potreros, pastos mejorados, cultivos)

32 m²

25 m² potrero — 1 m² estela 1/2 m² ganadería

2 m² arroz — prima 2 potreros (centro)

5 m² frijoles | prima 3 sorgo | pastera (cerano)
comecastro 2 frijoles (cerano) ganado

Datos personales (personas en la casa, mano de obra)

5 personas — 3 estudiantes

1 trabajador + 34 meses 4 diario mano de obra

Fin

Comunidad Nombre

Lepique Tomas Banegas Rosales

Yo Tomas Banegas Rosales mayor de edad, casado N° 0811195000153 concedo autorización para que se haga una descripción general sobre sistema de producción que yo manejo en mis actividades agropecuarias. - Para la cual le doy la siguiente información:

Area Tierra:

- 2 M2 Café (\$80002 ingresos/año anteriormente)
- 2 M2 Maíz (\$40002 ingresos otro consumo)
- 8 M2 Potreros (Bosque + Hiperbarrenia rufa + Pasto Natural)
- 3 Equinos
- 1 Vácuo
- 3 Cerdos
- 35 gallinas (10 huevos diarios)

Personales

Hijos 10

- 1 profesionales
- 3 estudiantes
- 2 Pequeño

Resto Trabajan agricultura familiar

Tomas Banegas B.

Comunidad Nombre
 Luquique Elvis castro

Yo Elvis Castro, mayor de edad soltero
 con identidad N° 1811187500025 afirmo
 que la información que brindo es de
 carácter voluntario y de colaboración a la
 persona que realiza la entrevista:

N° Cabezas vacunas 15
 equinos 6
 gallinas 20 (5 huevos diarios)

Area Tierra

Potreritos = 14 m² con pasto natural y Pin
 (2 lotes de 7 m² c/u).

Primera (Maiz = 5.5 m² (18000 ₡ + Consumo

Postera (Frijol) = 2 m² (Consumo

Café = 1 m² (Ingresos 2000 ₡

Pasto Mejorado = Brachiaria

Camerum

= King grass

= Caña Forrajera

5 m²

Datos = 3 adultos, 1 Profesional
 Personales = 4 1 adolescente estudiante

1 ELVIN Castro.P.

APPENDIX C
FORAGE EXPERT QUESTIONNAIRE

Dear colleague,

I am contacting you in your capacity as a tropical forage expert, and would greatly appreciate it if you could spend a few minutes on the following questionnaire. As you may be aware, I am currently working on my PhD, in conjunction with CIAT (International Center for Tropical Agriculture) in Cali, Colombia, and Curtin University of Technology in Perth, Western Australia. As part of my research, I am developing a software tool CaNaSTA (Crop Niche Selection in Tropical Agriculture). Please read the attached document for an overview of CaNaSTA.

There are two main sections to this survey. Firstly I would like your expert opinion on which crops would be suitable under different conditions. Please note this is different from the expert opinion gathered for SoFT (Selection of Forages for the Tropics). In this case I will describe a farmer's environment, and you are asked to recommend suitable species based on all information. I am not trying to gather comprehensive data, but rather a small amount of case studies with which to validate CaNaSTA.

In the second section I would like to gather your opinion on the potential of CaNaSTA, once it is fully developed and functioning. If you do not have time to complete the whole questionnaire, then I would still appreciate a partially completed questionnaire.

This is an electronic form. Please enter information electronically, save the completed form, and return it to me by email. I would greatly appreciate your cooperation in returning the completed survey to me by **Thursday 18 March 2004**. If you know of another forage expert who may be able to help, please pass on these documents.

Thank you for your time. Please contact me with any questions or additional comments.

Rachel O'Brien
r.obrien@cgiar.org
 10 March 2004

1 GENERAL INFORMATION

Name:

Position:

Company:

Field(s) of expertise:

Countries / regions of expertise:

Years of experience in forages (approx.):

Today's date:

Case 2: Farmer in San Dionisio-Wibuse, Nicaragua. 12°45'3"N, 85°49'8"W. Elevation is 430m, annual rainfall is 900mm, and the dry season is 5 months. Soil pH is moderately acid, soil texture is clay and soil fertility is medium. Farmer would like a forage species for pasture. The farmer is able to accept some risk.

A. Please list up to five forage species you would recommend in this situation:

- 1.
- 2.
- 3.
- 4.
- 5.

Comments:

B. Of the following species, would you recommend any of them in this situation (ignoring intended use)? Please indicate whether each species would be suitable, marginally suitable, or not suitable.

Species	Suitable	Marginally suitable	Not suitable
<i>Arachis pinto</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Brachiaria brizantha</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Centrosema pubescens</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Cratylia argentea</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Stylosanthes guianensis</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Comments:

C. Please indicate your familiarity with location:

☐ Very familiar ☐ Somewhat familiar ☐ Not familiar

Case 3: Farmer in El Corozo, Nicaragua, 12°46'42"N, 85°53'36"W. Elevation is 650m, annual rainfall is 800mm, and dry season is 6 months. Soil pH is moderately acid, soil texture is sandy loam and soil fertility is high. Farmer would like a forage species for cut and carry. The farmer is risk averse.

A. Please list up to five forage species you would recommend in this situation:

- 1.
- 2.
- 3.
- 4.
- 5.

Comments:

B. Of the following species, would you recommend any of them in this situation (ignoring intended use)? Please indicate whether each species would be suitable, marginally suitable, or not suitable.

Species	Suitable	Marginally suitable	Not suitable
<i>Arachis pinto</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Brachiaria brizantha</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Centrosema pubescens</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Cratylia argentea</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Stylosanthes guianensis</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Comments:

C. Please indicate your familiarity with location:

☐ Very familiar

☐ Somewhat familiar

☐ Not familiar

Case 4: Farmer near Flores, Honduras. 15°36'N, 87°15'W. Elevation is 70m, annual rainfall is 2606mm, and there is no dry season. Soil pH is acid, soil texture is loam and soil fertility is high. Farmer would like a species for pasture. The farmer is risk averse.

A. Please list up to five forage species you would recommend in this situation:

- 1.
- 2.
- 3.
- 4.
- 5.

Comments:

B. Of the following species, would you recommend any of them in this situation (ignoring intended use)? Please indicate whether each species would be suitable, marginally suitable, or not suitable.

Species	Suitable	Marginally suitable	Not suitable
<i>Arachis pinto</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Brachiaria brizantha</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Centrosema pubescens</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Cratylia argentea</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Stylosanthes guianensis</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Comments:

C. Please indicate your familiarity with location:

☐ Very familiar

☐ Somewhat familiar

☐ Not familiar

Case 5: Farmer in Esparza, Costa Rica. 9°59'N, 84°40'W. Elevation is 145m, annual rainfall is 2277mm, length of dry season is 5 months. Soil pH is moderately acid, soil texture is sandy loam and soil fertility is very high. Farmer would like a forage species for cut and carry and potentially live barriers. The farmer is able to accept some risk.

A. Please list up to five forage species you would recommend in this situation:

- 1.
- 2.
- 3.
- 4.
- 5.

Comments:

B. Of the following species, would you recommend any of them in this situation (ignoring intended use)? Please indicate whether each species would be suitable, marginally suitable, or not suitable.

Species	Suitable	Marginally suitable	Not suitable
<i>Arachis pinto</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Brachiaria brizantha</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Centrosema pubescens</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Cratylia argentea</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Stylosanthes guianensis</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Comments:

C. Please indicate your familiarity with location:

☐ Very familiar ☐ Somewhat familiar ☐ Not familiar

2. 1 For all cases

Did you refer to any literature in making the above recommendations?

- ☐ Yes. Please specify:
☐ No

3 PART II EXPERT OPINION – CANASTA-FORAGES

Please read the accompanying outline of CaNaSTA. You may have seen presentations or I may have discussed this work with you at some time (note: previously called GEMS). CaNaSTA-Forages is still under development and not yet available for user testing. However I would like to gather some opinions on the potential of CaNaSTA-Forages. Assuming CaNaSTA-Forages is fully developed and functioning as intended, please answer the following questions:

1. How familiar are you with CaNaSTA-Forages?

- ☐ Involved in development ☐ Seen presentations
☐ Discussed development ☐ Not familiar

Comments:

2. CaNaSTA-Forages will be expanded to the entire tropical world. Do you think CaNaSTA-Forages would be useful in your research?

- ☐ Yes ☐ Maybe ☐ No

Comments:

3. Do you think CaNaSTA-Forages will be used by the following people?

- | | |
|---|--|
| <input type="checkbox"/> Scientists at international research centres | <input type="checkbox"/> NGOs |
| <input type="checkbox"/> Scientists at national research centres | <input type="checkbox"/> National government |
| <input type="checkbox"/> Extension agents | <input type="checkbox"/> Farmers |
| <input type="checkbox"/> Others. Please specify: | |

Comments:

4. Do you think there is a need for this type of software?

- | | | |
|------------------------------|-------------------------------------|-----------------------------|
| <input type="checkbox"/> Yes | <input type="checkbox"/> Don't know | <input type="checkbox"/> No |
|------------------------------|-------------------------------------|-----------------------------|

Comments:

5. Do you think a spatial decision support system is an appropriate way to deliver forage information to farmers in the tropics?

- | | | |
|------------------------------|-------------------------------------|-----------------------------|
| <input type="checkbox"/> Yes | <input type="checkbox"/> Don't know | <input type="checkbox"/> No |
|------------------------------|-------------------------------------|-----------------------------|

Comments:

6. Which of the following do you think are necessary delivery modes?

- | | | |
|---|-----------------------------------|---|
| <input type="checkbox"/> CD-ROM | <input type="checkbox"/> Internet | <input type="checkbox"/> Print outs (maps and text) |
| <input type="checkbox"/> Other. Please specify: | | |

Comments:

7. Do you think CaNaSTA could work well with crops other than forages?

- | | |
|---|-----------------------------|
| <input type="checkbox"/> Yes. Give some examples: | <input type="checkbox"/> No |
| <input type="checkbox"/> Don't know | |

Comments:

8. Any other comments?