

SPATIAL ANALYSIS OF INVASIVE ALIEN PLANT DISTRIBUTION PATTERNS
AND PROCESSES USING BAYESIAN NETWORK-BASED DATA MINING
TECHNIQUES

by

WISDOM MDUMISENI DABULIZWE DLAMINI

submitted in accordance with the requirements
for the degree

DOCTOR OF PHILOSOPHY

in the subject

ENVIRONMENTAL SCIENCE

at the

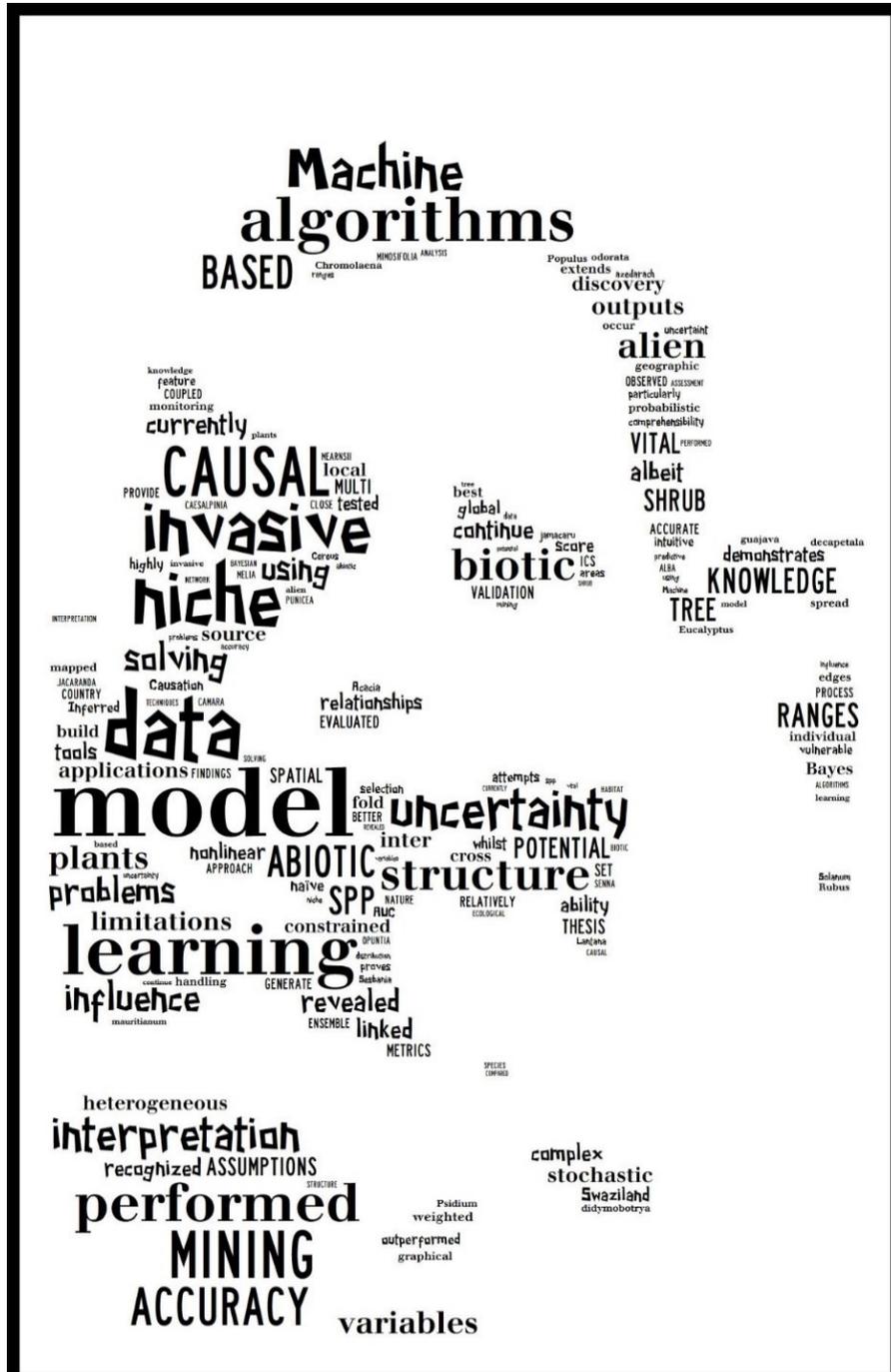
UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROF. WILLEM A.J. NEL
CO-SUPERVISOR: PROF. JIMMY HENDRICK
CO-SUPERVISOR: MR. MAARTEN JORDAAN
CO-SUPERVISOR: MR. JEAN-PIERRÉ LABUSCHAGNE

March 2016

DEDICATION

To the precious memory of my handsome and intelligent son, Andzile Betive Zanokuhle ‘Gadzimvelo’ Dlamini. We did it son. Daddy loves you!!!



DECLARATION

I Wisdom Mdumiseni Dabulizwe Dlamini hereby declare that the dissertation/thesis, which I hereby submit for the degree of Doctor of Philosophy – Environmental Science at the University of South Africa, is my own work and has not previously been submitted by me for a degree at this or any other institution.

I declare that the dissertation /thesis does not contain any written work presented by other persons whether written, pictures, graphs or data or any other information without acknowledging the source.

I declare that, where words from a written source have been used, the words have been paraphrased and referenced and where exact words from a source have been used, the words have been placed inside quotation marks and referenced.

I declare that I have not copied and pasted any information from the Internet, without specifically acknowledging the source and have inserted appropriate references to these sources in the reference section of the dissertation or thesis.

I declare that during my study I adhered to the Research Ethics Policy of the University of South Africa, received ethics approval for the duration of my study prior to the commencement of data gathering, and have not acted outside the approval conditions.

I declare that the content of my dissertation/thesis has been submitted through an electronic plagiarism detection program before the final submission for examination.



.....

Student Signature

30 March 2016

Date

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following individuals for their assistance and support towards the completion of this study:

- The Almighty God for the precious gift of life and granting me the favour, ability, inspiration, and direction to undertake and complete this study.
- My supervisor Professor Willem A.J. Nel and co-supervisors Professor Jimmy Hendrick, Mr. Maarten Jordaan and Mr. Jean-Pierré Labuschagne for their diligent and professional guidance throughout the study.
- My wife, Lencane for the meals and moral support that kept me going through the night at times.
- My boys, Asanda, Alwandze, Andzisiwe, Andzile and Andza, believed in me and I took strength from their motivation and intelligence. I strongly believe that my PhD will one day be an inspiration for theirs and even greater things too.
- The Senior Forestry Officer within the Ministry of Tourism and Environmental Affairs, Mr. Solomon Gamedze for the generous access to the national invasive alien plants mapping data from *Project A340 - Control of Invasive Alien Plant Species* commissioned by the Government of Swaziland through the Ministry of Tourism and Environmental Affairs.
- Private ecologist Kate Braun and botanist Linda Loffler to whom I am indebted for providing access to the data from the *Tree Atlas of Swaziland*.

ABSTRACT

Invasive alien plants have widespread ecological and socioeconomic impacts throughout many parts of the world, including Swaziland where the government declared them a national disaster. Control of these species requires knowledge on the invasion ecology of each species including how they interact with the invaded environment. Species distribution models are vital for providing solutions to such problems including the prediction of their niche and distribution. Various modelling approaches are used for species distribution modelling albeit with limitations resulting from statistical assumptions, implementation and interpretation of outputs.

This study explores the usefulness of Bayesian networks (BNs) due their ability to model stochastic, nonlinear inter-causal relationships and uncertainty. Data-driven BNs were used to explore patterns and processes influencing the spatial distribution of 16 priority invasive alien plants in Swaziland. Various BN structure learning algorithms were applied within the Weka software to build models from a set of 170 variables incorporating climatic, anthropogenic, topo-edaphic and biotic factors. While all the BN models produced accurate predictions of alien plant invasion, the globally scored networks, particularly the hill climbing algorithms, performed relatively well. However, when considering the probabilistic outputs, the constraint-based Inferred Causation algorithm which attempts to generate a causal BN structure, performed relatively better.

The learned BNs reveal that the main pathways of alien plants into new areas are ruderal areas such as road verges and riverbanks whilst humans and human activity are key driving factors and the main dispersal mechanism. However, the distribution of most of the species is constrained by climate particularly tolerance to very low temperatures and precipitation seasonality. Biotic interactions and/or associations among the species are also prevalent. The findings suggest that most of the species will proliferate by extending their range resulting in the whole country being at risk of further invasion.

The ability of BNs to express uncertain, rather complex conditional and probabilistic dependencies and to combine multisource data makes them an attractive technique for species distribution modelling, especially as joint invasive species distribution models (JiSDM). Suggestions for further research are provided including the need for rigorous invasive species monitoring, data stewardship and testing more BN learning algorithms.

Key terms: Bayesian network, data mining, directed acyclic graph, ecology, geographic information system, habitat, invasive alien plant, knowledge discovery, machine learning, species distribution model.

LIST OF ACRONYMS

AUC – area under the ROC curve
AUPRC – area under the precision-recall curve
BAN – Bayesian augmented naïve Bayes network
BIF – Bayesian interchange format
BN – Bayesian network
CPT – conditional probability table
CI – conditional independence
DAG – directed acyclic graph
EBK – Empirical Bayes Kriging
ENM – ecological niche model
GBN – general Bayesian network
GIS – geographic information system
GPS – global positioning system
GS – genetic search
HC – hill climbing
ICS – inductive causation
iSDM – invasive species distribution model
JiSDM – joint invasive species distribution model
JPD – joint probability distribution
JSDM – joint species distribution model
LAGD - look ahead in good directions
NB – naïve Bayes
PMAT - Probabilistic Map Algebra Tool
PPCI – posterior probability certainty index
QGIS – Quantum GIS
RHC – repeated hill climbing
ROC – receiver operating characteristic curve
SA – simulated annealing
SDM – species distribution model
TAN – tree augmented naïve
TS – tabu search
TSS – true skill statistic

VGI – volunteered geographic information

XML – eXtensible Markup Language

TABLE OF CONTENTS

Dedication.....	i
Declaration.....	ii
Acknowledgements	iii
Abstract.....	iv
List of Acronyms	vi
Table of Contents.....	viii
List of Figures.....	xi
List of Tables	xv
List of Appendices.....	xvi
Chapter 1 : Introduction.....	1
1.1 Background	1
1.2 Motivation.....	6
1.3 Research objectives and scope.....	10
1.4 Organization of the Thesis	12
Chapter 2 : Bayesian networks – A review of literature.....	14
2.1 Introduction.....	14
2.2 Bayes’ theorem	15
2.3 Bayesian network definition	16
2.4 Structure learning.....	18
2.4.1 Search-and-score approaches	19
2.4.2 Constraint-based search.....	21
2.5 Parameter learning	22
2.6 Bayesian network interpretation and reasoning.....	23
2.7 Applications in the species distribution modelling domain.....	29
Chapter 3 : Methods	43
3.1 Introduction.....	43
3.2 Study Area	43
3.3 Research process.....	46
3.4 Data acquisition and integration	50
3.4.1 Species distribution (target variable) data	50
3.4.2 Predictor (attribute) variables	53
3.5 Data pre-processing	53

3.5.1	Data cleaning	55
3.5.2	Transformation and conversion	56
3.5.3	Data integration	58
3.5.4	Discretization.....	61
3.6	Data balancing	63
3.7	Feature (variable) selection.....	65
3.8	Bayesian network learning.....	70
3.8.1	Structure learning	70
3.7.2	Parameter learning	75
3.7.3	Variable importance (sensitivity) analysis	76
3.8	Model evaluation	77
3.9	Visualization	81
Chapter 4 : Results.....		82
4.1	Introduction.....	82
4.2	Bayesian network learning algorithm performance	82
4.3	Learned Bayesian networks and predicted distributions	93
4.3.1	<i>Acacia mearnsii</i>	93
4.3.2	<i>Caesalpinia decapetala</i>	98
4.3.3	<i>Cereus jamacaru</i>	102
4.3.4	<i>Chromolaena odorata</i>	106
4.3.5	<i>Eucalyptus</i> species.....	110
4.3.6	<i>Jacaranda mimosifolia</i>	114
4.3.7	<i>Lantana camara</i>	118
4.3.8	<i>Melia azedarach</i>	123
4.3.9	<i>Opuntia</i> species.....	127
4.3.10	<i>Pinus</i> species.....	131
4.3.11	<i>Populus x canescens</i>	136
4.3.12	<i>Psidium guajava</i>	141
4.3.13	<i>Rubus</i> species.....	146
4.3.14	<i>Senna didymobotrya</i>	151
4.3.15	<i>Sesbania punicea</i>	155
4.3.16	<i>Solanum mauritianum</i>	160
4.4	General findings on learned Bayesian network models.....	164
4.4.1	Learned Bayesian network structures.....	164

4.4.2	Species distribution maps	165
4.4.3	Species distribution uncertainty.....	169
Chapter 5	: Discussion.....	174
5.1	Introduction.....	174
5.2	Bayesian network model development	174
5.3	Bayesian network model performance.....	179
5.4	Species distribution patterns and invasion processes.....	184
5.4.1	<i>Acacia mearnsii</i>	184
5.4.2	<i>Caesalpinia decapetala</i>	185
5.4.3	<i>Cereus jamacaru</i>	187
5.4.4	<i>Chromolaena odorata</i>	188
5.4.5	<i>Eucalyptus</i> species.....	189
5.4.6	<i>Jacaranda mimosifolia</i>	190
5.4.7	<i>Lantana camara</i>	191
5.4.8	<i>Melia azedarach</i>	193
5.4.9	<i>Opuntia</i> species.....	194
5.4.10	<i>Pinus</i> species.....	195
5.4.11	<i>Populus x canescens</i>	197
5.4.12	<i>Psidium guajava</i>	198
5.4.13	<i>Rubus</i> species.....	199
5.4.14	<i>Senna didymobotrya</i>	200
5.4.15	<i>Sesbania punicea</i>	201
5.4.16	<i>Solanum mauritianum</i>	202
5.5	General discussion on invasion patterns and processes.....	204
5.5.1	Invasion patterns and processes.....	204
5.5.2	Species distribution uncertainty.....	212
5.6	Applicability of Bayesian network-based data mining to species distribution modelling problems	215
Chapter 6	: Conclusions and Recommendations	219
6.1	Conclusions.....	219
6.2	Recommendations.....	222
References	224
Glossary	300

LIST OF FIGURES

Figure 2.1: An example of a Bayesian network with a sample conditional probability table.	23
Figure 2.2: A diverging or fork connection (adapted from Kjærulff and Madsen, 2013)...	25
Figure 2.3: A serial or causal chain connection (adapted from Kjærulff and Madsen, 2013).....	26
Figure 2.4: A converging or colliding connection (adapted from Kjærulff and Madsen, 2013).....	27
Figure 2.5: The number of publications produced between 1990 and 2015 focusing on Bayesian network-based species distribution modelling.	38
Figure 3.1: Location of the study area, highlighting the topography (source: own).	45
Figure 3.2: The research process followed in the study (EBK – Empirical Bayes Kriging, ED – Euclidean Distance, KDE – Kernel Density Estimation, source: own).	49
Figure 3.3: The re-ranking canonical algorithm (adapted from Bermejo <i>et al.</i> , 2012, p.39).....	67
Figure 3.4: Prediction performance and variable selection as a function of block (subset) size (source: own).....	69
Figure 4.1: Performance comparison (box plots) of all the BN learning algorithms using the logarithmic loss (source: own).	85
Figure 4.2: Performance comparison (box plots) using the area under the ROC curve (AUC) for all the BN learning algorithms (source: own).....	86
Figure 4.3: Performance comparison (box plots) using area true skill statistic (TSS) for all the BN learning algorithms (source: own).	87
Figure 4.4: Performance comparison (box plots) using Matthew’s correlation coefficient (MCC) for all the BN learning algorithms (source: own).	88
Figure 4.5: Scatter plots of model evaluation metrics plotted against species prevalence (source: own).	90
Figure 4.6: Box plots of the computation time (in seconds) for all the BN learning algorithms (source: own).	91
Figure 4.7: Plot of CPU time against species prevalence (left) and the number of selected variables (right)(source: own).	92
Figure 4.8: Plots of the number of selected features or variables against species prevalence (left) and the number of mean log loss against selected variables (right) (source: own). ...	93

Figure 4.9: A learned Bayesian network for <i>Acacia mearnsii</i> distribution.	94
Figure 4.10: Posterior probability of occurrence for <i>A. mearnsii</i> in Swaziland (derived from the BN in Figure 4.9).....	96
Figure 4.11: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>A. mearnsii</i> in Swaziland.	97
Figure 4.12: A learned Bayesian network for <i>Caesalpinia decapetala</i> distribution.	98
Figure 4.13: Posterior probability of occurrence for <i>A. mearnsii</i> in Swaziland (derived from the BN in Figure 4.12).....	100
Figure 4.14: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>C. decapetala</i> in Swaziland.	101
Figure 4.15: A learned Bayesian network for <i>Cereus jamacaru</i> distribution.	102
Figure 4.16: Posterior probability of occurrence for <i>C. jamacaru</i> in Swaziland (derived from the BN in Figure 4.15).....	104
Figure 4.17: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>C. jamacaru</i> in Swaziland.	105
Figure 4.18: A learned Bayesian network for <i>Chromolaena odorata</i> distribution.	106
Figure 4.19: Posterior probability of occurrence for <i>C. odorata</i> in Swaziland (derived from the BN in Figure 4.18).....	108
Figure 4.20: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>C. odorata</i> in Swaziland.	109
Figure 4.21: Learned Bayesian network for <i>Eucalyptus</i> species distribution.	110
Figure 4.22: Posterior probability of occurrence for <i>Eucalyptus</i> in Swaziland (derived from the BN in Figure 4.21).....	112
Figure 4.23: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>Eucalyptus</i> species in Swaziland.	113
Figure 4.24: A learned Bayesian network for <i>Jacaranda mimosifolia</i> distribution.	114
Figure 4.25: Posterior probability of occurrence for <i>J. mimosifolia</i> in Swaziland (derived from the BN in Figure 4.24).....	116
Figure 4.26: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>A. mearnsii</i> in Swaziland.	117
Figure 4.27: A learned Bayesian network for <i>Lantana camara</i> distribution.....	119
Figure 4.28: Posterior probability of occurrence for <i>L. camara</i> in Swaziland (derived from the BN in Figure 4.27).....	121

Figure 4.29: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>L. camara</i> in Swaziland.	122
Figure 4.30: A learned Bayesian network for <i>Melia azedarach</i> distribution.	123
Figure 4.31: Posterior probability of occurrence for <i>M. azedarach</i> in Swaziland (derived from the BN in Figure 4.30).	125
Figure 4.32: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>M. azedarach</i> in Swaziland.	126
Figure 4.33: A learned Bayesian network for <i>Opuntia</i> species distribution.	127
Figure 4.34: Posterior probability of occurrence for <i>Opuntia</i> species in Swaziland (derived from the BN in Figure 4.33).	129
Figure 4.35: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>Opuntia</i> species in Swaziland.	130
Figure 4.36: A learned Bayesian network for <i>Pinus</i> species distribution.	132
Figure 4.37: Posterior probability of occurrence for <i>Pinus</i> species in Swaziland (derived from the BN in Figure 4.36).	134
Figure 4.38: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>Pinus</i> species in Swaziland.	135
Figure 4.39: A learned Bayesian network for <i>Populus x canescens</i> distribution.	137
Figure 4.40: Posterior probability of occurrence for <i>P. x canescens</i> in Swaziland (derived from the BN in Figure 4.39).	139
Figure 4.41: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>P. x canescens</i> in Swaziland.	140
Figure 4.42: A learned Bayesian network for <i>Psidium guajava</i> distribution.	142
Figure 4.43: Posterior probability of occurrence for <i>P. guajava</i> in Swaziland (derived from the BN in Figure 4.42).	144
Figure 4.44: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>P. guajava</i> in Swaziland.	145
Figure 4.45: A learned Bayesian network for <i>Rubus</i> species distribution.	147
Figure 4.46: Posterior probability of occurrence for <i>Rubus</i> species in Swaziland (derived from the BN in Figure 4.45).	149
Figure 4.47: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>Rubus</i> species in Swaziland.	150
Figure 4.48: A learned Bayesian network for <i>Senna didymobotrya</i> distribution.	151

Figure 4.49: Posterior probability of occurrence for <i>S. didymobotrya</i> in Swaziland (derived from the BN in Figure 4.48).	153
Figure 4.50: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>S. didymobotrya</i> in Swaziland.	154
Figure 4.51: A learned Bayesian network for <i>Sesbania punicea</i> distribution.	156
Figure 4.52: Posterior probability of occurrence for <i>S. punicea</i> in Swaziland (derived from the BN in Figure 4.51).	158
Figure 4.53: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>S. punicea</i> in Swaziland.	159
Figure 4.54: A learned Bayesian network for <i>Solanum mauritianum</i> distribution.	160
Figure 4.55: Posterior probability of occurrence for <i>S. mauritianum</i> in Swaziland (derived from the BN in Figure 4.54).	162
Figure 4.56: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for <i>S. mauritianum</i> in Swaziland.	163
Figure 4.57: Box plots of the mean posterior probabilities from all the algorithms implemented for each species (source: own).	168
Figure 4.58: A plot of posterior probability against the posterior probability certainty index (source: own).	170
Figure 4.59: Box plots of the posterior probability certainty indices for all the species (source: own).	172
Figure 4.60: Scatter plots PPCI against prevalence and mean logarithmic loss (source: own).	173

LIST OF TABLES

Table 2.1: List of scientific applications of BN models in species distribution modeling..	31
Table 3.1: List of alien plant species studied and their characteristics (characteristics information derived from Henderson, 2007).....	52
Table 4.1: Mutual information for selected <i>Acacia mearnsii</i> predictor variables.	94
Table 4.2: Mutual information for selected <i>Caesalpinia decapetala</i> predictor variables. ..	99
Table 4.3: Mutual information for selected <i>Cereus jamacaru</i> predictor variables.	103
Table 4.4: Mutual information for selected <i>Chromolaena odorata</i> predictor variables. ..	107
Table 4.5: Mutual information for selected <i>Eucalyptus</i> species predictor variables.....	111
Table 4.6: Mutual information for selected <i>Jacaranda mimosifolia</i> predictor variables. .	115
Table 4.7: Mutual information for selected <i>Lantana camara</i> predictor variables.....	120
Table 4.8: Mutual information for selected <i>Melia azedarach</i> predictor variables.....	124
Table 4.9: Mutual information for selected <i>Opuntia</i> species predictor variables.	128
Table 4.10: Mutual information for selected <i>Pinus</i> species predictor variables.	133
Table 4.11: Mutual information for selected <i>Populus x canescens</i> predictor variables. ...	138
Table 4.12: Mutual information for selected <i>Psidium guajava</i> predictor variables.	142
Table 4.13: Mutual information for selected <i>Rubus</i> species predictor variables.....	148
Table 4.14: Mutual information for selected <i>Senna didymobotrya</i> predictor variables. ...	152
Table 4.15: Mutual information for selected <i>Sesbania punicea</i> predictor variables.	157
Table 4.16: Mutual information for selected <i>Solanum mauritianum</i> predictor variables. .	161

LIST OF APPENDICES

Appendix 1: List of variables (datasets) used in the study	274
Appendix 2: Observed distribution maps of all the species from the aerial survey and tree atlas datasets	281
Appendix 3: Performance of the Bayesian learning algorithms.....	286
Appendix 4: Variables selected to form the Markov blanket of all the species' distribution models.....	291
Appendix 5: Photographs showing alien plant invasion in Swaziland.....	293

CHAPTER 1 : INTRODUCTION

1.1 BACKGROUND

The spread of invasive species, driven mainly by human activities, is increasing worldwide (Butchart *et al.*, 2010) and poses potential problems not only to native biodiversity but also to economic development and human well-being (Chytry *et al.*, 2009; Pejchar and Mooney, 2009; Pyšek *et al.*, 2010; Vilà *et al.*, 2011; Skandrani *et al.*, 2014). The impacts of invasive species on biodiversity, ecosystem function and human welfare makes them key drivers of ecosystem change (Crowl *et al.*, 2008). The accelerating spread of invasive plant species threatens to displace native vegetation over large areas in South Africa, with significant impacts on water resources, catchments and the wildfire risks (van Wilgen *et al.*, 2012). A recent survey of alien plants in Swaziland revealed that 80% of the total land area is invaded (Kotzé *et al.*, 2010a) including large tracts of productive rangelands, natural ecosystems and existing and potential croplands. Kotzé *et al.* (2010a) estimated that it would cost approximately SZL665 million (US\$ 90 million) to be cleared at once. This figure represented approximately 2.2% of the country's GDP in 2010, money that could otherwise be used for other priority national development initiatives. However, this figure excludes the direct and indirect costs resulting from other associated impacts of these plants such as the loss of biodiversity and ecosystem services including reduced water availability, loss of pasture, among many others. The identified and possible impacts of these invasive alien plants point to the pressing need for the control and/or halting their spread.

The control of these invasive organisms is, however, one of the most complex tasks which requires a good understanding of the invasion processes and the key factors that determine or influence the observed invasion patterns (Gallien *et al.*, 2010; Strayer, 2012). This requires inventories or databases that can reveal the invasive alien plants' distribution patterns and provide critical information about each plant's invasion process (Fuentes *et al.*, 2013). Subsequently, there are several initiatives for collecting species distribution data from local to global scales in an effort to understand their distribution patterns and invasion dynamics (Danielsen *et al.*, 2005; Lavoie *et al.*, 2013). In the past few decades various institutions and

individuals have dedicated substantial resources into digitizing and collation of these data into digital species distribution atlases, resulting in large amounts of data that represent a largely untapped source of information for use in conservation biogeography (Elith and Leathwick, 2009; Richardson and Whittaker, 2010; Elith *et al.*, 2011; Marcer *et al.*, 2012). Such species occurrences have been recorded through a variety of approaches, ranging from systematic surveys to *ad hoc* records in the form of museum specimens, inventories, technical and scientific literature citations (Chapman, 2005; Higgins *et al.*, 2012; Marcer *et al.*, 2012; Lavoie *et al.*, 2013), including citizen science (Danielsen *et al.*, 2005; Dickinson *et al.*, 2010; Crall *et al.*, 2015). Coupled with this is the ever-increasing amount of data resulting from increased networking, continuous improvement in computational and data storage technology, sophisticated databases and the enormous expansion of automated data collection via diverse sensors and platforms (Lausch *et al.*, 2015).

Potentially useful information that is seldom made explicit or taken advantage of is often hidden in all these datasets. One such important piece of information is the current and historical distribution of invasive alien species, which is often used to develop species distribution models (SDMs) for risk analysis and effective control (Richardson and Whittaker, 2010; Stohlgren *et al.*, 2010; Jiménez-Valverde *et al.*, 2011). This information is also essential for maximizing the efficient use of limited financial resources (Nielsen *et al.*, 2008; Marcer *et al.*, 2012). Coupled with the increasing availability of species distribution data is a corresponding increase in the generation and availability of other ancillary geospatial information and earth science data obtained from remote sensing, global positioning and geographic information systems (GIS), amongst other related sources (Mennis and Guo, 2009). These multiple data sources should enable the identification and characterization of important interactions, processes and coincident spatial patterns at different scales of analyses. The duty of the scientist is to analyse the large volumes of data and derive sensible information by discovering the patterns that govern how the physical world functions, condensing and interpreting these in theories that can be used for prediction purposes (Witten *et al.*, 2011).

However, since these datasets typically come from disparate sources they predominantly indicate only the (often spatially biased) presence of species, or may show spatial aggregation derived from the sampling biases (Marcer *et al.*, 2012). Such data sets are often collected at

coarse scales and may be difficult and costly to geocode (Pressey, 2004). Nonetheless, given the observed trends in worldwide biodiversity loss and the need to address conservation problems, it becomes imperative to find tools and techniques that can make the best use of this existing information (Newbold, 2010; Venette *et al.*, 2010).

Species distribution and ecological niche modelling have become important tools for describing, understanding, and predicting the spatial and environmental distributions of species. Usage of these tools has increased drastically in recent years driven by rapid advances in computational power and technology including geospatial technologies and improvements in data collection and analysis techniques (Lobo *et al.*, 2010; Peterson *et al.*, 2011). SDMs are increasingly being used to predict species habitats including the establishment and spread of alien invasive species, also termed invasive species distribution models (iSDMs) (Elith and Leathwick, 2009; Václavík and Meentemeyer, 2009; Gallien *et al.*, 2010; Venette *et al.*, 2010; Zimmermann *et al.*, 2010, Hegel *et al.*, 2010; Gallien *et al.*, 2012). Terms such as “species distribution models” (SDM; Elith and Leathwick 2009; Franklin 2010), “ecological niche models” (ENM; Harrison, 1997; Peterson *et al.*, 1999) and “bioclimatic envelope models” (Araújo and Peterson, 2012) are widely used to refer to correlative summaries of species’ environmental associations and the relationships of those associations to their spatial distributions. There are many inconclusive arguments regarding these terms and their interpretation in environmental and geographic space, especially as they relate to the niche concept (Peterson, 2006; Elith and Leathwick, 2009; Franklin, 2010; Sillero, 2011; Araújo and Peterson, 2012; Peterson and Soberón, 2012; Warren, 2012).

In this study, the term SDM refers to a model of known species occurrences or distribution to generate hypotheses about species distributions, rather than modelling their ecological niche in its strictest sense. Modelling the processes that produce and shape species distribution patterns, transferring the causal factors in space, or interpreting the obtained patterns, requires some propositions about the ecology of the species, which is the domain of ecological niche modelling (Peterson and Soberón, 2012). Nevertheless, underlying all the modelling approaches is the need to uncover, describe, understand and predict the biogeographic and ecological relationships between species and the environments in which they are or likely to be found.

Similar to many biological systems, species-environment relationships are inherently complex, stochastic and non-linear which complicates the task of untangling of the underlying interactions and relationships. This complexity is further exacerbated by the fact that environmental data are typically polymorph (variable), incomplete (missing), noisy (difficult to observe) and, therefore, can be the source of errors (Gibert *et al.*, 2008). Furthermore, species interact through various mechanisms such as competition, predation, facilitation and mutualism, amongst others (Bascompte, 2009; Milns *et al.*, 2010; Wisz *et al.*, 2013). The relationships amongst biotic and abiotic factors that influence a species' distribution are themselves rarely deterministic and are often fuzzy in nature. Hence, the unravelling of the complex interactions and relationships is one of the fundamental challenges in ecology (Milns *et al.*, 2010; Boulangeat *et al.*, 2012; Wisz *et al.*, 2013).

Understanding the networks that form the systems is of growing importance for predicting and managing the potential spread of invasive alien organisms (Faisal *et al.*, 2010). Lawton (1999), as cited by (McMahon, 2005, p.833), claims that one of the reasons the field of community ecology has struggled to derive general laws that govern its components is that community components are highly interrelated and inferences from any one community are 'contingent' on conditions in that community. McMahon (2005) suggests that researchers must better quantify these inter-dependencies of the community into which a species invades so as to quantitatively describe the way components of a community interact with that species. Identifying these underlying patterns and processes requires novel approaches and methods that are capable of efficiently recovering and inferring the structure of these complex networks and acquiring, integrating and modelling massive quantities of diverse field data (Faisal *et al.*, 2010; Milns *et al.*, 2010; Hochachka *et al.*, 2012).

In the last decade, the species distribution modelling community has witnessed the appearance of new tools and methodologies from the fields of statistics and machine learning that have the potential to address the problems inherent in species distribution datasets. Numerous methods are now available and routinely applied for species distribution modelling albeit with considerable variations in both performance and spatial predictions (Elith and Leathwick, 2009; Grenouillet *et al.*, 2011). This variability in model performance and prediction is often attributed to the statistical foundations and the mathematical functions of those models. This includes the

assumptions used to describe the distribution of species in relation to geographical/environmental parameters and the data characteristics, among other causes (Guisan and Thuiller, 2005; Araújo and New, 2007; Elith and Leathwick, 2009; Grenouillet *et al.*, 2011; Dormann *et al.*, 2012). The differences in the theoretical foundations and outputs of SDMs may affect their usefulness for a variety of ecological applications.

Conventional statistical approaches provide widely accepted assessments of species-environment relationships and typically require large amounts of data collected with an appropriate experimental design (Millsbaugh and Thompson, 2009), resulting in failure to explicitly consider the underlying processes and uncertainty. Therefore, on the one hand, development of advanced techniques for obtaining field data is required to better understand the heterogeneous species-environment relationships to enhance the accuracy of SDMs and to deal with the underlying uncertainty (Beale and Lennon, 2012). On the other hand, species distribution data is collected at different spatial and temporal scales with various degrees of accuracy and quality, more so as geo-referenced data collection tools proliferate together with citizen science and volunteered geographic information (Kéry *et al.*, 2010; Yu *et al.*, 2010; Hochachka *et al.*, 2012; Crall *et al.*, 2015).

The volume, complexity and inherent uncertainty of the data from various sources is quickly transcending the limits of conventional analysis tools in terms of capacity and efficiency thereby limiting the usefulness of conventional techniques for extracting and describing spatial patterns. The significant growth in data collection and widespread use of multisource and multidimensional data have heightened the need for methods for dealing with biases and (semi)automated discovery of ecological knowledge. Most current predictive modelling methods use one source of knowledge, such as domain expert knowledge or field samples, to establish the relationship. Limited by the expertise of experts and poor representativeness of field samples, the extracted knowledge from these methods is usually not reliable. Multiple knowledge sources that are complementary to each other may enhance the quality of knowledge and thus improve geospatial prediction. On the other hand, the new web-based data sharing paradigm and volunteered geographic information (VGI) make diverse knowledge sources increasingly available and usable for geospatial prediction. As a result, a number of data mining and machine learning-based SDMs have emerged (Stockwell, 2006; Elith and Leathwick, 2009;

Václavík and Meentemeyer, 2009; Zimmermann *et al.*, 2010; Lorena *et al.*, 2011; Rangel and Loyola, 2012; Bhattacharya, 2013).

1.2 MOTIVATION

There remains a critical gap in the understanding of processes that induce observed invasion spatial patterns over a range of scales due to many studies focusing mainly on small-scale mechanisms of plant invasions whilst a few have examined large-scale spatial patterns (Pauchard and Shea, 2006). This can be partially attributed to the predominant use of correlative and regression-based SDMs and the lack of appropriate and reliable techniques that can explicitly model and decipher such processes from the observed patterns (Gallien *et al.*, 2010; Dormann *et al.*, 2012). Conventional SDMs are correlative in nature and may be accurate without capturing the essential causal species-environment relationships and ecological knowledge thus losing the niche theory in the statistics (Hirzel and Le Lay, 2008; Hortal *et al.*, 2012). Data mining and machine learning techniques have emerged as technologies to extract useful knowledge out of large and heterogeneous datasets although they can be implemented for much smaller datasets (Hochachka *et al.*, 2007).

Data mining is defined as the nontrivial process that innovatively attempts to obtain accurate, potentially valuable and easily understandable information or patterns (Fayyad *et al.*, 1996; Witten *et al.*, 2011). This includes multivariate geo-visualization in the case of mining of geographic data (Mennis and Guo, 2009). Data mining is generally predictive in nature and can be categorized into description, clustering, classification and regression tasks (Hastie *et al.*, 2009; Witten *et al.*, 2011). Data driven prediction attempts to find a correlated target function ($y = f(x)$) between predictor attributes (x) and the target attribute (y) by studying given sample data in order to accurately predict unknown or previously unobserved classification of samples. By relying on several parameters, the task is to divide the data into a multi-dimensional space to reveal general properties of concentrated data for building a predictive model that will obtain a consistent solution to a practical problem. In the case of a discrete target variable, the predictive model is a classification or clustering model, otherwise it is a regression model. Therefore, SDMs are considered classification and regression models, pointing to the potential application of data mining techniques, especially in cases where little prior knowledge exists of an ecosystem, species and/or when accurate predictions are the desired product (Hochachka *et*

al., 2007). Data mining can help in uncovering interesting and previously unknown but potentially useful patterns from which ecological processes may be inferred. However, the choice of the data model is a very crucial and complex task in data mining because the chosen model ought to represent the data precisely and should likewise be appropriate for the technique used (Mennis and Guo, 2009; Witten *et al.*, 2011).

The past two decades have seen a great deal of interest in probabilistic graphical models as data mining and modelling tools. These models, in particular Bayesian networks (BN) (Pearl, 1988), are able to represent the probabilities and logical structures of real world complex and non-linear systems in a compact way. BNs are directed acyclic graphical models that are used for modelling uncertain relationships amongst variables in a complex system and reasoning under uncertainty, where nodes represent random variables (discrete and/or continuous) and arcs (or links) represent direct causal and informative connections between them (Pearl, 1988; Korb and Nicholson, 2011). As such, BN techniques are increasingly being used in complex scientific application such as intrusion detection, system reliability analysis, medical diagnosis, clinical decision support, crime analysis, sensor validation, information retrieval, credit-rating, risk management, epidemiology, forensic science, robotics and establishing genome pathways, among other applications (Pourret *et al.*, 2008; Koski and Noble, 2012). For example, the major challenge for which BNs are used in genome analysis is to uncover gene/protein interactions and the key biological features of cellular systems given DNA hybridization arrays, which simultaneously measure the expression levels for thousands of genes (Friedman *et al.*, 2000; Markowitz and Spang, 2007; Smith, 2010; Su *et al.*, 2013). The usefulness of BNs in such complex areas of systems biology points to the potential usefulness of BNs to study the higher levels of biological organization such as biogeography and ecology where complexity, parameter estimation and stability analysis are common problems (Larjo *et al.*, 2013).

BN models are mostly useful for analysing and communicating causal assumptions that are not easily expressed using mathematical notation and for analysing the multivariate and complex relationships among variables (Pollino and Hart, 2008; Koski and Noble, 2012). Unlike many other ecological modelling approaches, BNs can utilize prior knowledge and can extract knowledge from large and heterogeneous datasets and represent this knowledge in the form of a probabilistic graphical model, providing a compact description of the given data and allowing

predictions for new data (Heckerman, 1997; Korb and Nicholson, 2011). This intuitive visual representation is useful in clarifying previously opaque assumptions or reasoning, which is characteristic of most other machine learning approaches that are difficult to interpret (Rangel and Loyola, 2012; Bhattacharya, 2013). This emphasizes the potential application of BNs in situations where the graph is constructed along causal principles, i.e. where parent variables are considered direct causes of a target or response variable. In addition, the graphical nature of BN models permits inference to be done relatively easier through the computation of posteriori probabilities for values of variables that were not seen or measured in the given field data. In addition, the posterior probabilities could be future values in a dynamical model.

Since BN tools are able to deal with the analytical challenge presented by intricate species-environment relationships, complex interaction networks can be inferred from species distribution data while simultaneously estimating the influence of covariates and uncertainty (Milns *et al.*, 2010; Beale and Lennon, 2012). A comprehensive understanding of any ecological system requires the connection of interacting variables through concerted research involving both empirical and modelling approaches. McMahon (2005) observes that BNs can determine and quantify the relationships and influences between the components of a natural system and can further provide dynamic inferential analysis of the learned parameters. Therefore, BNs should allow for better understanding and managing ecosystems despite their inherent complexity (Smith, 2010). Through abductive inference and use of Bayes' theorem, BNs reduce bias and provide a framework for assessing causality and examining attributable risk or probability of causes from combining data sources (Pearl, 2000; Pollino and Hart, 2008). This is critical for alien invasive species where such knowledge is needed to prioritize locations for early detection and control of invasion. BNs can also incorporate future monitoring data to update predictions based on new knowledge (Nyberg *et al.*, 2006; Aguilera *et al.*, 2011), which is relevant for species distribution monitoring especially invasive species that are not at equilibrium with their environment.

The real strength of BNs is in the application of Bayesian probability rules to consistently propagate the impact of evidence on the probabilities of uncertain outcomes. This, therefore, enables BNs to model uncertain events and arguments about them, hence the increasing interest in their application in the environmental and ecological sciences where high uncertainties exist

(Uusitalo, 2007; Hamilton *et al.*, 2009; Aguilera *et al.*, 2011; Maldonado *et al.*, in press). Many SDMs, including those focusing on invasive species, still poorly characterize and represent uncertainties associated with the predictive outputs which is problematic in the identification and prioritization of possible interventions or management actions (Elith and Leathwick, 2009; Beale and Lennon, 2012).

BNs have also been found valuable in spatial knowledge discovery and data mining (Buang *et al.*, 2006; Huang and Yuan, 2007; Johnson *et al.*, 2012a) where they have been found useful for spatial knowledge representation, spatial classification, spatial clustering, and spatial prediction (Huang and Yuan, 2007). This again makes BNs useful for spatial modelling (Uusitalo, 2007; Johnson *et al.*, 2012a) although their application for species distribution modelling is still limited. In recent reviews of software and techniques used for species distribution modelling by Joppa *et al.* (2013) and Ahmed *et al.* (2015), BNs do not feature prominently. Nevertheless, there have been attempts to apply BNs in species distribution-related problems and these include studies by McMahon (2005), Pullar and Phan (2007), Smith *et al.* (2007), Aguilera *et al.* (2010), Atzmanstorfer *et al.* (2007), Wilson *et al.* (2008), Murray *et al.* (2012), Grech and Coles (2010), Milns *et al.* (2010), Smith *et al.* (2011), Chen and Pollino (2012), Murray *et al.* (2012), Wilhere (2012), Douglas and Newton (2014), Gieder *et al.* (2014), Murray *et al.* (2014), amongst others. Most of these studies, however, have mainly been developing habitat suitability indices (Tantipisanuh *et al.*, 2014) rather than SDMs and, except for a limited few studies such as Milns *et al.* (2010), Alameddine *et al.* (2011) and Boets *et al.* (2015), they relied on expert knowledge and parametric approaches for manually constructing and parameterizing the networks.

Eliciting a BN model for a given application by domain experts, whilst very important, can be a time consuming and highly complex task that may result in bias especially where there are large datasets and disparate opinions (Alameddine *et al.*, 2011). The use of expert opinion has the potential to limit possible discovery of new ecological knowledge on species distributions. To this end, Strayer (2012) provides caution on the use of expert opinion in invasion ecology and recommends the replacement of such expert opinion with actual data or at least testing that opinion for reliability where firm, reliable or accurate answers are needed. On the same note, BN learning from data is maturing in many other disciplines but there are few such in ecological sciences. Hence, techniques that automatically learn BNs from data are indispensable especially

in cases where the invasion patterns and processes are least known and potentially complex. Boets *et al.* (2015) and Hamilton *et al.* (2015) have recently found that data-driven and combined expert- and data-driven models performed better than models based only on expert knowledge, thereby pointing to the need to explore data mining or machine learning techniques. The approaches for automating the process of learning the BN structure from data have been developed based on two main techniques, namely constraint based and score-based approaches both of which have shown to be effective (Jensen and Nielsen, 2007; Korb and Nicholson, 2011), although hybrid approaches are emerging (Daly *et al.*, 2011; Koski and Noble, 2012).

1.3 RESEARCH OBJECTIVES AND SCOPE

The main objective of this study is to investigate the spatial distribution patterns and invasion processes of selected alien plant species in Swaziland using BN models. The study tests the data-driven application of BN models using score-and-search and constraint-based algorithms for extracting landscape level patterns and ecological knowledge on the spatial distribution of 16 priority invasive alien plants. Various BN algorithms are implemented using a machine learning or data-driven approach to model the causal factors and underlying processes that produce and shape the distribution of the observed spatial patterns. This approach allows for handling the species distribution data from disparate sources to derive spatially explicit probabilistic models that will contribute to a better understanding of alien plant invasion patterns and processes at the local and landscape scales to inform policy and design appropriate control strategies.

Specifically, the study aims to:

1. Collate existing information on the distribution of selected (priority) invasive alien plants in Swaziland.
2. Develop and test BN-based methodologies for probabilistic species distribution modelling using selected invasive alien plants in Swaziland.
3. Identify the key predictors or causal factors influencing alien invasive plant distribution patterns using BN-based data mining.

4. Use the resultant BN model structures to infer and explain the processes driving the observed alien plant species invasion patterns for better planning and control.

Using a knowledge discovery or data mining approach and in the pursuit of these study aims, the study aims to answer the following scientific questions:

1. What is the current spatial distribution of the selected invasive alien plant species in Swaziland?
2. Which BN structure learning approaches can best model and describe the spatial patterns in the distribution of the selected invasive alien plants in Swaziland?
3. What are the main factors driving the alien plant species invasion patterns in Swaziland?
4. What are the processes likely influencing the observed distribution of each alien plant species' relationships with the causal factors?

The contributions of this study are both in the methodologies and their applications to species distribution modelling and analysis. Specifically, this study's contribution is in the following areas:

- i. The novel development and application of BN structure learning techniques for species distribution modelling within a data mining framework.
- ii. The development of approaches for handling the problem of spatial data integration and uncertainty in species distribution modelling using data-driven BN approaches and algorithms.
- iii. The modelling and better understanding of the interactions between the species and biotic and abiotic variables (both discrete and continuous) using the graphical models in order to infer the invasion processes and relationships amongst causal factors.

1.4 ORGANIZATION OF THE THESIS

The thesis is organized into a logical sequence of chapters. Chapter 1 is an introduction to the thesis wherein the problem of alien plant invasion and the motivation for this study is presented. A brief synthesis of current approaches and challenges in species distribution modelling will be provided, concluding with the objectives and contribution of this study.

Chapter 2 introduces the theoretical background to BNs focusing on structure and parameter learning, causal analysis and inference. A concise review, including a critique, of past and current applications of BN techniques to species distribution modelling problems is also presented.

Chapter 3 presents in detail the methodology used in the study focusing on data collection, integration, pre-processing and the entire data mining process flow up to model validation and knowledge presentation and visualization. BN structure and parameter learning together with the algorithms and approaches used in this thesis are provided and explained in detail based on the theoretical foundation provided in Chapter 2.

Chapter 4 is dedicated to the detailed presentation of the results with emphasis on the efficacy and validity of the various algorithms used for the problem under investigation, which is alien invasive plant species distribution modelling. The performance and outputs of the different BN learning methods are presented in detail including the modelled distribution patterns and learned graphical models. The discovered graphical relationships and learned causal models are provided.

An in-depth discussion of findings and the observed spatial patterns is provided in Chapter 5 focusing on the resulting BN structures to infer and explain the discovered species invasion knowledge. In addition, this chapter simultaneously consolidates, discusses and explains the observed patterns in relation to the learned BN structures, the findings presented in the preceding chapter and general species distribution modelling. The implications of the findings for species distribution modelling and alien invasive plant control and management are articulated in this chapter.

Chapter 6 is the final chapter where general conclusions and limitations of the study are presented. The key contributions of this thesis are summarized and issues for further research in the area of BN applications to species distribution modelling, especially with respect to alien invasive plants, are proposed. The chapter concludes by providing key invasive alien plant policy and management recommendations and suggestions for future research.

CHAPTER 2 : BAYESIAN NETWORKS – A REVIEW OF LITERATURE

2.1 INTRODUCTION

There is abundant literature on BN learning and inference and these include, but not limited to, Pearl (1988), Neapolitan (2004), Cowell *et al.*, (2007), Jensen and Nielsen (2007), Pourret *et al.* (2008), Koski and Noble (2009), Darwiche (2009), Korb and Nicholson (2011) and Kjærulff and Madsen (2013). Daly *et al.* (2011), Flores *et al.* (2012), Koski and Noble (2012) and Bielza and Larrañaga (2014) provide detailed reviews of contemporary approaches to and issues on learning BNs from data, the former focusing on the widely used discrete BN classifiers. As briefly highlighted in the preceding chapter, BNs provide an efficient representation for expressing joint probability distributions (JPDs) and for inference. Over the last three decades, BNs have gained popularity representation for encoding uncertain knowledge in many domains. Subsequently, several techniques for learning BNs from data have been developed, most of which have been shown to be remarkably effective for many data analysis problems. In this chapter, a general theoretical background of BNs is briefly provided with a focus on the techniques for extracting and encoding knowledge from data. The process of structure and parameter learning from data is briefly explained including the underlying assumptions. The chapter is not meant to be an exhaustive review of BN theory but rather introduces the theoretical basis for the approaches used in this study. Examples of BN reasoning and some of the general rules that govern the way direct and induced independencies are expressed in a BN model are provided. The chapter ends with a comprehensive review of BN applications in species distribution modelling with a focus on identifying gaps in the structure and parameter learning methods currently used.

2.2 BAYES' THEOREM

There have been two general contending views on probability: the classical interpretation, which views probability as a physical property of the world (e.g., the probability that a dice will land in one of its 6 faces), and the views expressed by Thomas Bayes (Bayes, 1764) and Pierre Simon de Laplace (de Laplace, 1820). The latter views express probabilities as subjective degrees of belief, where the probability of an event is interpreted as a person's *degree of belief* in that event (Heckerman, 1997; Korb and Nicholson, 2011). Bayesian probability, therefore, is a property of the individual who assigns the probability, while the classical probability of an event is referred to as the true or physical probability.

The Bayesian philosophy to probability and statistics is important for understanding BN techniques. The general assertion of Bayesian philosophy is that in order to understand human opinion, constrained by ignorance and uncertainty, probability calculus is the most important tool for representing appropriate strengths of belief. The origin of Bayesian thinking lies in the interpretation of Bayes' theorem or law: given two events h and e such that $P(e) \neq 0$ and $P(h) \neq 0$, then:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)} \quad (2.1)$$

The theorem asserts that the probability of an event (or hypothesis) h conditioned upon some other event (or evidence) e is equal to its likelihood $P(e|h)$ times its probability prior to any evidence $P(h)$, normalized by dividing by $P(e)$ so that the conditional probabilities of all hypotheses sum to 1. The term $P(h|e)$ is often known as the *conditional probability* or posterior (or *a posteriori*) probability of h given e while $P(e|h)$ is the likelihood of e given h . The term $P(h)$ is the prior or marginal probability of h . Bayes' formula (equation 2.1) is fundamental to many contemporary machine learning techniques. Given the conditional probability formulation, it is now possible to define what it means for events to be conditionally independent. The events h and e are independent if $P(h|e) = P(h)$ and $P(e|h) = P(e)$. It follows then that if both $P(h)$ and $P(e)$ are positive, then both $P(h|e)$ and $P(e|h)$ imply the other. This notion of conditional (in)dependence is fundamental to BNs and the interpretation of

probabilistic relationships. In the following subsequent subsections, BNs are defined and the basic tenets of how BNs are learned from data are explained.

2.3 BAYESIAN NETWORK DEFINITION

Judea Pearl (Pearl, 1982; Pearl, 1988), formally introduced BNs including the term itself, in the 1980s although probabilistic graphical models in general have been in use for the last 50 years. BNs are also known as recursive graphical models (Lauritzen, 1995), Bayesian belief networks (Cheng *et al.*, 1997), belief networks (Darwiche, 2002), causal probabilistic networks (Jensen *et al.*, 1990), causal networks (Heckerman, 2007), influence diagrams (Shachter, 1986), to name some of the used terminologies. However, the term “Bayesian network” has become the predominant description of this type of graphical model and it is the term used in this thesis.

A BN is essentially a graphical representation of a probability distribution over a set of variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, $n \geq 1$. Formally, a BN consists of two parts, $B = \langle G, \theta \rangle$, where G is a directed network structure in the form of a directed acyclic graph (DAG), and θ is a set of the local probability distributions for each node/variable, conditional on each value combination of the parent nodes. The graphical component, G makes BNs a class of probabilistic graphical models for reasoning under uncertainty, where the nodes represent variables (which can be discrete and/or continuous) and the arcs represent direct (and sometimes causal) connections and dependencies between the linked variables. Those variables that are not linked directly in the graph are conditionally independent of each other. The second part of the network, θ represents the conditional probability distributions, which model the quantitative strength of the connections or dependencies between variables. These are represented through conditional probability tables (CPTs), allowing probabilistic beliefs to be updated automatically as new information becomes available. The local probability distributions contain a parameter $\theta_{x_i|\Pi_{x_i}} = P_B(x_i|\Pi_{x_i})$ for each possible value x_i of X_i , given each combination of the direct parent variables of X_i , Π_{x_i} of Π_{X_i} , where Π_{X_i} denotes the set of direct parents of X_i in G . The network G then represents the following JPD:

$$P_G(X_1, \dots, X_n) = \prod_{i=1}^n P_G(X_i | \Pi_{X_i}) = \prod_{i=1}^n \theta_{X_i | \Pi_{X_i}} \quad (2.2)$$

This is the chain rule, which states that for a given set of variables \mathbf{X} , the BN specifies a unique JPD $P_G(\mathbf{X})$ given by the product of all the CPTs specified in the BN. This is one of a number of important features of BNs. For this class of graphical models, the graph G is restricted to be acyclic because the BN represents a minimal set of dependencies coded in a particular factoring (equation 2.2) of a JPD. The JPD might have multiple mathematically equivalent factorings that represent the same dependence and conditional independence relationships. This suggests that one probability distribution can be represented with equal validity by an equivalence class, the collection of BNs representing the same JPD with differences only in the direction of (some of) their arcs. The fact that each parameter $\theta_{x_i|\Pi_{x_i}} = P_G(x_i|\Pi_{x_i})$ in a discrete BN can be specified independently implies that the statistical relationship between a child and its parents is denoted by arbitrary combinatorics, which simplifies the modelling of non-linear, stochastic and non-additive relationships. This ability to model statistical relationships without the need to specify the form of the dependency is one of the reasons that discrete BNs have been used in so many different domains such as the applications in this thesis.

The dual nature of a BN, therefore, makes the learning process a two-stage activity consisting of structure learning and parameter learning, the main purpose of which is to graphically summarize conditional independence relations. The most challenging task is the former and research in this direction is growing because of its enormous usefulness as much for end-user applications as for the learning of causal networks in many domains. Typically, datasets contain variables that can be either discrete or continuous, and while BNs can handle them, the limitations are too restrictive. The most widely used solution is to discretize the continuous variables into pre-determined bins or using some thresholding criteria. However, discretization implies capturing a coarser view of the original distribution resulting in some loss of information (Landuyt *et al.*, 2013; Aguilera *et al.*, 2010). Hence, methods to simultaneously handle both continuous and discrete data have been proposed and these include the Conditional Gaussian networks (Lauritzen, 1992; Lauritzen and Jensen, 2001), the Mixture of Truncated Exponentials (Moral *et al.*, 2001), the Mixtures of Polynomials (Shenoy and West, 2011) and the Mixtures of Truncated Basis Functions models (Langseth *et al.*, 2012). However, the development of empirical discretization methods for predictor variables that are ecologically relevant and statistically rigorous is on-going (Lucena-Moya *et al.*, 2015) making discrete BNs even more

robust. An overview of the prevalent approaches and techniques used for learning discrete BN structures is presented in the following section. In the subsequent subsections, approaches to learning the parameters given the structure are briefly highlighted.

2.4 STRUCTURE LEARNING

The field of BNs covers a range of problems and techniques of data analysis and probabilistic reasoning, where data is collected on a large number of variables and the aim is to factorize the distribution, represent it graphically and exploit the graphical representation (Koski and Noble, 2012; Kjærulff and Madsen, 2013). Creating the BN structure is an attempt to develop an accurate graphical model for the data or problem being solved. There are generally two approaches to constructing BNs, although some applications integrate the two (Darwiche 2009; Korb and Nicholson, 2011). The first approach, which is largely subjective, is also called knowledge representation whereby the modeller uses domain or expert knowledge about cause and effect within the system to structure the graphical model and calculate the probabilities. Alternatively, this information may be derived from other formal knowledge sources such as blueprints, flow charts, or diagrams. This approach is useful in cases where data is limiting.

The second approach to constructing BNs is called machine learning or learning from data, which is the domain of this thesis. In data mining problems, the interest is in the search for relationships among a large number of variables. BNs are suited to this task because the graphical model efficiently encodes the JPD for a large set of variables. Ideally, the resulting BN structure should be able to effectively handle new instances of the data, sampled from the same underlying distribution. When building BNs from prior knowledge alone, the resulting probabilities are Bayesian whereas when learning BNs from data, the probabilities will be true or physical and their values may be uncertain (Heckerman, 1997). Learning BNs from data is computationally intensive and without restrictive assumptions, the task is non-deterministic polynomial-time (NP)-hard (Chickering, 1996). The primary difficulty with learning the structure that best represents the JPD that most closely encodes the dependencies and probability parameters matching those in the data is that the number of possible structures grows super-exponentially with the number of variables. The number of possible combinations $f(n)$ of DAGs of n variables is estimated using the recursive formula (Robinson, 1977):

$$f(n) = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i). \quad (2.3)$$

Therefore, it would take exponential time to learn the true BN structure if more variables are included in the data (i.e. if n in equation 2.3 is increased). For instance, $f(n=3) = 25$, $f(n=5) = 29281$, $f(n=10) \approx 4.2 \times 10^{18}$ and there are currently no methods available that would allow learning of the BN structure from data in polynomial time. Hence, approximate search techniques are required to identify good models (Chickering, 1996; Heckerman *et al.*, 1995; Jensen and Nielson, 2007; Korb and Nicholson, 2011). When using these techniques, the problem of learning a network from data is an optimization problem where the goal is to find a probabilistic model that maximizes the posterior probability of the network given a dataset of instances drawn from a multivariate JPD. Based only on data, the BN structure learning techniques discover a suitable equivalence class, either by finding a DAG within the equivalence class or by finding the essential graph (Koski and Noble, 2012).

The structure learning task consists of finding an appropriate BN given a data set D over a set of variables, \mathbf{X} . In the data mining approach, both the structure of a network and the CPTs are learned from data using one or more of several available algorithms. In general, learning the structure or topology of the network through machine learning or data mining uses two approaches: scoring-based and constraint-based approaches (Korb and Nicholson, 2011; Koski and Noble, 2012). In addition, dynamic programming, genetic algorithms, model averaging and evolutionary approaches and hybrid approaches have emerged as alternative approaches and are gaining popularity (Daly *et al.*, 2011; Koski and Noble, 2012). Hybrid algorithms combine constraint-based scores with search-and-score approaches. This may be done through various approaches such as the MMHC algorithm, L^1 -Regularisation, Gibbs sampling. The score-based and constraint-based approaches, which are used in this study, are briefly defined and discussed in the following subsections.

2.4.1 Search-and-score approaches

The scoring-based learning algorithms seek a structure that maximizes a scoring function, e.g. the Bayesian Information Criterion or Minimum Description Length (MDL) (Cooper and

Herskovits, 1992; Heckerman *et al.*, 1995). Scoring-based learning approaches consider the search as an optimization problem where the exploration of the network structures space is guided by a statistical metric that quantifies how each network models the data (Cooper and Herskovits, 1992). This approach aims to find the structure of the network by searching through a set of potential models and finding the network structure that has the highest score. Using a score function, a score is assigned to each potential network that has been calculated using the provided data. As noted before, the search space for BNs consists of a super exponential number of structures (equation 2.3). However, there is no efficient solution to finding the optimal structure using a score-based approach. Hence, heuristic search techniques are employed to find the best network in a given time. Several algorithms are used to search through the candidate network structures, returning the network with the highest score. These include Greedy Equivalence Search, Markov Chain Monte Carlo (MCMC), Optimal Reinsertion, and Sparse Candidate, amongst several others (Koski and Noble, 2012). It is important that an appropriate score function and a good search algorithm are used in order to return a network structure that is reasonably close to the ‘true’ network. Since the score function is applied on the entire network structure, search-and-score approaches are not severely affected by individual failures like constraint-based approaches. This makes them more flexible in the way variables are dependent on the data.

Two approaches to network scoring are normally implemented, namely local and global score metrics. When using local score metrics learning a network structure is considered an optimization problem where a quality measure of the given network structure needs to be maximized, given the training data. The quality measure may be based on a Bayesian approach, minimum description length, information or other metrics. These metrics have the practical property that the network score can be decomposed as the sum (or product) of the score of the individual nodes. This property allows for local scoring, hence they are referred to as local search methods. On the other hand, global score metrics use a natural way to measure the performance of a BN on a given data set by predicting its future performance through estimating expected utilities, such as classification accuracy (Witten *et al.*, 2011).

2.4.2 Constraint-based search

Constraint-based learning approaches use statistical tests to find dependency relations among the variables, which are then mapped to the BN structure (Spirtes *et al.*, 2000). These methods carry out tests on the so-called triples $(X, Y, S$, where X and Y are variables and S is a subset of variables), to decide whether X is conditionally independent of Y given another variable Z . The results of these tests become the constraints and a DAG that satisfies as many of the constraints as possible is then selected. The DAG structure encodes the conditional independence relationships among the nodes according to the d-separation concept (Pearl, 1988). This then suggests possible learning of the BN structure by identifying the conditional independence relationships among the nodes. Using predetermined statistical tests (such as mutual information) or score metrics, the conditional independence relationships among the attributes can be established and used as constraints to construct a BN. Hence, these are aptly called conditional independence-based algorithms (Spirtes *et al.*, 2000; Cheng *et al.*, 1997).

The conditional independence approach views the BN as a set of independencies where conditional dependence and independence is tested in the data to find an equivalence class of networks. The network is then slowly built up from the dependency tests between variables in the set of data. Constraint-based methods are however, subject to the performance of the individual independence tests that can determine the resultant overall BN structure and its accuracy. For instance, a set of variables that may seem independent of each other due to randomness may actually be related. The result from the dependency tests influences the structure learning process and ultimately the learned network structure for the data.

The conditional independence methods mainly stem from the goal of uncovering the causal structure based on the assumption that there is a network structure representing the independencies in the distribution given the data. Subsequently if a (conditional) independency is identified in the data between two variables then there would be no arc between those two variables or nodes and vice versa. Once locations of edges are identified, the direction of the edges is assigned such that conditional independencies in the data are properly represented.

2.5 PARAMETER LEARNING

After learning the BN structure, which consists of a DAG and specifications of the conditional probabilities corresponding to the factorization, the network may be used for computation of conditional probabilities. Although learning the parameters in a BN is an important task in itself, it is significant in the context of learning the BN structure. This is because many structure learning algorithms, particularly the search-and-score approaches, estimate parameters as part of the structure learning process. This does not imply that in learning a structure, parameters need to be explicitly represented and learned. However, it means that when scoring a network, an implicit parameterization is given.

The parameters that are learned in a BN depend on the assumptions that are made about how the learning is to proceed. For example, in the case of maximum likelihood learning, the parameters could be the actual probabilities in the CPT attached to each node. In the case of BN classifiers, these can be generative, discriminative or hybrid (generative-discriminative) in nature (Bielza and Larrañaga, 2014). Generative parameter learning involves learning a model of the JPD using Bayes's rule to compute the posterior probability of the class (target) variable. On the other hand, discriminative BNs directly learn the posterior probability of the class variable, which is the distribution used for classification. Hence, generative parameter learning maximizes the log-likelihood or a related function, whereas discriminative parameter learning maximizes the conditional log-likelihood (Bielza and Larrañaga, 2014).

Other parameter learning algorithms include the popular and efficient Lauritzen and Spiegelhalter's (1988) Aalborg algorithm which is useful when there is a large number of discrete variables, where each clique on the junction tree is relatively small (Koski and Noble, 2012). Other algorithms include Bayesian Model Averaging, Bayesian bootstrapping, and the simple Bayesian estimator. In a Bayesian setting, the parameters are used to specify a conditional density, which in turn models the probabilities in a CPT. Fitting parameters to a model, has been frequently attempted from the point of view of statistical machine learning. Good background material on this subject is given by Spiegelhalter and Lauritzen (1990), Neapolitan (2004), Buntine (1996) and Heckerman *et al.* (1995) and Bielza and Larrañaga (2014).

2.6 BAYESIAN NETWORK INTERPRETATION AND REASONING

To illustrate the process of interpreting a Bayesian network, one can consider an abstract situation in the field, a domain abstracted to five variables. *A* (representing the event that the land surface cover is natural), *B* (representing the event that the area is sandy soil), *C* (representing the event that another invasive plant species is present), *D* (representing the event that livestock density is high or low), and *E* (representing whether an invasive plant species of interest is present in a specified locality). A family metaphor is often used to describe a BN structure: a node is a parent of a child if there is an arc from the former to the latter. In the same way, if there is a directed chain of nodes, one node is an ancestor of another if it appears earlier in the chain; otherwise it is a descendant of another node if it comes later in the chain. In Figure 2.1, the native plant species *C* node has two parents, sandy soil and land cover, while sandy soil is an ancestor of both livestock density and invasive plant species. Similarly, invasive plant species is a child of livestock density and descendant of sandy soil and land cover.

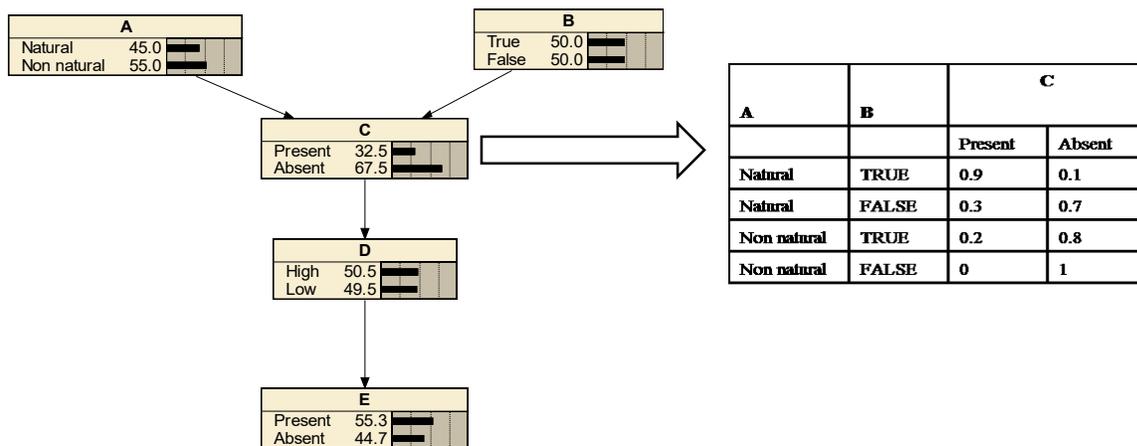


Figure 2.1: An example of a Bayesian network with a sample conditional probability table.

The BN topology captures qualitative relationships between variables whereby two nodes are connected directly if one depends on, affects or causes the other, whilst the arc indicates the direction of the dependence or effect. Intuitively, the BN structure in Figure 2.1 means that *C* depends on *A* and *B* but *A* and *B* are independent. Another implied statement is that *A* and *D* become independent once the value of *C* is known or fixed. In general, the interpretation of a BN is based on a set of statistical conditional independence statements that are implied by its

structure. Under certain assumptions, it includes statements of dependence among variables. From a set of axioms described in Pearl (1997), the entire set of independence relations that are implied by a particular BN model can be produced.

However, determining the independence relations that are entailed by the DAG from these statements can be cumbersome because it requires that they be used repeatedly until the desired relation is proved or otherwise. An equivalent approach that is frequently used is to learn those independencies from the structure of a BN model using the rules of dependence separation (*d*-separation) (Pearl, 1995, see section 2.5) which is significantly easier than using the set of axioms. In the domain of Figure 2.1 where all variables are binary, each row of each CPT records the probabilities of that variable taking the value “true” or “false” for a particular combination of values (“true” or “false”) of its parents. For example, given that there is natural land cover and sandy soil, the probability that an invasive plant species is “present” is 0.9.

In addition to the family analogy, other commonly used terms come from the “tree” analogy (even though BNs in general are graphs rather than trees): any node without parents is called a root node, while any node without children is called a leaf node. Any other node (non-leaf and non-root) is called an intermediate node. Given a causal understanding of the BN structure, the root nodes represent original causes, while leaf nodes represent the final effects. However, that is only true under certain assumptions, the most important being:

1. whether there are any common unobserved (latent or hidden) causes (variables) of two or more observed variables in the domain. If there are no such variables, the property of causal sufficiency is said to hold true.
2. whether it is possible, given causal sufficiency, for more than one BN structure to fit the constraints that have been observed in the domain. These constraints are statistical independencies that are observed in the data. Only one of these networks can be the ‘true’ underlying generative model that embodies the real cause-effect relationships that govern the data-generating mechanisms of the domain.

In this example, the causes sandy soil and land cover are root nodes, while the effect invasive alien plant is a leaf node. By convention, for easier visual examination of BN structure, networks

are usually laid out so that the arcs generally point from top to bottom. This means that the BN “tree” is usually depicted upside down, with roots at the top and leaves at the bottom.

Furthermore, BNs can be used as causal models using the usual cause-effect interpretation when the above assumptions hold. A causal network is attained when the DAG along which the probability distribution factorizes is considered to have a causal interpretation, i.e. the parents of a variable are those that have a direct causal effect on a variable (Koski and Noble, 2012). In this case, the conditional probabilities are the basic building blocks to constructing probability distributions even for larger systems. In a BN, there are three ways in which two variables with no direct connection between them can be connected via a third; the diverging (fork), serial (chain) and converging (colliding) connections, respectively. These have clear interpretations when the BN has been derived from causal principles (Daly *et al.*, 2011, Kjærulff and Madsen, 2013 and Koski and Noble, 2012).

The diverging or fork connection is illustrated in Figure 2.2 where the variable C is a common cause. A probability distribution over the variables A, B, C that factorizes according to the BN in Figure 2.2 is expressed as

$$P(A, B, C) = P(C)P(A|C)P(B|C) \quad (2.3)$$

It should be noted though that A and B are conditionally independent given C , but A and B are not, or at least not necessarily. If a causal interpretation is valid, then the variable C is a common cause. There is an association between A and B , i.e. they are not independent of each other. However, this association may be explained fully through the state of the hidden variable C .

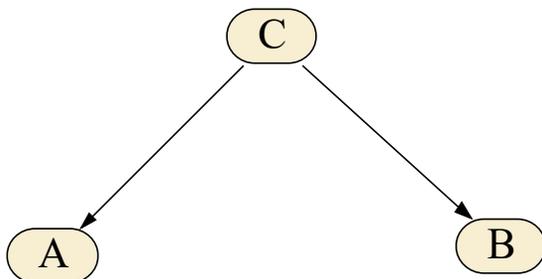


Figure 2.2: A diverging or fork connection (adapted from Kjærulff and Madsen, 2013).

The diverging connection can be summarized as follows: with no evidence on C , evidence on A will affect the belief about the state of B and vice versa. However, with hard evidence on C , evidence on A will not affect the belief about the state of B and vice versa. This is the basis for the concept of conditional independence.

The serial or chain connection is illustrated in Figure 2.3. This describes a situation where the association between A and B is only through C ; A has a causal influence on C , which in turn has a causal influence on B . A probability distribution over (A, B, C) that factorizes along the graph in Figure 2.3 may be written as:

$$P(A, B, C) = P(A)P(C|A)P(B|C) \quad (2.4)$$

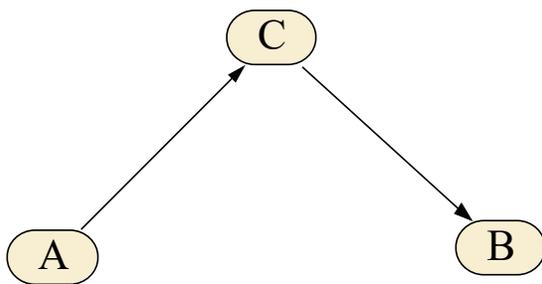


Figure 2.3: A serial or causal chain connection (adapted from Kjærulff and Madsen, 2013).

In the same way as the diverging connection, A and B are conditionally independent given C , but A and B are not, or at least not necessarily. In the case of a serial (causal chain) connection with no hard evidence on C , evidence on A will affect the belief about the state of B and vice versa. However, with hard evidence on C , evidence on A will have no effect on the belief about the state of B and vice versa.

The converging connection is illustrated in Figure 2.4. Here A and B both have a causal influence on C . This corresponds to a factorization:

$$P(A, B, C) = P(A)P(B)P(C|A, B) \quad (2.5)$$

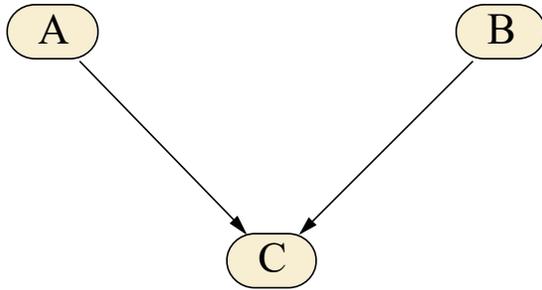


Figure 2.4: A converging or colliding connection (adapted from Kjærulff and Madsen, 2013).

For variables that factorize according to a collider, the properties are the opposite; A and B are not conditionally independent given C , but A and B are; at least they are not necessarily conditionally independent. If there is information on the state of C , then there will be a flow of information between A and B . This can be interpreted using the classic example: that of $A =$ ‘burglary’, $B =$ ‘earth tremor’ and $C =$ ‘alarm’. An earth tremor and a real burglary can both set off the burglar alarm. If the burglar alarm rings, the information that there has been an earth tremor in the area will reduce one’s fear that there may have been a real burglary. Hence, information passes between A and B only if there is information on C . For a converging connection with no evidence on C or any of its descendants, information about A will not affect the belief about the state of B and vice versa. However, with (possibly soft) evidence on C or any of its descendants, information about A will affect the belief about the state of B and vice versa.

The property of converging connections, $A \rightarrow B \leftarrow C$, that information about the state of A (C) provides an explanation for an observed effect on B , and hence confirms or dismisses C (A) as the cause of the effect, is often referred to as the “explaining away” effect or as “intercausal inference” (Koski and Noble, 2012). The ability to perform such intercausal inference is unique for graphical models, and is one of the key differences between automatic reasoning systems based on probabilistic networks and systems based on, for example, production rules where there would be a need for dedicated rules for taking care of intercausal reasoning. The principle of the common cause states that if two variables are probabilistically dependent, then either one causes the other (directly or indirectly) or they have a common ancestor (Korb and Nicholson, 2011)

Two nodes A and B are said to be d -separated by a set S if all paths between A and B , have either a diverging or serial connection in S , or a converging connection not in S , with none of its descendants in S . The BN is said to be ‘faithful’ when d -separation statements for the DAG and independence statements for the probability distribution are equivalent. An ordering of the variables gives a factorization and a corresponding BN. An efficient factorization is one where the d -separation statements of the BN represent as much of the independence structure as possible. If there is a causal structure such as one derived using the constraint-based approach, then this provides a natural ordering of the variables, a factorization and an efficient BN with a causal interpretation.

Another important feature for a node in a BN is its Markov blanket (Pearl, 1988). The Markov blanket of a node X in a BN consists of the set of parents and children and parents of children of the variable. An important characteristic of the Markov blanket is that X is d -separated from the remaining variables in the network by its Markov blanket. Hence, the Markov blanket of the target variable Y is the set of variables S that make Y conditionally independent of the other variables in the network, given the set of variables S , i.e.

$$p(y|\mathbf{x}) = p(y|\mathbf{x}_S),$$

where \mathbf{x}_S denotes the projection of \mathbf{x} onto the variable set S . Consequently, the Markov blanket of Y is the only information needed to predict its behaviour. A probability distribution p is faithful to a DAG representing a BN if, for all triples of variables (X, Y, S) , they are conditionally independent with respect to p and if they are d -separated in the DAG.

2.7 APPLICATIONS IN THE SPECIES DISTRIBUTION MODELLING DOMAIN

The ability of BNs to model complex and uncertain systems has triggered an upsurge in applications in the biological sciences mainly in areas of bioinformatics such as inferring cellular networks, classification, data integration, genetic data analysis, modelling protein signalling pathways and systems biology (Neapolitan, 2009). This has made them attractive for applications in ecology over the past two decades as evidenced by the number of studies in literature (Varis and Kuikka, 1999; McCann *et al.*, 2006; Uusitalo, 2007; Henriksen and Balebo, 2008; Aguilera *et al.*, 2011; Barton *et al.*, 2012; Landuyt *et al.*, 2013). However, BN modelling has had limited application in species distribution modelling where the goal is to relate observed species occurrence (target variable) to environmental (attribute) variables in order to predict distributions over the entire area of interest.

Earlier applications have used Bayes' theorem to combine relationships between observed data and individual predictive factors with prior probabilities of presence to produce probability surfaces for species (Aspinall, 1992; Aspinall and Veitch, 1993; Royle *et al.*, 2002). Wikle (2002, 2003), Wikle and Royle (2004) and Clark *et al.* (2004) recently presented examples of full Bayesian hierarchical modelling applied to individual plant or bird species. However, since these approaches use a contingency table approach and only carry over point estimates from the data stage to the generation of predictions, they are not directly comparable to BN models. Several assessments of wildlife habitats have used BNs (e.g., Raphael *et al.*, 2001, Lee and Irwin, 2005, McNay *et al.*, 2006, Smith *et al.*, 2007). Ropero *et al.* (2014) present BNs as a tool to solve different problems in species distribution models such as classification, characterization and regression. As discussed in Section 1.2, most BN applications have focused on developing habitat suitability indices mainly based on expert opinion and elicitation.

The Web of Science database and other websites such as Google Scholar and Scopus were searched for BN applications by using the keywords “Bayesian network” or “Bayesian belief network” and “species distribution model”, “ecological niche model” or “habitat suitability”. Only papers published between January 1990 and December 2015 were considered and these were screened to determine whether they sufficiently dealt with species distribution or habitat modelling within the scope of this study. The analysis specifically focused on several

characteristics of the model development process: the approach used to develop the DAGs and learn the parameters or populate the CPTs, the species or taxa studied, the approach used to select predictor variables, whether a spatial output/map was produced and the BN software package used. Table 2.1 provides a summary of the 88 studies that used BNs for habitat modelling or species distribution modelling with a view to highlight gaps on the use of data-driven or machine learning approaches.

The findings in Table 2.1 indicate that, although BNs evolved in the late 1980s, their use for species distribution modelling started in the late 1990s and only gained momentum in the past decade (Figure 2.5). Hence, it can be seen that BN use in this domain is relatively new and growing. The full potential of BN applications to species distribution modelling problems remains largely untapped (Ropero *et al.*, 2014). However, it is also encouraging to note that a wide variety of organisms have been studied using BNs covering both fauna and flora, including microorganisms. The widespread use of the software Netica (Norsys Software Corporation, 2014a), which is widely used for BN applications in the environmental sciences, is also evident with 61% of the studies utilizing it. This is widely attributed to the user-friendly graphical user interface and the ease of manually constructing and parameterizing BNs. However, the major limitation of this software is the lack of structure learning algorithms. This is evidenced by the fact that in all the applications using this software, only 8% of the DAGs were learned solely from data, the rest were constructed manually using expert knowledge, literature or a combination of the two. It is evident, therefore, that the software used determined the approach used to develop the BN structures. An overview of the studies indicates that a majority (90%) used the converging or colliding arc topology whereby the linkages between variables (depicted as parent nodes) and the prediction (child) node were pre-determined typically following the guidelines prescribed by Marcot *et al.* (2006).

Table 2.1: List of scientific applications of BN models in species distribution modeling¹.

Author	Species/Taxa	Structure learning	Parameter learning	Variable selection	Map output	Software
Adriaenssens <i>et al.</i> (2004)	Macroinvertebrates	E	D	E	No	Matlab
Alameddine <i>et al.</i> (2011)	Phytoplankton/Algae	D	D	D	Yes	Hugin
Allan <i>et al.</i> (2012)	Macroinvertebrates	E	D	E	No	Netica
Altartouri and Jolma (2013)	Common reed (<i>Phragmites australis</i>)	D	D	D	Yes	Not stated
Aps <i>et al.</i> (2009)	Birds and seals	E	E/M	E	No	Not stated
Ayre <i>et al.</i> (2014)	Colorado river cutthroat trout (<i>Oncorhynchus clarkii pleuriticus</i>), Rio Grande cutthroat trout (<i>Oncorhynchus clarkii virginalis</i>)	E	E/D/L	E	No	Netica
Ban <i>et al.</i> (2015)	Corals	D/E	D/E	E	Yes	Netica
Bashari and Hemami (2013)	Wild sheep (<i>Ovis orientalis</i>)	E/L	E	E/L	No	Netica
Boets <i>et al.</i> (2015)	Macroinvertebrates (gammarids)	D/E	D	E	Yes	Netica/R
Borsuk <i>et al.</i> (2004)	Algae	E/L	E/M	E/L	No	Analytica
Borsuk <i>et al.</i> (2006)	Brown trout (<i>Salmo trutta</i>)	E	D/L/E	E	No	Analytica
Burkhardt-Holm (2008)	Brown trout (<i>Salmo trutta</i>)	E/L	D/M	E/L	No	Not stated

¹ (E – expert/domain knowledge (including stakeholder knowledge), L - literature, M - model simulations, D - empirical data)

Author	Species/Taxa	Structure learning	Parameter learning	Variable selection	Map output	Software
Chan <i>et al.</i> (2012)	Barramundi (<i>Lates calcarifer</i>) and sooty grunter (<i>Hephaestus fuliginosus</i>)	E	E/D	E	No	Netica
Chen and Pollino (2012)	Giant freshwater crayfish (<i>Astacopsis gouldi</i>)	E	D/E	E/L	Yes	Netica
Copps <i>et al.</i> (2007)	Fish	E/L	E/M	E/L	Yes	Not stated
Douglas and Newton (2014)	Four plant species, 2 butterfly species, 1 Orthoptera, 1 fungus species	E/L	E	E	Yes	Hugin
Falke <i>et al.</i> (2015)	Bull trout (<i>Salvelinus confluentus</i>)	E/L	D/L	E/L	No	Netica
Fu <i>et al.</i> (2015)	16 vegetation species, 13 waterbird species, 4 fish groups	E/L	D/M	E/L	No	ICMS
Gawne <i>et al.</i> (2012)	Fish (<i>Retropinna semoni</i> , <i>Hypseleotris</i> spp., <i>Macquaria ambigua</i> , <i>Cyprinus carpio</i>)	E	E/D	E	No	Netica
Gibbs (2007)	Birds	E/L	D/L/M	E/L	No	Not stated
Gieder <i>et al.</i> (2014)	Piping plover (<i>Charadrius melodus</i>)	E/L	D	E	Yes	Netica
Goudarzi <i>et al.</i> (2015)	Persian fallow deer (<i>Dama mesopotamica</i>)	E/L	E/L	E/L	No	Netica
Grech and Coles (2010)	Coastal seagrass	E	D	E/L	Yes	SamIam
Haas (1991)	Aspen (<i>Populus</i> spp.)	L	L	L	No	Not stated
Hamilton <i>et al.</i> (2007)	<i>Lyngbya majuscula</i>	E	D/E/M	E	No	Netica
Hamilton <i>et al.</i> (2015)	Giant freshwater crayfish (<i>Astacopsis gouldi</i>)	E/D	E/D	E/L	No	Netica
Helle <i>et al.</i> (2011)	Grey seal (<i>Halichoerus grypus</i>), common eider (<i>Somateria mollissima</i>), blue mussel (<i>Mytilus trossulus</i>), Baltic herring (<i>Clupea harengus membras</i>), prickly saltwort (<i>Salsola kali kali</i>), scarab beetle (<i>Aegialia arenaria</i>)	E/L	E/L/D	E/L	No	Hugin

Author	Species/Taxa	Structure learning	Parameter learning	Variable selection	Map output	Software
Howes <i>et al.</i> (2010)	Noisy Miner (<i>Manorina melanocephala</i>)	E	D	E	No	Netica
Huang <i>et al.</i> (2013)	Pinewood nematode (<i>Bursaphelenchus xylophilus</i>)	D	D	E	Yes	Matlab
Jay <i>et al.</i> (2011)	Pacific walrus (<i>Odobenus rosmarus divergens</i>)	E	E/L	E	No	Netica
Jellinek <i>et al.</i> (2014)	Reptile (Class Reptilia) and beetle species (Order Coleoptera)	E	E/D	E/D	No	Netica
Johnson <i>et al.</i> (2010a)	Cheetah (<i>Acynonyx jubatus</i>)	E	E	E	No	Hugin/IBN DC (UML)
Johnson <i>et al.</i> (2010b)	<i>Lyngbya majuscula</i>	E	D/E/M	E	No	Netica/Hugin
Kuikka <i>et al.</i> (1999)	Baltic cod (<i>Gadus morhua</i>)	E	D/E/M	E	No	Hugin
Laws and Kesler (2012)	Guam Micronesian kingfishers (<i>Todiramphus cinnamominus cinnamominus</i>)	E/L	D/L	L	No	Netica
Lee (2000)	Bull trout (<i>Salvelinus confluentus</i>)	E	E/M	E	No	Netica
Lecklin <i>et al.</i> (2011)	Terrestrial plants, fish, birds, mammals, littoral macrofauna, macrophytes.	E/L	E	E/L	No	Hugin
Lehmkuhl <i>et al.</i> (2001)	Elk (<i>Cervus elaphus</i>), mule deer (<i>Odocoileus hemionus</i>), white-tailed deer (<i>Odocoileus virginianus</i>)	E/L	D/E/L	E/L	Yes	Netica
Liedloff <i>et al.</i> (2013)	Barramundi (<i>Lates calcarifer</i>), Black Bream (<i>Hephaestus jenkinsi</i>), catfish (<i>Arius sp.</i>), Freshwater Mussels (<i>Velesunio sp.</i>), Long-necked Turtle (<i>Chelodina sp.</i>), Sawfish (<i>Pristis sp.</i>),	E	D/E	E	No	Netica
Liu <i>et al.</i> (2013)	<i>Cyprinus carpio</i>	E/L	M	L	No	Hugin
Liu <i>et al.</i> (2015)	Pheasant-tailed jacana	E/L	E/M	E/L	No	Netica
MacCracken <i>et al.</i> (2013)	Pacific walrus (<i>Odobenus rosmarus divergens</i>)	E/L	E/M	E/L	No	Netica
Mantyka-Pringle <i>et al.</i> (2014)	Macroinvertebrates and fish	E/L	D/E/M	E	No	Netica

Author	Species/Taxa	Structure learning	Parameter learning	Variable selection	Map output	Software
Marcot (2006)	Fungus (<i>Bridgeoporus nobillissimus</i>)	E	E/D	E	No	Netica
Marcot <i>et al.</i> (2001)	Fish and vertebrates	E/L	E/D	E/L	No	Netica
Marcot <i>et al.</i> (2012)	Chinook salmon (<i>Oncorhynchus Tshawytscha</i>), coho salmon (<i>O. kisutch</i>), winter steelhead (<i>O. mykiss</i>)	E	E	E	No	Netica
Martin <i>et al.</i> (2015)	Buffel grass (<i>Cenchrus ciliaris</i>)	E/L	E	E	Yes	Netica
McNay <i>et al.</i> (2006)	Mountain caribou (<i>Rangifer tarandus caribou</i>)	E	E/D	E	Yes	Netica
McNay <i>et al.</i> (2011)	Grizzly (<i>Ursus arctos</i>), caribou (<i>Rangifer tarandus</i>), wolverine (<i>Gulo gulo</i>), fisher (<i>Martes pennanti</i>), Spruce Grouse (<i>Falcapennis canadensis</i>), marten (<i>Martes americana</i>), Lewis's Woodpecker (<i>Melanerpes lewis</i>), red squirrel (<i>Tamiasciurus hudsonicus</i>), badger (<i>Taxidea taxus</i>), ermine (<i>Mustela erminea</i>)	E	D/M	E/L	Yes	Netica
Meineri <i>et al.</i> (2015)	Vascular plants (<i>Actaea spicata</i> , <i>Convallaria majalis</i> , <i>Hepatica nobilis</i> and <i>Carex digitate</i>)	D/E	D/M	L	No	R (bnlearn)
Mello <i>et al.</i> (2013)	Soybean	D	D	E/L	Yes	R (BayNeRD)
Milns <i>et al.</i> (2010)	Birds	D	D	E	No	Banjo
Murray <i>et al.</i> (2012)	Lippia (<i>Phyla canescens</i>)	E	D/E	E	Yes	Netica
Murray <i>et al.</i> (2014)	European rabbit (<i>Oryctolagus cuniculus</i>)	E/L	E	E/L	Yes	Netica
Murty <i>et al.</i> (2009)	Mosquitoes (<i>Culex tritaniorhynchus</i> , <i>Culex psuedovishmii</i> , <i>Culex vishmii</i> , <i>Culex gelidus</i> , <i>Culex quinquefasciatus</i>)	D/E	D	E	No	JEBNET
O'Brien (2006)	Coffee and Cowpea (<i>Vigna unguiculata</i>)	D	D	E/L	Yes	CaNaSTA
Pellika <i>et al.</i> (2005)	Forest grouse, small predators, large predators, ungulates, mountain hare	E	E	E	No	FC BeNe

Author	Species/Taxa	Structure learning	Parameter learning	Variable selection	Map output	Software
Peterson <i>et al.</i> (2008)	Westslope cutthroat trout (<i>Oncorhynchus clarkii lewisi</i>) and brook trout (<i>Salvelinus fontinalis</i>)	E	E/D	E/L	No	Netica
Peterson <i>et al.</i> (2013)	Bull trout (<i>Salvelinus confluentus</i>)	E	M	E	Yes	Netica
Pollino <i>et al.</i> (2007)	Swamp Gum (<i>Eucalyptus camphora</i>)	E/L	D	E/L	No	Netica
Pullar and Phan (2007)	Koala	E	D	E	Yes	Netica
Qian and Miltner (2015)	Ephemeroptera (mayfly), Plecoptera (stonefly), and Trichoptera (caddisfly)	E/L	D/M	E/L	No	WinBUGS
Raphael <i>et al.</i> (2001)	Thirty one vertebrates (primarily pygmy nuthatch <i>Sitta pygmaea</i> , sage grouse <i>Centrocercus</i> spp., and wolverine <i>Gulo gulo</i>)	E	D	E	Yes	Netica
Rehr <i>et al.</i> (2014)	Eelgrass (<i>Zostera marina</i>)	E	E/M	E	No	Netica
Renken and Mumby (2009)	Common Caribbean macroalgae (<i>Dictyota</i> spp.)	E/L	D	E/L	No	Netica
Rieman <i>et al.</i> (2001)	Six salmonid fish species	E/L	E/D	E/L	Yes	Netica
Rowland <i>et al.</i> (2003)	Wolverine (<i>Gulo gulo</i>)	E/L	E/D	E/L	No	Netica
Rüger <i>et al.</i> (2005)	Euphratica poplar (<i>Populus euphratica</i> , syn. <i>ariana</i>)	E/L	E/D	E/L	Yes	Not stated
Shenton <i>et al.</i> (2011)	Fish (Australian grayling and River blackfish)	E	D/E/M	E	No	Netica
Shenton <i>et al.</i> (2014)	Fish (Australian grayling)	E	D/E/M	E	No	Netica
Murray <i>et al.</i> (2014)	European rabbit (<i>Oryctolagus cuniculus</i>)	E/L	E	E/L	Yes	Netica
Smith <i>et al.</i> (2007)	Julia Creek dunnart (<i>Sminthopsis douglasi</i>)	E/L	D/E	E/L	Yes	Netica

Author	Species/Taxa	Structure learning	Parameter learning	Variable selection	Map output	Software
Smith <i>et al.</i> (2012)	Parkinsonia (<i>Parkinsonia aculeata</i>)	E	E	E	Yes	Netica
Stafford <i>et al.</i> (2015)	<i>Nucella lapillus</i> , <i>Osilinus lineatus</i> , <i>Patella vulgata</i> , <i>Littorina littorea</i> , <i>Ulva spp.</i> , <i>Coarallina officinalis</i> , <i>Fucus vesiculosus</i> , <i>Cthamalus</i> , <i>Semibalanus</i>	E/L	E/M	E/L	No	MS Excel
Steventon and Daust (2009)	Mountain pine beetle	E	D/E/M	E	No	Netica
Suring <i>et al.</i> (2011)	Birds, mammals, amphibians, reptiles (primarily wolverine (<i>Gulo gulo</i>) and northern goshawk (<i>Accipiter gentiles</i>))	E	E/M	E/L	No	Netica
Tantipisanuh <i>et al.</i> (2014)	Fifty vertebrates	E	E/L	E/L	Yes	Netica
Tattari <i>et al.</i> (2003)	Plant, insect and bird species	E	E	E	No	FC BeNe
Ticehurst <i>et al.</i> (2007)	Fish, threatened fauna and flora	E	D/E/M	E	No	ICMS
Trifonova <i>et al.</i> (2015)	Fish, birds, mammals, zooplankton, phytoplankton	D	D	E/L	Yes	Matlab
Uusitalo <i>et al.</i> (2011)	Eleven fish species	E/L	D	E	No	Hugin
van Klinken and Murray (2011)	Parthenium (<i>Parthenium hysterophorus</i>)	E	M	E	Yes	Netica
Van Klinken <i>et al.</i> (2015)	Chilean needle grass (<i>Nassella neesiana</i>)	E/L	E/M	E	Yes	Netica
Varis and Kuikka (1999)	Fish	E/L	E/M	E/L	No	Not stated
Vilizzi <i>et al.</i> (2013)	Fish (golden perch <i>Macquaria ambigua</i> , carp gudgeon <i>Hypseleotris spp.</i> , Australian smelt <i>Retropinna semoni</i>) and common carp <i>Cyprinus carpio carpio</i> .	E	E	E/L	No	Netica
Voie (2003)	Large blue butterfly (<i>Maculinea arion</i>) and other threatened species	E/L	E/M	E	No	Netica

Author	Species/Taxa	Structure learning	Parameter learning	Variable selection	Map output	Software
Wilhere (2012)	Florida scrub-jay (<i>Aphelocoma coerulescens</i>)	L	M/D	L	No	Netica
Wilson <i>et al.</i> (2008)	Pacific giant salamander (<i>Dicamptodon tenebrosus</i>), tailed frogs (<i>Ascaphus truei</i>), southern torrent salamander (<i>Rhyacotriton variegatus</i>), and Columbia torrent salamander (<i>Rhyacotriton kezeri</i>)	E	M/D	E	No	WinBUGS

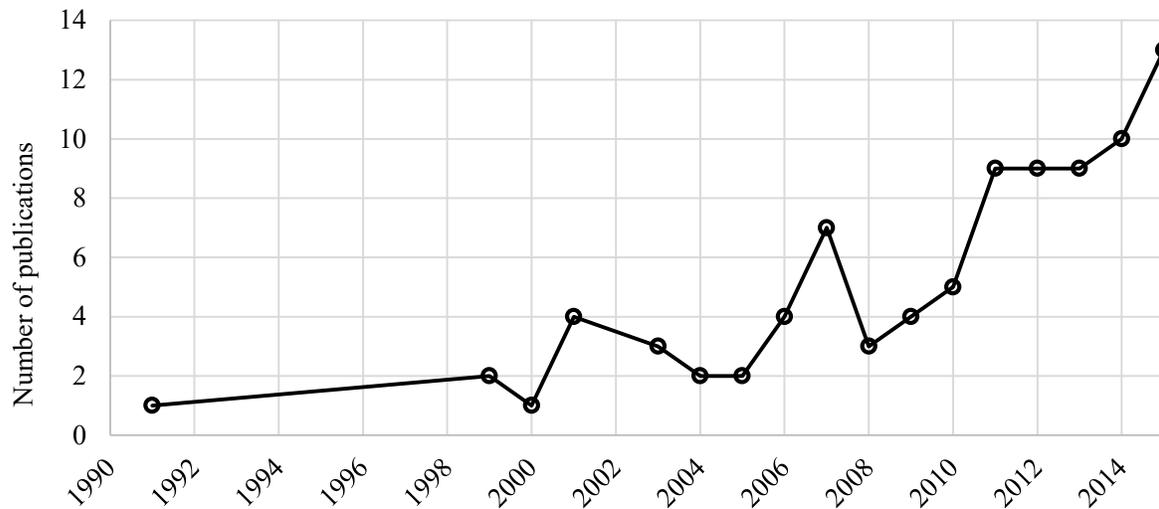


Figure 2.5: The number of Bayesian network-based species distribution modelling publications produced between the years 1990 and 2015.

Nevertheless, the tree augmented naïve (TAN) structure learning algorithm (Friedman *et al.*, 1997) was added in recent updates to the Netica software. Furthermore, Fienen and Plant (2015) recently extended Netica’s capabilities through an open-source Python package CVNetica which enables the software package to perform cross-validation and to read, rebuild, and learn BNs from data. This presents an opportunity to experiment with objective data-driven learning of BN structures. Hugin (Hugin, 2014) is another software that is used in the domain although less frequently. Additional to the functions available in Netica, Hugin also offers structure learning from data. The choice of variables is predominantly (>90%) determined *a priori* from expert knowledge and/or literature review of each species’ ecology. Only 2% of the studies selected variables through feature selection or automated methods. This is, however, a common practice in the entire species distribution modelling domain where the predictor variables are often chosen based on existing knowledge of a species.

Recent reviews (Varis and Kuikka, 1999; McCann *et al.*, 2006; Uusitalo, 2007; Henriksen and Balebo, 2008; Newton, 2009; Aguilera *et al.*, 2011; Barton *et al.*, 2012; Landuyt *et al.*, 2013) indicate that most of the applications of BNs were applied within the context of expert and/or stakeholder-driven inference tools for decision support or risk analysis. As such, most of those applications have used deductive methods that rely on qualitative data or knowledge elicitation

from experts or stakeholders for structure construction and parameterization. These applications are largely influenced by the methodological guidance provided by Marcot *et al.* (2006). This is often attributed to the fact that, for most of the studies, there are small quantities of replicated data needed for data mining (Newton *et al.*, 2007; Hamilton *et al.*, 2015; Martin *et al.*, 2015). As a result, the deductive process generally relies more on subjective information that may potentially limit the model by only utilizing factors that are comprehensible to the individual at that point in time. Smith *et al.* (2012) suggest that expert knowledge may be necessary to define both the key variables of habitat suitability and the species-environment relationships in order to increase model reliability and applicability in new situations and for new or recently introduced species.

The current habitat suitability modelling approaches tend to limit the number of variables used and the number of discretized states thereby oversimplifying the complexity of ecological systems and limiting their robustness. Limiting variables to those historically known to experts or those traditionally used limits the ability of the field to take advantage of opportunities for new ecological knowledge that could be derived from the wide variety of environmental and other data sets which so characterizes the high volume (big) data era. The expert and stakeholder-driven approach is difficult and time consuming, among other limitations highlighted earlier, especially when considering the increasing volume and heterogeneity of data. Complex models that explore larger sets of nonlinear features and interactions are appropriate for generating hypotheses about underlying ecological processes that would not often be identified with simpler models (Merow *et al.*, 2014). Hence, there is growing interest in automated or data-driven methods for learning BNs from the vast amounts of data now available.

However, Ahmed *et al.* (2015) observed that most traditional SDM scientists lack the skills on the use programming languages such as Visual Basic, MATLAB, Fortran, C, C++, Java, C#, # and Python. The two software packages (Netica and Hugin) continue to dominate despite the fact that there is an increasing availability of powerful open-source software such as R (R Development Core Team, 2008) for species distribution modelling as observed by Ahmed *et al.* (2015) and Joppa *et al.* (2013). Other BN learning packages such as Weka, Bayeserver, Analytics, Agenarisk, amongst a few others listed on Kevin Murphy's webpage (Murphy, 2014)

were not used in the studies under review. In all the papers reviewed, there are less than a handful of notable publications taking advantage of the power of open source software platforms such as R and Python even though several packages based on these platforms offer a number of robust BN learning algorithms. The R package BayNERD (Mello *et al.*, 2013), has a built-in capability to undertake species distribution modelling and automates the process of model development although the BN structure is still learned manually from expert/domain knowledge. However, this package is evolving to include parallel processing, automated structure learning and spatial features (neighbourhood information) in the probability computation (de O. Silva *et al.*, 2014).

Estimation of the probability of occurrence of a species, conditional on environmental or explanatory variables, is one of the fundamental tasks of any SDM (Phillips and Elith, 2013). Concerning parameter learning, there is a similar trend from the reviewed BN applications where the usual model development protocol is applied wherein experts and stakeholders are first consulted to develop a basic DAG structure (Landuyt *et al.*, 2013). This is then followed by the use of experts to populate the CPTs. Only in situations where data is sufficient are the CPTs populated directly from empirical data alone predominantly using the expectation-maximization (EM) technique and structural equation modelling based on observed species-environment relationships. From the studies, only 22% learned parameters solely from data whilst almost two thirds used a combination of data, expert knowledge and simulations. Insufficiency of data is often cited as the main reason for using expert knowledge in estimating CPTs (Aguilera *et al.*, 2011; Landuyt *et al.*, 2013).

The use of expert knowledge makes the conventional BN-based models vulnerable to criticism as subjective or ‘unscientific’ (Landuyt *et al.*, 2013) and such an approach is seen as producing unreliable estimates of the probability of a presence of a species. The general use of fewer variables, which is mainly necessitated by the need to minimize the CPT in the absence of data, may further bring questions on the validity of the developed expert-driven BN models which could be a serious limitation to their use as SDMs. A very recent practical example is the International Union for the Conservation of Nature (IUCN)’s criticism of Armstrup *et al.* (2008, 2010) BN-based polar bear models as ‘utterly unsuitable for scientifically estimating future populations’ (Polar Bear Specialist Group, 2014).

The use of pre- or expert selected variables is likely to limit ecological knowledge discovery and the advancement of species distribution knowledge especially for alien species that are encountering new and changing environmental conditions. However, this limitation is valid for most correlative species distribution modelling approaches which still use previous knowledge in the selection of variables for model formulation. Merow *et al.* (2014) note that scientists do not always have this prior understanding of ecological systems. Domisch *et al.* (2013) observed that the choice of variables, whilst it might not necessarily influence model performance, affects the spatial projections of habitat suitability. Since model performance and spatial output are not necessarily always similar, habitat suitability projections require careful consideration.

Spatial output is key in species distribution modelling or habitat suitability mapping as it forms the basis for understanding the geographic outcome of species-environment interactions. Moreover, spatially explicit models facilitate the application of BN models at the desired scale or level of analysis such as the grid cell (Johnson *et al.*, 2012b; Morgan *et al.*, 2012). Applications with no spatial outputs used BNs as decision support tools (systems) which may limit the visualization of geographical areas of intervention. Only 34% of the reviewed literature produced explicit mapping outputs for visualization of the BN models. This observation confirms the findings from the reviews by Newton (2009), Aguilera *et al.* (2011), Barton *et al.* (2012) and Landuyt *et al.* (2013) who found that very few of the BN applications were spatially explicit. This relatively low rate of spatially explicit BN outputs may be attributed to two main reasons: the traditional predominant use of BNs simply as qualitative graphical decision supports tools and the limited integration or coupling of BNs to GIS software. The lack of programming skills limits the ability to couple or embed BN packages with(in) GISs.

Nevertheless, a few software packages have recently emerged which directly apply or embed BNs on spatial data. These include:

- a) BayNERD (Mello *et al.*, 2013) which is based on R,
- b) Geo-Netica (Norsys Software Corporation, 2014b) which integrates the Netica package and ArcGIS (Environmental Systems Research Institute, 2011),
- c) the Probabilistic Map Algebra Tool (PMAT) (Landuyt *et al.*, 2015) which is a plugin that interfaces Netica and Quantum GIS (QGIS Development Team 2012); and

d) QuickScan (Verweij *et al.*, 2014) which links Netica to ArcGIS.

However, with the exception of PMAT, none of them has an interface to cartographically represent the uncertainties associated with the BN outputs. Even these packages were not specifically developed for species distribution modelling, they have built-in capabilities to do so albeit with the appropriate data preparation and pre-processing required by each software package. Further integration of BNs and GIS is required to enable the incorporation of neighbourhood dependencies and spatial interactions in the DAG, an area that has yet to be fully explored (Ames and Anselmo, 2008; Giretti *et al.*, 2012; Johnson *et al.*, 2012b, Morgan *et al.*, 2012; Landuyt *et al.*, 2013; Landuyt *et al.*, 2015). This highlights the need for stronger collaboration between the disciplines of software development, machine learning and ecology as is the case in other fields such as bioinformatics.

This thesis aims to demonstrate an inductive approach which relies on empirically-derived spatial relationships between the observed occurrence locations of invasive alien plants and factors that make up the physical environment (anthropogenic, topographic, climatic, etc.). The quantitative relationships between the invasive alien plant occurrence and those environmental factors are defined using BNs and each occurrence probabilistically mapped in geographic space. This is critical when the study domain consists of numerous interacting variables. The advantage of this methodological approach is that, through harnessing the processing speed of modern computers, algorithms can be implemented to empirically discover both known and unknown relationships between factors and invasion events as described in the next chapter.

CHAPTER 3 : METHODS

3.1 INTRODUCTION

This chapter outlines the methodology used in the execution of the study. Firstly, a description of the country's biophysical environment is provided to elucidate the key climatic, topographic and geological features of Swaziland. The scientific method is followed within a data mining framework and this is represented through in-depth descriptions of the steps followed from data collection, through data analysis to final model outputs and evaluation. The identification of the possible predictor variables, their collection from various sources and their integration is explained. Data pre-processing is also undertaken to clean, transform (convert) and format the data for analyses in the data mining and GIS software. This is performed together with data balancing to deal with species prevalence and feature selection, which seeks not only to select variables that have maximum relevance and minimum redundancy but also to select the most predictive interacting variables whilst reducing the data dimensionality. The selected features are then used to learn the structure and parameters of the BN which best represents each invasive plant's distribution from the given data. The search-and-score and constraint-based structure and parameter learning techniques described in the previous chapter are employed to model and predict the distribution of the species under investigation. The learned models are then spatially and graphically visualized and evaluated using a suite of metrics to assess their performance.

3.2 STUDY AREA

The Kingdom of Swaziland, located in the southern African region, is landlocked and covers an area of approximately 17,364 km². The country is bounded by South Africa in the north, west and south and by Mozambique on the east (Figure 3.1). It is characterized by divergent geomorphology and altitude that ranges from approximately 40m to 1860m above sea level. According to Wilson (1982), the western half of the country is typically composed of the igneous and metamorphic rocks of the Archean basement complex, whereas the Lowveld and Lebombo plateau are underlain by sedimentary Karoo formations. The dominant rock type in the western upland areas is granite with subordinate metamorphosed sedimentary rocks of the Onverwacht group and other metamorphic rocks (gneiss and quartzite). The most commonly occurring lithological formations of the central part of the country is the Ngwane gneiss,

followed by granites and granodiorites, with shale subordinate. Sandstones, claystones, coal and other sedimentary rocks of the Karoo Ecca series, together with subordinate dolerite intrusions, dominate the eastern lowlands. Karoo basalts (basic volcanic rock), which may be up to 5km thick also occur in the area. The Lebombo plateau consists of the youngest Karoo rock type of volcanic. These underlying geological formations largely determine the topography and ultimately the soil, climate and vegetation patterns in the country. The topographic and altitudinal variations result in diverse climatic conditions, ranging from sub-humid and temperate in the west to semi-arid and warm in the east. Dlamini (2011) observed that the vegetation units and habitats found in the country are closely associated with the geology as well as climate.

Climatically, the country is subtropical with summer rains (concentrated within the period from October to March) and distinct seasons. Mean annual rainfall, which varies considerably from year to year, ranges from an average of about 1,500 mm in the western part decreasing with altitude down to 500mm in the southeast where drought is an inherent feature. Conversely, mean annual temperature varies from 17°C in the northwest rising up to 22°C in the southeast although there are variations caused by localized topographic features, particularly mountain ranges, hills and valleys.

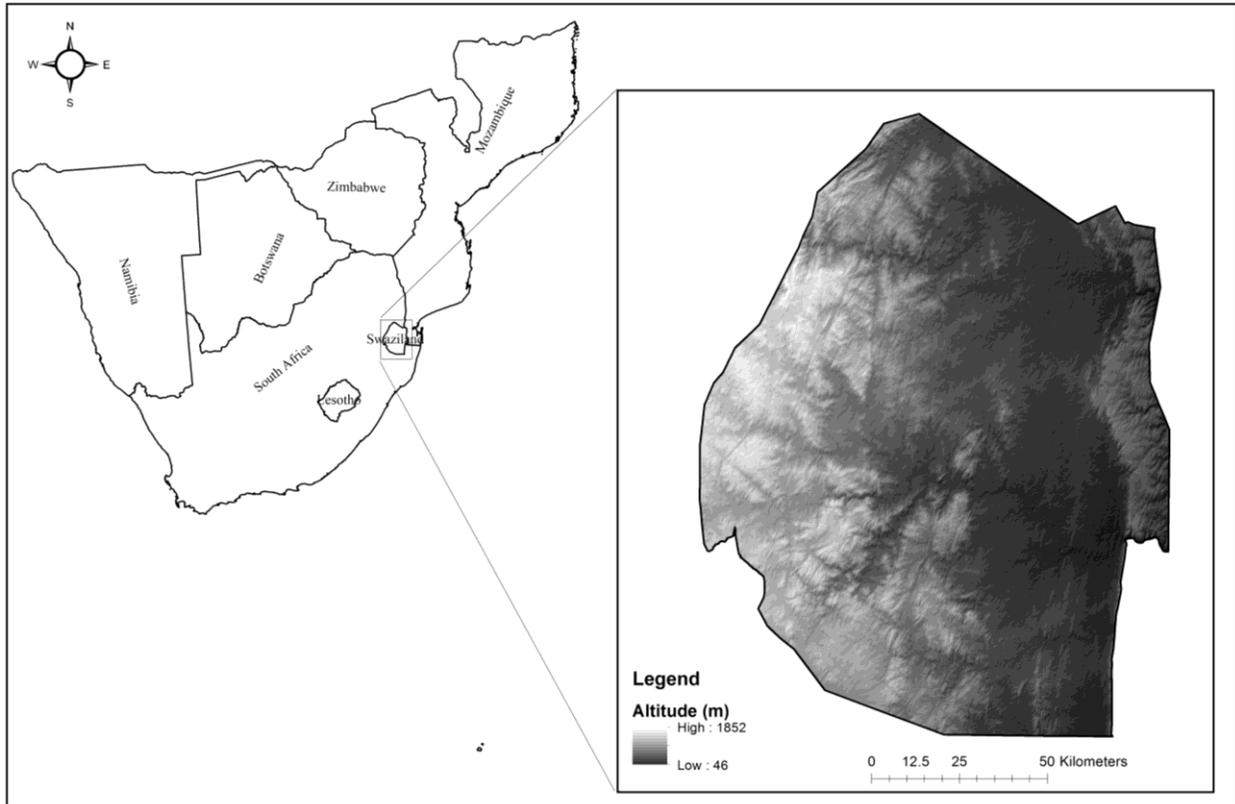


Figure 3.1: Location of the study area, highlighting the topography (source: own).

The climatic and physiographic complexity supports a high diversity of vegetation types and ecosystems. Hence, the country forms part of the Drakensberg Afromontane Regional System and the Maputaland–Pondoland-Albany Region, both of which support high concentrations of endemic taxa and very high biodiversity. Although the country’s knowledge of its biodiversity is still at a developmental stage, these centres and other country’s ecosystems harbour many species that include a variety of mammals (127), birds (500 species), reptiles (111 species), amphibians (44 species), and fishes (57 species), among many other taxa (Monadjem *et al.*, 2003). Current records indicate that there are over 3,400 species of higher plants in Swaziland, representing 771 genera in 135 families. This includes over 700 species of trees and shrubs, and about 3,000 shrubs, herbs and grasses. This complexity in the biotic and abiotic features of the country results in intricate topo-climatic, ecological and socio-economic interactions at the landscape level. Such a heterogeneous landscape with steep environmental gradients provides

a good opportunity to test the effectiveness of BNs in studying species distribution (as encapsulated in Section 1.3).

The geology, coupled with the climate, invariably influences the various land capabilities and agro-ecological zones with differentiated suitability for varying land uses such as human settlement, grazing, wildlife, irrigation agriculture, livestock ranching, and subsistence agriculture, amongst others (Rommelzwaal and Dlamini, 1994; pers. obs.). These land uses are found under communal Swazi National Land (about 52% of total land area held in trust by the King) and Title Deed Land, which constitutes about 47% of total land area (Rommelzwaal and Vilakati, 1994). Crown (government) land and concession land are two other minor categories, which cover less than a percent of the country's surface area. The land tenure system influences the spatial patterns and types of land uses, settlements and subsequently the land cover and ecological patterns and processes in the country.

3.3 RESEARCH PROCESS

This study generally followed the scientific method within the context of the knowledge discovery in databases (KDD) process (Fayyad *et al.*, 1996), for which data mining provides the methods and tools for extracting the knowledge from data. It is important to choose an appropriate knowledge representation for the knowledge discovery task. In this research, this task was done using Bayesian networks (BNs) (Pearl, 1988) because they are able to intuitively and graphically represent the probabilistic structures of complex and non-linear systems and are able to represent uncertain relationships amongst variables (Pearl, 1988; Korb and Nicholson, 2011). Ultimately, the aim was to integrate the spatial data objects with known invasive alien plant species presence or absence and use those to develop and find a model that best represents the probabilistic relationships between the variables. Furthermore, the developed model was used to predict the species' occurrence even in areas where it is either unknown or not yet observed. The challenge is to capture patterns present in the geographic space thereby producing specific knowledge to understand the alien plant species invasion processes in Swaziland.

The spatial positions of species presences or absences captured during the species mapping process implicitly indicate the relationships between different species and the environmental and geographical space in which they occur. Hence, when species distributions are mapped, the

multiple data layers are implicitly integrated. The basic idea of extracting ecological knowledge from these distribution maps is to reverse this mapping process. Therefore, the relationships between species and the environment can be revealed through a knowledge discovery approach by analysing distribution maps together with the environmental data captured using GIS.

Figure 3.2 depicts the workflow followed in the study in order to collect, collate and convert the multi-source, multidimensional spatial data into useful ecological knowledge of each species' invasion patterns and processes. The interconnected steps could be iteratively looped between any two steps to refine the relation between data and the species distribution patterns.

These steps are defined as follows:

- Data acquisition: all the spatial and non-spatial datasets, including the species distribution data and predictor variables, are acquired.
- Data cleaning: a process of removing noisy and inconsistent data,
- Data integration: a process of combining the multiple data sources,
- Pre-processing and feature selection: a process of transforming or consolidating the data into forms appropriate for the mining task and selecting the data which is relevant to the analysis task,
- Data mining: this is the essential process where the BN techniques will be learned from data and applied in order to extract species distribution patterns,
- Spatial prediction and pattern evaluation: this is the process that identifies and explains the common geographic and environmental patterns representing knowledge on each species' presence or absence based on the BN models,
- Spatial knowledge presentation: here spatial and graphical visualization and knowledge representation techniques are used to present the mined ecological knowledge on each species.

One of the contributions of this study, as indicated in Section 1.3, is to explore knowledge integration for geospatial predictive modelling specifically focusing on the integration of different knowledge sources, existing thematic maps together with field data collected with different sampling strategies, for the prediction of invasive alien plant species. Existing thematic maps, which constitute the environmental variables, serve as documented environmental and geographical knowledge. The integration of knowledge from these datasets is of great

importance for both practical and theoretical ecological applications. Various approaches have been used to integrate BNs and GIS (Walker *et al.*, 2004; Ames and Anselmo, 2008; Johnson *et al.*, 2012b; Morgan *et al.*, 2012).

The approach used in this study was based on a GIS-BN interaction wherein BNs are used to combine GIS raster layers so as to account for uncertainty as illustrated by Stassopoulou *et al.* (1998). Building on the objective BN modelling approach of Aitkenhead and Aalders (2009), this study used the framework proposed by Morgan *et al.* (2012) to derive both the BN structure and parameters from the spatial datasets and to propagate the BN outputs back into the GIS for spatially explicit analysis and visualization. Furthermore, the evidence and uncertainty were propagated through the BN to generate and update the probability of occurrence maps of each species. Hence, the BNs were used to combine layers of GIS data for each grid cell to create an environmental and geographical feature space and to account for the inherent uncertainty. The extracted species distribution patterns, together with the learned BN structures, were then examined and interpreted to derive ecological knowledge on the underlying invasion processes.

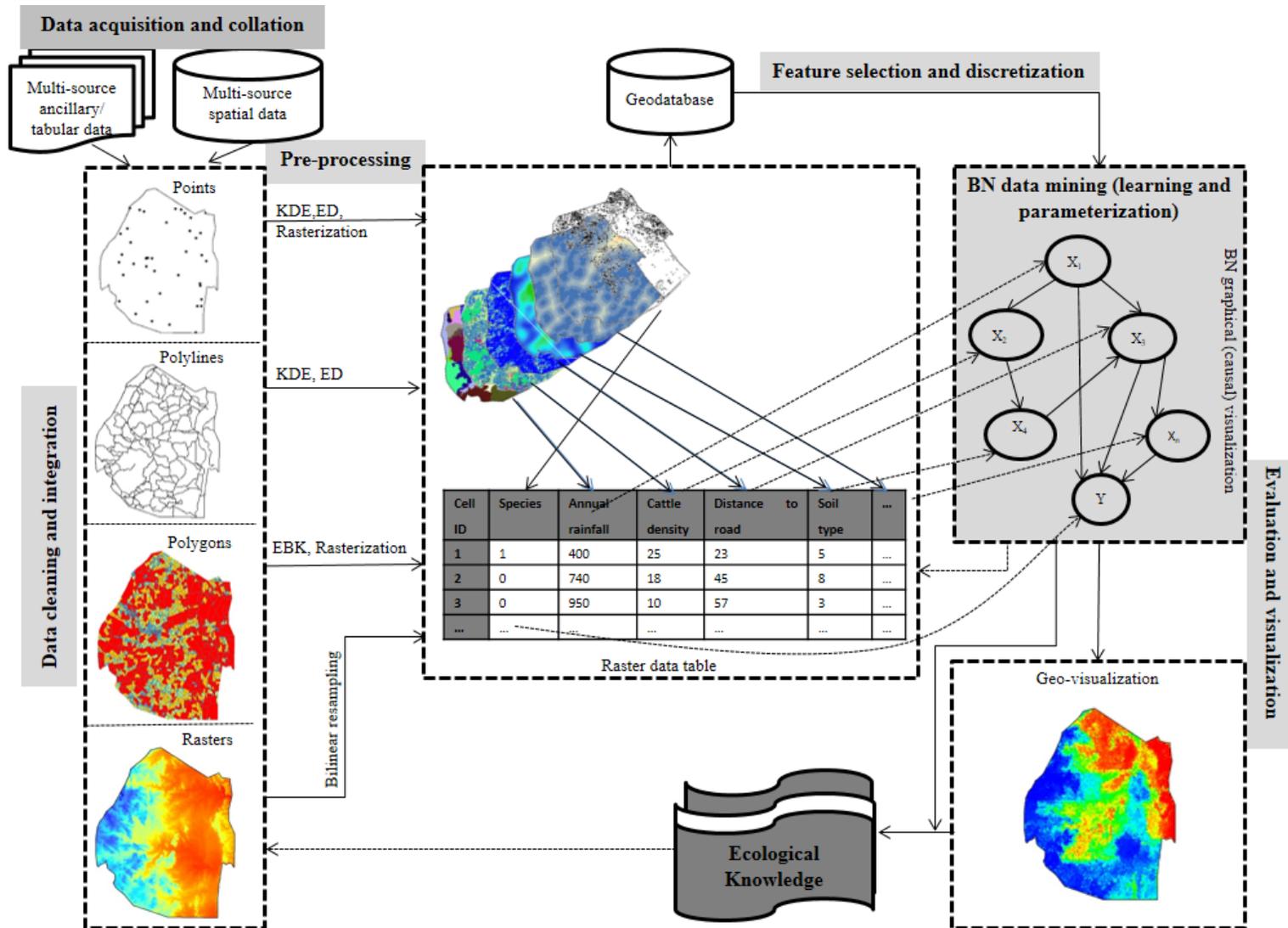


Figure 3.2: The research process followed in the study (EBK – Empirical Bayes Kriging, ED – Euclidean Distance, KDE – Kernel Density Estimation, source: own).

3.4 DATA ACQUISITION AND INTEGRATION

3.4.1 Species distribution (target variable) data

To discover knowledge on species-environment relationships, the first task was to collate existing species occurrence data. Natural historians, museums, herbariums, individual scientists, amongst others, have recorded species occurrence information for many years in the country and around the world through various means such as ad hoc and systematic collections. As the interest in biodiversity conservation increases, both governmental and non-governmental institutions are investing considerable resources into digitizing of data on species (including alien and invasive) into digital species distribution atlases and making them publicly available, where necessary and possible. Such databases present a wealth of information on species distribution and an indispensable asset for science and conservation. Such data is available in Swaziland albeit in different formats and from various sources.

In recognition of the problem of invasive alien plant species and their subsequent declaration as a national disaster in 2005, the government of Swaziland commissioned a stratified national aerial survey of these plants. The aerial survey was conducted throughout the country during the period between 17 May and 18 September 2009 and coordinates were collected using global positioning system (GPS) receivers at an average 50m altitude supplemented by road surveys (Kotzé *et al.*, 2010a; Kotzé *et al.*, 2010b). The period was selected to ensure maximum detectability of each of the species of interest and hence it coincided with the time of year during which the target species could effectively be identified through contrast differences in colour and flowering from aerial observation (Kotzé *et al.*, 2010a).

A complementary road survey was also carried out during the same period whereby 2 000 points randomly selected along the road network were surveyed to serve as an independent quality control of the aerial survey data and for evaluating a methodology for future monitoring (Kotzé *et al.*, 2010a). The resulting database, which was acquired for this study, was very comprehensive and is suitable for further exploration through modelling techniques. All presence and absence points for *Acacia mearnsii* De Wild., *Chromolaena odorata* (L.) R.M. King & H. Rob., *Caesalpinia decapetala* (Roth) Alston, *Cereus jamacaru* De Candolle,

Eucalyptus spp., *Jacaranda mimosifolia* D. Don., *Lantana camara* L., *Melia azedarach* L., *Opuntia* spp. L., *Populus x canescens* (Aiton) Sm., *Psidium guajava* L., *Rubus* spp., *Senna didymobotrya* (Fresen.) Irwin and Barneby, *Sesbania punicea* (Cav.) Benth., and *Solanum mauritianum* Scop. were extracted and collated from the database. This dataset is especially suitable for this study as the rate of false absences should be negligible due to the high sampling effort combined with extensive expert effort. These species, briefly described in Table 3.1, are on Henderson's (2007) list of prominent invaders in southern African forest, grassland and savanna biomes.

The final database consisted of 18,066 detailed points of all the species studied. Hence, there was no need to create simulated background, virtual or pseudo-absences, which may be biased even though they are often required especially for techniques or algorithms that require presence-absence data. This is important to consider because true-absence data are a critical ingredient for accurate calibration and ecologically meaningful assessment of SDMs that focus on predictions of actual distributions of invasive species (Václavík and Meentemeyer, 2009).

An independent database from the Swaziland Tree Atlas (Loffler and Loffler, 2005) was similarly acquired and integrated. This database was derived from fieldwork conducted in 585 plots spread throughout the country between 1999 and 2004 thereby consisting of another set of presence/absence data albeit with lower spatial sampling but with more certainty of identification. Hand-held GPS coordinates were obtained from each plot wherein the tree species found within 2 km transects and radius were recorded. The field work was conducted sporadically throughout the six-year period so as to cover as many flowering, fruiting and growing seasons as possible (Loffler and Loffler, 2005), thus increasing confidence in the identification of species. The data has been updated with few recent observations albeit collected at lower ad hoc survey intensities and on an *ad hoc* basis (Loffler, 2013: Personal Communication). The Tree Atlas data contained more than 26,000 presence/absence records from the 585 plot locations for all the invasive plant species studied. However, since the Tree Atlas data used a different sampling approach, not all the grids from the aerial survey were covered. Hence, those grids that had no records database were represented as missing data or unknown labels. Since the two datasets were collected using different sampling approaches, it

is of interest in this study to do a preliminary analysis of the differences in depicting the spatial distribution of each of the species under investigation.

Table 3.1: List of alien plant species studied and their characteristics (characteristics information derived from Henderson, 2007).

Species	Common name	Growth form	Origin
<i>Acacia mearnsii</i>	Black wattle	tree	Southern Temperate (Australia)
<i>Caesalpinia decapetala</i>	Mauritius thorn	shrub/climber	Tropical (Asia)
<i>Cereus jamacaru</i>	Queen of the night	tree/shrub	Tropical (Americas)
<i>Chromolaena odorata</i>	Triffid weed	shrub	Tropical (Americas)
<i>Eucalyptus</i> spp.	Eucalyptus	tree	Tropical (Australia)
<i>Jacaranda mimosifolia</i>	Jacaranda	tree	Tropical (Americas)
<i>Lantana camara</i>	Lantana	shrub	Tropical (Americas)
<i>Melia azedarach</i>	Syringa	tree	Tropical (Australia)
<i>Opuntia</i> spp.	Sweet prickly pear	tree/shrub	Tropical (Americas)
<i>Pinus</i> spp.	Pine	tree	Tropical (Americas)
<i>Populus x canescens</i>	Poplar	tree	Northern Temperate
<i>Psidium guajava</i>	Guava	tree/shrub	Tropical (Americas)
<i>Rubus</i> spp.	Bramble	shrub	Northern Temperate
<i>Senna didymobotrya</i>	Peanut cassia	tree/shrub	Tropical (Americas)
<i>Sesbania punicea</i>	Brazilian glory pea	tree/shrub	Tropical (Americas)
<i>Solanum mauritianum</i>	Bugweed	tree/shrub	Tropical (Americas)

The Tree Atlas data was supplemented with data from the atlas of alien invasive plant species (Braun and Dlamini, 2005). This database consisted of all known alien invasive species of non-domestic/agricultural plants occurring in Swaziland. This dataset is a collection and collation of alien and invasive species data that was existent at that time primarily sourced from literature,

protected area institutions, title deed farmers, Water Resources Branch, the then Ministry of Agriculture and Cooperatives and private consultants who were agreeable carrying out countrywide fieldwork research to fill in data gaps (Braun and Dlamini, 2005). This resulted in the preparation of distribution maps at 8th degree grid level together with accompanying photographs and a report on notable alien invasive species. The database includes confirmed records (herbarium specimens), visual records and suspected occurrences and some records that were simply classified by a verbal location. For quality control purposes, only records with actual observation and confirmed records as depicting a species' presence were considered. The distribution maps of all the invasive alien plant species under investigation are shown in Appendix 2: Observed distribution maps of all the species from the aerial survey and tree atlas datasets.

3.4.2 Predictor (attribute) variables

The questions of what makes an invader successful is central in the study of invasion (Elton, 1958). The identification of the key and relevant factors that constrain the range of species is one of the most important goals in ecology, biogeography and species distribution modelling (González-Salazar *et al.*, 2013). Discovery of ecologically meaningful species-environment relationships embedded within existing distribution data requires the choice of relevant factors or variables that effectively describe each species' ecology. The choice of these predictor variables is very important because the chosen variables are closely related to the hypotheses on the mechanisms regulating species distributions and will influence invariably model performance and final interpretation (Syphard and Franklin, 2009; Meier *et al.*, 2010; Syphard and Franklin, 2010; Hortal *et al.*, 2012). Merow *et al.* (2014) suggest that, if data are available, increasing the number of predictors ensures a more accurate understanding of the drivers of distributions. Such variables are also useful for testing some of the most important hypotheses in invasion biology (Jeschke, 2014).

Invasive alien plant species distribution is influenced by a variety of factors such as environmental gradients, climate, land use, land cover, species interactions, and a range of anthropogenic and contingent factors. Whilst traditional SDMs, especially those implemented at the regional level, have typically included only climatic variables there is now an increasing recognition of the importance of other factors at the appropriate scale. González-Salazar *et al.*

(2013) observe that model predictability is higher when integrating both biotic and abiotic variables. This leads to a fuller and more comprehensive understanding of each species' niche within the landscape. Studies have recognized that the effects of various explanatory factors on species distribution may vary with spatial (and temporary) scale (Gelfand *et al.*, 2003).

The final dataset primarily comprised of proximal (and to a lesser extent distal) climatic, topographic, edaphic, disturbance, biotic and socio-economic and anthropogenic variables that are potentially useful for modelling the distribution of invasive plants. The basic climatic data used included the 19 bioclimatic data layers from the Worldclim current conditions (1950-2000) dataset (Hijmans *et al.*, 2005) obtained in 30 arc-second (~880m) resolution grids. The Worldclim dataset was selected because of its wide use in SDMs and it provides comparable data across different eco-climatic regions of the world. Edaphic parameters are also important on plant growth and survival although they are rarely incorporated as predictors in plant species distribution models (Thuiller, 2013; Beauregard and de Blois, 2014). Socioeconomic and anthropogenic variables are important in explaining the landscape level patterns of invasive alien plant distributions (Allen *et al.*, 2013). Variables such as housing (Gavier-Pizarro *et al.*, 2010), infrastructure and physical factors, socioeconomic factors, primarily population (Allen *et al.*, 2013), help to explain plant invasion. Human-mediated dispersal through vectors such as livestock and transportation mechanisms and their relationships with the physical environment are also useful for understanding and managing pathways of dispersal including invasion (Auffret *et al.*, 2014).

Hence, these were included in this study to capture the human influence. Other datasets such as the poverty and livestock data were obtained in tabular form and subsequently geocoded and linked to corresponding polygon or point localities from gazetteers for mapping and further spatial processing. The spatial and non-spatial data used were obtained from various institutions and government agencies in various data formats that were not only structure-specific (e.g., raster vs. vector based spatial data, object-oriented vs. relational models, different spatial storage and indexing structures), but also vendor-specific formats (e.g., Environmental Systems Research Institute (ESRI), Stata, Excel spreadsheet, etc.) and at different resolutions. Ultimately, a large set of 170 key predictor variables consisting of biotic and abiotic variables, including socio-economic and propagule pressure variables were identified and collected from

various sources (Appendix 2). The presence-absence data of all the other invasive alien plant species were included as well in order to capture possible biotic interactions and co-occurrences, whether direct or indirect.

While most of the chosen predictors have been used in modified form in other studies, their choice is intentionally objective and diverse in order to learn more about their influence on the species distribution. A contribution of this study (Section 1.3) is to present and evaluate a framework in which all the factors associated with different data types can be integrated into BN-based data mining models that provide precise predictions of the distributions and a fuller understanding of the underlying invasion patterns and processes of the given alien plant species. Since the ESRI shapefile and grid are the de facto and widely used vector and raster GIS data formats, respectively, all the datasets were imported, converted and stored in these formats thus facilitating their integration.

3.5 DATA PRE-PROCESSING

3.5.1 Data cleaning

Data cleaning is an important step in data pre-processing and it involves removing noise or outliers and eliminating invariant or redundant representations of the data. It is known that mapping processes, whether manual or automated, are not only time-consuming, but may be inconsistent and error-prone. However, few species distribution modelling studies explicitly pay detailed attention to data cleaning. Each of the environmental data sets listed in Appendix 1: List of variables (datasets) used in the study was explored in detail and all records within each data set were checked and all noisy and inconsistent data points were removed. This included removal of double allocation cases where the same is repeated in the same observation point, verification of misspelling and inconsistent species nomenclature in cases where elements of both common and scientific names were simultaneously used for the same species.

Improving the quality of georeferencing is another very important and crucial step in increasing the utility of SDMs. Whilst there is notable progress with regards to automated georeferencing through various technologies, determining coordinates by consulting detailed maps, field notes in museum archives, and the original collectors and experts allows for improved coordinates. A few georeferencing errors were found in the aerial survey data and these were in the form of

inconsistencies in the Loffler and Loffler (2005) and the Braun and Dlamini (2005) data where some records had no coordinates and these were removed. This uncertainty in position is common with museum, herbaria, survey and opportunistically observed data and is a result of inaccuracies in location and georeferencing (Naimi *et al.*, 2014). This was easier to deal with in cases where nearby locations had similar attribute values to the original location.

Proper georeferencing of all species occurrence records allows researchers to circumvent the serious problems posed by the effects of sampling bias across geographical and environmental space (Anderson, 2012). Nevertheless, there were very close geographical matches between all the datasets. The aerial survey data was also explored in detail and formatted into presence-absence format followed by a rigorous cleaning process to remove noisy and inconsistent data. Similarly, values that were anomalously high or low were removed and replaced as missing and those grid cells or features that had missing values such as NODATA values were retained because the BNs are able to handle.

3.5.2 Transformation and conversion

Most spatial modelling and simulation tools are designed around the representation of space either as continuous spatial information in the form of grids of regular cells, or as a set of vector geometries representing the shape of well delimited objects (Castets *et al.*, 2014). This spatial data duality on the representation of space and time is a long-standing issue in GIS studies (Goodchild *et al.*, 2007). However, the choice of a format depends primarily on the spatial scale. In many environmental modelling situations though, the ability to simultaneously use both forms in a seamlessly integrated modelling design is desirable. The ultimate focus of this study was on the spatial domain of species distribution using raster or grid cell algebra, which is generally efficient for spatial analysis and modelling (Bolstad, 2012; Morgan *et al.*, 2012). The raster data model has become the primary spatial data source for analytical modelling with GIS and is well suited to the quantitative analysis of numerous data layers. This spatial data format enables the integration of data from different sources within a consistent modelling framework, wherein each grid cell contains a set of the predictive variables and the target variable.

Whilst vector data use is relatively fast and efficient, particularly for mapping single species, rasterization or gridding is often necessary for the analyses of data across different species.

Hence, as part of the data pre-processing, all vector (point and polygon) datasets containing continuous and discrete variables were converted to raster formats through various spatial analysis techniques that conform to Tobler's first of law of geography² (Tobler, 1970, p 236). For those point and polygon datasets (e.g. census data, livestock/dipping tank locality) data that required conversion to continuous surface data, the empirical Bayesian Kriging (EBK) interpolation method (Pilz and Spöck, 2007) was employed. This method automates the task of building a valid kriging model by automatically calculating parameters through a process of subsetting and simulations, unlike other kriging methods that require manual adjustment of parameters to receive accurate results. Moreover, the EBK method accounts for the error introduced by estimating the underlying semivariogram, thereby taking the uncertainty of semivariogram estimation into account (Pilz and Spöck, 2007).

Density estimates from applicable point and line data such as roads, rivers and settlements were done using the non-parametric kernel density estimation (KDE) method (Silverman, 1986). The KDE approach uses a kernel function to fit a surface to each point or polyline, indicating the intensity of individual observations over a geographical area. Hence, points (e.g. settlements) or lines (e.g. roads or streams) near and/or within the centre of a search radius or bandwidth were weighted more than those further away. A bandwidth of 5km was used for this analysis, which provided a balance between parsimony and surface smoothness while offering relevance to the empirical and domain knowledge of the variables in question. When using KDE the kernel weighting varies according to the distance from the point or line as the intensity estimated and the stipulated bandwidth or search radius. Additionally, proximity data was derived by calculating Euclidean distances, which are the straight-line distances from each cell centre to the closest object of interest, such as urban area, roads, or a river. All other vector data containing discrete variables, e.g. soil types or vegetation types, and the presence absence data of species were simply converted to raster data.

To obtain a uniform grid dataset from the heterogeneous sources, all raster datasets were converted to a standardized mesoscale 30 arc-second (~880m x 880m) grid cell or raster resolution using the GIS software ArcGIS version 10.2 (Environmental Systems Research

² "Everything is related to everything else, but near things are more related than distant things".

Institute, 2011). In the end, the entire study area consisted of an array of 22,687 grid cells each containing a total of 170 predictor variables and the class variable. This array of semantically and geographically consistent grid cells or rasters formed an n -dimensional spatial data cube. This became the basis of the species distribution (hypothesis) node wherein the state of each grid cell represented a species-environment relationships hypothesis.

The ability to make the geographical linkage, via the raster data, back to the GIS for visualization and interpretation was the final process in the knowledge discovery process. Furthermore, integrating the collected spatial data with BNs at the resolution of the grid cells provided an opportunity to perform probabilistic analysis (Ames and Anselmo, 2008), which could then be created by transferring the results of the BN-derived probabilistic map algebra back into GIS. The raster dataset was then exported into comma separated value (*.csv) files for further processing and BN modelling using the open-source data mining software, Waikato Environment for Knowledge Analysis (WEKA) version 3.7.12 (Hall *et al.*, 2009; Witten *et al.*, 2011).

3.5.3 Data integration

Analysis of integrated data makes it necessary to resolve several issues to ensure geographical comparability and uniformity. Within a GIS, the different data sets will constitute several separate layers whose overlay is possible only if their geographic components (x,y) use the same projection system. The datasets collected were from diverse national and thematic origins and were produced in diverse projection systems, most often conforming to the geographical specificities of the country. Given this frequent heterogeneity and the usual broad geographic scale used in the context of multidisciplinary research, it is often recommended to work with a universal projection system.

Some of the data sets lacked projection information and most of them had no metadata and as such required in-depth and visual expert examination of the dimensions of each map to infer on the possible projection system and units used. Although modern GIS packages can interpret and transform projections on-the-fly, all the datasets were re-projected and stored in the Transverse Mercator-based Lo31 reference system which is the standard coordinate reference system used by the Surveyor-General's Office in Swaziland. This projection system, which uses the South

African Survey Grid Zone 31 (Transverse Mercator - South orientated) as its projection, allowed for the easier and efficient calculation of some of the variables, e.g. distances, in metric units and ensured that all components are consistently presented and processed.

The second issue that needed to be addressed in order to achieve correct data integration was scale. Scale is a central concept to describe any phenomena with a geographical dimension on the Earth' surface and in the modelling of environmental patterns and processes (Joost *et al.*, 2010). In this context, scale means spatial resolution or the fineness of distinctions recorded in the data, i.e. the size of the cell in a grid or the size of a pixel. Geographic objects, and even spatial processes in the context of plant species invasion, are continuous in scale, but the interpretation of their behaviour has to rely on discrete steps or levels defining the scale of interest. Between these levels, a continuum of entities, features and processes is observed and joined (Marceau, 1999). The chosen thresholds are specific to organization levels in the scale hierarchy of natural features and processes studied, and are defined by the elements to be described and analysed. The data integration process was inevitably confronted with several kinds of geographical objects corresponding to several organization levels, and it was difficult to determine a common scale of interest, i.e. the best possible scale of analysis given the heterogeneity of scales that had to be dealt with. This problem directly addressed the sensitivity of analytical results to the definition of the chosen spatial scale and units.

Integrating different data sets in a GIS inevitably presents a multi-scale problem, although the complexity will vary. The consequences are that, once the scale of analysis is selected, generalization and data aggregation problems will occur in the processing and the analysis of data and cause unavoidable uncertainties. These uncertainties may be propagated through the process flow chain to the final prediction of species distribution. For instance, socio-economic data (human population, poverty indices, etc.) were collected at enumeration area level at a scale of 1:5000 while the climatic data were collected with a grid resolution of approximately 1 km. The land cover information was obtained at a 10 m resolution (derived from SPOT 4 imagery) and Shuttle Radar Topography Mission (SRTM) topographic data at a 3-arcsecond (~90 m) resolution (Rabus *et al.* 2003). This heterogeneity illustrates very well the challenge of integrating multisource data sets, the potential problems related to the overlay operation, and all issues arising when comparing and analysing relationships between integrated thematic layers.

All the layers or variables influence the distribution of invasive plant species at different scales. Furthermore, while some data sets may be most appropriate at the 1km² scale, they are less relevant at a 1ha scale. The complexity of carrying out comparisons in this interdisciplinary and multiscale context, and especially inferring processes from patterns, means that this process requires extreme care.

Error propagation, resulting from the combination of several heterogeneous data layers within a GIS or from rasterizing vector data, can produce significant noise that affects the interpretation of results. These uncertainties in the data may stem from measurement and sampling errors, environmental variability or incomplete knowledge of system behaviour, as well as positional error, feature classification error, resolution, attribute error, data completeness, currency, and logical consistency (Kraak and Ormeling, 2011). However, techniques such as the BNs are better suited for such cases as they deal with data integration better and are able to characterize such uncertainty in the form of local probability distributions (Laskey *et al.*, 2010).

The entire study area, therefore, consisted of an array of 22,687 grid cells each containing a total of 170 predictor variables and the class (response) variable. This array of semantically and geographically consistent grid cells or rasters formed an *n*-dimensional spatial data cube which was exported to comma separated value (csv) files for use in data analysis and data mining software. The integrated data enhances usefulness for final input into models and knowledge discovery including user experience with the species occurrence data and the ability to further explore such data through advanced data visualization tools. This ranges in complexity from having visualization of statistics about the data through more complex and custom visualizations to integrating with a number of interactive predictive distribution models. The large sample also increases the high signal to noise ratios, thereby making it relatively easier to evaluate the strength of the species–environment pattern in the presence of complex processes (Merow *et al.*, 2014).

The integrated data becomes the basis for the species distribution (hypothesis) node wherein each state in each grid cell will represent a different hypothesis specific to the species–environment relationships defined in the models. The approach to be used in the eventual species distribution modelling is based on a GIS-BN interaction wherein BNs will be used to combine the GIS raster layers in order to account for uncertainty (Stassopoulou *et al.*, 1998).

This maintains the ability to make the geographical linkage, via the raster data, back to the GIS for visualization and interpretation, which is the final process in the knowledge discovery process. Integrating the collected spatial data with BNs at the resolution of the grid cells also provides an opportunity to perform probabilistic analysis (Ames and Anselmo, 2008), which can then be created by transferring the results of the BN-derived probabilistic map algebra back into GIS.

3.5.4 Discretization

Discretization of data is one of the major pre-processing steps which has been studied for the past two decades in discrete BN modelling. The purpose of the discretization process is primarily to transform continuous variables into a finite number of discrete values or interval valued features in order to improve classification performance, whilst facilitating the induction process of a classifier and/or enhancing the interpretability of the learned models (Kaya *et al.*, 2011; García *et al.*, 2013; Velikova *et al.*, 2013). Moreover, it is important the selected discretization technique should be appropriate for the underlying marginal and conditional distributions of the variables (Alameddine *et al.*, 2011). Additionally, optimal discretization provides finer partitioning for the regions where the variable distribution changes rapidly while allocating wider intervals for areas that are relatively flat (Kozlov and Koller, 1997).

Discretization of data values is primarily used to compromise between the averaging out of noise, accuracy of the model and complexity/accuracy of the model/parameter learning. In the context of BNs, discretization is used as method to re-examine the probabilistic parameters using both supervised and unsupervised approaches. Alameddine *et al.* (2011) observed that the binning process has profound influence on the generated BN structure and performance, which invariably affects the model's usefulness and ability to satisfactorily describe the given data (Myllymaki *et al.*, 2002).

The widely used unsupervised methods for discretization which do not use class information are equal frequency binning and equal width binning which determine the bin boundaries by first sorting the data on ascending values and subsequently dividing the sorted data into equally sized or ranged bins, respectively. Various supervised methods, which take into account class or target variable information, have been developed. One of the widely and successfully used

methods, which was selected for this study, is the minimal description length (MDL) criteria (Fayyad and Irani, 1993). This algorithm searches for cut-off points that minimize the class variable entropy given each predictor by selecting bin boundaries based on the minimization of the class information entropy (Fayyad and Irani, 1993; Fernandes *et al.*, 2010). The class entropy of a (sub)set S is defined as:

$$Ent(S) = -\sum_{i=1}^k p(C_i, S) \log_2(p(C_i, S)) \quad (3.1)$$

where $p(C_i, S)$ represents the proportion of instances in S with class C_i and k is the number of classes. For each candidate cut point T of a feature A , a weighted average is calculated of the entropy of the two subsets S_1 and S_2 created by the cut point:

$$E(A, T; S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \quad (3.2)$$

where $|\cdot|$ is the cardinality of a given set (i.e. the number of the configurations that the members of a given set S can take), and S_1 and S_2 are the subsets of the split samples for the left and right part of S , respectively. The candidate cut-off point for which this function is minimal is selected and the process is repeated on the subclasses to create multiple bins and using the MDL criterion as a stopping criterion to avoid too many bins. Partitioning or the cut-off point is accepted if and only if:

$$Gain(A, T; S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A, T; S)}{N}$$

$$Gain(A, T; S) = Ent(S) - E(A, T; S) \quad (3.3)$$

$$\Delta(A, T; S) = \log_2(3^Z - 2) - [Z Ent(S) - Z_1 Ent(S_1) - Z_2 Ent(S_2)]$$

where N is the number of the samples in S , Z is the number of the classes in the dataset and Z_1 and Z_2 are the numbers of the classes present in S_1 and S_2 , respectively.

This method has been shown to be effective compared to a number of discretization methods because it tends to obtain a good trade-off between the number of intervals and accuracy (García *et al.*, 2013). These random discrete or discretized continuous variables become uncertain quantities that can take a discrete number of mutually exclusive and exhaustive values. The discretized continuous variables could then be used along with discrete or nominal variables to

strengthen a model's ability to predict each species' occurrence. The Kononenko criterion with efficient split encoding (Kononenko, 1995), which includes an adjustment to the MDL when multiple attributes are discretized, was applied to correct for potential bias the entropy measure has towards an attribute with many values.

3.6 DATA BALANCING

An exploratory analysis of the species occurrences from the dataset showed that most of the instances, except for *L. camara*, *C. odorata* and, to a lesser extent, *A. mearnsii*, consisted of a large number of absence instances and very small percentages of species presences resulting in class imbalance. This imbalance can greatly influence most modelling approaches (Santika, 2011; Bean *et al.*, 2012; Hanberry *et al.*, 2012; Johnson *et al.*, 2012a). Referred to as 'prevalence' in species distribution modelling, the class imbalance problem is an important issue to address because it often limits the capability of conventional algorithms to classify or predict the cases of interest such as species presence (Johnson *et al.*, 2012a). Models developed from imbalanced data tend to ignore the minority (presence) class of interest albeit with high predictive accuracy. This could potentially lower the true positive (presence) rates thus undermining the main objective of detecting and predicting species occurrences or presences.

To solve this problem, data balancing techniques used in machine learning are a promising tool albeit their limited application in species distribution modelling. Johnson *et al.* (2012a) provided the first rigorous evaluation of the class imbalance problem of species distribution modelling performance using various metrics. It has been observed that the SDM performance depends on prevalence, the complexity of the concept represented by the data, the training dataset size, and the classifier used (Japkowicz and Stephen, 2002; Johnson *et al.*, 2012a). Upon testing various data balancing approaches using selected species and the review of relevant literature, the best performing and appropriate technique was the spread subsample technique (Hall *et al.*, 2009) and cost-sensitive learning (Domingos, 1999). In this study, a hybrid approach was used whereby both approaches were implemented. Firstly, the spread subsample technique was implemented before feature selection and discretization and thereafter cost-sensitive structure learning of the BN and their parameters. This was done to ensure that the selected features and discretization points are not biased by the more frequent species absence instances whilst ensuring that the misclassified instances receive further penalization. This was also necessary

considering that the species under investigation are invasive (i.e. have increasing presences) and not in equilibrium with their environment.

Data balancing in SDMs, although seldomly applied, has been observed to have positive results. For example, Drummond and Holte (2003) found that down-sampling outperformed up-sampling when using a decision tree learner. Evans and Cushman (2009) found down-sampling to perform well when mapping the presence of four conifer species. McCarthy *et al.* (2005) found cost-sensitive learning to have a slight advantage over down-sampling and up-sampling in very large datasets (greater than 10,000 instances) when using random forests. The spread subsample technique is a down-sampling technique that produces a random subsample of the majority class for each fold during the training stage, thereby minimizing the effect of species prevalence or class imbalance. A maximum 1:1 ratio of the minority (presence) to the often prevalent (absence) class frequencies was specified, resulting in a uniform spread and the corresponding recommended prevalence of 0.5 (McPherson *et al.*, 2004). McPherson *et al.* (2004) demonstrated that a 1:1 ratio of presence to absence observations optimally balances between omission and commission errors in model predictions.

Cost sensitive learning, on the other hand, applies a cost-matrix which indicates the cost or penalty for misclassifying any particular data sample (Domingos, 1999; Liu and Zhou, 2006). Various configurations were tested and eventually opted for a $\begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$ matrix which tended to optimize both commission and omission errors. This matrix essentially stipulates that the cost or penalty of false species absence predictions is twice that of false species presence predictions for any model. Incrementing the correct classification percentage of species presences implied slightly decreasing the classification accuracy of absences. This is, however, not much of a concern for invasive organism SDMs because such absences could be uncertain due to the processes shaping species distributions or methodological issues e.g. low detectability or sampling bias (Lobo *et al.*, 2010). This approach was applied to all the BN learning algorithms within the WEKA environment. Despite recent observations to the contrary (Freeman *et al.*, 2012), the down-sampling and cost-sensitive learning approach used improved the performances of all models.

3.7 FEATURE (VARIABLE) SELECTION

Important tasks in ecological studies include knowledge discovery from predictor variables that determine species distribution and developing species distribution predictions based on those variables (Johnson *et al.*, 2012a). The appropriate selection of the most relevant attribute (predictor) variables is vital to identifying ecologically meaningful relationships that provide the most accurate predictions under biological invasions (Barbet-Massin and Jetz, 2014). Furthermore, Merow *et al.* (2014) argue that building models with an appropriate amount of complexity is critical for robust ecological inference and that researchers must constrain model complexity based on attributes of the data, study objectives and an understanding of how these interact with the underlying biological processes. Araújo and Guisan (2006) suggest that greater focus be given to the relative explanatory power and causality or ecological basis for choosing each predictor used in SDMs. Merow *et al.* (2014) also suggest that more proximal variables are preferred as these represent the resources and direct gradients that influence species distributions. However, few species distribution modelling studies pay attention to the selection of variables for inclusion into models save for using expert knowledge.

The task here was to select, from the 170 predictor variables, those variables that influence each species' occurrence. In order to reduce the dimensionality of the multi-dimensional feature space and to remove redundant, irrelevant or noisy variables, feature selection was performed as part of the process before the actual classification. Feature selection was used to select a reduced set of features from a large initial set with the aim of producing a set that best contributes to distinguishing areas where a species may occur or not occur. The feature selection process also helps to improve both classification efficiency and scalability by speeding up the computation time, whilst improving data quality and increasing the accuracy of the resulting model (Krishnapuram *et al.*, 2004).

For this process, a method that could efficiently achieve a high degree of dimensionality reduction through finding the optimal set of ecologically meaningful features or variables that have more predictive information was required. For this process, a hybrid approach was used which iteratively alternated between filter ranking construction and wrapper feature subset selection. This approach takes advantage of the strengths of both techniques whilst minimizing the disadvantages of either. The filter ranking technique, a re-ranking-based feature subset

selection method, is based on the algorithm developed by Bermejo *et al.* (2012) (see Figure 3.3). This method works incrementally at both the attribute level and block or set of attributes level, taking into account the selected subset (S) in previous blocks or subsets. A univariate filter measure is used to rank the attributes, and then an incremental filter-wrapper algorithm is applied but only over the first feature subset, i.e. over the first ranked attributes. An initial subset of attributes is selected followed by the computation of new ranking over the remaining attributes but taking into account the already selected subset. The filter-wrapper algorithm is run again over the first block of features with the new ranking. This process is iterated until there is no change in the selected subset. Bermejo *et al.* (2012) show that the number of re-ranks is very small thereby greatly reducing wrapper evaluations (hence reducing computation time) without decreasing the accuracy of the output obtained.

From Figure 3.3, it is apparent that this algorithm requires, on top of the training dataset, the specification of (1) a selection algorithm (2) a stop criterion, (3) a block size and (4) the re-ranking algorithm. The feature evaluation algorithm used was the correlation-based feature selection (CFS) filter method (Hall, 1998). The CFS technique evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. The CFS method, which can be applied to continuous and discrete variable problems, uses an evaluation heuristic that prefers subsets of features which are highly correlated with or highly predictive of the class variable while having low correlation amongst themselves. The CFS approach can handle missing values as well whereby the counts for missing values are distributed across other values in proportion to their frequency (Hall, 1998). This method, which assumes conditional independence amongst the features given the class, has been found effective in selecting non-redundant predictors that are likely to be independent of each other without reducing overall classification accuracy (Fernandes *et al.*, 2010).

In	\mathbf{T} training set, M filter measure, C classifier, B block size
Out	S // The selected subset
1	list $R = \{\}$ // The ranking, best attributes first
2	for each predictive attribute A_i in \mathbf{T}
3	$Score = M_T(A_i, \text{class})$
4	insert A_i in R according to $Score$
5	$sol.S = \emptyset$ // selected variables
6	$sol.eval = null$ // data about the wrapper evaluation of $sol.S$
7	$B =$ first block of size B in R // B is ordered
8	remove first B variables from R
9	$sol = \text{IncrementalSelection}(\mathbf{T}, B, C, S)$
10	$continue = \text{true}$
11	while $continue$ do
12	$R' = \{\}$
13	for each predictive attribute A_i in R
14	$Score = M_T(A_i, \text{class} S)$
15	insert A_i in R' according to $Score$
16	$R = R'$
17	$B =$ first block of size B in R // B is ordered
18	remove first B variables from R
19	$sol' = \text{IncrementalSelection}(\mathbf{T}, B, C, S)$
20	if($sol.S == sol'.S$) //no new feature selected
21	then $continue = \text{false}$
22	else $sol = sol'$
23	return ($sol.S$)

Figure 3.3: The re-ranking canonical algorithm (adapted from Bermejo *et al.*, 2012, p.39)

The stopping criterion is determined dynamically and stops when analysing a new block or feature subset does not produce any change in the selected subset, i.e. it provides the same variable subset received as seed (Bermejo *et al.*, 2012). Hence, the number of attributes to consider is decided dynamically and is dependent on the progression of the selection process. The variable subset size or block size should provide an optimum balance between freedom of choice for the wrapper algorithm and usefulness in order to take advantage of using re-ranking. Although Bermejo *et al.* (2012) suggest value between 30 and 50, a block size of 15 features was chosen after doing some sensitivity analysis of the BN learning algorithms on all the species (Figure 3.4). Figure 3.4 illustrates that both prediction performance increases with block size up to a value of 15 after which the number of variables selected increases sharply (thereby increasing computational cost and potential variable redundancy) while the prediction performance increases slightly due to model overfit. Once the subsets of candidate features are

selected, they were then scored using a metric function which measures a feature's ability to discriminate the classes (presence/absence) in data. For this purpose, Peng *et al.*'s (2005) maximum relevance - minimum redundancy (mRmR) algorithm was used because it has the dual optimization goal of maximizing relevance and minimizing redundancy. The mRmR algorithm often gives more accurate and stable performance (Yun *et al.*, 2007).

Although the mRmR algorithm typically uses mutual information as a measure of the mutual dependence between two variables, this study utilized symmetric uncertainty for evaluating both redundancy and relevancy of each variable. The symmetric uncertainty between two variables X, Y is defined as (Press *et al.*, 1988):

$$SU(X, Y) = 2 \left[\frac{IG(X, Y)}{H(X) + H(Y)} \right],$$

where $IG(X, Y)$ is the information gain between X and Y (expressed as $H(X) - H(X|Y)$), $H(X)$ and $H(Y)$ are the entropies of X and Y , respectively. The symmetric uncertainty metric is the appropriate metric of correlations between features for this study because it compensates for the information gain's bias toward features with more values and normalizes its values to the range $[0, 1]$. A value of 1 indicates that knowledge of the value of either one completely predicts the value of the other and the value 0 indicating that X and Y are independent (Yu and Liu, 2003).

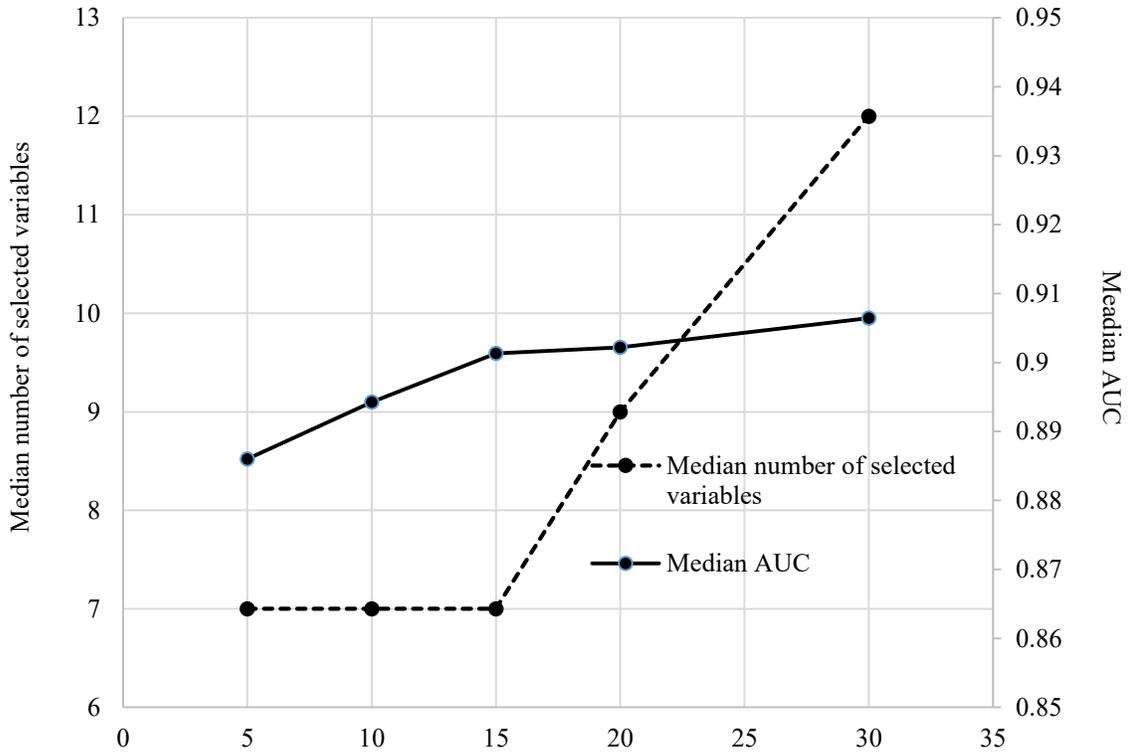


Figure 3.4: Prediction performance and variable selection as a function of block (subset) size (source: own)

The variables were then ranked, according to their symmetric uncertainty, in decreasing order and split into blocks of 15 variables after which a search is run. The block size of 15 variables also ensured that a minimal subsets is obtained whilst avoiding either under- or over-fitting. The mRmR algorithm then selects those variables that have the highest relevance (correlation) with the target variable and are minimally redundant, i.e., selects variables that are maximally predictive and dissimilar to each other. Maximal Relevance D is to search a set of features S satisfying:

$$\max D(S, y_i), D = \frac{1}{S} \sum_{x_i \in S} SU(x_i, y_i), \quad (3.4)$$

where $SU(x_i ; y_i)$ means the symmetric uncertainty between feature x_i and class y_i . The mRmR uses the symmetric uncertainty between feature variables as a measure of the redundancy of each feature. The following condition finds the minimal Redundancy feature set R :

$$\min R(S), D = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} SU(x_i, x_j), \quad (3.5)$$

where $SU(x_i, x_j)$ means the symmetric uncertainty between features x_i and x_j . The criterion that combines the above two conditions is hence called minimal-Redundancy and maximal-Relevance (mRmR) and has the following form to optimize D and R simultaneously:

$$\max \Phi(D, R), \Phi = (D - R) \quad (3.6)$$

The mRmR approximation repeatedly adds the feature or variable that has the best ratio between relevance and redundancy to the already selected features. Theoretically, the optimal subset of features can be found through an exhaustive search of all possible subsets. However, an exhaustive search of the feature space needs to search all of 2^n possible subsets of n features, which is almost impractical in the case of large number of features (1.49658×10^{51} for this study with 170 features). Hence, a search procedure that is easier to implement is required to find the optimal subset of features. The geometric particle swarm optimization (PSO) algorithm with bit-flip mutation (Eberhart and Kennedy, 1995; Moraglio *et al.*, 2007) was used for this purpose. Through PSO, the potential solutions evaluate and compare themselves to others, and imitate the behaviour of those that are regarded as more successful in the search for an optimal solution.

3.8 BAYESIAN NETWORK LEARNING

3.8.1 Structure learning

The nature of BNs makes the learning process a structure learning and parameter learning process, the purpose of which is to summarize conditional independence relations probabilistically and graphically (see Sections 2.4 and 2.5). The structure G of the BN is first learned followed by the estimation of the parameters, Θ . The task for this study was to find BN structures and the corresponding parameters purely from data. In effect, the attempt was to learn a DAG structure that encodes the key features of the dependence structure between the variables of the given data set, when presented with a given number of complete or incomplete instantiations. The direct influence relationships between variables represented by arcs are the hypotheses about each species distribution.

In this thesis, the motivation for building the BNs was to probe the geographical and environmental space within which the different species exist, with the primary aim of finding the dependency and possible causal structure among the different variables that influence each species' distribution and invasion patterns. Thereafter, a parameterization of the resulting BN

structures is undertaken to make them usable for making probabilistic predictions, inference and graphical visualization of the results. Due to the complexities associated with finding and applying a robust test to detect dependencies, both scoring-based and constrained-based approaches were tested in this study. In using the constraint-based approach, the aim was to apply the concept of conditional independence and directional-separation (d-separation) to build the BN. As defined in section 2.4, the scoring based-approaches aim to search for network structures using an adequate heuristic and to assign a score that penalizes model complexity to each structure. Both the local and global scoring metrics were implemented.

In addition to the naïve Bayes (NB) (Duda and Hart, 1973), the score-based BN structures learned for all the 16 species were the Tree-Augmented Naive Bayes (TAN) (Friedman *et al.*, 1997), General BN (GBN) and Bayesian Augmented Naïve (BAN) networks. The BANs were learned with different search strategies namely Greedy Hill Climber (HC) (Buntine, 1996), K2 (Cooper and Herskovits, 1992), Look ahead in good directions (LAGD) Hill Climber (Holland *et al.*, 2008), Repeated Hill Climber (RHC) (Buntine, 1996), Simulated Annealing (SA) (Kirkpatrick *et al.*, 1983) and Tabu Search (TS) (Glover, 1989). The search strategies were implemented in both the local and global search-based BNs except for the LAGD hill climber, which was only available for the local search implementation. These network topologies and search strategies differ in the trade-offs made between network structure, computational complexity and structural richness as briefly described below. In the NB case, the class node is a parent to all the parent nodes and there are no arcs between the attribute nodes, i.e. it assumes that all the variables are conditionally independent. The TAN relaxes the conditional independence assumption of the NB by adding tree-like dependencies among the variables where the tree is formed by calculating the maximum weight-spanning tree using the Chow and Liu (1968) algorithm.

Greedy Hill Climber (HC) (Buntine, 1996): In this algorithm, all of the possible solutions to a given problem are represented as a three-dimensional landscape. The HC follows the graph from node to node, always increasing the value of the solution, until a local maximum is reached. This BN learning algorithm uses a hill climbing algorithm to add, delete and reverse arcs without fixed ordering of variables and the search is not restricted by an order on the variables.

K2 (Cooper and Herskovits, 1992): The K2 is a score-based greedy search algorithm for learning BNs from data. It maximizes the probability of an optimal graph topology, given a dataset, by using a Bayesian score to rank different graphs. The algorithm adds and deletes arcs and is restricted by an order on the variables.

Look ahead in good directions (LAGD) hill climbing (Holland *et al.*, 2008): This algorithm uses a generalization approach that calculates in advance k steps about the chosen scoring function. The LAGD offers a new class of parameterized algorithms including the configurable number of look ahead steps k and the number of calculated good operations per each look ahead step.

Repeated Hill Climber (RHC) (Buntine, 1996): This algorithm searches for BN structures by repeatedly generating a random network and applying to it the hill climbing algorithm mentioned above. This is done until the best network is returned. The advantage of this algorithm is that when the HC algorithm gets stuck at a node, a new node is chosen at random and the HC process is restarted. This is repeated k times and the algorithm returns the best maximum found.

Simulated Annealing (SA) (Kirkpatrick *et al.*, 1983): This is a general-purpose combinatorial optimization algorithm. The algorithm is inspired by the process of annealing metal in order to harden it. The basic idea of the algorithm is to assign to the problem a temperature (a control parameter) and consider the cost of a solution as an energy level. The solution then corresponds with the state of the metal: as the temperature is lowered, the solution becomes more defined, with less moves or states available to it to change to.

Tabu Search (TS) (Bouckaert, 1995): This is another hill climbing algorithm which continues the search after reaching the local optimum by choosing a move that makes the least reduction in the score of the network. However, the TS keeps a list of recently performed operations in memory and does not consider it to prevent a cycle of repetitive operations. The TS algorithm then returns the best network whilst traversing the search space.

Additional to the search procedure, the Bayes score metric (Heckerman *et al.*, 1995) was used as the scoring function. This metric provides an *a posteriori* probability that the learned BN structure is the true model of the underlying data and generally avoids the problem of model

overfitting. A structure prior $P(G)$ is the prior probability on different graph structures and $P(\theta_G | G)$ is the parameter prior that puts a probability on different choice of parameters once the graph is given. Thus by Bayes rule:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \quad (3.7)$$

where $P(D)$ is the normalizing factor that is the same for all networks. The Bayesian score is hence defined as:

$$score_B(G : D) = \log P(D|G) + \log P(G) \quad (3.8)$$

The term $P(D|G)$ takes into account uncertainty over the parameters which ultimately helps avoid the overfit problem and is calculated as follows:

$$P(D|G) = \int_{\theta_G} P(D|\theta_G, G)P(\theta_G | G)d\theta_G, \quad (3.9)$$

where $P(D|\theta_G, G)$ is the likelihood of the data given the BN $\langle G, \theta_G \rangle$ and $P(\theta_G | G)$ is the prior distribution over different parameter values for the network G . Since the unknown parameters are marginalized out, this is called the marginal likelihood of the data, given the BN structure. Unlike the maximum likelihood, which provides optimistic scoring, the marginal likelihood (also known as the marginal likelihood of the model) provides a more realistic value of the score. This is achieved by measuring the *expected* likelihood through the integration of $P(D|\theta_G, G)$ over different parameter values of θ_G (Heckerman *et al.*, 1995) and it is not affected by a specific BN structure (Bouckaert, 1995). For the global score-based structure learning the leave one out cross-validation (LOO-CV), which selects training sets simply by taking the data set D and removing the i th record for the training set D'_i (Bouckaert *et al.*, 2014), was implemented. The validation or test set consist of the i^{th} single record. The accuracy of the classifier was then assessed using the estimated presence probability of each species.

Thereafter, a Markov blanket (Pearl, 1988) was applied over the target node to ensure that every node in the network is a parent or child of a sibling of the prediction node. As explained in the previous chapter, the Markov blanket of the prediction node consists of its direct parents, its direct successors, and all of its direct parents' direct successors within the given BN (Pearl, 1988; Bouckaert *et al.*, 2014). This implies that variables within the Markov blanket have the

greatest influence on the distribution of each species. In order to guarantee the impartiality of the analyses, the number of parents for any node was set to a value of 100,000 which imposes no restrictions on the number of parents a node could have. This allowed for the learning of Bayesian Augmented Naïve (BAN) networks, which extend the TAN by allowing the attributes to form an arbitrary graph instead of just a tree (Friedman *et al.*, 1997).

For the constraint-based approach, the conditional independence (CI) algorithm and Inferred-Causation (ICS) algorithms (Verma and Pearl, 1992) were applied. The CI algorithm tests whether variables x and y are conditionally independent given a set of variables. This is done by comparing a DAG structure with arrows $\forall z \in Z \rightarrow y$ is compared with one with arrows $\{x \rightarrow y\} \cup \forall z \in Z \rightarrow y$ (Bouckaert *et al.*, 2014). The ICS algorithm takes into account latent variables and produces a network with undirected, unidirected and bidirected arcs (Daly *et al.*, 2011). A variant of the Spirtes, Glymour, Scheines (SGS) algorithm (Spirtes *et al.*, 1993) and proposed by Pearl and Verma (1991) and Verma and Pearl (1992), it differs from previous approaches in that it first generates an undirected graph that models dependencies between variables, as opposed to using the complete undirected graph. The ICS algorithm first finds a skeleton (the undirected graph with arcs if there is an arrow in network structure) followed by directing all the arcs in the skeleton to get a DAG. Starting with a complete undirected graph, the ICS attempts to find conditional independencies $\langle X, Y | Z \rangle$ in the data (Bouckaert *et al.*, 2014). For each pair of nodes X, Y , sets S starting with cardinality 0 are considered, then 1 up to a pre-determined maximum. Furthermore, the set S is a subset of nodes that are neighbours of both x and y and if an independency is identified, the arc between X and Y is removed from the skeleton. The first step in directing arrows is to check for every configuration $X-s-Y$ where X and Y not connected in the skeleton whether s is in the set S of variables that justified removing the arc between X and Y . If s is not in S , the direction $X \rightarrow s \leftarrow Y$ can be assigned. Thereafter, a set of graphical rules given in Verma and Pearl (1992) is applied to direct the remaining arrows. Hence, one of the key characteristics of the ICS algorithm is that it is optimized for recovering the causal structure as opposed to finding the optimal classifier (Bouckaert *et al.*, 2014). Additionally, the maximum cardinality which determines the largest subset of Z to be considered in conditional independence tests $\langle X, Y | Z \rangle$ was set to 15 to match the block size set during the feature-selection process. The bigger number increases flexibility in the number of

potential interacting variables. The Bayesian score metric and the Markov blanket correction were similarly applied to both the CI and ICS algorithms.

3.7.2 Parameter learning

After learning the structures, the CPTs for each node were computed using the simple probability estimator (Bouckaert *et al.*, 2014) which estimates the conditional probabilities by directly computing the relative frequencies of the associated combinations of the attribute values in the training data.

$$P(x_i = k \mid pa(x_i) = j) = \frac{N_{ijk} + N'_{ijk}}{N_{ij} + N'_{ij}}, \quad (3.10)$$

where N is the number of records or instances in the dataset D , N_{ij} ($1 \leq i \leq n$, $1 \leq j \leq q_i$) is the number of records in the dataset D for which the parent set of x_i ($pa(x_i)$) takes its j th value, N_{ijk} ($1 \leq i \leq n$, $1 \leq j \leq q_i$, $1 \leq k \leq r_i$) is the number of records in the dataset D for which $pa(x_i)$ takes its j th value and for which x_i takes its k th value, N'_{ij} represents the choice of priors on counts and N'_{ijk} is the alpha parameter ($0 \leq N'_{ijk} \leq 1$). This parameter can be interpreted as the initial count on each value. A value of 0 reduces the estimate to a maximum likelihood estimate. The parameters r_i ($1 \leq i \leq n$) is the cardinality of x_i , i.e. the number of different values to which x_i can be instantiated and q_i is the cardinality of $pa(x_i)$ in the BN structure G . The value of q_i is the product of cardinalities of the nodes in $pa(x_i)$, $q_i = \prod_{x_j \in pa(x_i)} r_j$. Hence, if $pa(x_i) = \emptyset$, then $q_i = 1$.

During the training stage all nodes are observable i.e. in addition to the predictor variables, the BN is populated with the observed species occurrence information. The alpha parameter N'_{ijk} was specified in the estimation of the CPT for each of the BN models. This was set to 0.5 to avoid bias as verified through some preliminary analysis which was undertaken to ascertain the performance of different values of α_i across all the algorithms. The BN parameters were learned on the selected set of feature vectors, resulting in the induction of conditional probabilities of the attributes given the class variable (presence or absence of a species). Hence, it was possible

to efficiently compute occurrence probabilities of each of the 16 species in a structure given the attributes from the selected explanatory variables.

Since BNs require estimation of prior (unconditional) probabilities, these were estimated from the data by estimating the proportion of instances in each category or class. For example, if mean annual rainfall was a node discretized into <200mm, 200 – 400mm, 400 – 800mm and >=800mm, then the prior probabilities were simply the total number of grid cells/data instances in each range divided by the total number of grid cells in the study area. The BN models were trained on the selected set of feature vectors, resulting in the induction of conditional probabilities of the attributes given the class variable (presence or absence of a species). Hence, posterior occurrence probabilities of each of the 16 species were efficiently computed in a BN structure given the attributes from the selected explanatory variables. The learned BN then used exact inference (Pearl, 1988) from which causal and inter-causal reasoning could then be performed.

An ensemble distribution map of all the BN algorithms was developed by calculating the median values. The ensemble model aims to minimize the variability in the predictions of potentially under- and over-fitted models and between-model variance (Araújo and New, 2007; Marmion *et al.*, 2009; Stohlgren *et al.*, 2010).

3.7.3 Variable importance (sensitivity) analysis

The relative importance of each of the selected variables for each BN model was measured by calculating its influence within the BN model of each species. This was done through computing and ranking the mutual information or entropy reduction (Pearl, 1988) between the target (species occurrence) node and the environmental variable. For this purpose, the DAGs learned through the best performing algorithms were used for each species. Given a probability distribution, p defined over two sets of variables X and Y , the mutual information between X and Y , which is measured in bits, is given as:

$$I(X, Y) = \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \quad (3.11)$$

where $p(x)$ and $p(y)$ are the probability densities of X and Y , and $p(x, y)$ is the joint probability density. This can also be expressed in terms of entropy as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \quad (3.12)$$

where $H(X)$ and $H(Y)$, are the entropies of X and Y , respectively, and $H(X,Y)$ is the joint entropy of X and Y .

Given a joint distribution $Q(Y,X)$ of two variables X and Y , the symmetric mutual information attempts to extract the relevant information that Y contains about X and is a very good measure of the average number of bits needed to convey the information X contains about Y and vice versa. As such, the mutual information is able to detect additional non-linear dependencies and correlations among variables that are undetectable using conventional measures (Guyon and Elisseeff, 2003). It ranges from zero when the variables are independent, and attains its maximum when one variable is a deterministic function of the other. Hence, the mutual information was used to reveal the relative influence of each variable on the spatial distribution of each of the species. This information is useful when providing the ecological interpretation of the observed graphical models and maps. The sensitivity analysis also helps in validating the obtained relative variable influences on species distribution with the observed spatial patterns and domain knowledge.

3.8 MODEL EVALUATION

Model evaluation is a very important part of the data mining process that has received attention in species distribution modelling in the recent past with various metrics or criteria being proposed. Notwithstanding the use of different metrics and their differences in interpretability and given the complexity of the modelled species-environment relations, the most robust modelling approaches are likely those that attempt to match the realized model with ecological knowledge (Elith and Leathwick, 2009; Mouton *et al.*, 2010). The interest in this study was to assess the ability of a learned BN model to predict or correctly reproduce the observed spatial patterns in the distribution particularly the presence of the invasive alien plants under investigation. Discrimination capacity and reliability are preferred criteria for assessing model performance and hence metrics that focuses on both were preferred. In this case, the aim was

not necessarily to test if the model is accurate in terms of both omission and commission errors (since there is now information on the true potential distribution of the invasive alien plants), but rather testing each model's ecological relevance and usefulness. For the purpose of this study, the usefulness criteria was that the model successfully predicts species presence in a high proportion of test localities (i.e. known occurrences) whilst not predicting an excessively large proportion of the study area as suitable.

Ten runs of 10-fold cross-validation were performed on each BN classifier to obtain its prediction accuracy thereby ensuring that the final calibration of every model used all of the data available in making predictions. *K*-fold testing is more reliable with large data sets and is one of the recommended approaches for evaluating BN model prediction performance (Marcot, 2012). A BN structure is evaluated by estimating the network's parameters from the training set and the resulting BN's performance determined against the validation set. The average performance of the BN over the validation sets in turn provides a metric for the quality of the network. Moreover, 10-fold cross-validation has been found to be the right number to get the best estimate of error in addition to supporting theoretical evidence (Witten *et al.*, 2011).

The logarithmic loss (Morgan and Henrion, 1990; Marcot, 2012) was selected to evaluate model discrimination performance because of its suitability and reliability for tasks where posterior probability values are an important consideration (Morgan and Henrion, 1990; Marcot, 2012). The logarithmic loss is an evaluation metric whose value is only determined by the probability of the outcome that actually occurs (Cowell *et al.*, 1993). This was calculated using the following equations (Morgan and Henrion, 1990; Pearl, 1988):

$$\text{Logarithmic loss} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{ij} \left[-\ln p_{ij} \right], \quad (3.13)$$

where n is the number of cases or instances in the test set, m is the number of class labels or states, \ln is the natural logarithm, y_{ij} is 1 if observation i is in class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j .

The logarithmic loss, which has scores between 0 and infinity, is a cross-entropy estimate that measures the additional penalty for using an approximation instead of the true model. A value of zero indicates the lowest penalty and the best performance whereby the network's probability

distribution totally matches the true distribution. The models were also tested using other metrics such as the area under the precision-recall curve (AUPRC) (Davis and Goadrich, 2006), Matthew's correlation coefficient (MCC) (Baldi *et al.*, 2000), the area under the receiver operating characteristic (ROC) curve (AUC) (Hand 1997) and the True Skill Statistic (Allouche *et al.*, 2006). Whilst the AUPRC and AUC range from 0 to 1, the MCC values range from 1 for a perfect prediction to -1 for the worst performance while values close to 0 indicate a model that performs randomly (Baldi *et al.*, 2000). The use of multiple metrics is recommended due to the varying responses of metrics to data characteristics in particular prevalence (Jarnevich *et al.*, 2015).

Corrected two-tailed resampled t-tests (Nadeau and Bengio, 2003) were conducted at the 1% ($p = 0.01$) significance level to compare each individual algorithm performance against the conditional independence (CI) model. The BN model was set as the baseline model as its performance tended to be close to the median of the results. The 1% significance level was preferred in order to increase confidence in the conclusions on algorithm performance. Based on the performance metrics, this stage involved scoring the number of statistically significant model differences in performance (wins, ties or losses) between each model and the baseline model.

BN model predictions are, by their interpretation, probabilistic assessments each species occurring at each grid cell and the computed probabilities represent the likelihoods that a particular species will occur taking into account model, parameter and structural uncertainties associated with the predictions. In order to evaluate, interpret and communicate the degree of certainty in the model outputs, the information theory-based posterior probability certainty index (PPCI) proposed by Marcot (2012) was calculated. Ranging from 0 to 1, the PPCI is an entropy-based metric for which higher values denote greater certainty in outcome predictions (Marcot, 2012). The PPCI is an adaptation of the classic evenness index (Hill, 1973) which has long been used to measure the relative distribution of species' abundances in a community. The concept of posterior probability distributions, which consist of p_i probability values among m number of class labels, is extended where p_i ranges $[0,1]$ and

$$\sum_{i=1}^m p_i = 1 \quad (3.14)$$

The PPCI is then calculated as $(1 - J')$, where

$$J' = \frac{H'}{H'_{max}}, \quad (3.15)$$

and

$$H' = \sum_{i=1}^m p_i L, \quad (3.16)$$

where

$$L = \begin{cases} \ln p_i, & p_i > 0 \\ 0, & p_i = 0 \end{cases}, \quad (3.17)$$

and $H'_{max} = \ln(m)$. J' , which is a measure of entropy or uncertainty in information theory, normalizes the metric proportional to m , so that the degree of certainty of posterior probability distributions can be compared among outcomes with different numbers of states m (Marcot, 2012). In this study, only two states are considered: the presence and absence of a target species within a grid cell. The PPCI values were visualized for each of the species in each grid cell throughout the study area.

The four measures identified by Han and Kamber (2006) for characterizing the resultant models were noted for each model, in addition to the forementioned metrics. These were:

- *Speed*: the computational costs involved in generating and using the given model and a set of predictor variables. Hence, the time taken to compute each model was considered.
- *Robustness*: the ability of the model to make correct predictions given noisy data or data with missing values.
- *Scalability*: the ability to construct the model efficiently given large amounts of data.
- *Interpretability*: although subjective and difficult to assess, this refers to the level of understanding and insight that is provided by the model.

3.9 VISUALIZATION

The essence of this study was to implement data mining to produce prediction maps from the BNs that indicate variable (inter-)dependencies with a view to discover alien plant invasion patterns and knowledge. This visualization of findings or knowledge obtained from data mining could then be produced in visual forms such as the graphical BN models and maps. The BNs were stored in Extensible Markup Language (XML)-based BayesNet Interchange Format (BIF) for viewing and exploration using other BN software. The visual displays help give users a clear impression and overview of the data relationships and processes. Subsequently, the BN model outputs for each species at each grid cell within the study area were linked to the open source Quantum GIS (QGIS Development Team, 2012) for geo-visualization. The spatial outputs in the form of maps can then be simultaneously visualized together with the BNs for easier conceptualization and interpretation of the graphical and geographical relationships between the selected variables and the mapped distributions of each species.

This chapter outlined in detail the process used in the study from data collection through data analyses to model evaluation. The results from the analyses are visualized and presented in detail in the next chapter.

CHAPTER 4 : RESULTS

4.1 INTRODUCTION

This chapter presents the findings of the study specifically the collated invasive alien plant data, the learned BN structures and corresponding maps showing the model parameters. The BN modelling process produced two main components: a qualitative component and a quantitative component. The qualitative components of the BN models are the DAGs with nodes representing the variables selected during the feature selection stage. The qualitative components of the BN model are shown as DAGs with nodes representing the influential and/or causal factors selected for each species and links representing the causal influences between the linked variables. As stated before, the influence relationships between variables constitute hypotheses about each species' distribution. The quantitative aspects of the BN models are represented by the CPTs in each node. The predictive performances of each model and for each species, represented by the posterior probabilities of the target variable, are compared using various metrics described in the preceding chapter. The BNs learned using the best performing algorithm and accompanying prediction, ensemble and uncertainty maps are presented for each species. The data from the aerial survey and tree atlas are also overlaid on the prediction maps for a visual comparison.

4.2 BAYESIAN NETWORK LEARNING ALGORITHM PERFORMANCE

Both the constraint-based and score-based approaches automated the process of structure learning very well and show promise in retrieving the BN structure from the species distribution and environmental data. Despite the differences in BN structures, all the learned BNs seem to model the data almost equally well, which can be gathered from their predictive accuracies. In all the runs and for all the species, both the locally and globally scored genetic search algorithms were computationally intensive and in all the runs and folds, the calculations could not be completed even after 24-hour runs. These were stuck in local maxima during the searching and learning process. Learning the BN structures using the genetic search algorithm was, therefore, computationally intractable even though their performance were expected to be comparable or

equivalent to the other algorithms. Hence, these were excluded from further analyses. Similarly, the repeated hill-climbing algorithms and simulated annealing algorithms with global scoring could not be run to completion for high prevalence species such as the *C. odorata*, *Pinus* species and *L. camara*.

Generally, there were predominantly more ties and few statistically significant differences ($p < 0.01$) in the performance of the BN learning algorithms. Hence, the best performing algorithms for each species were solely based on numerical magnitude. The performance of all the algorithms is given in detail in Appendix 3. Figure 4.1 shows the logarithmic loss (log loss) values for all the algorithms and species. The high dimensionality and, for some species, high imbalance of the data was very well handled by the BN learning algorithms to produce log loss values ranging from a mean of 1.166 to a minimum of 0.25 and a mean of 0.679. The constraint-based ICS algorithm was more robust with relatively lower log loss values for most species. This indicates that the causal structure with converging or colliding arc topology results in parameters that are a better representation of the observed (field) data.

The TAN and the hill-climbing algorithms learned with local scoring together with the K2 and simulated annealing algorithms learned with global scoring also performed relatively well. A few algorithms produced log loss values slightly exceeding unity whilst the conditional independence (CI) algorithm and the naïve Bayes (NB) were the worst performers, implying that the conditional independence assumption results in a mismatch between posterior probability distribution and the data. Whilst performance varied between species, the findings point to the fact that interaction amongst the variables results in probability distributions that had better match with the field data. This is demonstrated in Sections 4.3 and 4.4.

The AUC values were also high for all the BN algorithms across all species ranging from 0.807 for ICS models of *L. camara* to 0.994 for globally-scored simulated annealing and hill-climbing models of *J. mimosifolia* (mean = 0.930). The results in Figure 4.2 show that, when considering the AUC values for all the species, there were few statistically significant differences in the performance of all the BN learning algorithms compared to the NB model and the CI which were generally outperformed for all species but *P. x canescens* (Figure 4.2). Hence, there was no straightforward superior algorithm when considering the AUC.

Considering the true skill statistic (TSS), values ranged from 0.480 for the ICS model of *L. camara* to 0.945 for an ICS model of *P. x canescens*, averaging at 0.767 (Figure 4.3). These figures are generally indicative of the strong predictive power of the BN models. Besides the slight inferiority of the ICS algorithm, no algorithm shows outright performance dominance. Overall, the MCC values averaged at 0.42 with values ranging from a low of 0.094 for the ICS-based model of *P. x canescens* to a maximum of 0.699 for the global search-based simulated annealing and repeated hill-climbing models of *M. azedarach* (Figure 4.4). However, in all the runs both locally and globally scored BNs achieved better accuracies.

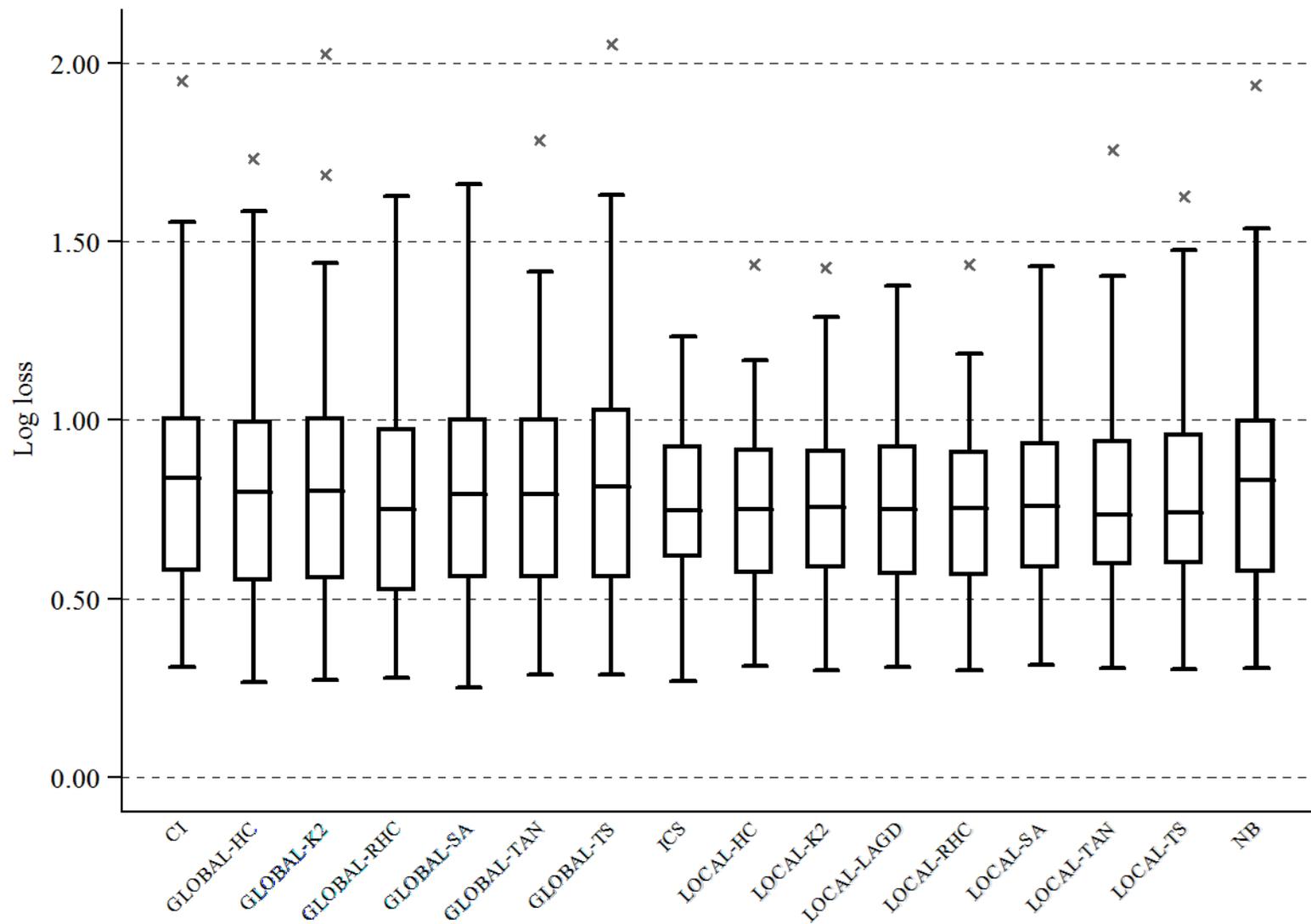


Figure 4.1: Performance comparison (box plots) of all the BN learning algorithms using the logarithmic loss (source: own).

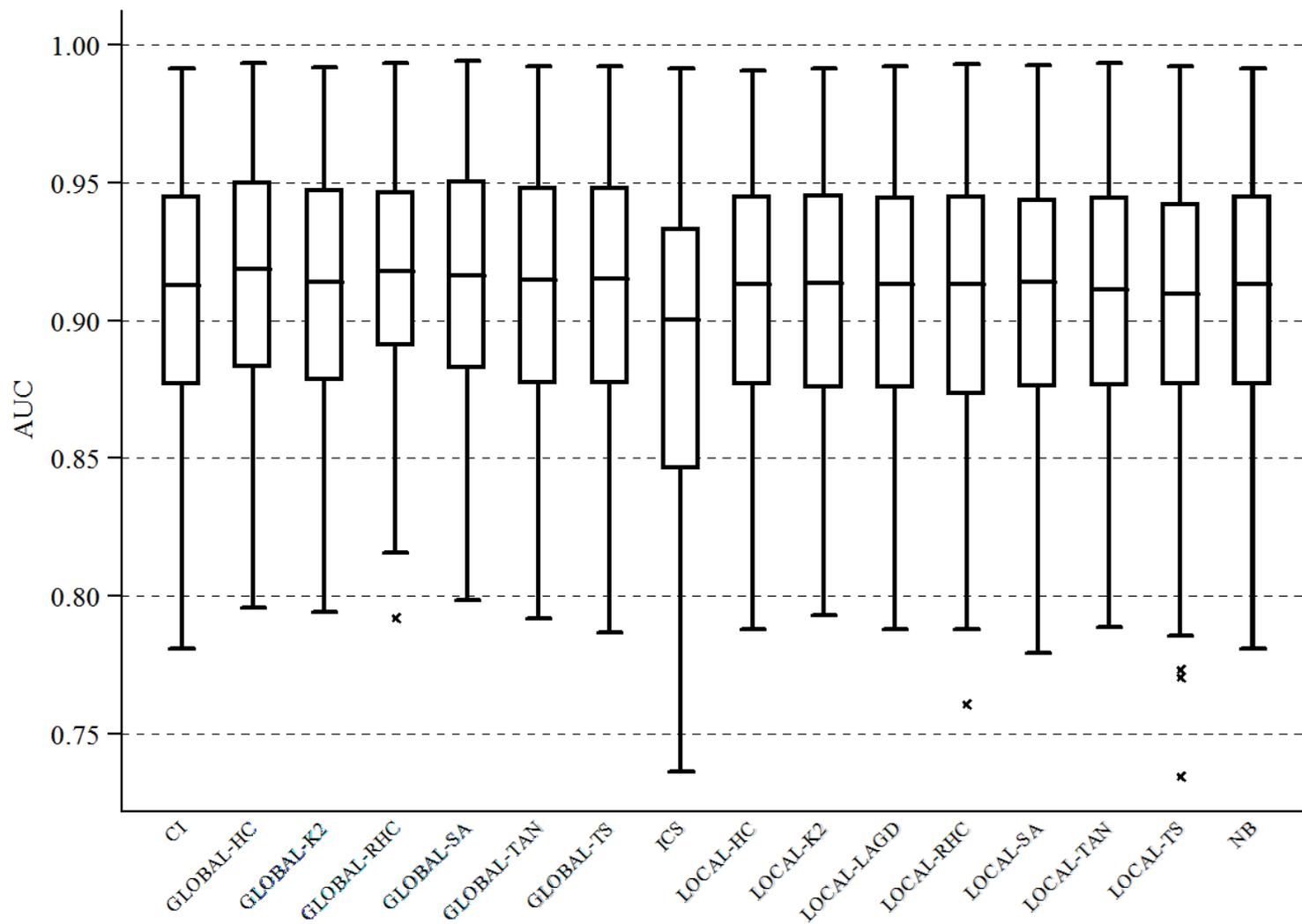


Figure 4.2: Performance comparison (box plots) using the area under the ROC curve (AUC) for all the BN learning algorithms (source: own).

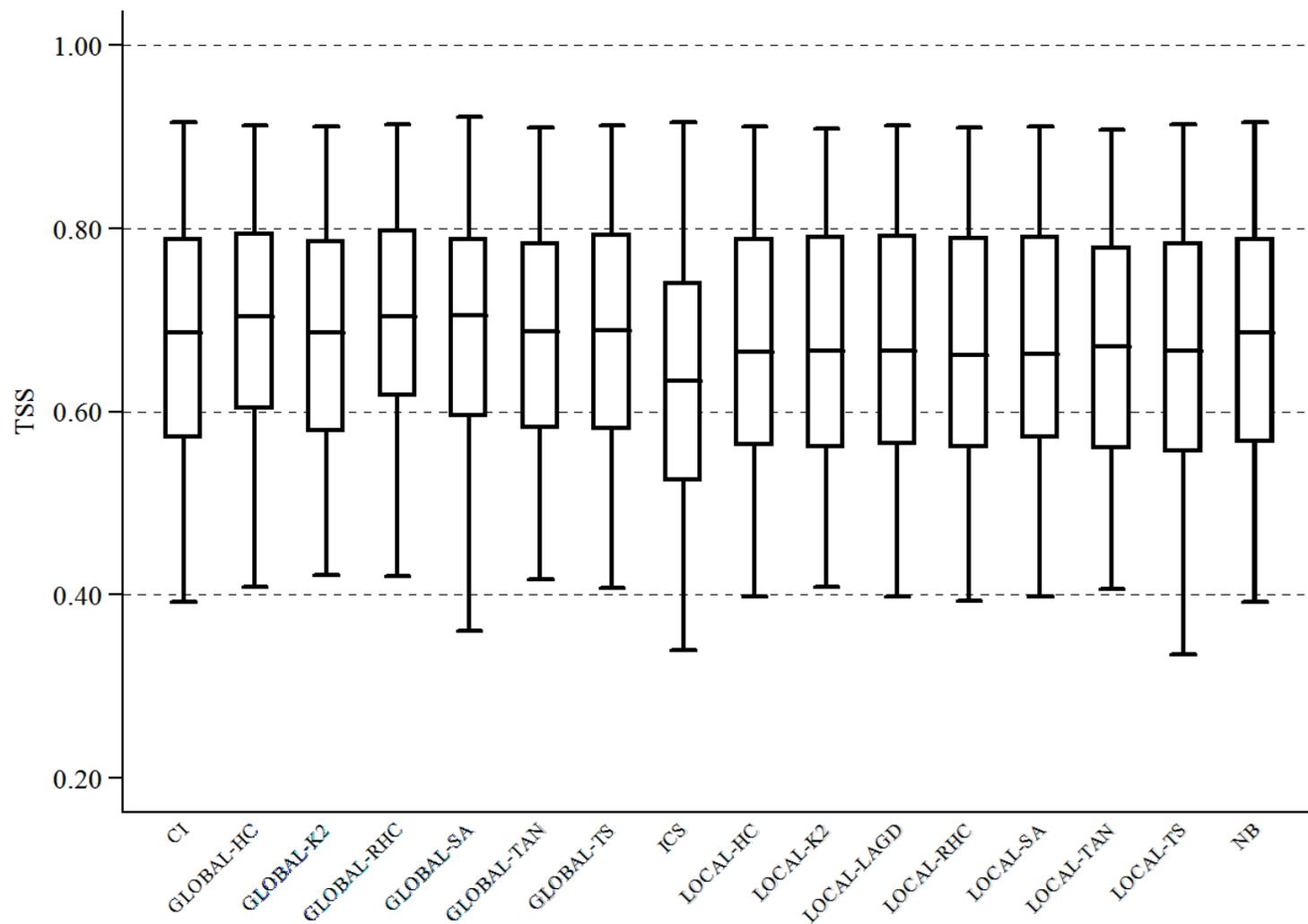


Figure 4.3: Performance comparison (box plots) using area true skill statistic (TSS) for all the BN learning algorithms (source: own).

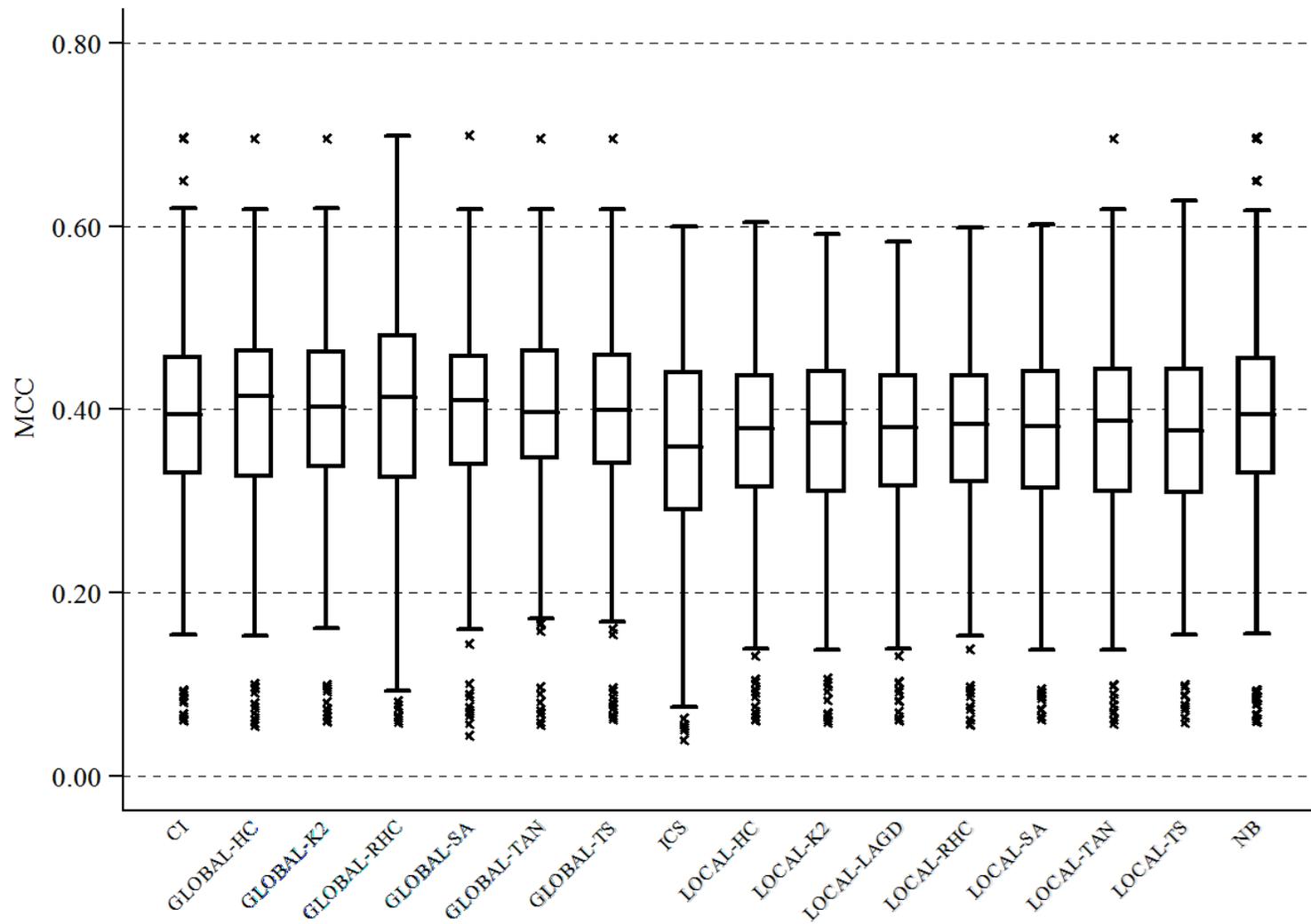


Figure 4.4: Performance comparison (box plots) using Matthew's correlation coefficient (MCC) for all the BN learning algorithms (source: own).

Considering Figure 4.5, it is evident that the AUC, AUPRC and TSS are significantly and negatively correlated to species prevalence ($R^2 = 0.519$, $R^2 = 0.909$ and $R^2 = 0.474$, respectively) suggesting the sensitivity of these metrics to imbalanced data. These relationships indicate that the ranges of rare or limited distribution species are more predictable than those of more common or widely distributed species. Similarly, there was a significant and positive correlation between the root mean square error (RMSE) and species prevalence ($R^2 = 0.335$) which implies that model reliability is sensitive to prevalence. The only insignificant relationships were between the MCC ($R^2 = 0.046$) and log loss ($R^2 = 0.160$) which showed very weak positive correlations. These two metrics are, therefore, relatively less sensitive to species prevalence and may be considered unbiased and suitable for measuring model performance.

The results indicate that, on average, globally scored GBNs, specifically the hill-climbing searches, were more robust than locally scored BNs and constrained-based algorithms when the same parameter estimation procedures are used. However, when considering the threshold-independent and probability-based logarithmic loss metric, the constraint-based ICS algorithm performed relatively better. The causal structure derived from the ICS algorithm, which has the collider or converging BN topology, is also easier to interpret in terms of the direction of the arcs and their relationship with the target variable.

As expected, learning the naïve Bayes and locally-scored TAN models was done relatively faster than the rest of the other algorithms (Figure 4.6). These scores were based on a computer with 16GB RAM and 2GHz processing speed (with overclocking capability up to 3GHz). The computation times ranged from 0.297 seconds for a locally scored TAN-based *P. x canescens* model to a maximum of 1349.141 seconds for a globally scored repeated hill-climbing model of *S. mauritanum*. The mean computation time was 68.171 seconds or just over a minute. Generally, the globally-scored algorithms mainly those using the hill climbing search and simulated annealing were computationally intensive compared to the locally scored algorithms whilst the naïve Bayes-based and the constraint-based algorithms were the most computationally efficient.

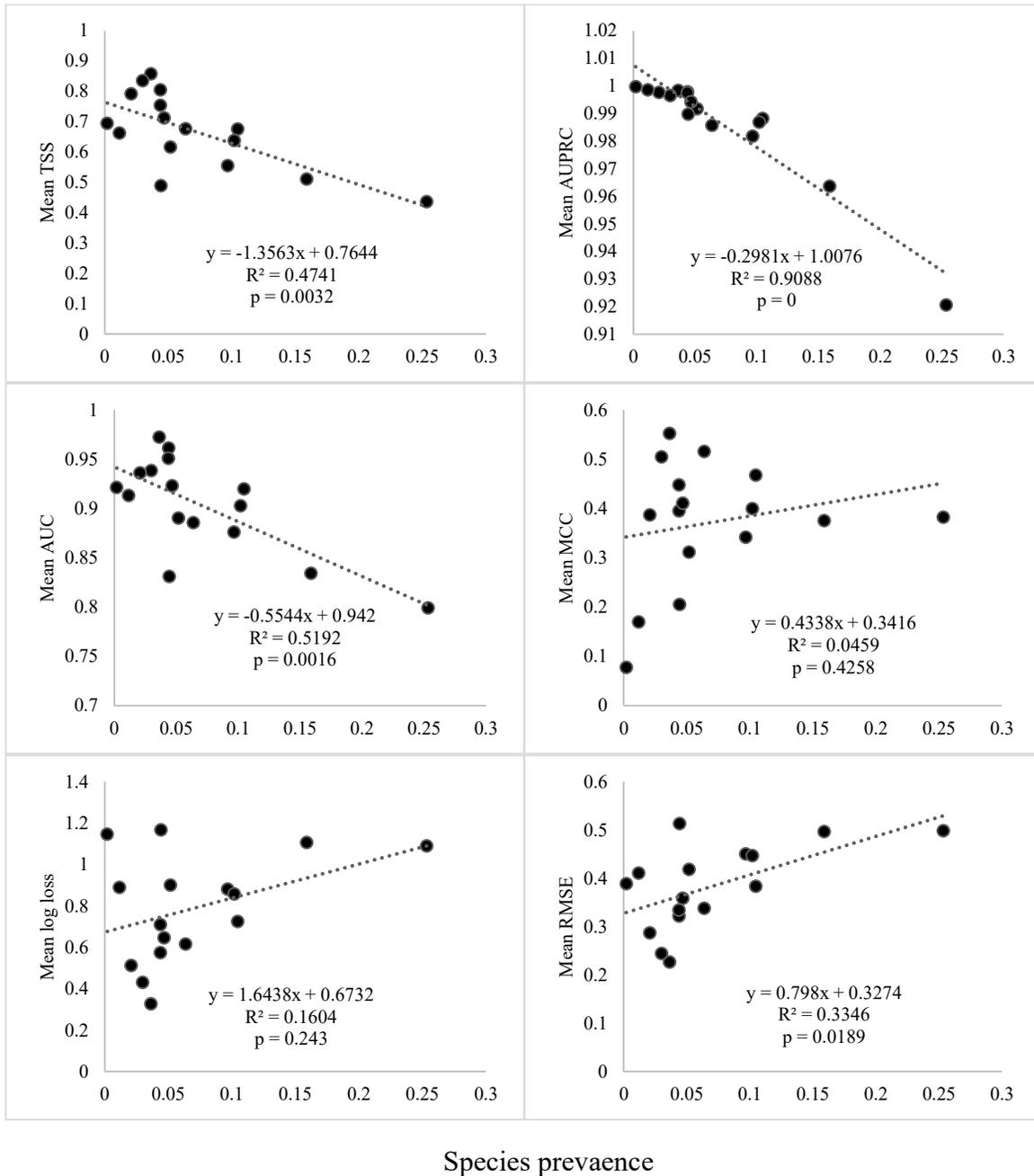


Figure 4.5: Scatter plots of model evaluation metrics plotted against species prevalence (source: own).

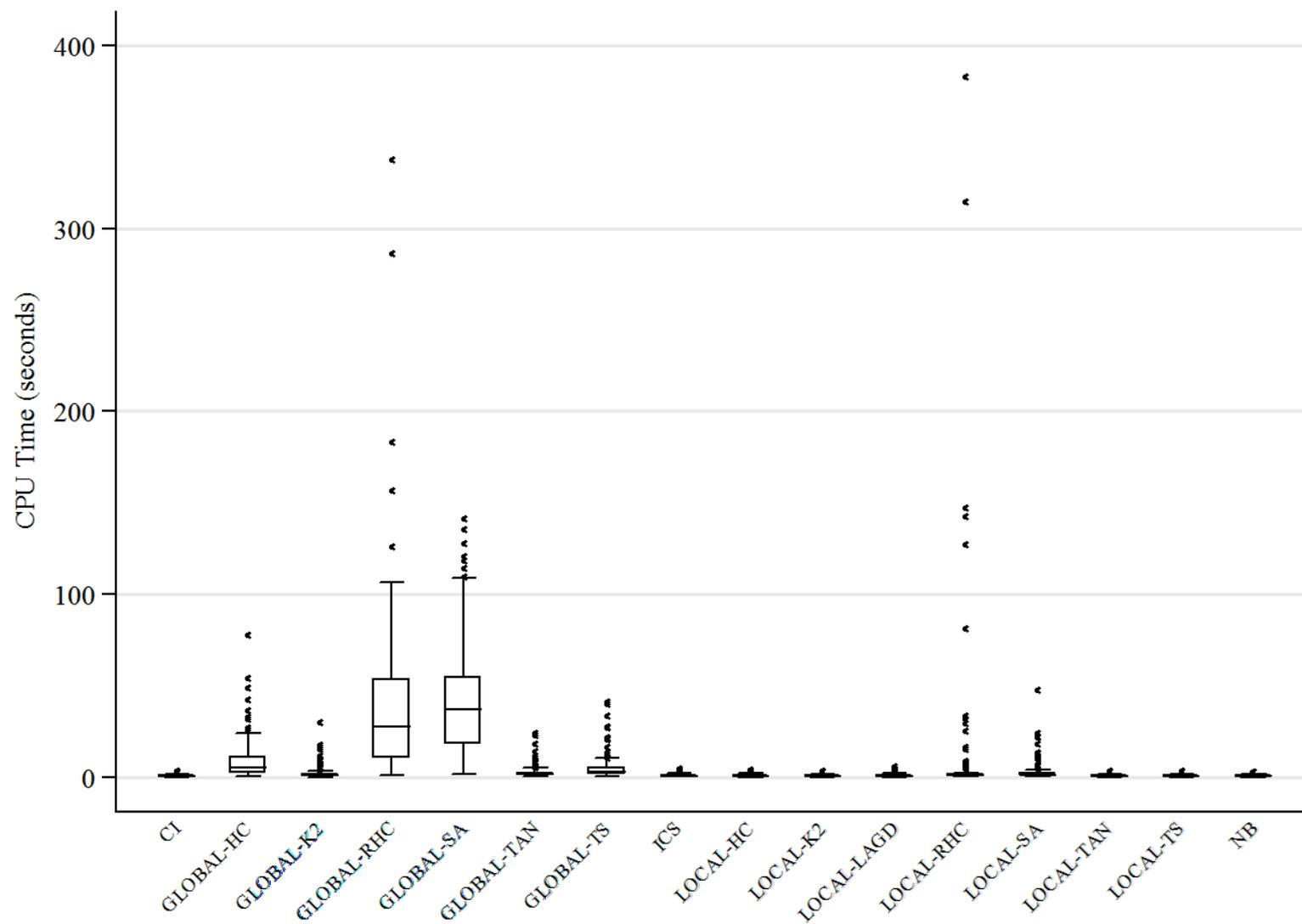


Figure 4.6: Box plots of the computation time (in seconds) for all the BN learning algorithms (source: own).

When examining the effect of prevalence against computation time, it is apparent that there is a significant relationship (Figure 4.7). This implies that it takes relatively longer to learn BN structures and compute parameters for species with a broader niche. However, model complexity, as measured by the number of variables included in the model, did not influence the computation time. As such, the number of nodes or variables within the BN model did not affect the learning time.

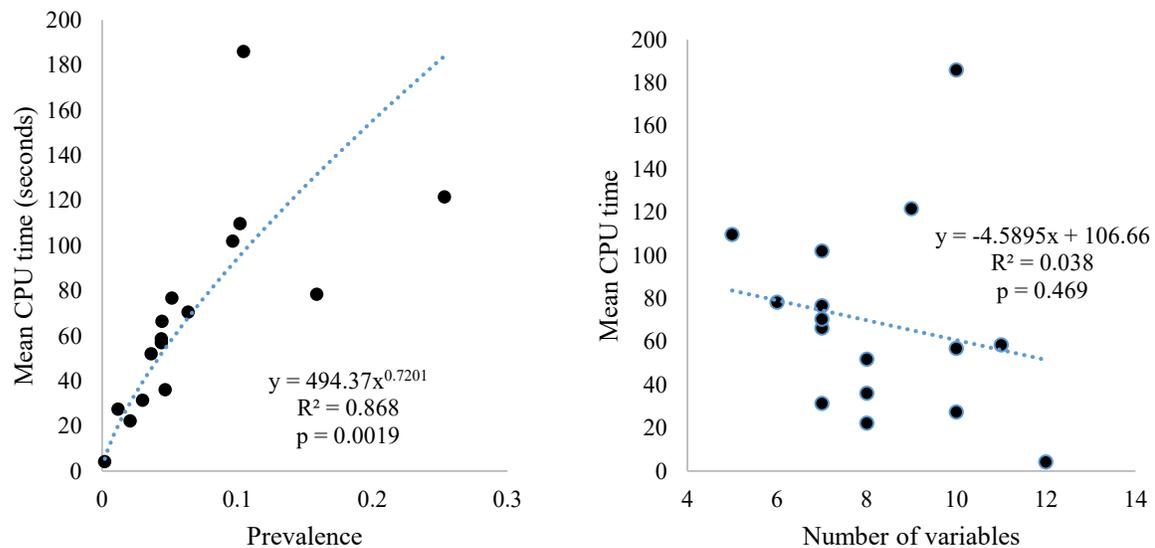


Figure 4.7: Plot of CPU time against species prevalence (left) and the number of selected variables (right)(source: own).

Figure 4.7 indicates that there was no significant correlation between the feature selection process, hence the number of variables selected, and species prevalence. This indicates that the feature selection algorithm, hence the choice of variables used, was not influenced by a species' geographic spread or rarity but its ecology. Similarly, the number of variables used in the models had no effect on the performance of the BN learning algorithms (Figure 4.8). This implies that model complexity, as determined by the number of model variables, did not affect the BN models' ability to produce posterior probabilities that match the field data.

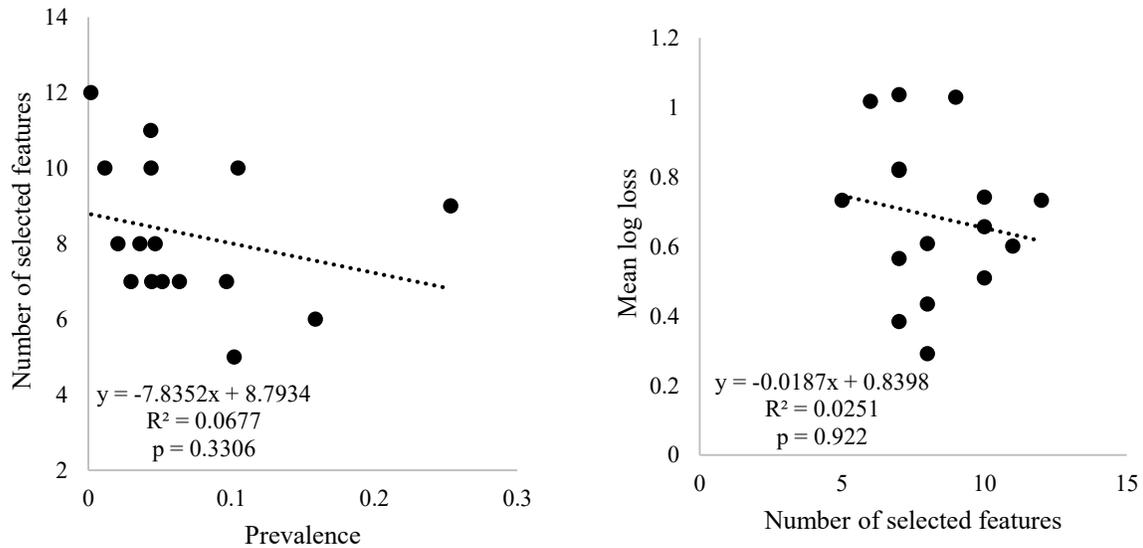


Figure 4.8: Plots of the number of selected features or variables against species prevalence (left) and the number of mean log loss against selected variables (right) (source: own).

4.3 LEARNED BAYESIAN NETWORKS AND PREDICTED DISTRIBUTIONS

4.3.1 *Acacia mearnsii*

Five variables were selected as key determinants for the spatial distribution of *A. mearnsii*, all of which had direct arcs to the target node as learned through the globally scored TAN algorithm (Figure 4.9). The TAN algorithm, therefore, was more scalable and better maximized the Bayesian scoring function given the *A. mearnsii* distribution data.

The selected variables were the mean temperature of coldest quarter, human population density and the presence of *J. mimosifolia*, *Rubus* species and *P. x canescens*. These variables were within the *A. mearnsii* Markov blanket and had complex interdependencies amongst themselves and the target species. The species is restricted to areas with mean temperature of coldest quarter less than 15.2°C. The BN models reveal that *A. mearnsii* occurs in populated areas (human population density > 5 people/km²).

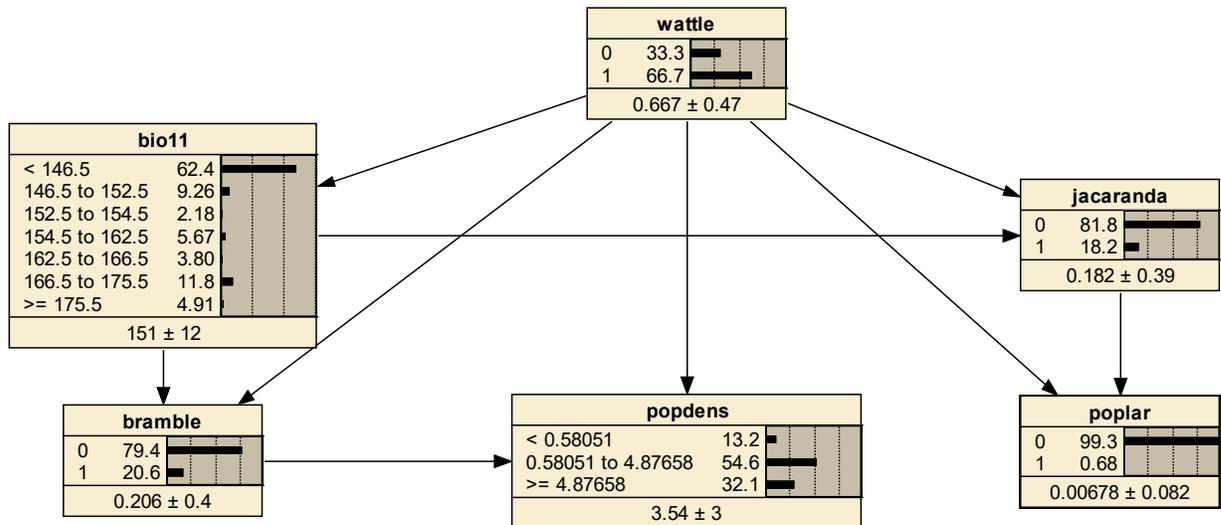


Figure 4.9: A learned Bayesian network for *Acacia mearnsii* distribution.

The mutual information values (Table 4.1) indicate that the mean temperature of coldest quarter is the strongest predictor of *A. mearnsii* distribution. This perhaps indicates that the distribution pattern of this species is limited by its tolerance to relatively low temperature conditions or its non-tolerance to higher temperatures. This was followed by its association with *Rubus* species, *J. mimosifolia* and human population density, the latter factor being a possible indicator of human activities as dispersal or propagation agents. The commensalism of this species with humans potentially drives its spread to climatically suitable habitats. The presence of *Populus x canescens*, although important, had relatively lesser influence on the species' distribution.

Table 4.1: Mutual information for selected *Acacia mearnsii* predictor variables.

Variable	Mutual Information
bio11	0.29489
bramble	0.10581
jacaranda	0.0929
popdens	0.04509
poplar	0.00276

The selected variables are probabilistically interdependent resulting in the predicted distribution shown in Figure 4.10. The dominant constraint of mean temperature of the coldest quarter is evidenced by the confinement of *A. mearnsii* to the western (coldest) part of the country. The ensemble model shows similar patterns (Figure 4.11) whilst the PPCI map indicate the confidence in the predictions based on the field data.

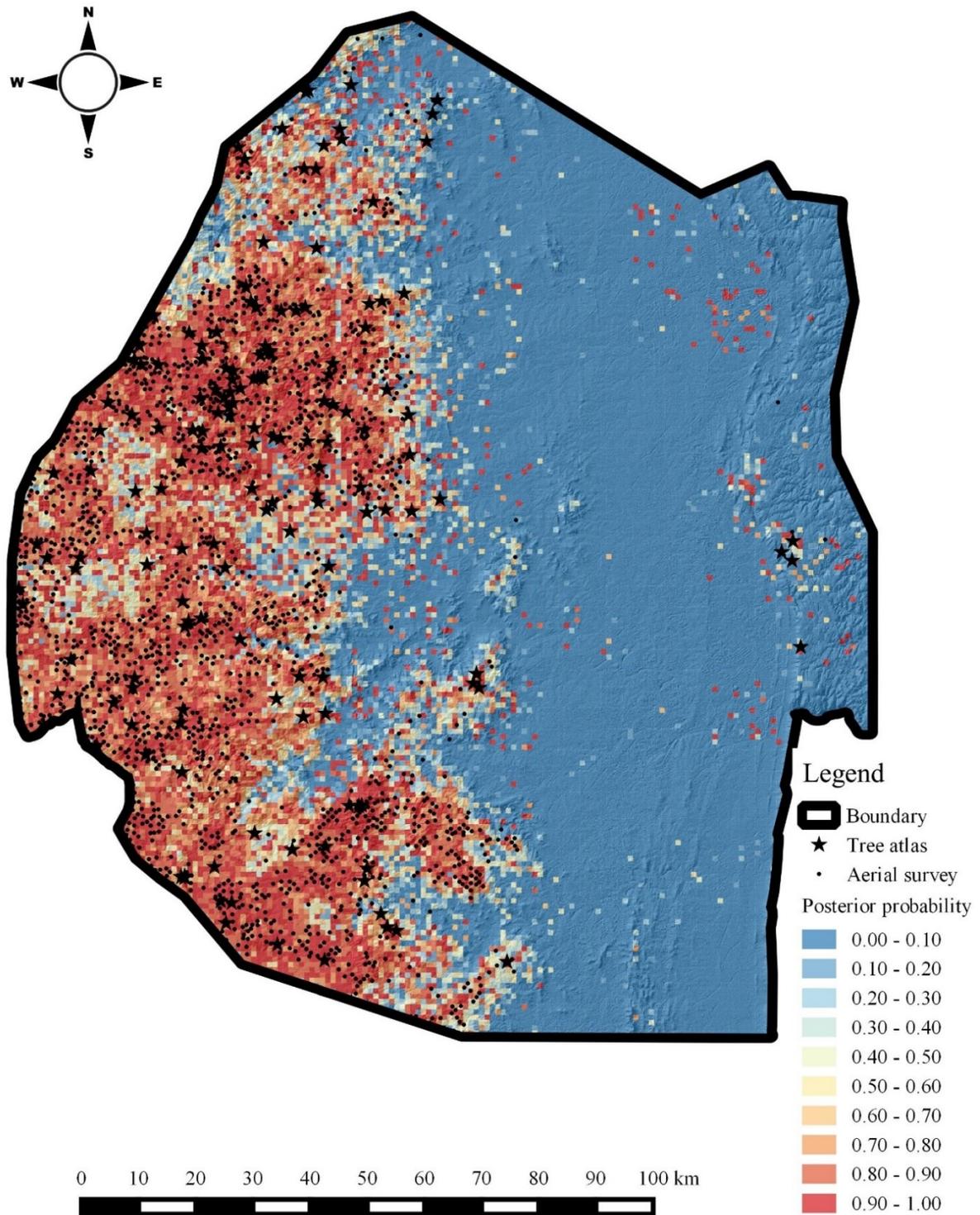


Figure 4.10: Posterior probability of occurrence for *A. mearnsii* in Swaziland (derived from the BN in Figure 4.9).

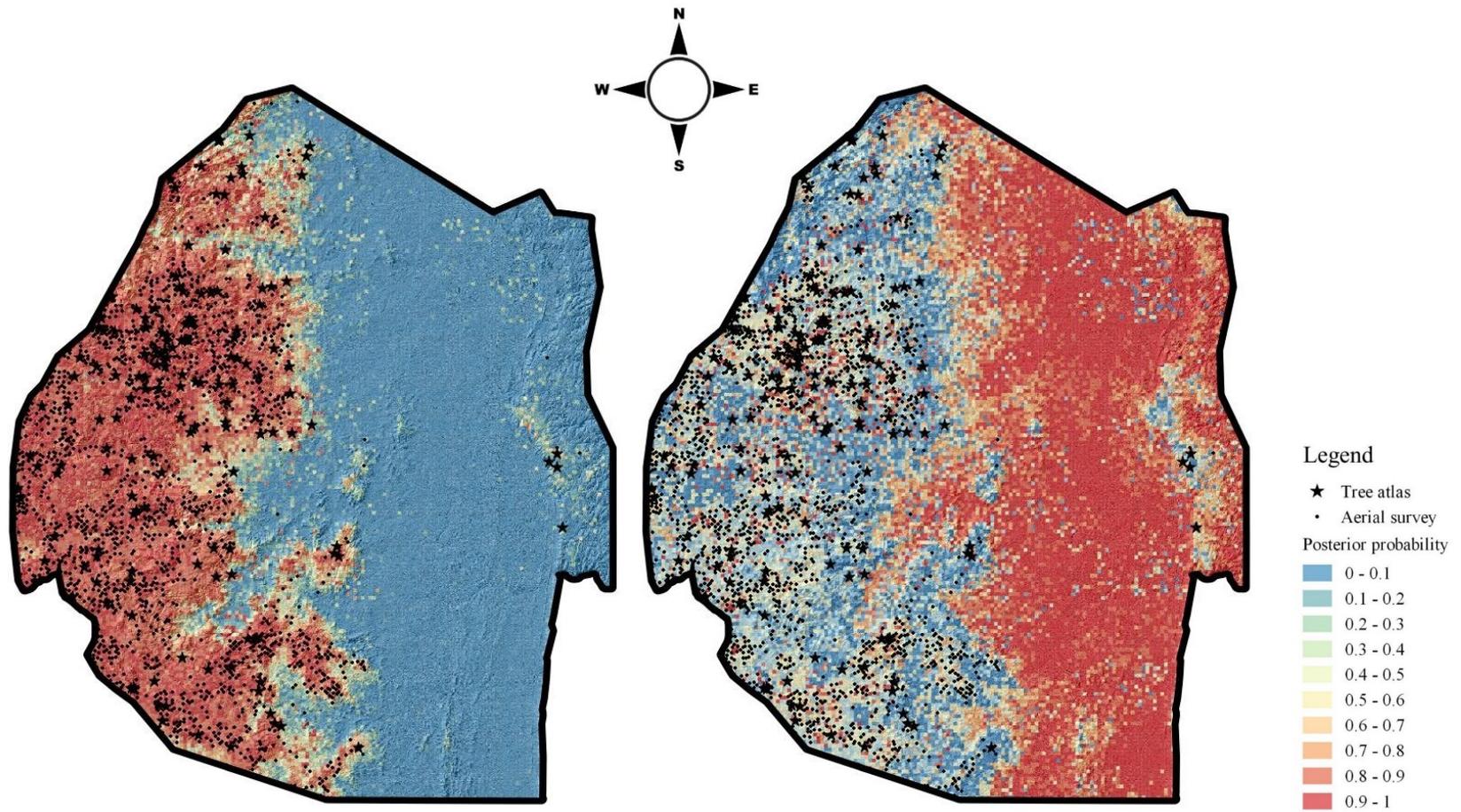


Figure 4.11: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *A. mearnsii* in Swaziland.

4.3.2 *Caesalpinia decapetala*

The ICS algorithm provided the best match between *C. decapetala* probability distributions and field data resulting in the low logarithmic loss. Hence, as shown Figure 4.12, the distribution of *C. decapetala*, was found to be determined by seven interdependent variables that were within the Markov blanket, namely the minimum temperature of the coldest month, human population density, land surface curvature, soil bulk density at 100-200cm depth and the presence of *L. camara*, *S. punicea* and *S. didymobotrya*. All the variables had direct arcs to the target variable whilst the minimum temperature of the coldest month was the root node.

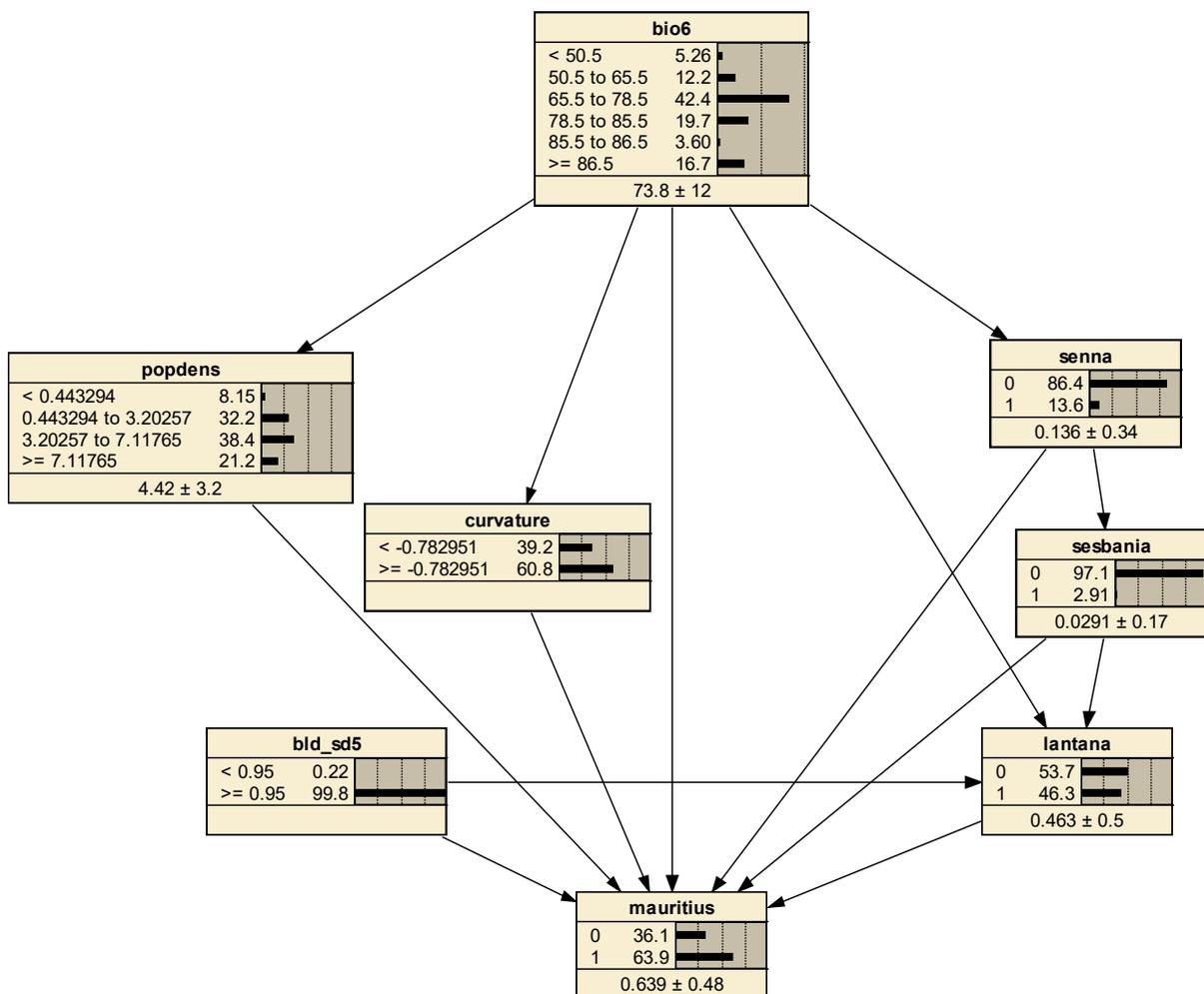


Figure 4.12: A learned Bayesian network for *Caesalpinia decapetala* distribution.

The discretization indicates optimum minimum temperature ranges between 5 and 8.5°C. The land surface curvature values indicate that *C. decapetala* occurs on hillsides and along drainage lines. Ultimately, the minimum temperature of the coldest month was the most important factor limiting the distribution of *C. decapetala* occurrence in Swaziland (Table 4.2). This was followed by human population density and the presence of *L. camara*, indicating possible biotic interactions between the two species. Human use of this species is the key driver of invasion in the country aided by biotic (possibly facilitative) interactions with other species such as *L. camara* and resource availability. The presence of *S. punicea* had the least influence relative to other factors.

Table 4.2: Mutual information for selected *Caesalpinia decapetala* predictor variables.

Variable	Mutual information
bio6	0.13669
popdens	0.05607
lantana	0.04903
curvature	0.02259
senna	0.00335
bld_sd5	0.0002
sesbania	0.00011

Figure 4.13 provides evidence of the influence of the minimum temperature of the coldest quarter on *C. decapetala* distribution as indicated by the BN in Figure 4.12. The influence of human population density manifests itself in the central part of the country, where there is high population density. The ensemble map in Figure 4.14 shows similar patterns whilst the PPCI map highlights areas of uncertainty particularly in uninvaded areas that are closer to where the species occurs.

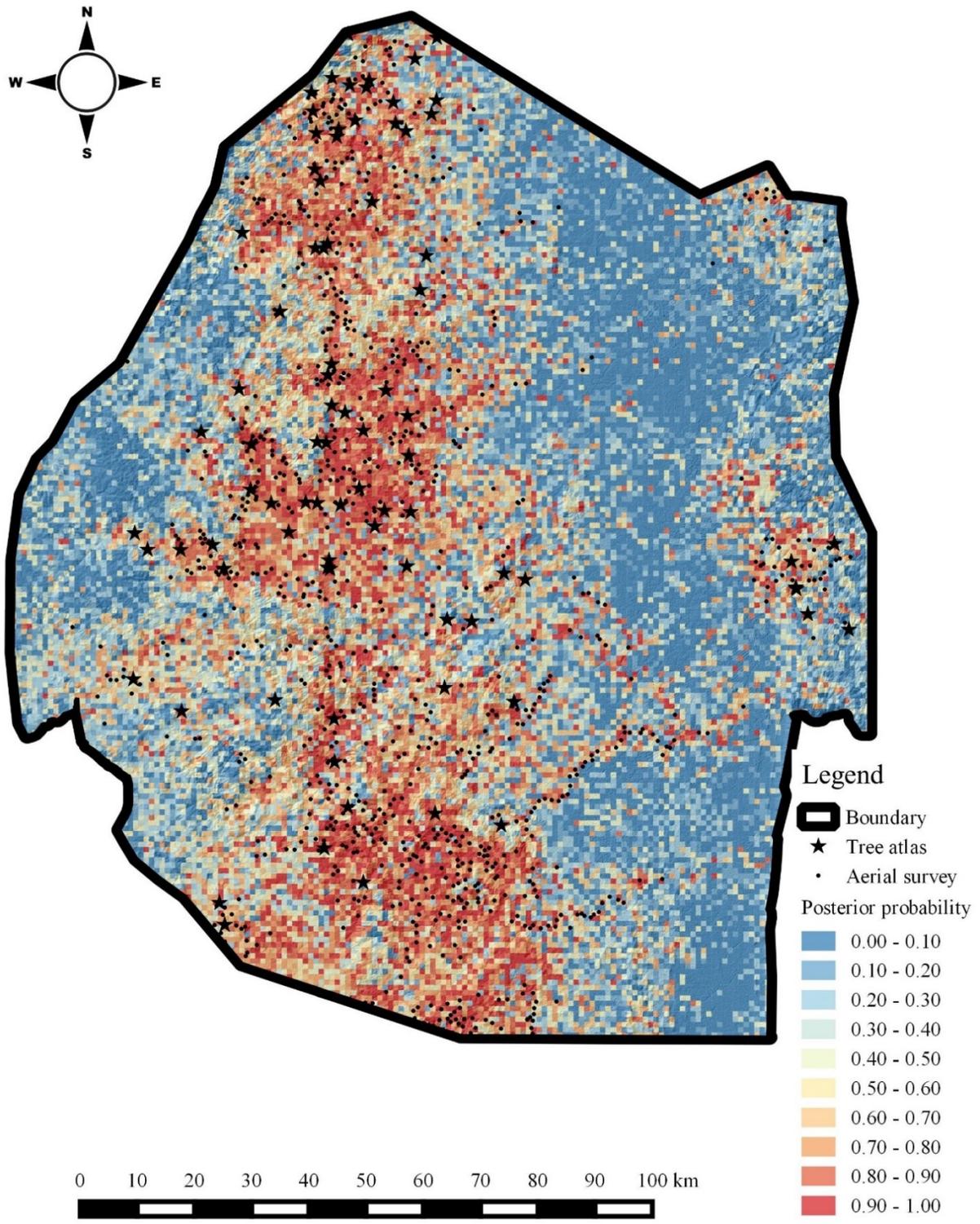


Figure 4.13: Posterior probability of occurrence for *C. decapetala* in Swaziland (derived from the BN in Figure 4.12).

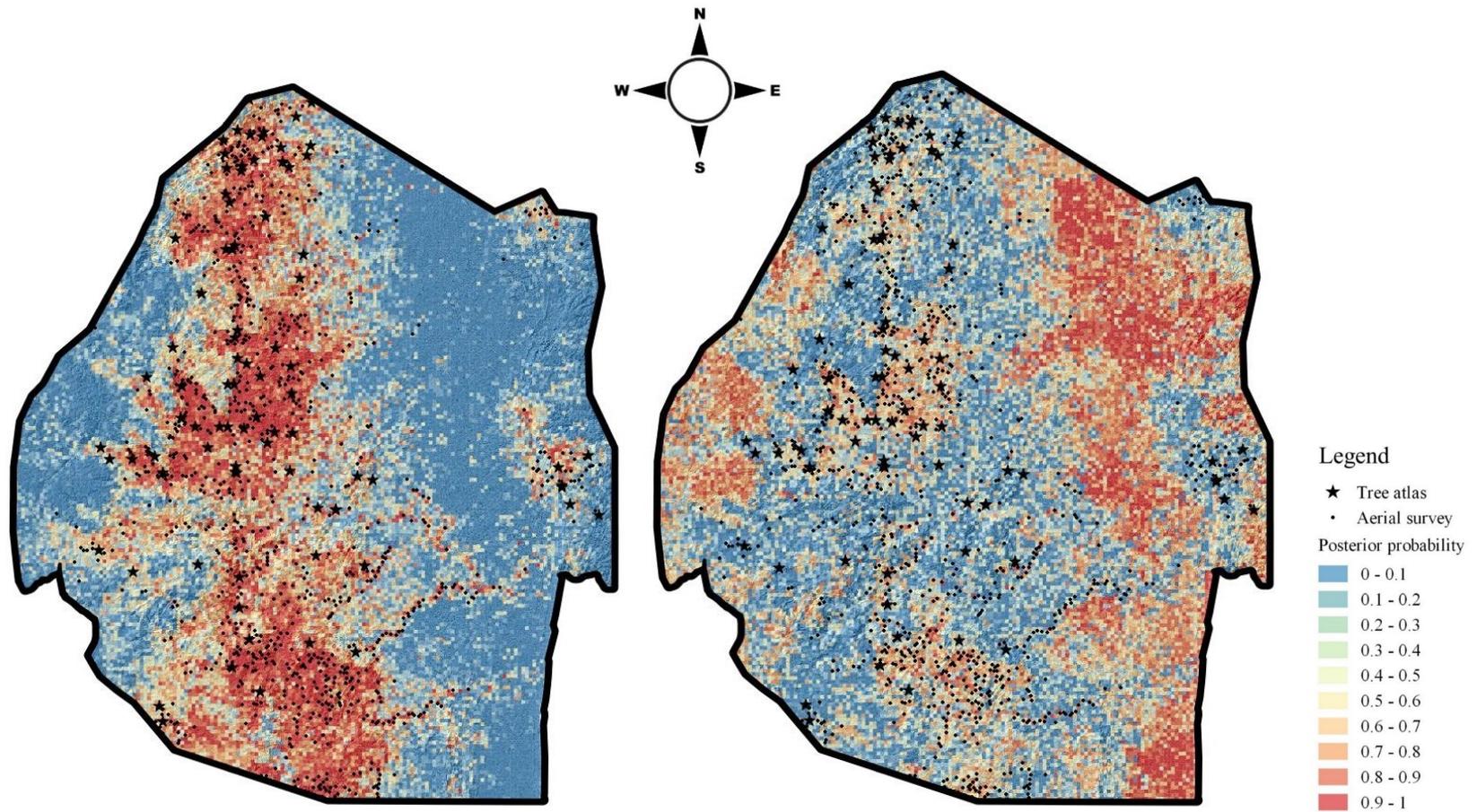


Figure 4.14: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *C. decapetala* in Swaziland.

4.3.3 *Cereus jamacaru*

Proximity to tourism sites and human disturbed areas, precipitation seasonality, precipitation of driest quarter and the occurrences of *S. mauritianum*, *J. mimosifolia*, *Opuntia* species and *M. azedarach* were the eight variables selected as important in determining the distribution of *C. jamacaru*. The best performing GBN structure was obtained through the simulated annealing algorithm learned through global scoring (Figure 4.15). All the variables were directly linked to the target variable in addition to interlinkages amongst themselves.

The models indicate that areas with low precipitation seasonality (coefficient of variation < 6.75) are preferred by the species as well as those areas with drier summer conditions (precipitation < 70mm). Areas near tourism areas (<50km) and human-disturbed areas (< 25km) are found to be vulnerable to invasion.

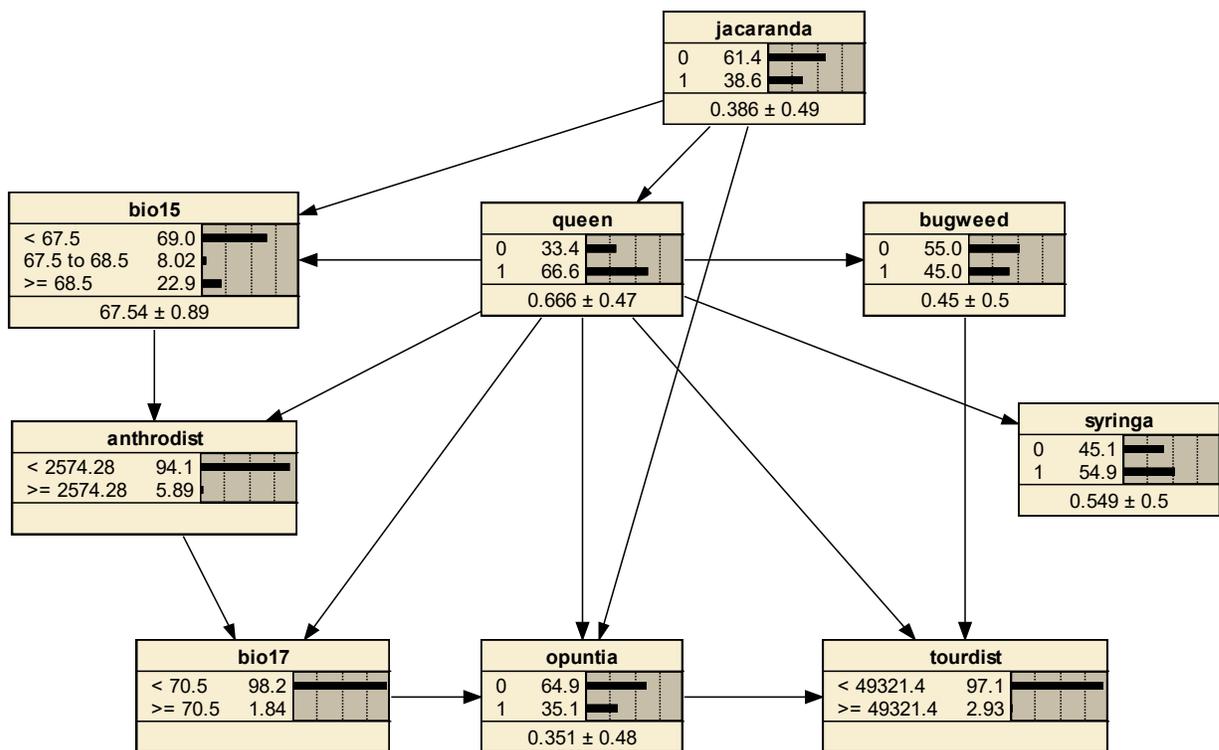


Figure 4.15: A learned Bayesian network for *Cereus jamacaru* distribution.

When considering the mutual information from the BN in Figure 4.15, the occurrence of the three species in particular *M. azedarach* were the more influential predictors of *C. jamacaru* distribution in Swaziland (Table 4.3). The proximity to human-disturbed sites had relatively lesser influence. Nevertheless, it suffices to say that based on the derived model, the occurrence of *C. jamacaru* is driven by human activity and constrained by precipitation regimes.

Table 4.3: Mutual information for selected *Cereus jamacaru* predictor variables.

Variable	Mutual information
syringa	0.41454
jacaranda	0.26531
opuntia	0.22754
bugweed	0.20979
bio15	0.04312
tourdist	0.02594
bio17	0.01174
anthrodist	0.01017

The predicted distribution of *C. jamacaru* occurrence is shown in Figure 4.16 and the ensemble shown in Figure 4.17. This distribution concurs with the BN in Figure 4.15 and the relative influence of the variables in Table 4.3. Although the PPCI values are generally high, prediction uncertainties exist in localities closer to where the species is currently observed. These represent areas where the species was predicted to occur but was not observed.

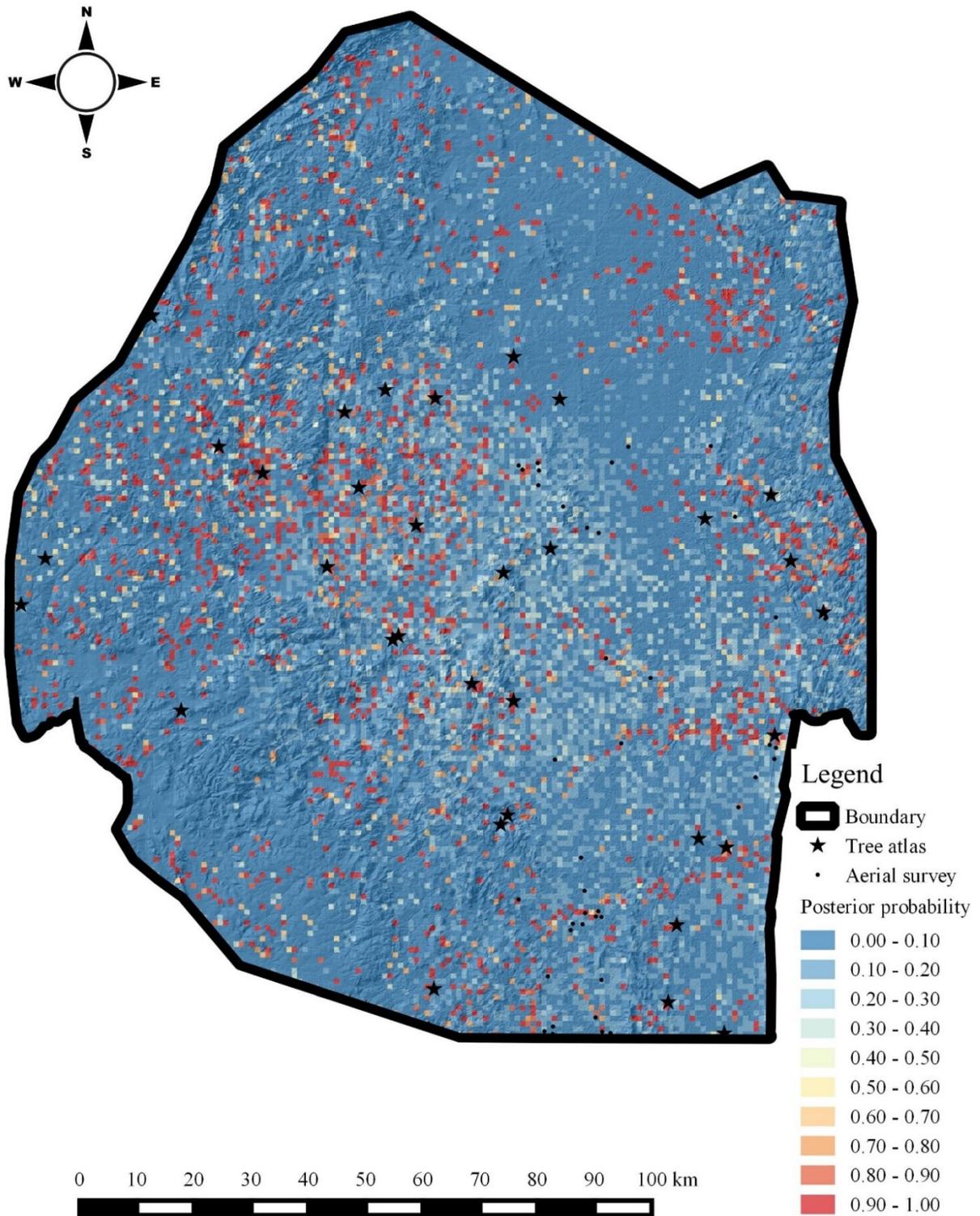


Figure 4.16: Posterior probability of occurrence for *C. jamacaru* in Swaziland (derived from the BN in Figure 4.15).

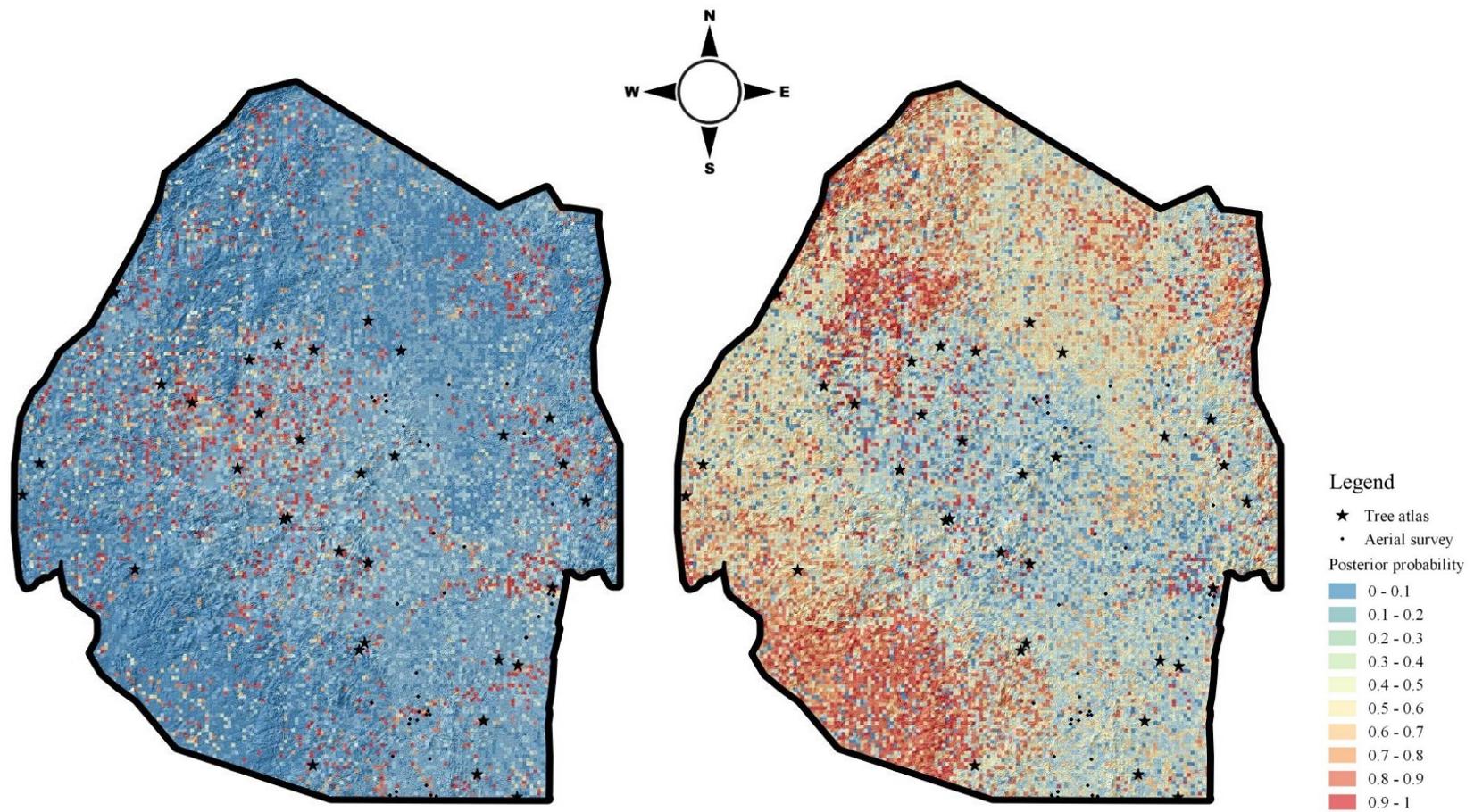


Figure 4.17: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *C. jamacaru* in Swaziland.

4.3.4 *Chromolaena odorata*

The hill-climbing algorithm learned with global scoring marginally outperformed all other algorithms for *C. odorata*, resulting in a GBN structure with six interacting variables having direct arcs to the target variable (Figure 4.18). The distribution of *C. odorata* in Swaziland is primarily determined by the minimum temperature of the coldest month, land surface curvature, land surface form, percentage of population with access to electricity and possible interactions with *C. decapetala* and *L. camara*.

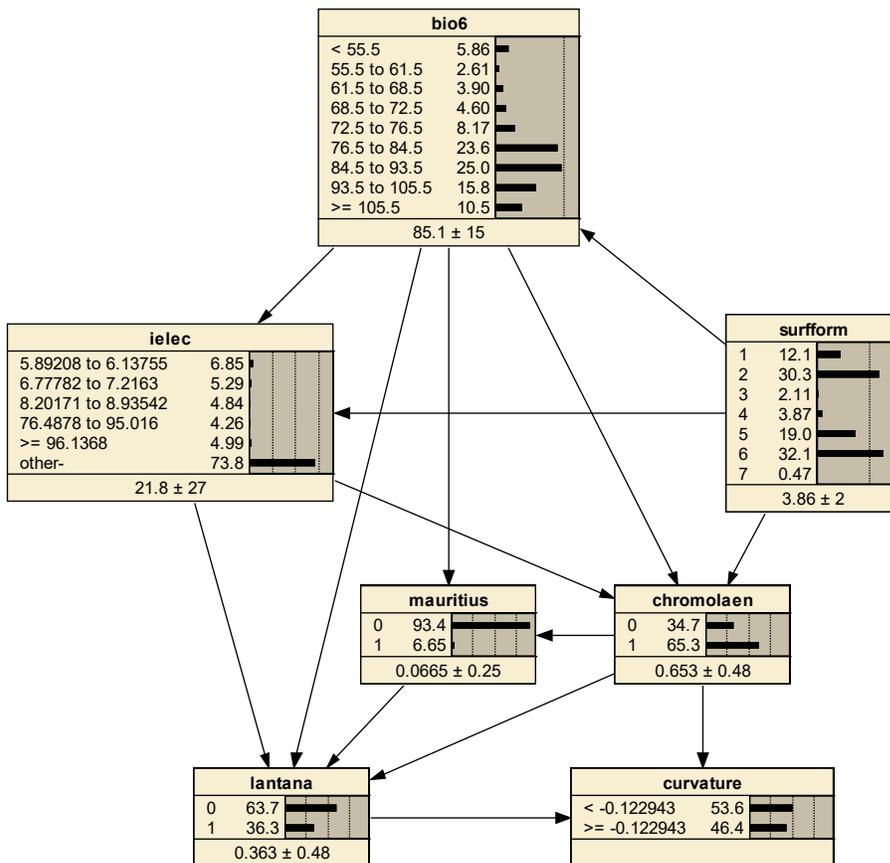


Figure 4.18: A learned Bayesian network for *Chromolaena odorata* distribution.

A minimum temperature of the coldest month of 7.2°C is found to be threshold for *C. odorata* distribution. Areas with a high percentage of people with access to electricity are also highly invaded by the species although the relationship is non-linear. Hills or low mountains and breaks

(drainage lines) are similarly prone to invasion as equally elucidated by the species' occurrence relationship with surface form and land surface curvature.

The percentage of people with access to electricity and the minimum temperature of the coldest month are the strongest predictors of *C. odorata* occurrence whilst land surface form has the least influence of the six variables (Table 4.4). Hence, the socioeconomic conditions of an area, which likely determine the extent and nature of human use of environmental resources, is the key facilitator of *C. odorata* invasion. This is constrained by low temperature conditions, resource availability, and biotic (probably facilitative/mutualistic) interactions with *L. camara* and *C. decapetala*.

Table 4.4: Mutual information for selected *Chromolaena odorata* predictor variables.

Variable	Mutual information
ielec	0.15692
bio6	0.09207
lantana	0.01074
curvature	0.00532
mauritus	0.00545
surfform	0.00194

The strong influence of the interacting and dominant factors shown in Table 4.4 and Figure 4.18 result in the predicted spatial distribution in Figure 4.19. The ensemble model in Figure 4.20 confirms this spatial pattern. The high prediction uncertainty areas in Figure 4.20 are found in highly suitable areas near currently invaded areas.

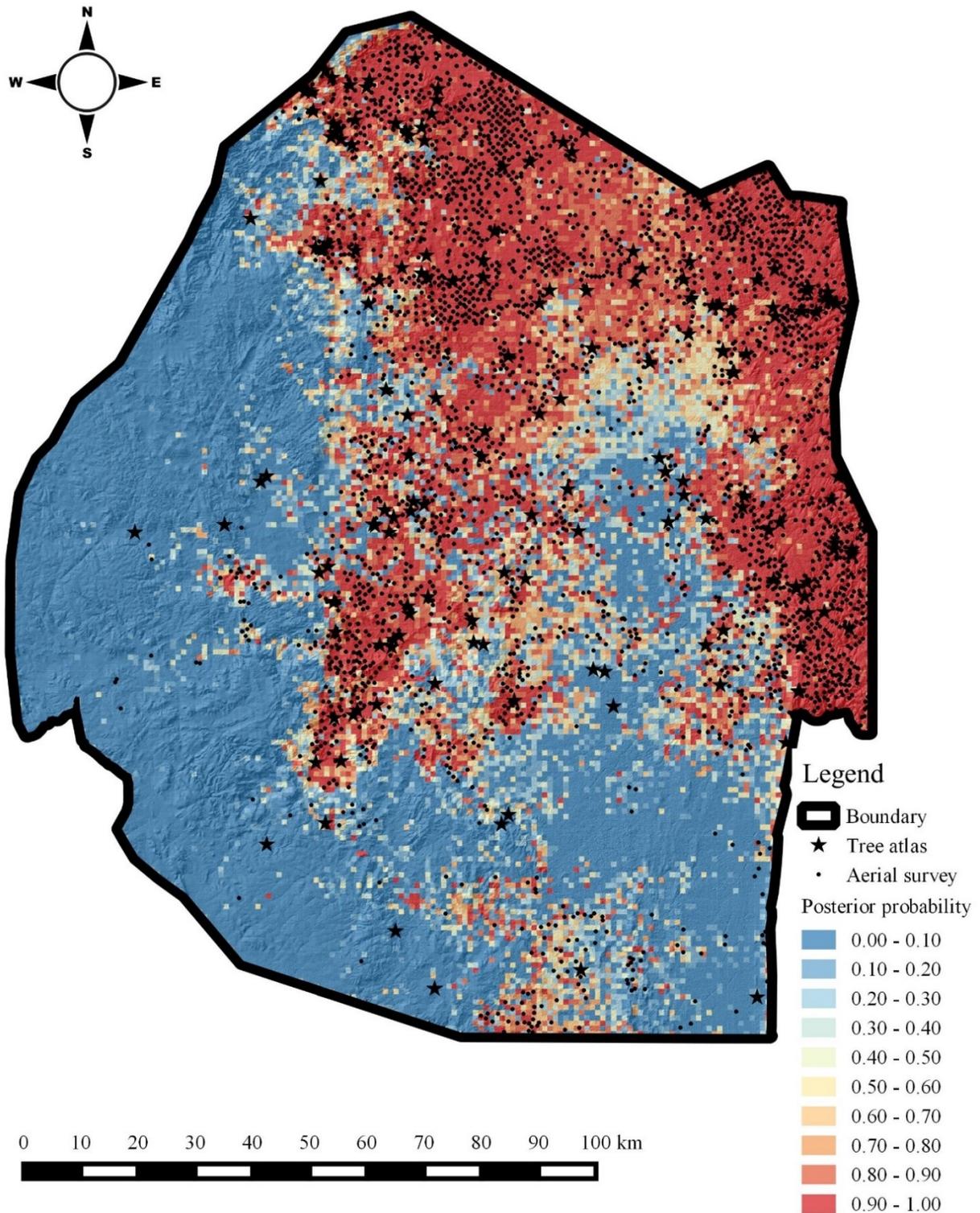


Figure 4.19: Posterior probability of occurrence for *C. odorata* in Swaziland (derived from the BN in Figure 4.18).

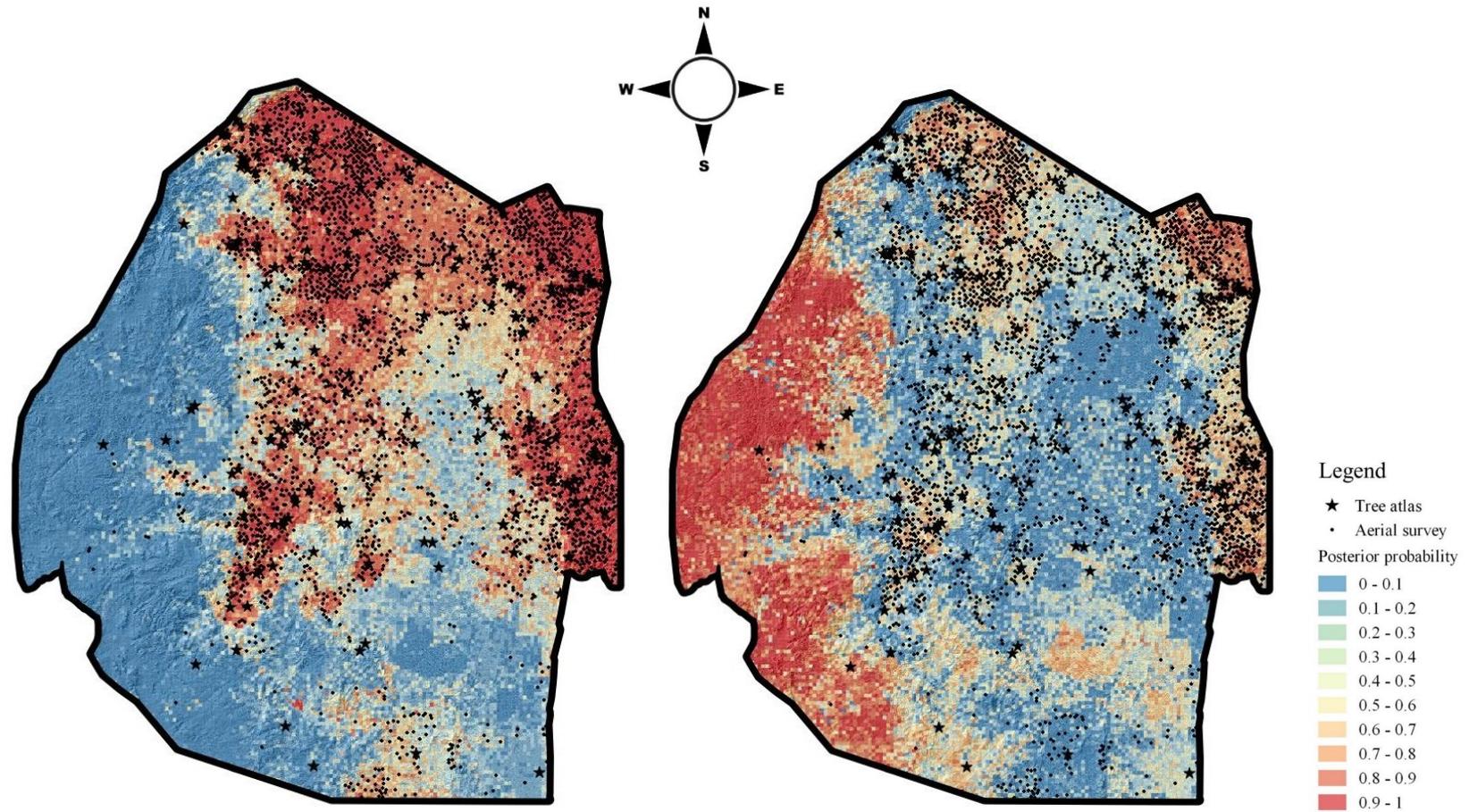


Figure 4.20: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *C. odorata* in Swaziland.

4.3.5 *Eucalyptus* species

The derived models indicate that the distribution of *Eucalyptus* species is predominantly governed by the aridity index, land use, proximity to major roads, slope aspect, proximity to rivers. There are also associations, co-occurrences with *S. mauritianum* and *A. mearnsii*. The best performing algorithm was the repeated hill climbing algorithm learned through global scoring which highlights the interdependencies between all the seven variables (Figure 4.21). The resultant BN reveals a structure with colliding or converging arcs and conditional dependencies between the variables.

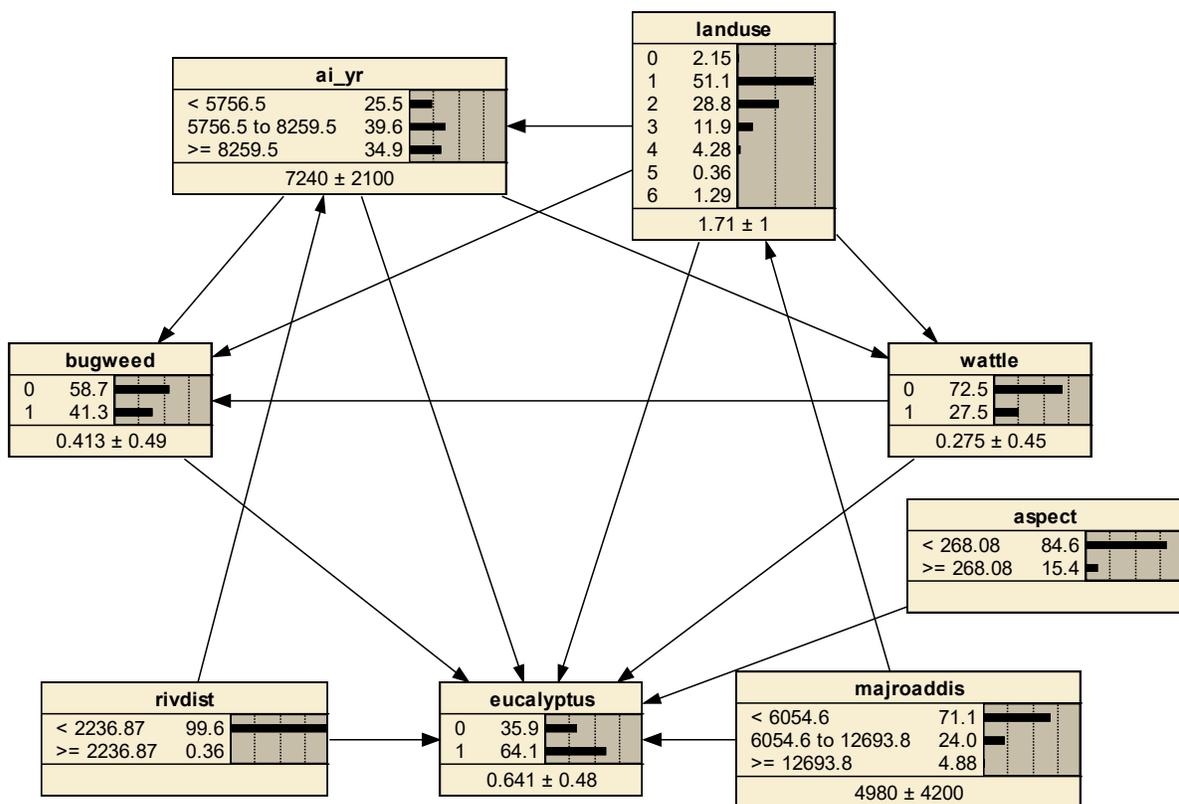


Figure 4.21: Learned Bayesian network for *Eucalyptus* species distribution.

Areas with aridity indices above 0.58 as well as areas within 2.2km from rivers and streams were found suitable. Similarly, the BN indicates the influence of land use (primarily plantation

forestry) and areas within 6km from major roads. The relations with slope aspect (<268°) indicate that *Eucalyptus* species grow predominantly on west-facing hill slopes.

The sensitivity analysis indicates that aridity is the strongest predictor of *Eucalyptus* occurrence followed by the presence of *S. mauritianum* and land use (Table 4.5). Proximity to rivers and slope aspect had the least influence compared to the rest of the variables. *Eucalyptus* is commensal with humans and hence its establishment and spread is primarily through propagation within earmarked land parcels, facilitated by road infrastructure and constrained by resource (moisture) availability.

Table 4.5: Mutual information for selected *Eucalyptus* species predictor variables.

Variable	Mutual information
ai_yr	0.168
bugweed	0.13778
landuse	0.1136
wattle	0.05764
majroaddis	0.03021
aspect	0.0051
rivdist	0.00074

Figure 4.22 and Figure 4.23 show the spatial predictions of the BN in Figure 4.21. In conformity to the mutual information values in Table 4.5, the wetter higher elevation areas are predicted to be more suitable for *Eucalyptus* establishment albeit with high uncertainty in some areas (Figure 4.23). Low PPCI values are predominantly in the uninvaded areas that have a suitable niche.

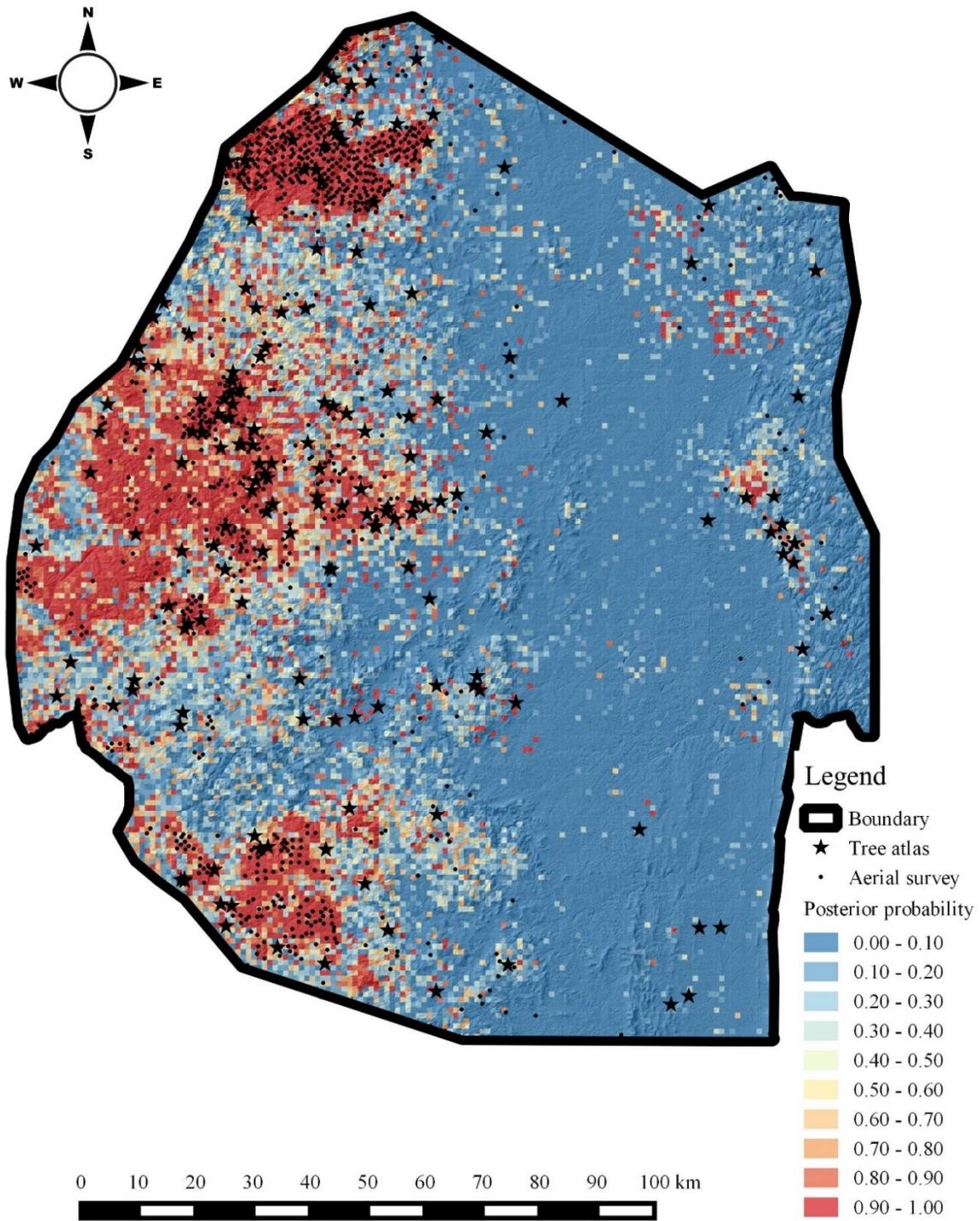


Figure 4.22: Posterior probability of occurrence for *Eucalyptus* in Swaziland (derived from the BN in Figure 4.21).

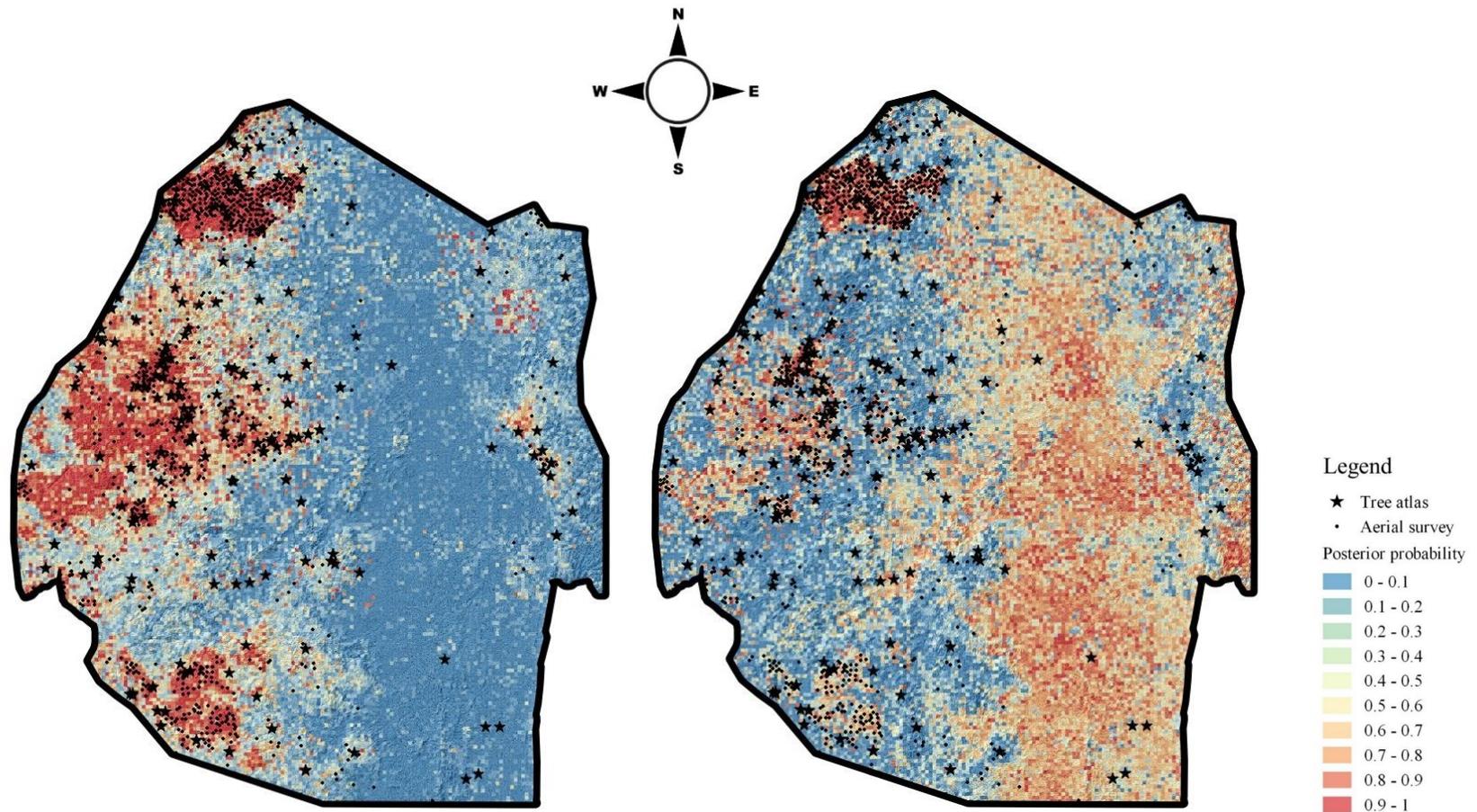


Figure 4.23: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *Eucalyptus* species in Swaziland.

4.3.6 *Jacaranda mimosifolia*

Simulated annealing with global scoring was the best performing algorithm resulting in the BAN structure in Figure 4.24. As revealed, the spatial distribution of *J. mimosifolia* is strongly determined by human settlement density and temperature annual range, which were the only abiotic variables. The other variables were associations with *P. guajava*, *Eucalyptus* species, *S. mauritianum*, *Opuntia* species, *M. azedarach* and *A. mearnsii*.

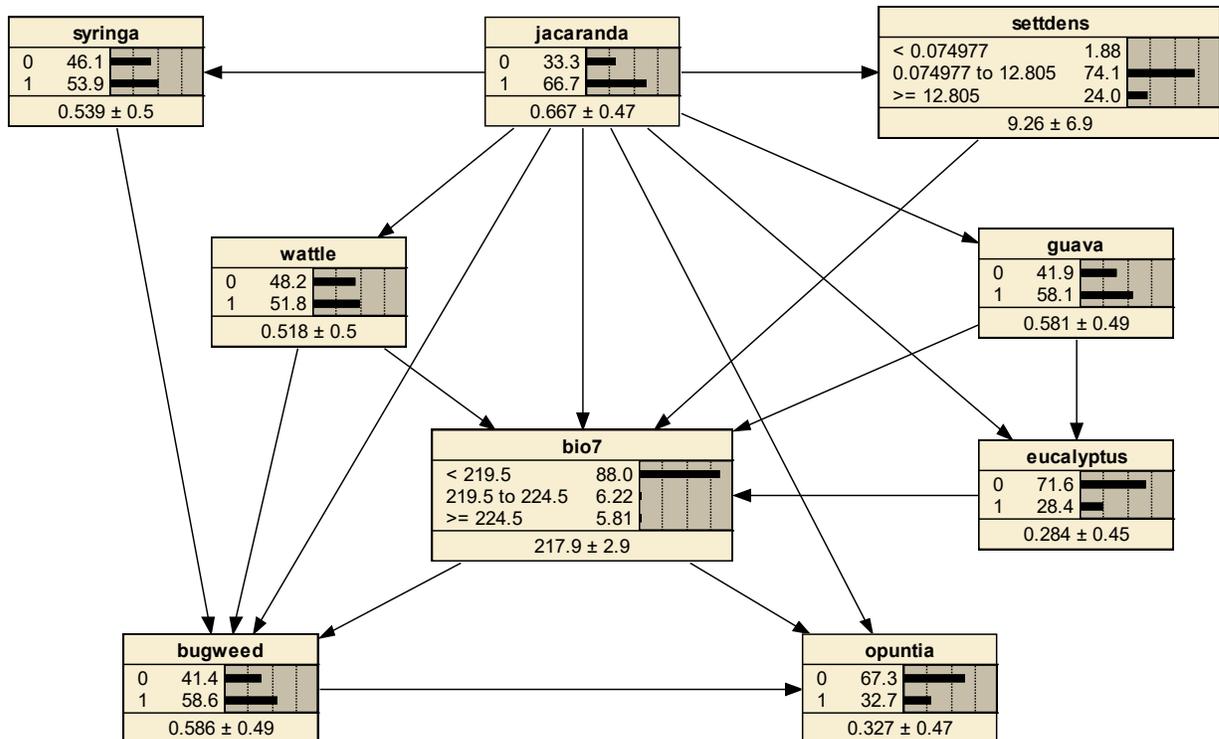


Figure 4.24: A learned Bayesian network for *Jacaranda mimosifolia* distribution.

Areas with temperature annual range below 22°C restrict this species' distribution. Posterior probabilities are similarly high in areas where human settlement densities exceed 13 settlements/km². The invasion process of this species is, therefore, facilitated by humans and constrained by temperature range. Table 4.6 indicates that the occurrence of *J. mimosifolia* is strongly associated with *P. guajava*, *S. mauritianum*, *M. azedarach* and *A. mearnsii* while the settlement density had relatively the least influence.

Table 4.6: Mutual information for selected *Jacaranda mimosifolia* predictor variables.

Variable	Mutual information
guava	0.4258
bugweed	0.40161
syringa	0.40131
wattle	0.35252
opuntia	0.20338
eucalyptus	0.11714
bio7	0.06109
settdens	0.04988

The BN in Figure 4.24 coupled with the relative influence of each factor, as shown in Table 4.6, results in the spatial predictions in Figure 4.25 (see also Figure 4.26). There is an observed widespread distribution of *J. mimosifolia* albeit with localized invasions closer to human settlements. The PPCI values in Figure 4.26 indicate high certainty in the predictions throughout the country save for highly suitable areas which are currently uninvaded.

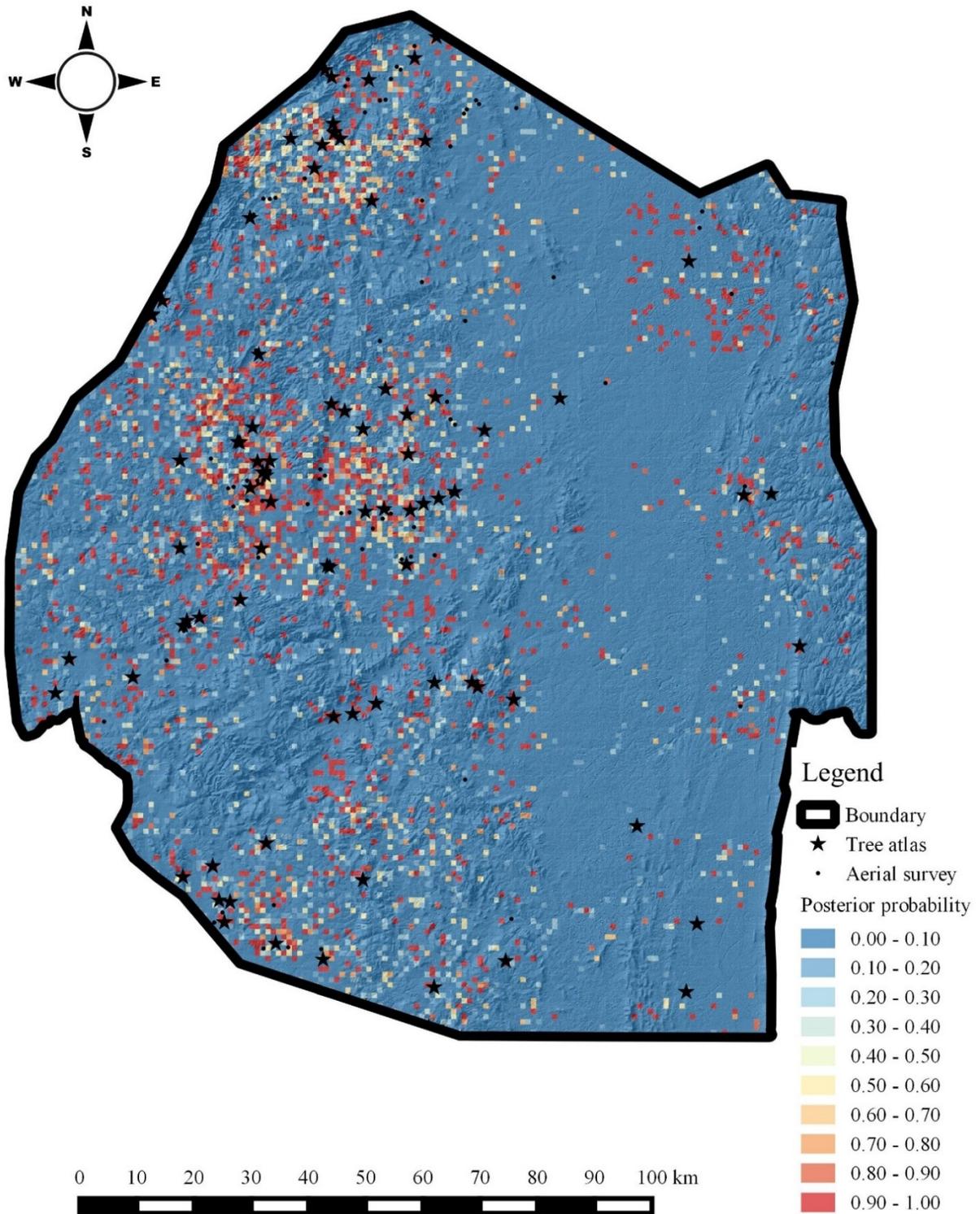


Figure 4.25: Posterior probability of occurrence for *J. mimosifolia* in Swaziland (derived from the BN in Figure 4.24).

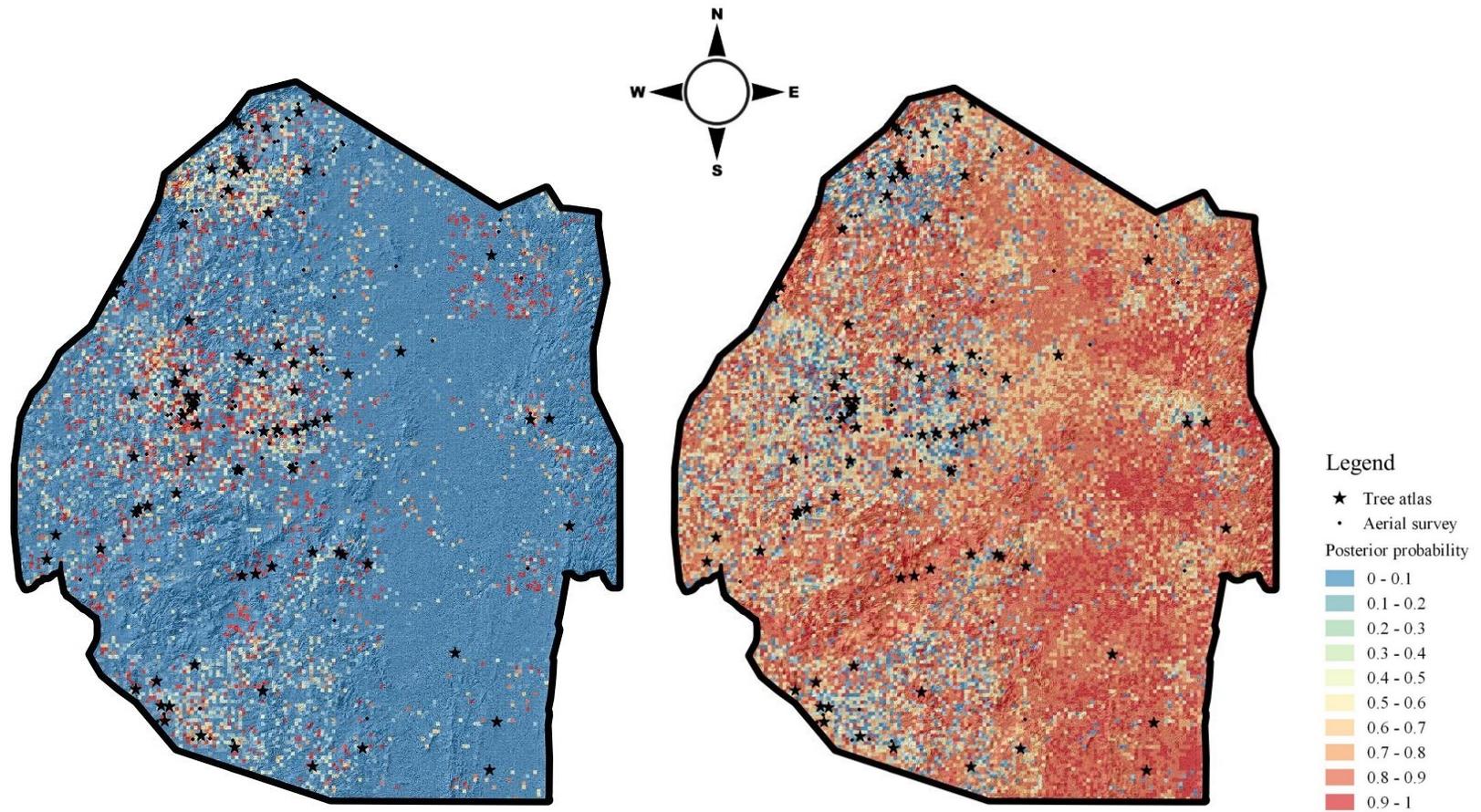


Figure 4.26: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *A. mearnsii* in Swaziland.

4.3.7 *Lantana camara*

The foremost determinants of *L. camara* distribution in Swaziland are precipitation seasonality, precipitation of the wettest month, human population density, alien plant density, slope aspect, the presence of *C. decapetala* and *C. odorata*, and the fraction of sand and silt at 5-15cm and 30-60cm depths, respectively. In Figure 4.27 is the structure learned using the best performing algorithm, the locally scored repeated hill climbing algorithm wherein all the predictor variables have arcs to the target variable. Interestingly, the presence of *L. camara* was also observed to be an important factor in the distribution of *C. decapetala* and *C. odorata*.

Optimal conditions include precipitation of the wettest month between 99.5 to 134.5mm and areas with precipitation coefficient of variation between 58.5 and 67.5%. There is also an observed preference for slope aspect less than 285.5°. Other highly susceptible areas include those where human population density exceeds 2.2 people/km², soils with low silt content (< 11.5 g/kg) at 30-60cm depth and high sand content (>56.5g/kg) at 5-15cm.

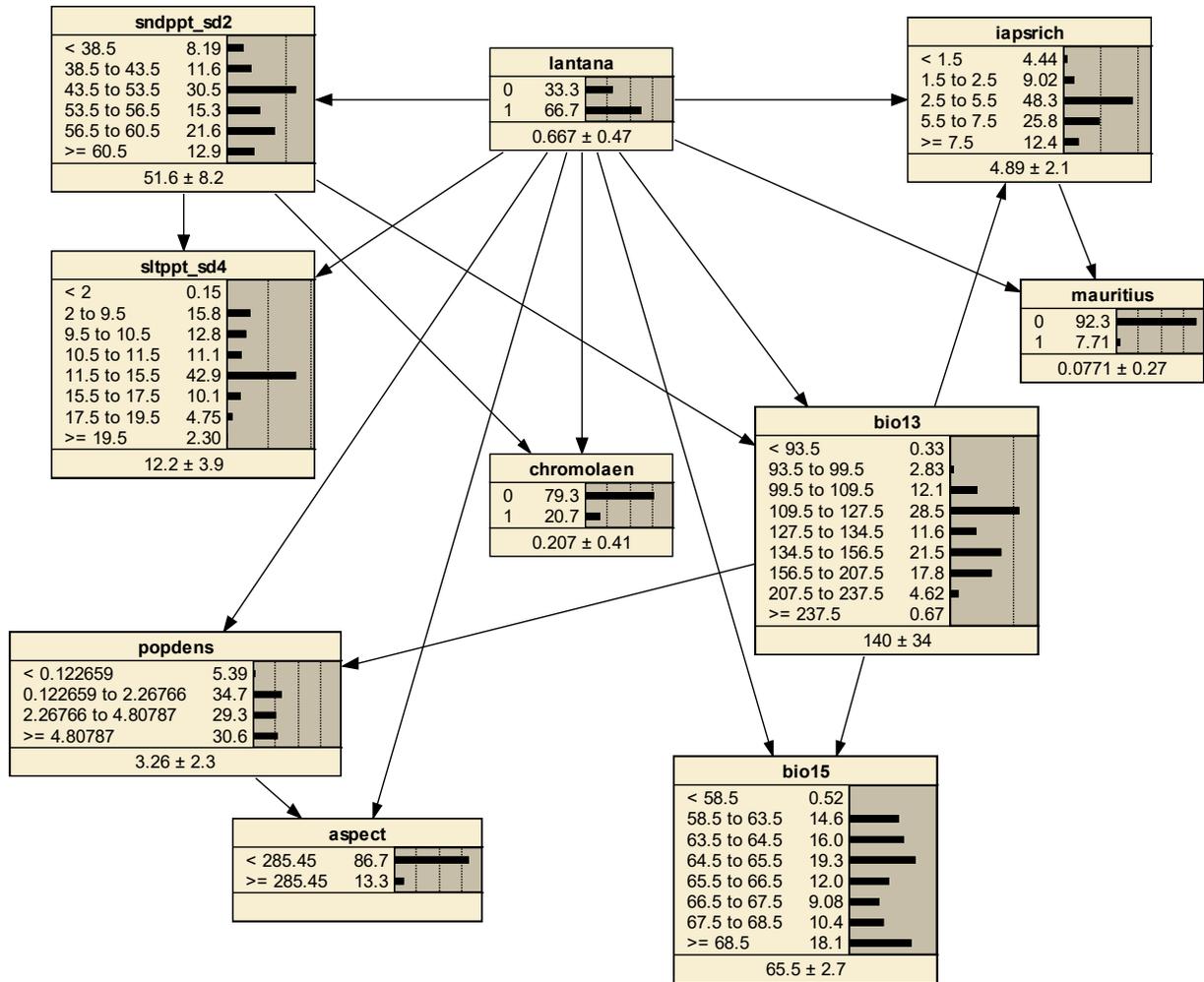


Figure 4.27: A learned Bayesian network for *Lantana camara* distribution.

The strong influence of bioclimatic variables in *L. camara* distribution is substantiated by the mutual information values in Table 4.7 where precipitation of the wettest quarter and precipitation seasonality were the strongest predictors followed by the fraction of sand and silt at 5-15cm and 30-60cm depths, respectively. The slope aspect was relatively the weakest predictor. The *L. camara* invasion process is, therefore, complex and determined by resource availability, biotic (mutualistic and facilitative) interactions with other invasive plants and mediated by human activity.

Table 4.7: Mutual information for selected *Lantana camara* predictor variables.

Variable	Mutual information
bio15	0.11114
bio13	0.05704
iapsrich	0.04404
sltppt_sd4	0.03818
popdens	0.03333
sndppt_sd2	0.03455
mauritiu	0.01702
chromolaen	0.01385
aspect	0.00212

The complex spatial distribution of *L. camara* is shown in Figure 4.28. The ensemble of all the algorithms shows a similar spatial pattern resulting from the complex interplay of the factors in Table 4.7, particularly precipitation seasonality. As expected, prediction uncertainty was high in those areas where the habitat was suitable but the species was either absent or rarely observed (Figure 4.29). These are primarily the areas bordering those that are currently invaded.

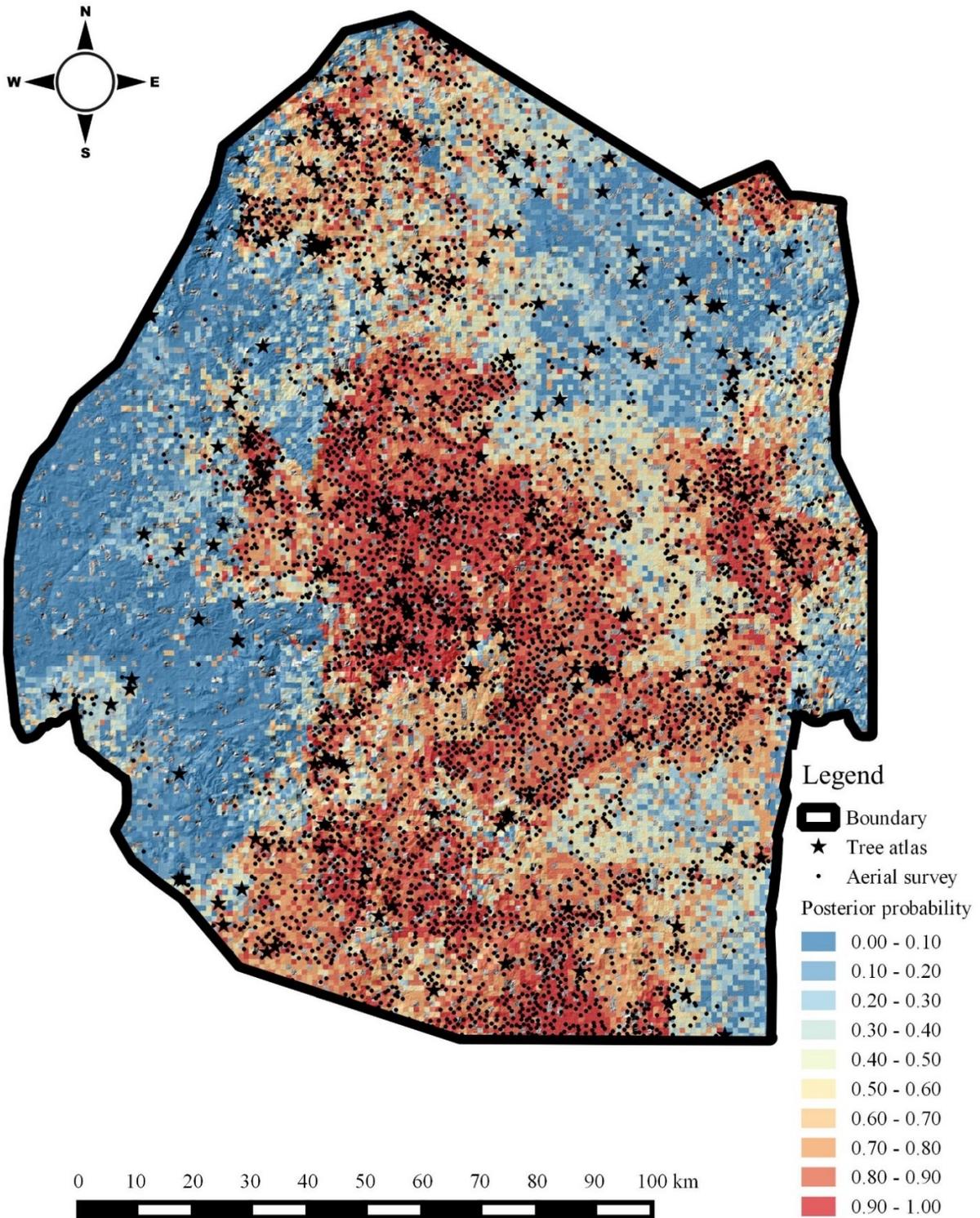


Figure 4.28: Posterior probability of occurrence for *L. camara* in Swaziland (derived from the BN in Figure 4.27).

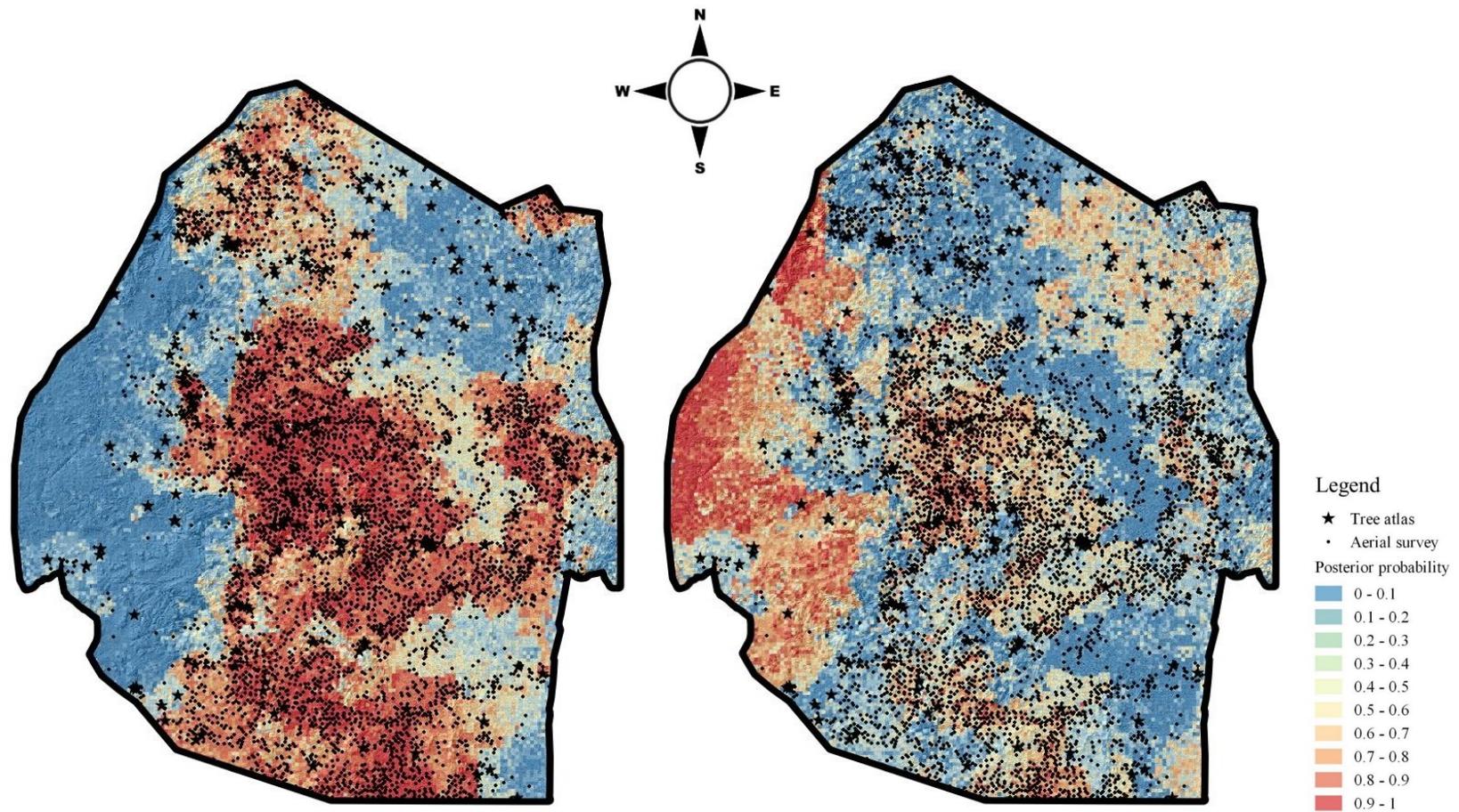


Figure 4.29: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *L. camara* in Swaziland.

4.3.8 *Melia azedarach*

The spatial distribution of *M. azedarach* is controlled by temperature seasonality, solar radiation duration, August soil water content, and the occurrence of other invasive plant species namely *S. didymobotrya*, *Opuntia* spp., *C. jamararu*, and *J. mimosifolia*. A BAN structure with arcs pointing to the target variable was learned with the globally scored K2 algorithm (Figure 4.30).

Late winter soil water content less than 57.5mm and a temperature coefficient of variation less 35.2 indicates the *M. azedarach*'s preference for seasonally dry conditions. Areas with direct radiation duration values below 4268 are also suitable habitat.

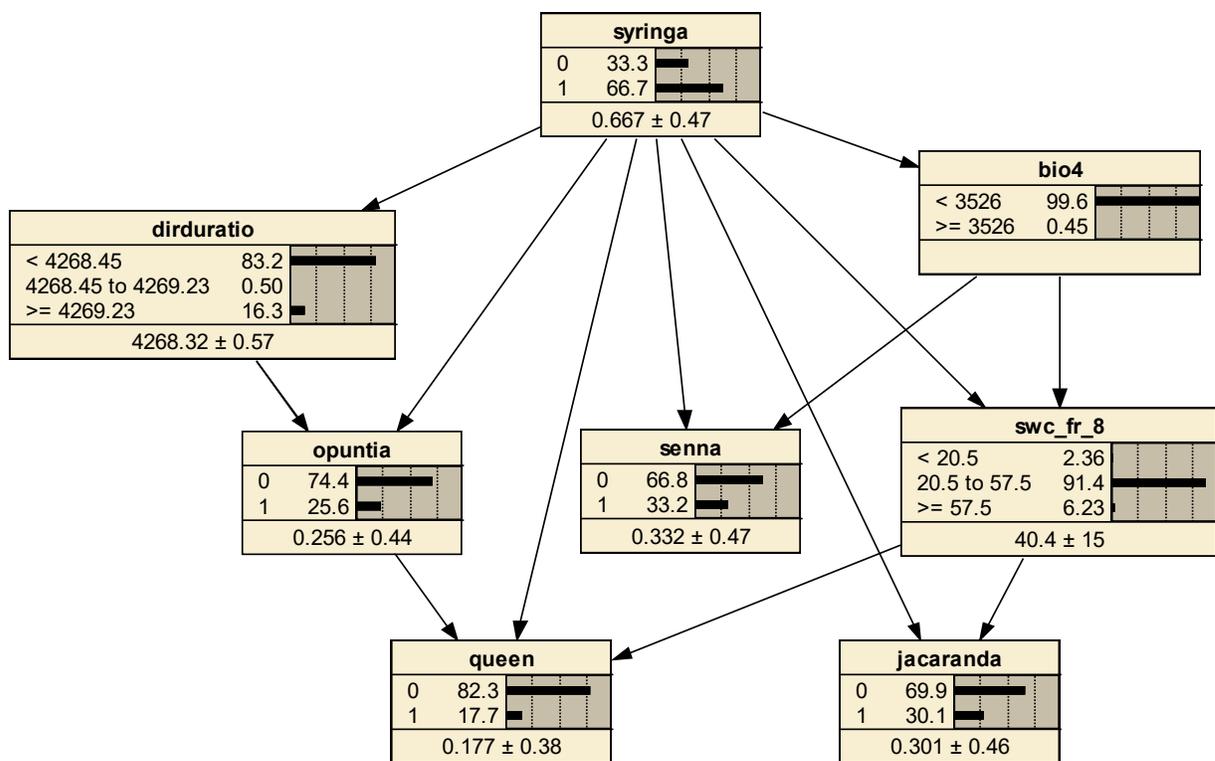


Figure 4.30: A learned Bayesian network for *Melia azedarach* distribution.

The mutual information values highlight the fact that *M. azedarach* distribution is primarily associated with other invasive alien plants such as *S. didymobotrya*, followed by *J. mimosifolia*, and *Opuntia* species (Table 4.8). Temperature seasonality had the least influence compared to

the other variables. The results, therefore, indicate that resource availability, temperature seasonality and possible biotic interactions or associations drive *M. azedarach* invasion.

Table 4.8: Mutual information for selected *Melia azedarach* predictor variables.

Node	Mutual information
senna	0.20342
jacaranda	0.19601
opuntia	0.15772
queen	0.09511
dirduratio	0.01173
swc_fr_8	0.01205
bio4	0.00592

These determining factors interact and are probabilistically dependent as depicted in Figure 4.30 resulting in the predicted distribution in Figure 4.31 and Figure 4.32. The co-occurrence with *S. didymobotrya* is evidenced by the high probabilities near watercourses, as is the co-occurrence with *J. mimosifolia* near human populated areas. However, there are high prediction uncertainties throughout the country except where the species was not observed.

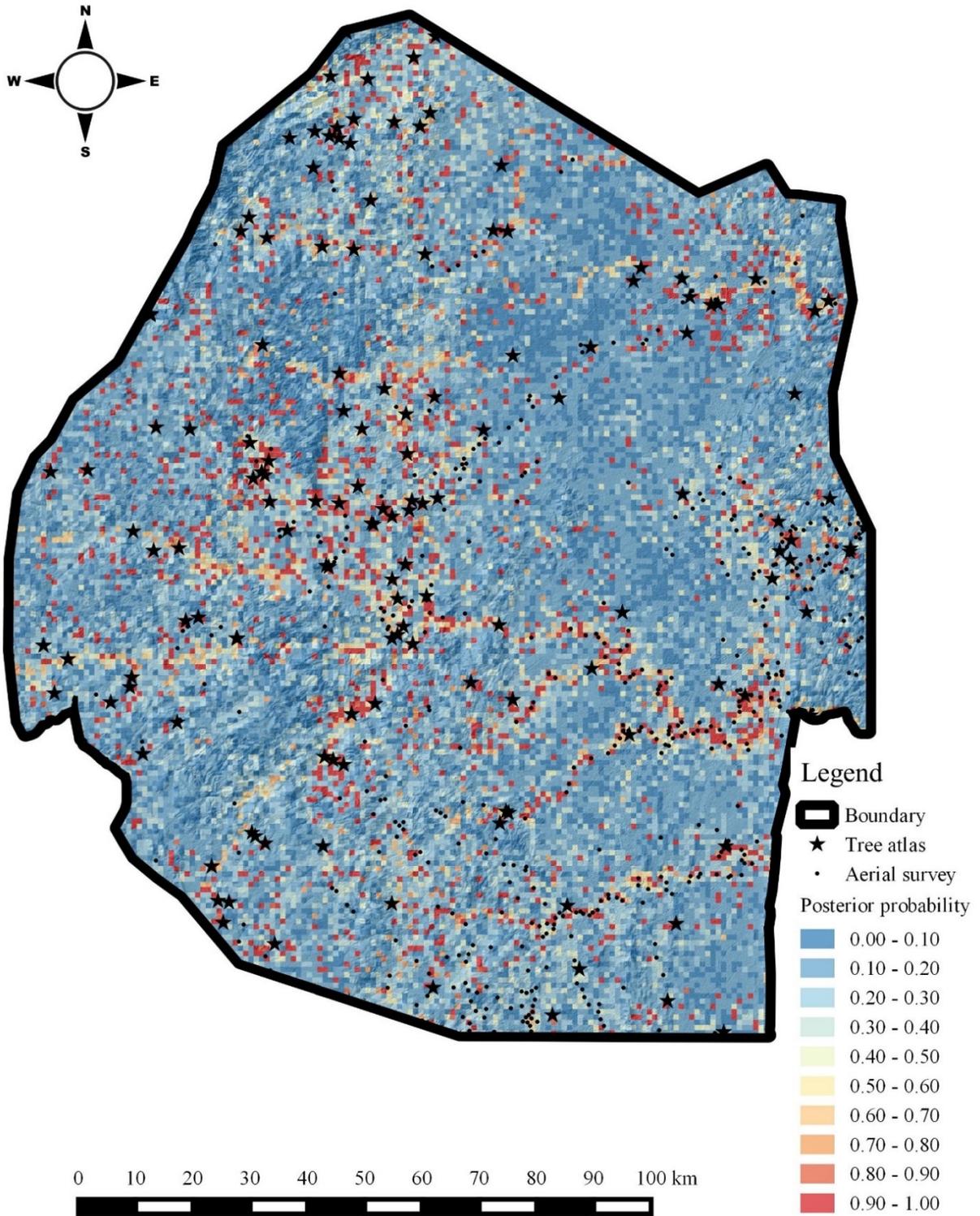


Figure 4.31: Posterior probability of occurrence for *M. azedarach* in Swaziland (derived from the BN in Figure 4.30).

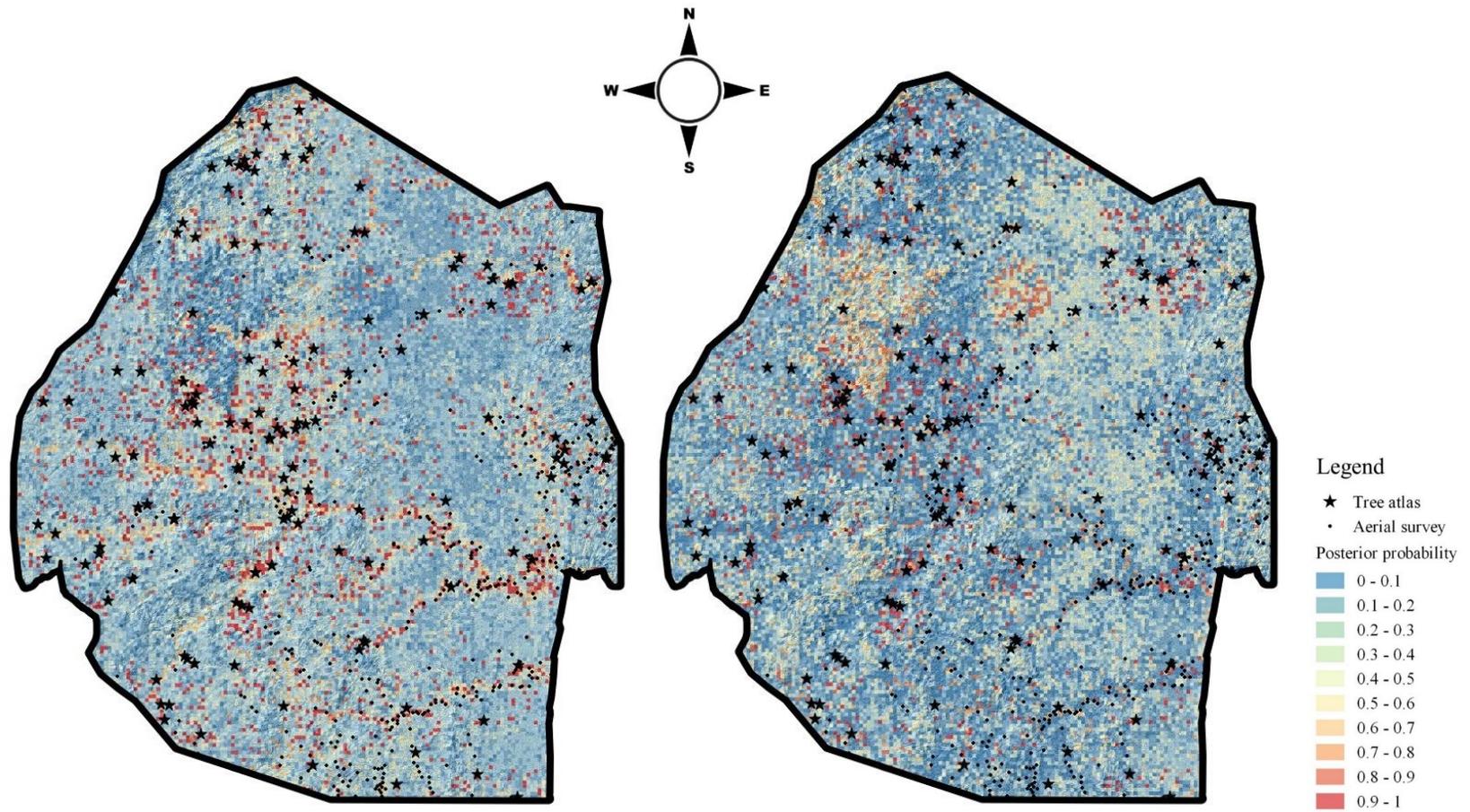


Figure 4.32: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *M. azedarach* in Swaziland.

4.3.9 *Opuntia* species

Proximity to rivers, May actual evapotranspiration and proximity to major roads were the main abiotic predictors of *Opuntia* species occurrence together with associative relationships with *P. guajava*, *C. jamacaru*, *J. mimosifolia*, and *M. azedarach*. The best performing algorithm, the tabu search with global scoring, derived a GBN structure with all the variables having direct arcs to the target variable (Figure 4.33).

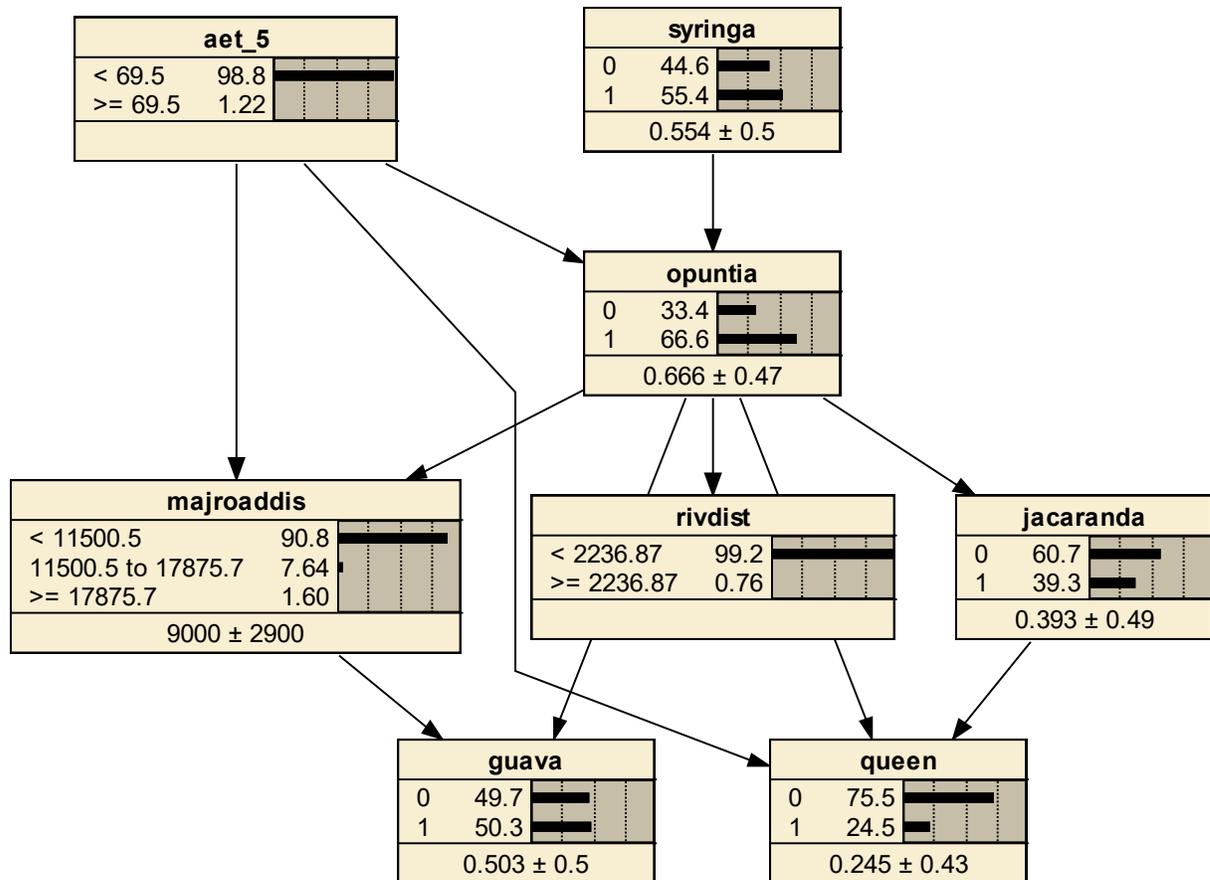


Figure 4.33: A learned Bayesian network for *Opuntia* species distribution.

Higher posterior probabilities of *Opuntia* species occurrence were predicted in areas with actual evapotranspiration values lower than 69.5mm. Areas within 2km from rivers are likewise important for this species as are areas within 11km from major roads.

The occurrence of *M. azedarach* had the strongest predictive power on *Opuntia* species occurrence followed by that of *P. guajava* and *J. mimosifolia* (Table 4.9). May actual evapotranspiration had relatively weaker predictive power compared to the other variables. Nevertheless, resource availability is a key driver of *Opuntia* species invasion facilitated by human activity through transportation routes whilst associations with other species provided strong prediction information.

Table 4.9: Mutual information for selected *Opuntia* species predictor variables.

Variable	Mutual information
syringa	0.42754
guava	0.28772
jacaranda	0.26583
queen	0.13551
majroaddis	0.02996
rivdist	0.00563
aet_5	0.00515

The BN in Figure 4.33 predicted the spatial distribution of *Opuntia* species as shown in Figure 4.34. The ensemble prediction of all the algorithms can be seen in in Figure 4.35. The spatial patterns are a result of the relative influences of the factors including co-occurrence with the species shown in Table 4.9. The influence of roads and rivers is predominantly evident. However, there is moderate uncertainty spread throughout the country whilst high certainty is found in areas where the species is predicted to be absent.

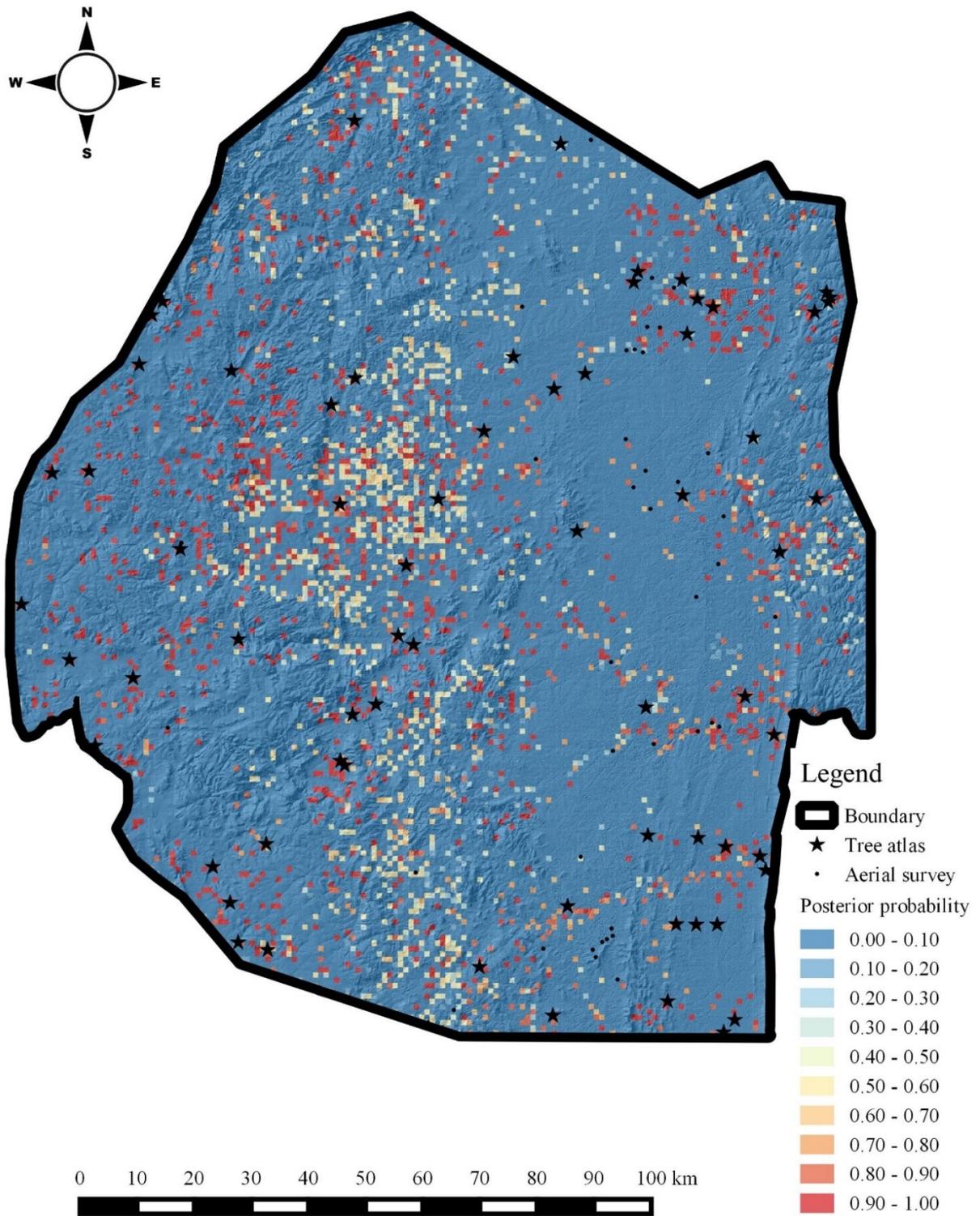


Figure 4.34: Posterior probability of occurrence for *Opuntia* species in Swaziland (derived from the BN in Figure 4.33).

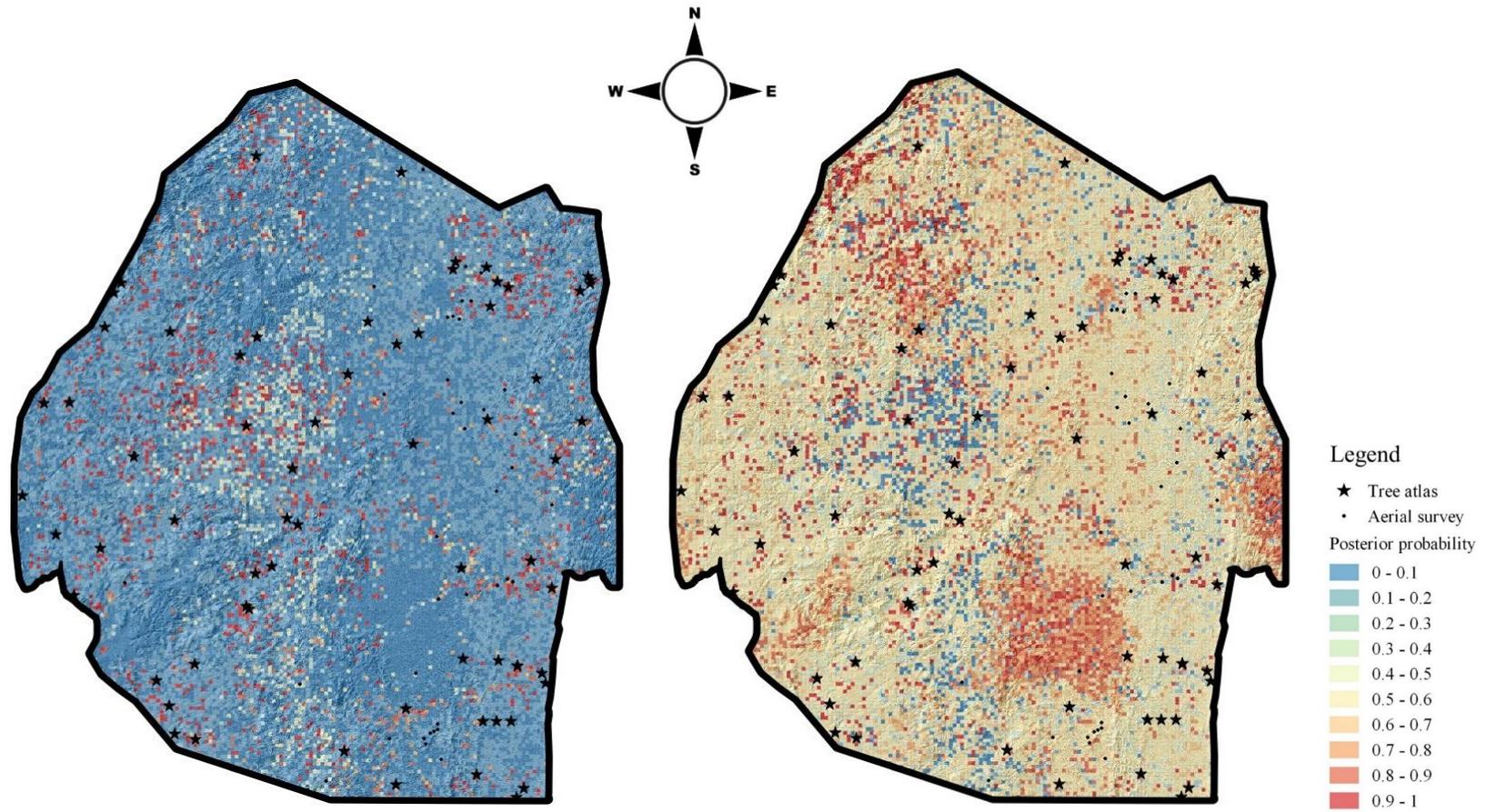


Figure 4.35: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *Opuntia* species in Swaziland.

4.3.10 *Pinus* species

Invasive plant and tree species richness, mean temperature of the wettest quarter, April actual evapotranspiration, cattle density, land cover, land use, river/stream density, proximity to major rivers, slope aspect and the presence of *S. mauritianum* were the most relevant and least redundant predictors of *Pinus* species. The best performing algorithm was the TAN structure learned through local scoring (Figure 4.36).

As expected, land used for plantation forestry was highly probable to harbour *Pinus* species. West-facing slopes (192 -272° aspect) are also preferred *Pinus* species establishment sites as well as well-drained areas, i.e. those within 1.5 and 19km from perennial rivers and high stream/river density (0.48 to 0.91km/km). Mean temperatures of the wettest quarter below 21.5°C provided optimum conditions. Similarly, high cattle density areas (>300), areas with 4 to 5 other invasive species as well as areas with moderate to high tree species rich areas (78 to 239) have high posterior probabilities of *Pinus* species occurrence.

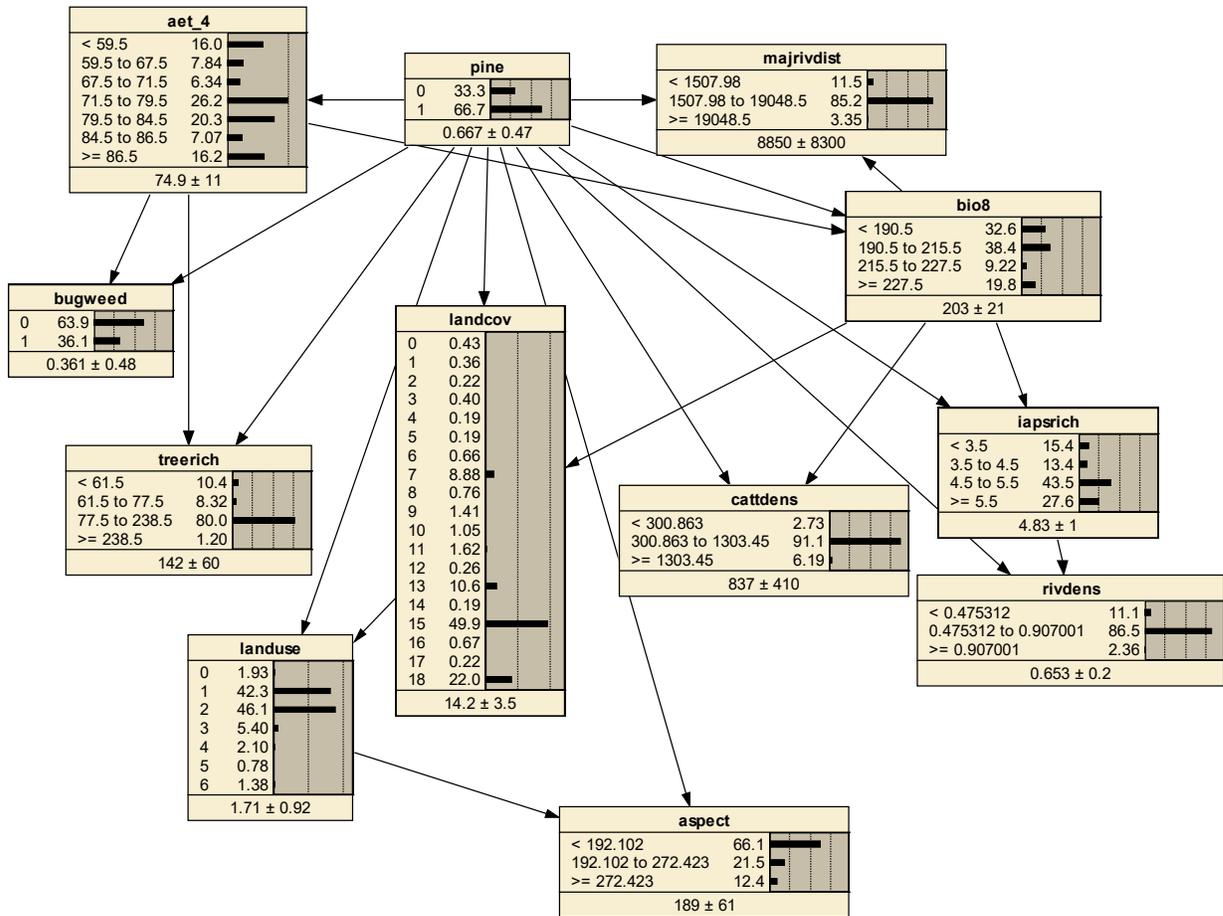


Figure 4.36: A learned Bayesian network for *Pinus* species distribution.

April actual evapotranspiration was the foremost determinant of *Pinus* species distribution in Swaziland (Table 4.10). This was closely followed by mean temperature of the wettest quarter, land cover and land use. Slope aspect and river/stream density were the least influential of this set. Hence, human land utilization (commensalism with humans) and resource availability are the key drivers of *Pinus* species invasion, constrained by temperatures in the wettest season.

Table 4.10: Mutual information for selected *Pinus* species predictor variables.

Variable	Mutual information
aet_4	0.44513
bio8	0.41134
landcov	0.39384
landuse	0.29131
iapsrich	0.15401
bugweed	0.13522
treerich	0.10021
cattdens	0.06375
majrivdist	0.04749
rivdens	0.03958
aspect	0.01655

The influence of these factors within the BN in Figure 4.36 results in the prediction maps shown in Figure 4.37. The influence of land use and bioclimatic variables is evidenced by the high posterior probabilities in the cooler western part of the country, especially within plantation forestry areas. The ensemble of all the algorithms exhibits similar patterns (Figure 4.38). However, prediction certainty is low in areas outside the plantations where conditions are still suitable but the species was rarely observed (Figure 4.38).

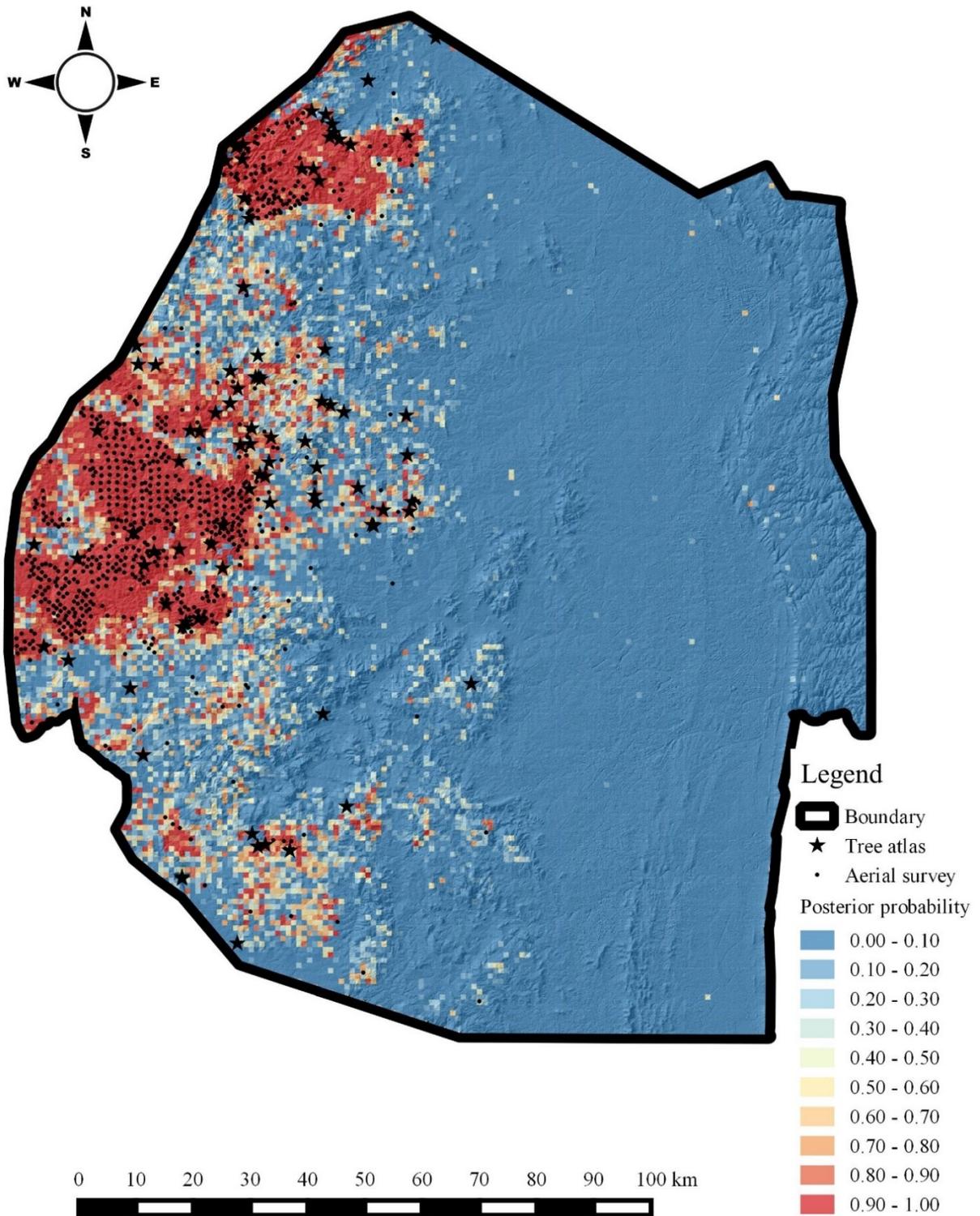


Figure 4.37: Posterior probability of occurrence for *Pinus* species in Swaziland (derived from the BN in Figure 4.36).

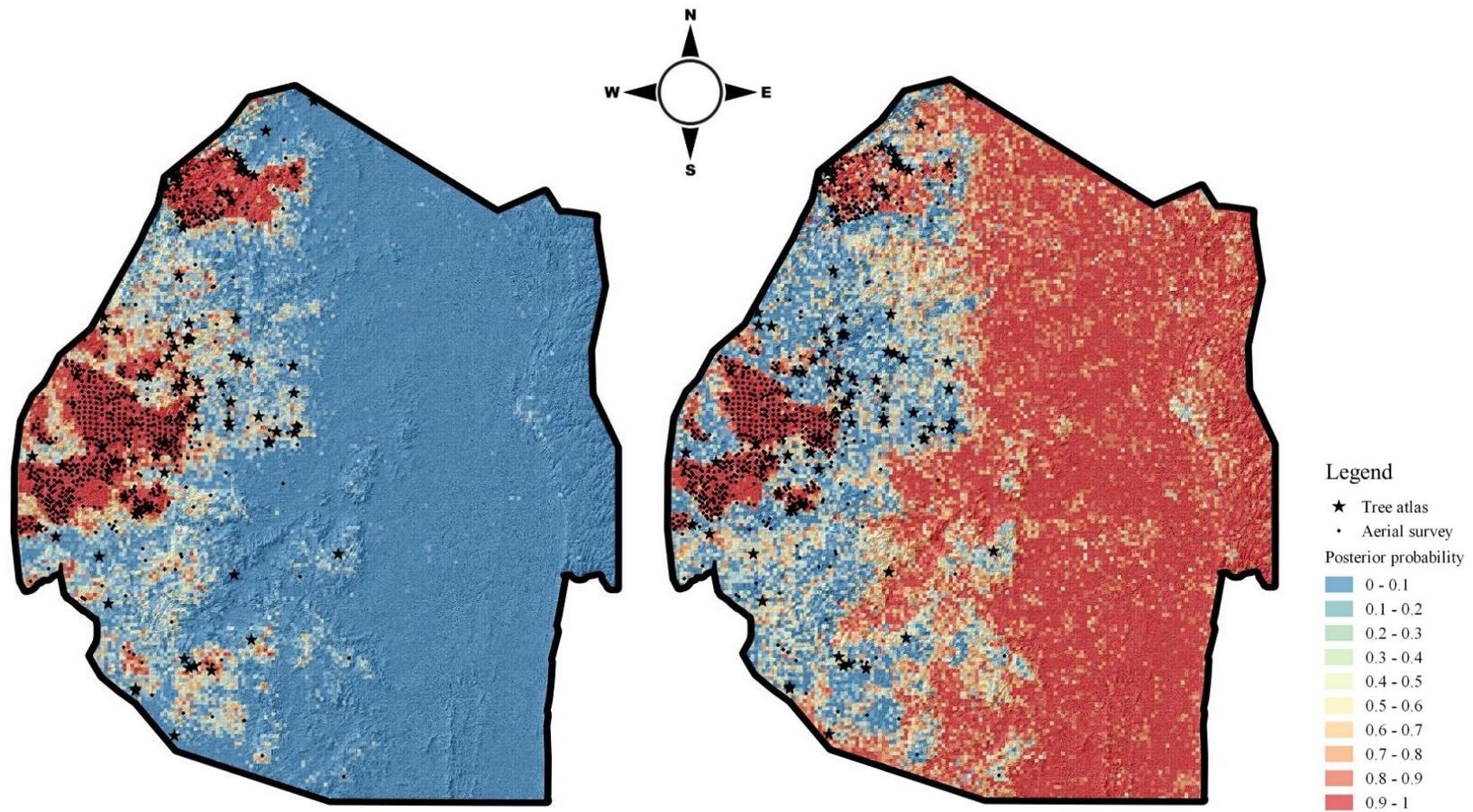


Figure 4.38: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *Pinus* species in Swaziland.

4.3.11 *Populus x canescens*

The distribution of *P. x canescens* seemed more complex and its pattern could be elucidated with 11 variables learned through the locally scored TAN algorithm (Figure 4.39). The selected variables were proximity to rivers, proximity to tourism sites, land cover fragmentation, road density, human settlement density, March potential evapotranspiration, number of frost days, cation exchange capacity at 15-30cm depths, soil bulk density, sand and coarse fragments fraction at 30-60cm depths, silt fraction at 60-100cm depth, and the occurrence of *S. punicea*.

The species invades areas with more than six frost days per year in areas within 1km from a river or stream. Similarly, areas of moderate March (winter) potential evapotranspiration (117.5 to 130.5mm) within highly fragmented land surface cover (Shannon index > 0.77) are suitable for this species. The BN model reveals *P. x canescens*' preference for soils with moderate bulk density (1.25 to 1.35kg/m³) at 30 to 60cm depths, silt content higher than 12.5g/kg at 60 to 100cm depths and a high proportion of coarse fragments (6.5 cm³/cm³). Human influence is evidenced by the high occurrence probabilities in areas with relatively high settlement densities (>4 homesteads/km²), high road density (> 1.35km/km²) and within 34km from tourism attractions.

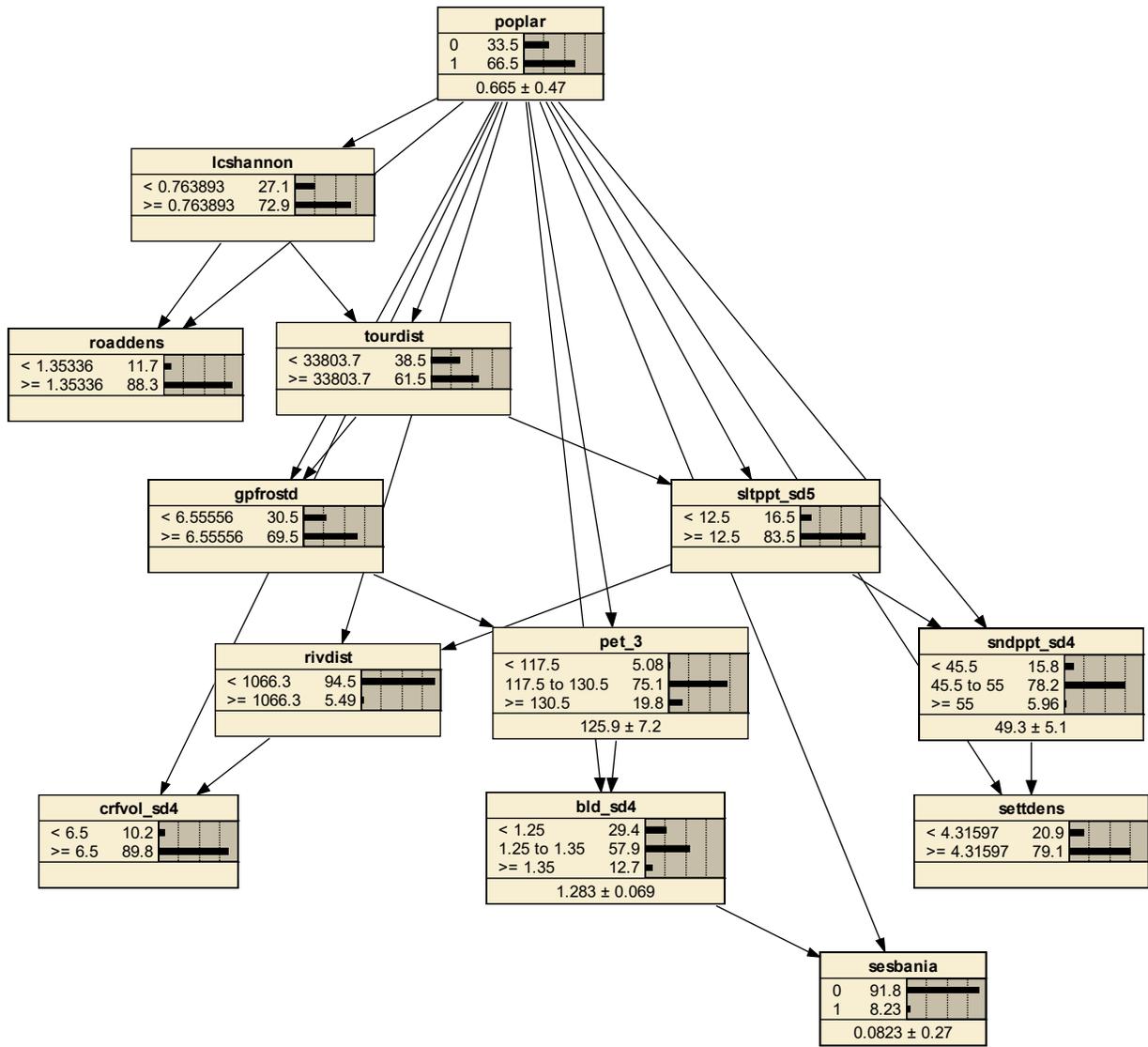


Figure 4.39: A learned Bayesian network for *Populus x canescens* distribution.

The number of frost days, proximity to tourism attractions and March potential evapotranspiration were strong predictors of *P. x canescens* occurrence (Table 4.11). Conversely, the presence of *S. punicea* and proximity to rivers were the weakest predictors. Hence, the *P. x canescens* invasion process is primarily driven by human activity and is restricted by bioclimatic conditions and resource availability.

Table 4.11: Mutual information for selected *Populus x canescens* predictor variables.

Node	Mutual information
gpfrostd	0.4958
tourdist	0.39074
pet_3	0.37327
sndppt_sd4	0.29368
lcs Shannon	0.19016
settdens	0.15396
sltppt_sd5	0.15036
roaddens	0.13525
bld_sd4	0.12885
crfvol_sd4	0.10667
rivdist	0.03513
sesbania	0.005

The interaction of these variables shown in Table 4.11 and Figure 4.39 reveals the predicted spatial distribution in Figure 4.40. The influence of frost occurrence and proximity to tourism facilities is apparent in the maps. The high posterior probabilities in the southwestern part of the country is a result of an interplay of these factors (see Figure 4.41). Whilst most of the country was predicted with high certainty (Figure 4.41), uninvaded areas located closer to currently invaded areas have low certainty predictions.

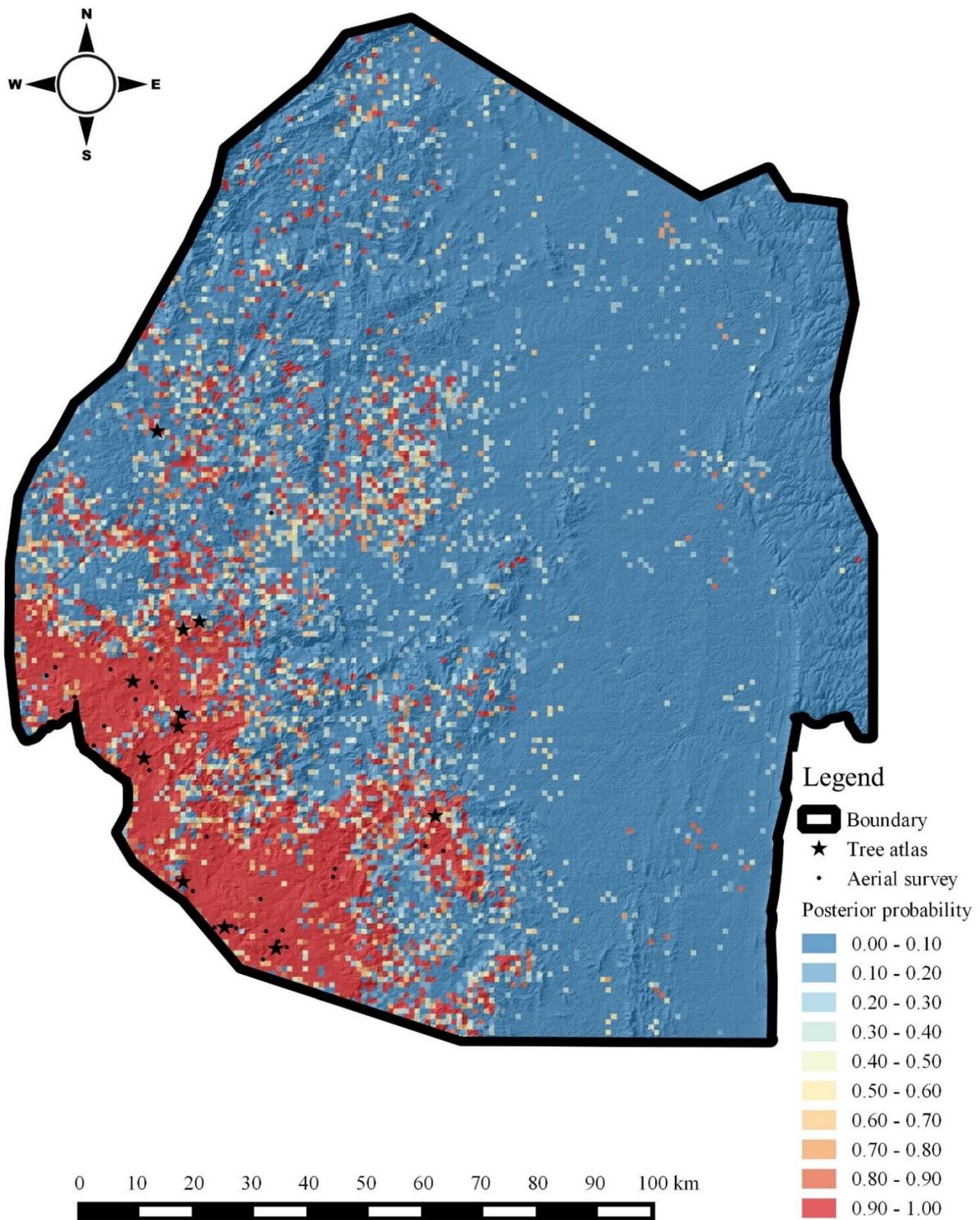


Figure 4.40: Posterior probability of occurrence for *P. x canescens* in Swaziland (derived from the BN in Figure 4.39).

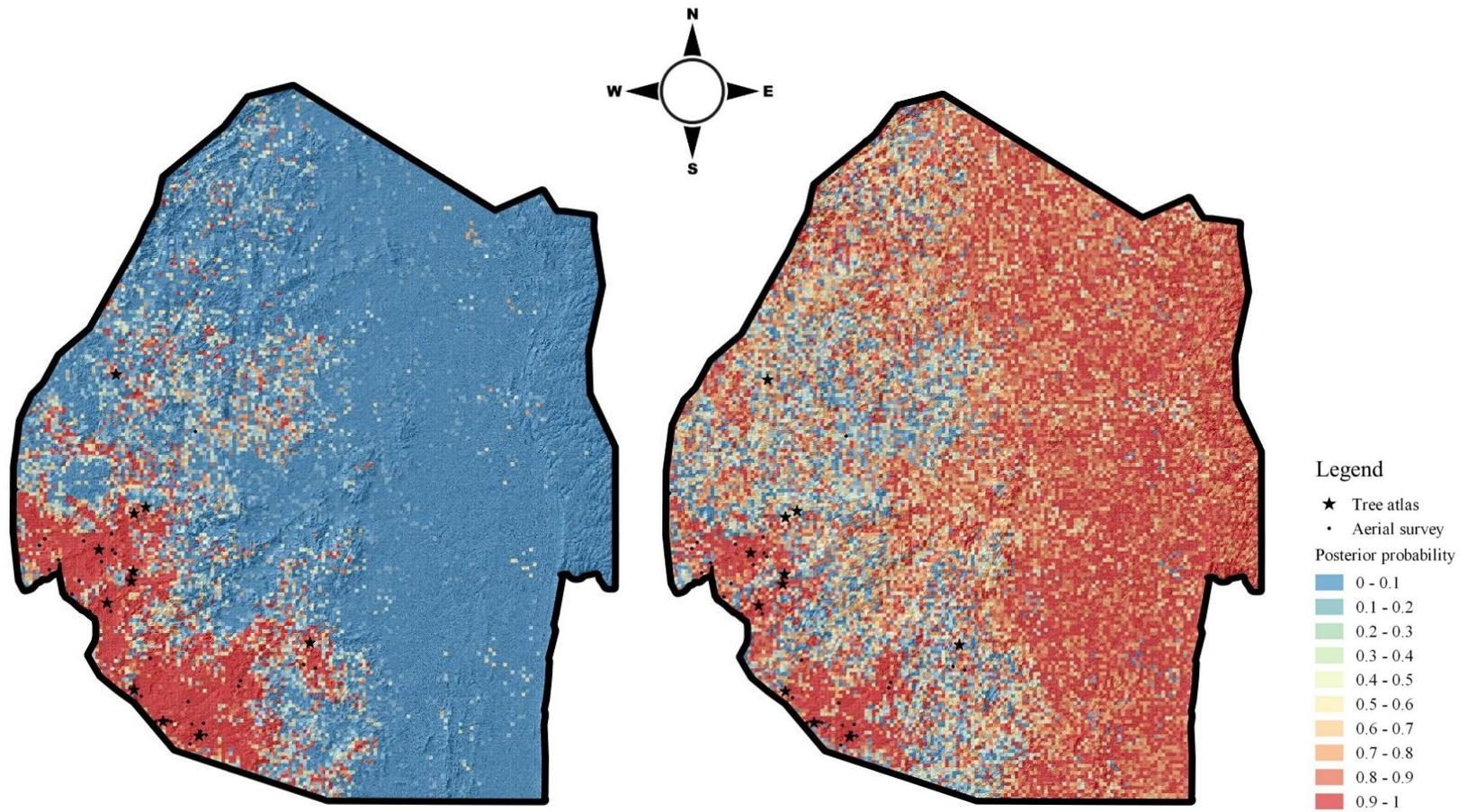


Figure 4.41: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *P. x canescens* in Swaziland.

4.3.12 *Psidium guajava*

The spatial distribution of *P. guajava* is determined by August (late dry season) actual evapotranspiration, minimum temperature of the coldest month, land cover fragmentation (Shannon index), proximity to main electricity lines, and the occurrences of *Opuntia* species, *S. didymobotrya* and *J. mimosifolia* all of which had direct arcs (Figure 4.42). The ICS algorithm outperformed all other algorithms highlighting the interactions amongst the variables.

Areas within 1.5km from electric power lines and those characterized by high fragmentation of the land surface cover (Shannon index > 0.34) have high *P. guajava* occurrence probabilities. Furthermore, minimum temperatures averaging more than 8.4°C and moderate August actual evapotranspiration (> 29.9 to 53.5mm) provide suitable habitat.

The sensitivity analysis points to the distribution of *P. guajava* in Swaziland being largely associated with *J. mimosifolia* and *S. didymobotrya* and regulated by the minimum temperature of the coldest month (Table 4.12). Proximity to electrical power supply lines and land cover fragmentation have relatively less influence. Nevertheless, electric infrastructure, human land utilization and resource availability are the key drivers of *P. guajava* invasion and this is constrained by low temperatures.

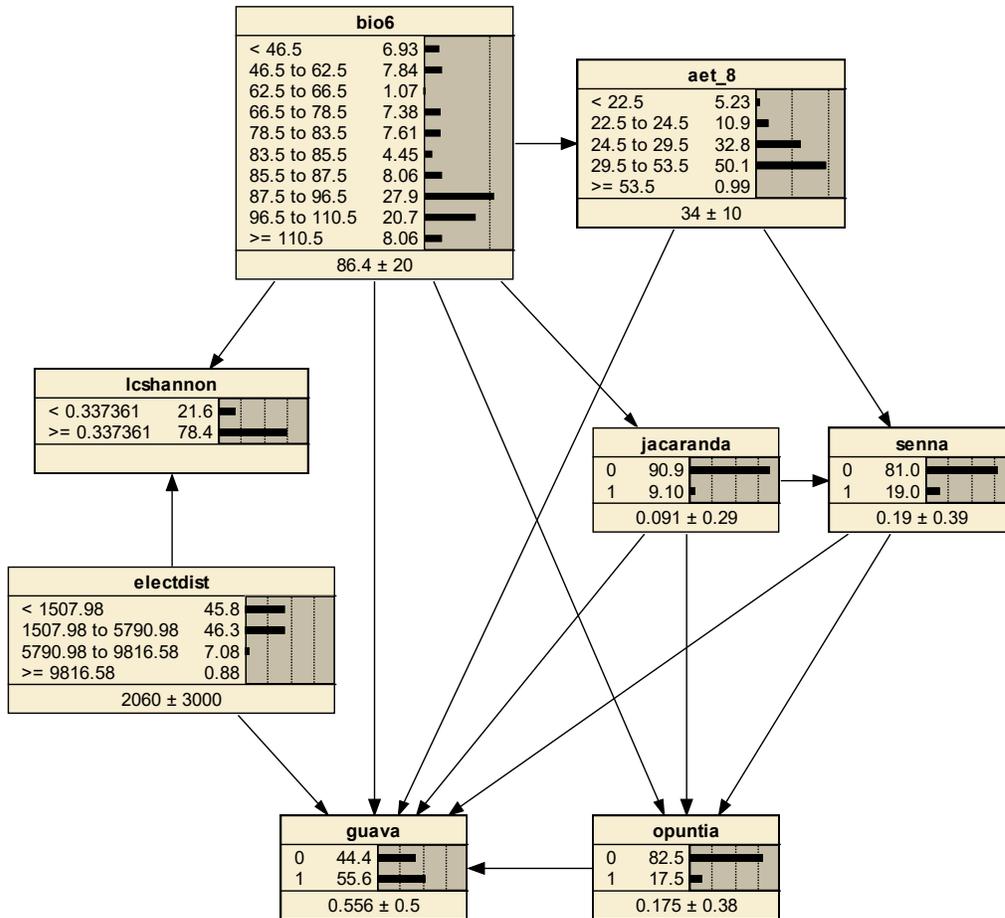


Figure 4.42: A learned Bayesian network for *Psidium guajava* distribution.

Table 4.12: Mutual information for selected *Psidium guajava* predictor variables.

Variable	Mutual information
jacaranda	0.12029
senna	0.10771
bio6	0.1024
opuntia	0.06851
aet_8	0.06275
electdist	0.03307
lshannon	0.01905

The predicted spatial distribution of *P. guajava* as determined by the BN in Figure 4.42 is shown in Figure 4.43. The influence of the minimum temperature of the coldest month is mainly largely responsible for the observed spatial pattern as is the influence of the co-occurrences with *J. mimosifolia* and *S. didymobotrya*. The effect of these factors results in the concentration of high probability areas in the central part of the country, areas that are moderately cold, highly populated and human disturbed. Figure 4.44 shows the predictions from the ensemble of all the algorithms and the accompanying PPCI values. It is evident that high prediction uncertainties exist in areas where *P. guajava* is likely to occur but currently not observed.

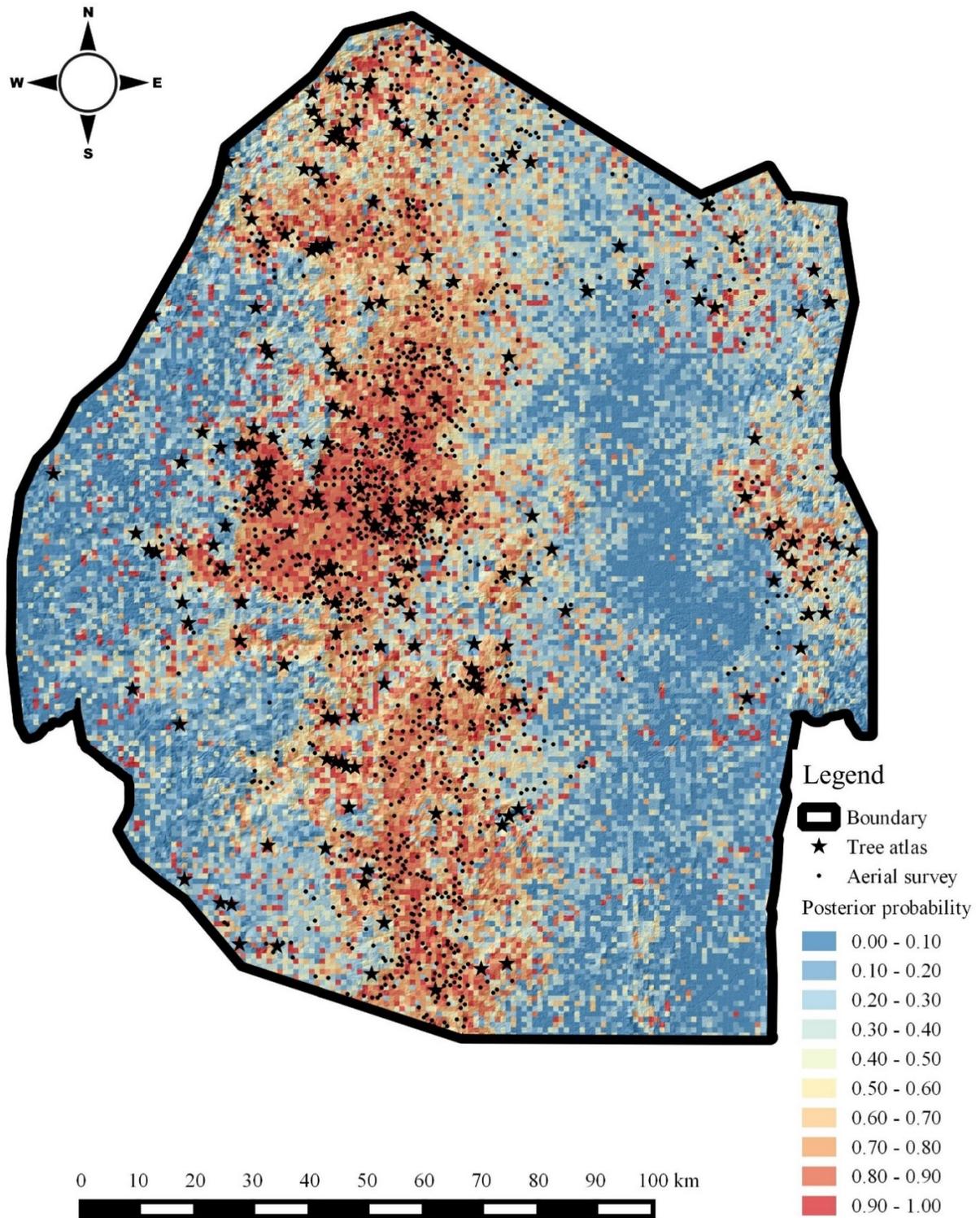


Figure 4.43: Posterior probability of occurrence for *P. guajava* in Swaziland (derived from the BN in Figure 4.42).

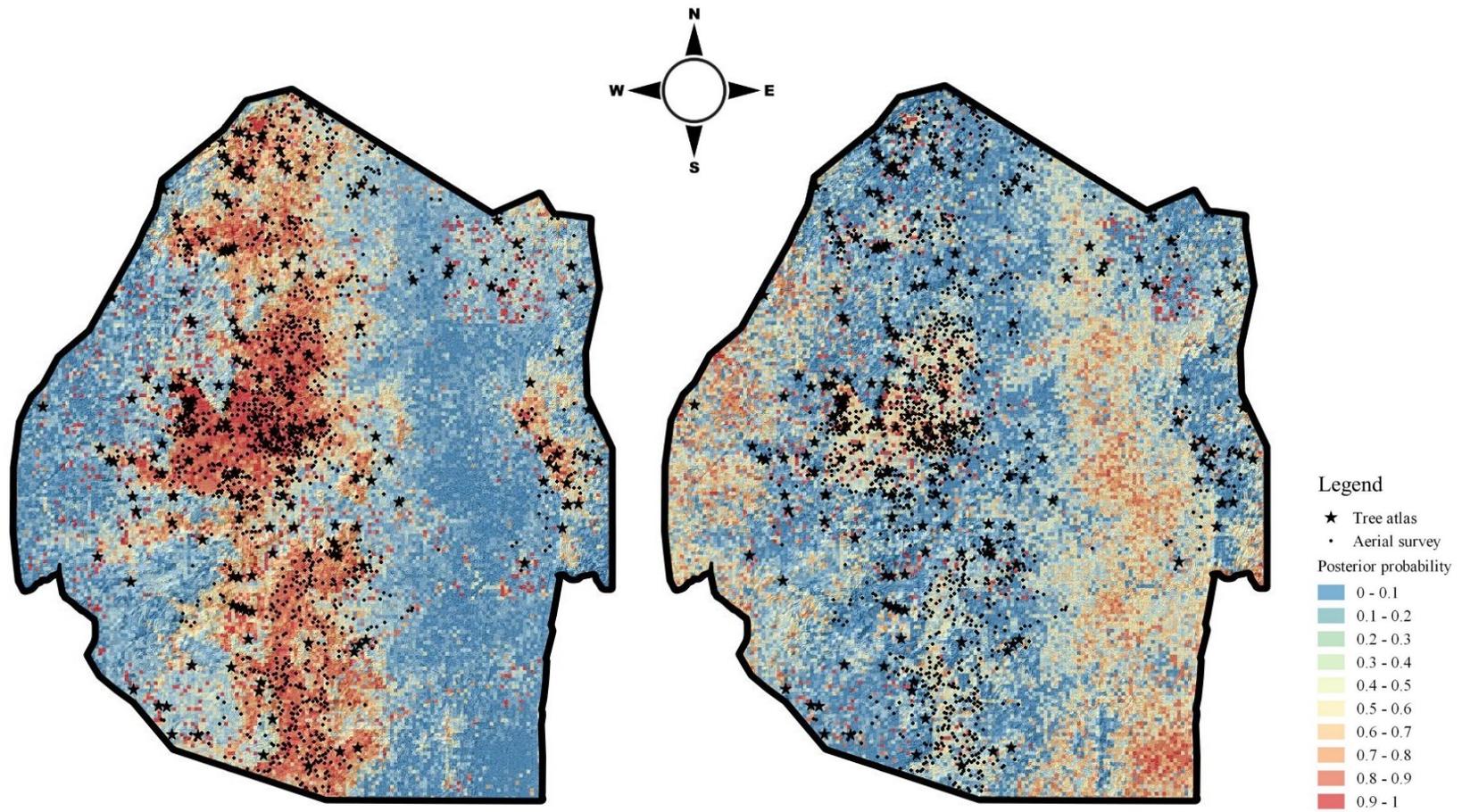


Figure 4.44: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *P. guajava* in Swaziland.

4.3.13 *Rubus* species

The BN in Figure 4.45 indicates that the key determinants of *Rubus* species distribution in Swaziland are precipitation seasonality, minimum temperature of the coldest month, proximity to tourism routes, proximity to major roads, stream/river density, tree species richness, and the presence of *S. mauritianum*, *Eucalyptus* species, *Populus* species, and *A. mearnsii*. The ICS algorithm, which attempted to create a causal structure of arcs linked to *Rubus* species occurrence, out-performed all other algorithms.

Rubus species distribution is limited to areas with minimum temperatures of the coldest month lower than 7.5°C and precipitation seasonality between 63.5 to 69%. High stream/river density (>0.49km/km²) and moderate tree species richness (102 to 253) provide a suitable niche for this species. Furthermore, the species is most likely to occur in areas that are within 9km from major roads as well as those within 25km from major tourism routes.

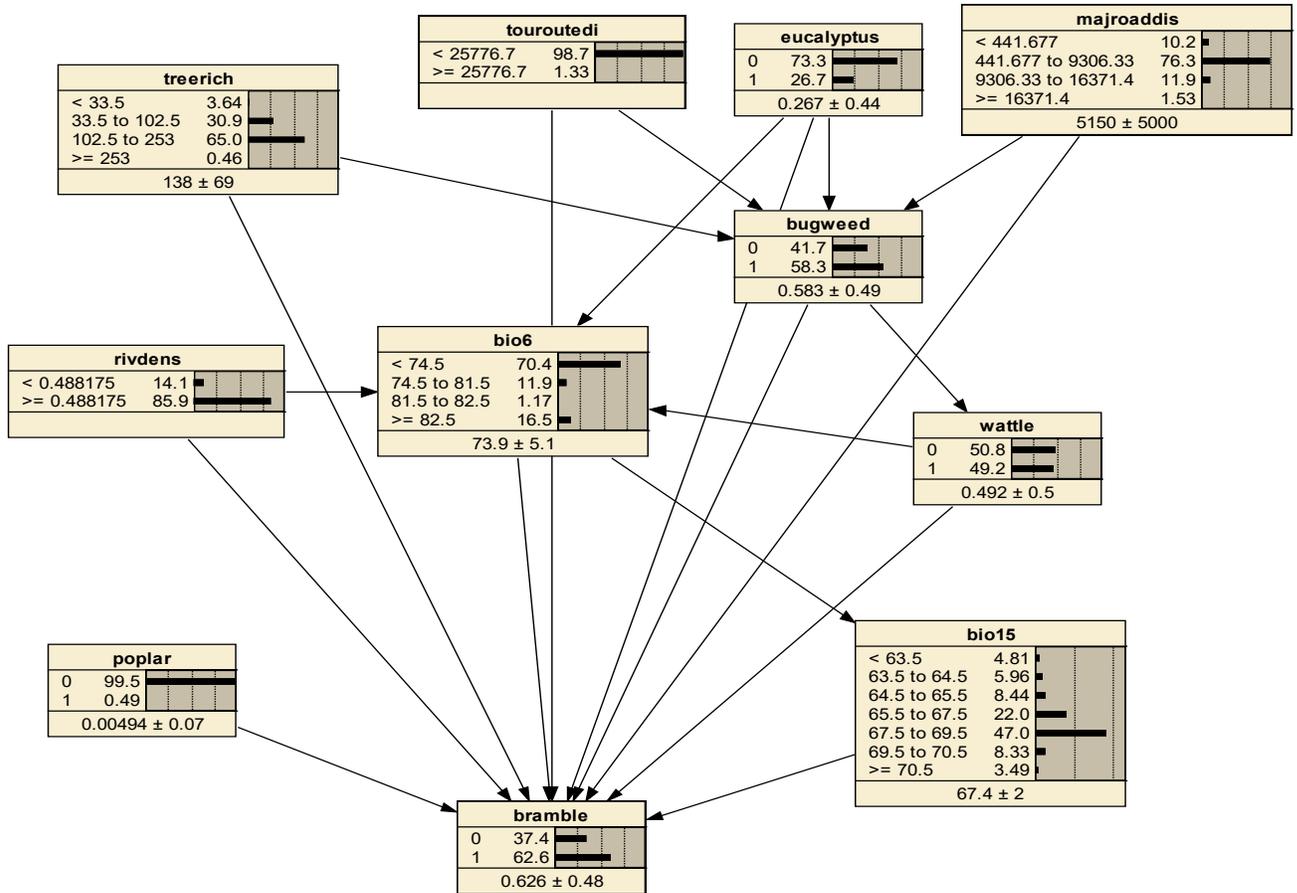


Figure 4.45: A learned Bayesian network for *Rubus* species distribution.

The sensitivity analysis (Table 4.13) indicates a strong association of *Rubus* species with *S. mauritianum* and *A. mearnsii*. The influence of the minimum temperature of the coldest month and the occurrence of *A. mearnsii* is likewise notable. Tree species richness and the occurrence of *P. x canescens* had the least influence. Hence, the invasion of this species can be described as being facilitated by human activity and regulated by climate and biotic interactions.

Table 4.13: Mutual information for selected *Rubus* species predictor variables.

Variable	Mutual information
bugweed	0.14621
wattle	0.13808
bio6	0.10289
bio15	0.0536
eucalyptus	0.02322
treerich	0.01233
majroaddis	0.00704
rivdens	0.00405
touroutedi	0.00083
poplar	0.0002

Figure 4.46 is the prediction map showing the posterior probabilities of *Rubus* species occurrence conditioned on the key determining factors as shown in Figure 4.45. Of note is the restriction of the species to the high rainfall areas and cooler western half of the country with isolated incursions towards the eastern part of the country. The PPCI values in Figure 4.47 indicate generally high prediction certainty especially in areas where there was a good correlation between probabilities and observed species occurrence.

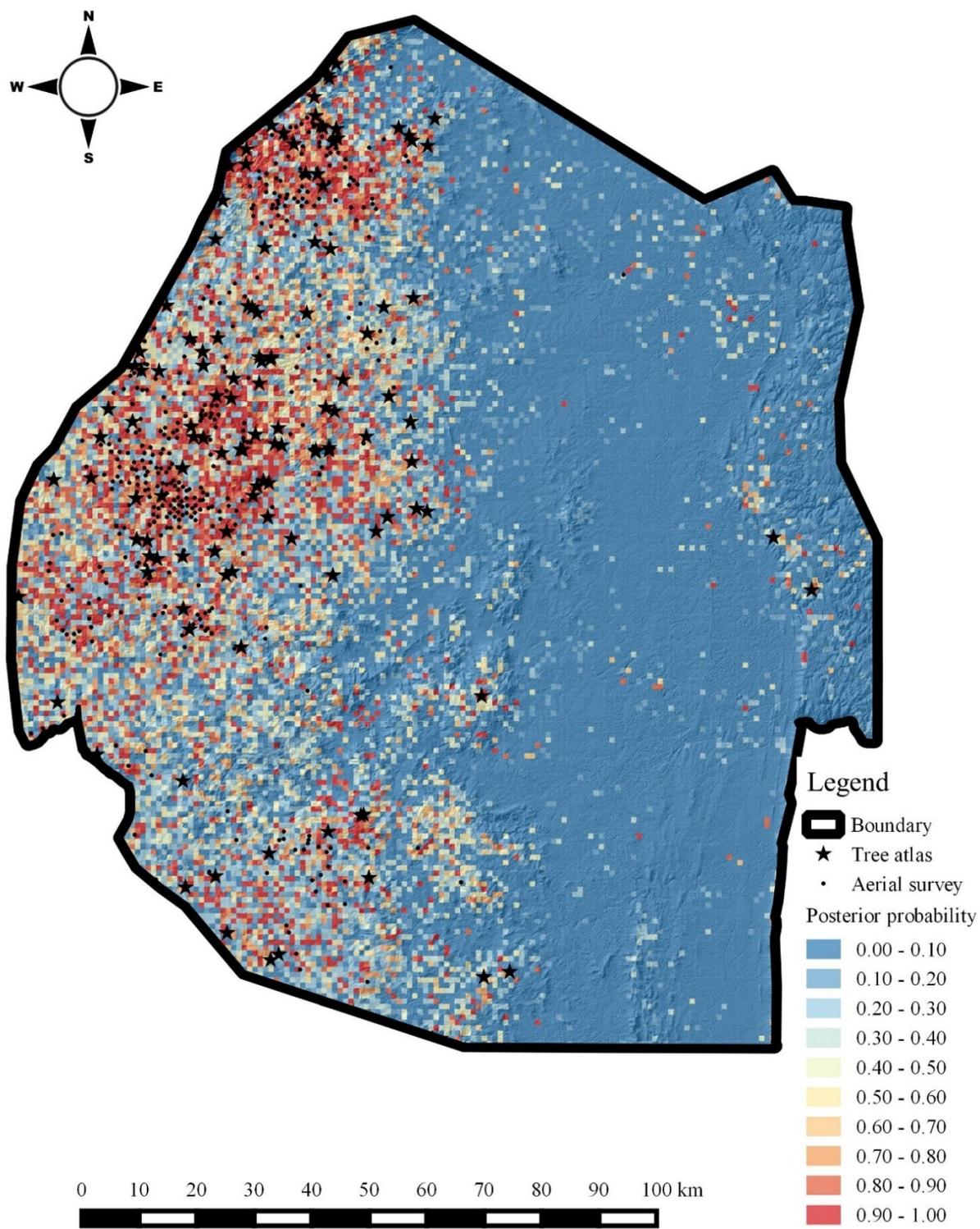


Figure 4.46: Posterior probability of occurrence for *Rubus* species in Swaziland (derived from the BN in Figure 4.45).

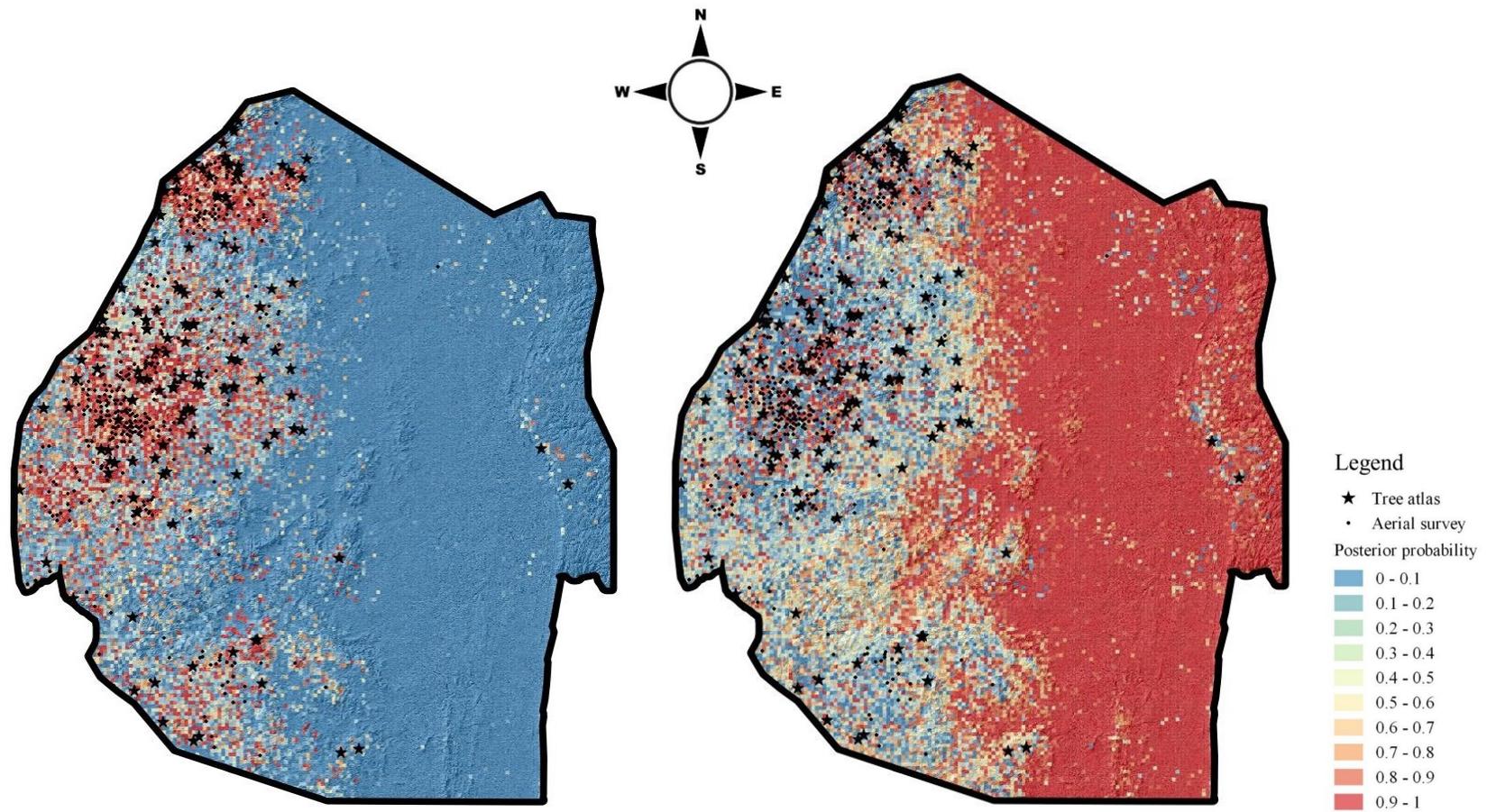


Figure 4.47: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *Rubus* species in Swaziland.

4.3.14 *Senna didymobotrya*

The main determinants of *S. didymobotrya* distribution in Swaziland were found to be the number of frost days, invasive plant species richness, land cover fragmentation, proximity to water sources, clay fraction at 15-30cm depth, and the presence of *S. punicea*, *M. azedarach*, and *P. guajava*. The best performing K2 algorithm learned with global scoring resulted in a BAN structure where all the predictor variables had direct arcs to the target variable in addition to interactions amongst themselves (Figure 4.48).

Areas within less than 2km from surface water sources as well as soils with clay content less than 41.5g/kg at 5-15cm depths were found to be highly suitable habitat. It also seems that land cover fragmentation (Shannon index > 0.34) promotes invasion and that *S. didymobotrya* tolerates exposure to frost for up to 25 frost days per year. The co-occurrence with other invasive species is affirmed by the high occurrence probabilities in areas with more than five other such species in particular *S. punicea*, *M. azedarach*, and *P. guajava*.

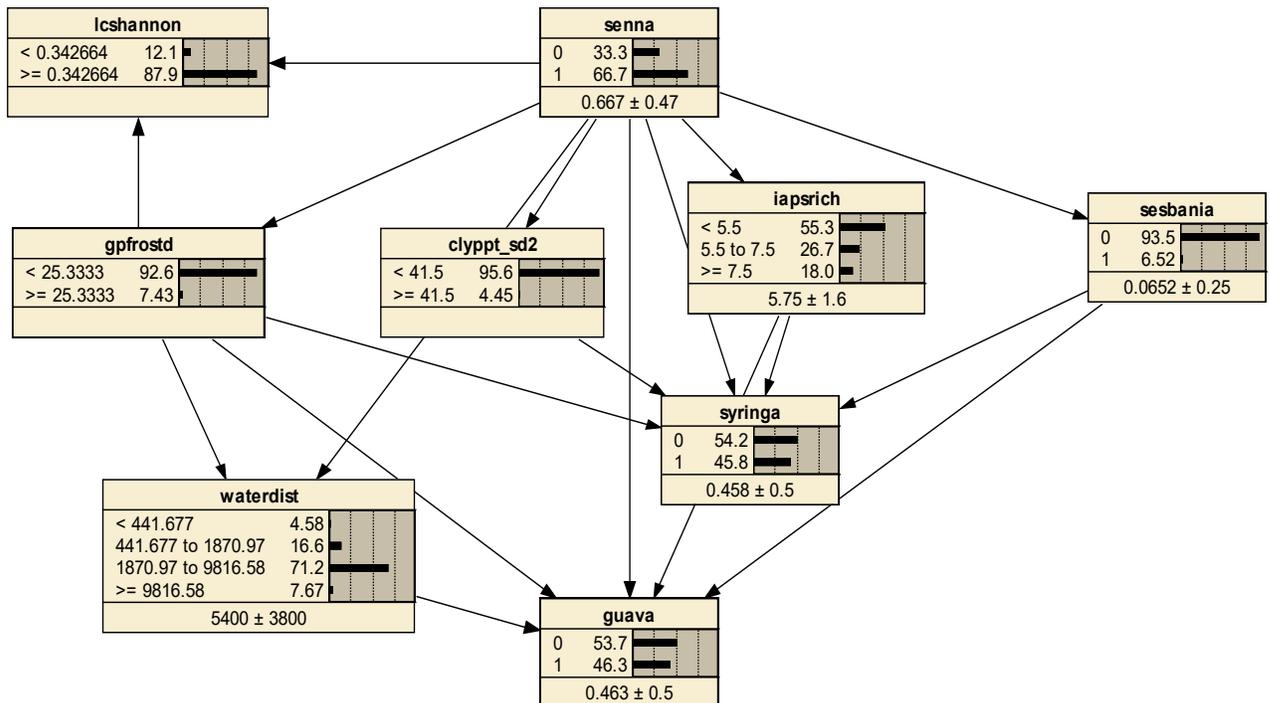


Figure 4.48: A learned Bayesian network for *Senna didymobotrya* distribution.

S. didymobotrya is strongly associated with *M. azedarach* distribution followed by *P. guajava* (Table 4.14). Clay fraction at 15-30cm depths had a relatively lesser influence. Therefore, *S. didymobotrya* invasion in the country is driven by human land use, and is regulated by frost occurrence, resource availability and biotic interactions.

Table 4.14: Mutual information for selected *Senna didymobotrya* predictor variables.

Node	Mutual information
syringa	0.27507
guava	0.22840
gpfrostd	0.04263
iapsrich	0.03821
waterdist	0.03032
lcshannon	0.02632
sesbania	0.02352
clyppt_sd2	0.01328

The co-occurrences with *M. azedarach* and *P. guajava* is evidenced by the high posterior probabilities in the central part of the country (Figure 4.49 and Figure 4.50). However, the most apparent pattern can be attributed to the influence of water sources, including rivers. The low occurrence probabilities to the west are a result of higher frost occurrence frequency. Neighbouring the high posterior probability areas are those areas predicted with low certainty due to the absence of *S. didymobotrya* in those suitable areas (Figure 4.50).

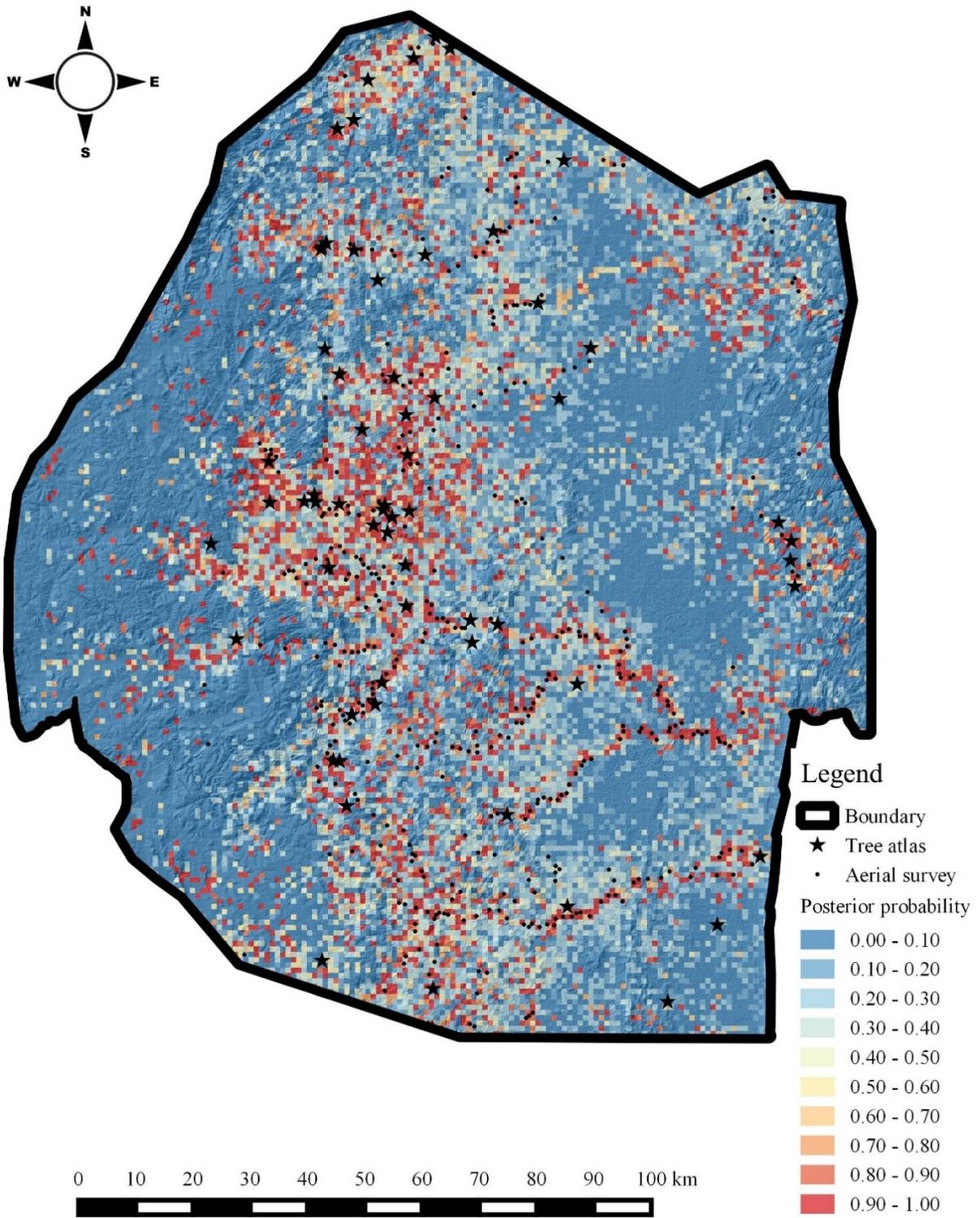


Figure 4.49: Posterior probability of occurrence for *S. didymobotrya* in Swaziland (derived from the BN in Figure 4.48).

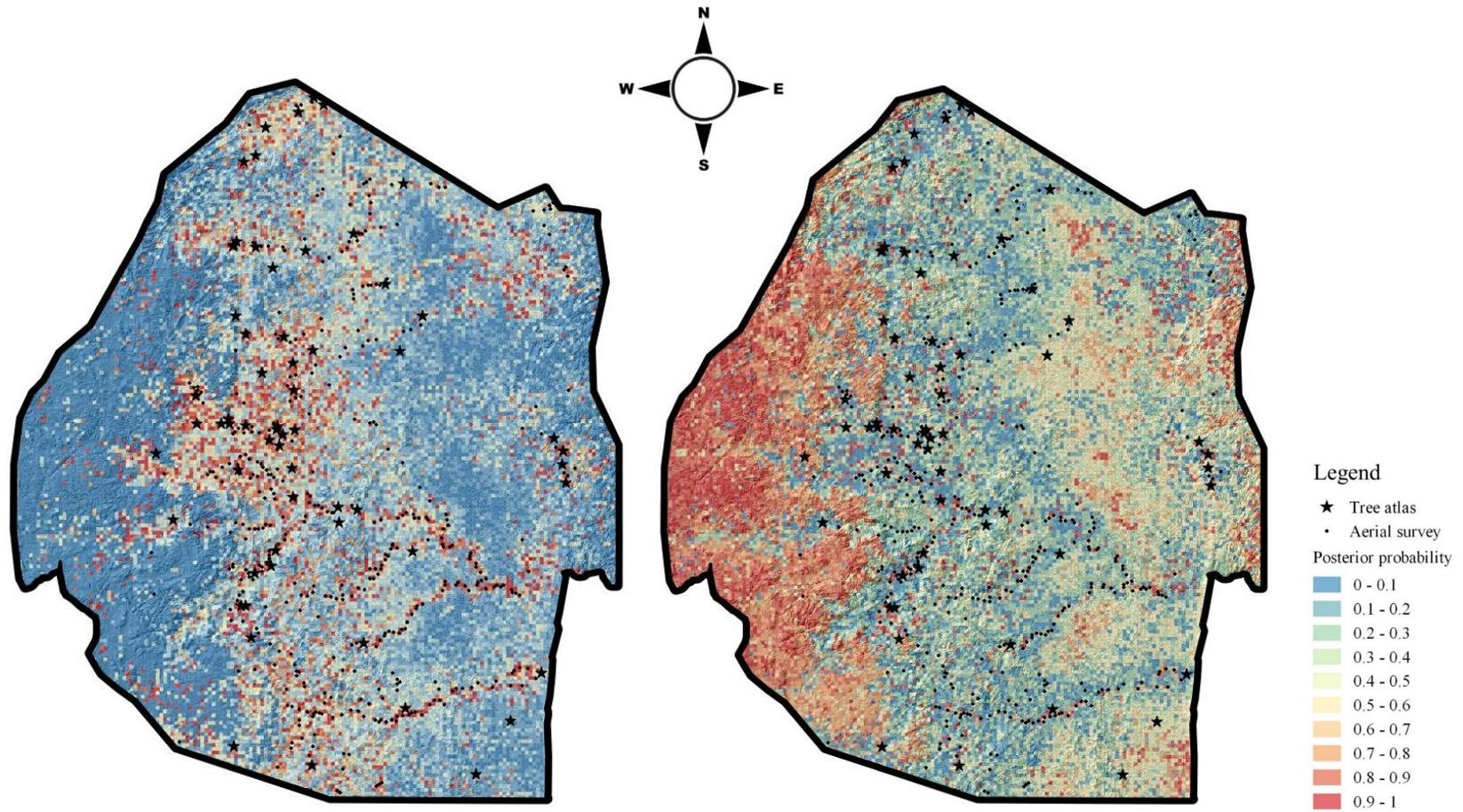


Figure 4.50: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *S. didymobotrya* in Swaziland.

4.3.15 *Sesbania punicea*

The distribution of *S. punicea* in the country was found to be strongly regulated by proximity to water sources, proximity to major rivers, land cover fragmentation, number of frost days, invasive alien plant species richness, soil clay fraction at 15-30cm depth, and the occurrences of *M. azedarach*, *P. guajava* and *S. didymobotrya*. The hill climbing algorithm learned with local scoring achieved the best prediction performance resulting in a GBN structure. This BN indicates that most of the variables are conditionally independent and inter-linked in their influence to *S. punicea* occurrence (Figure 4.51).

Areas within 1km from major (perennial) rivers and water sources (<3km) including those with low topographic index (<-0.64) were found to be suitable habitat for *S. punicea*. This species is found near human settlements especially where settlement densities exceed 13 homesteads/km². Moreover, areas with low fire frequencies (< one fire in five years) and high bulk density soils (>1.25 kg/m³) are also preferred by this plant. The species co-occurs and is associated with at least five other invasive plants including *M. azedarach*, *P. guajava* and *S. didymobotrya*.

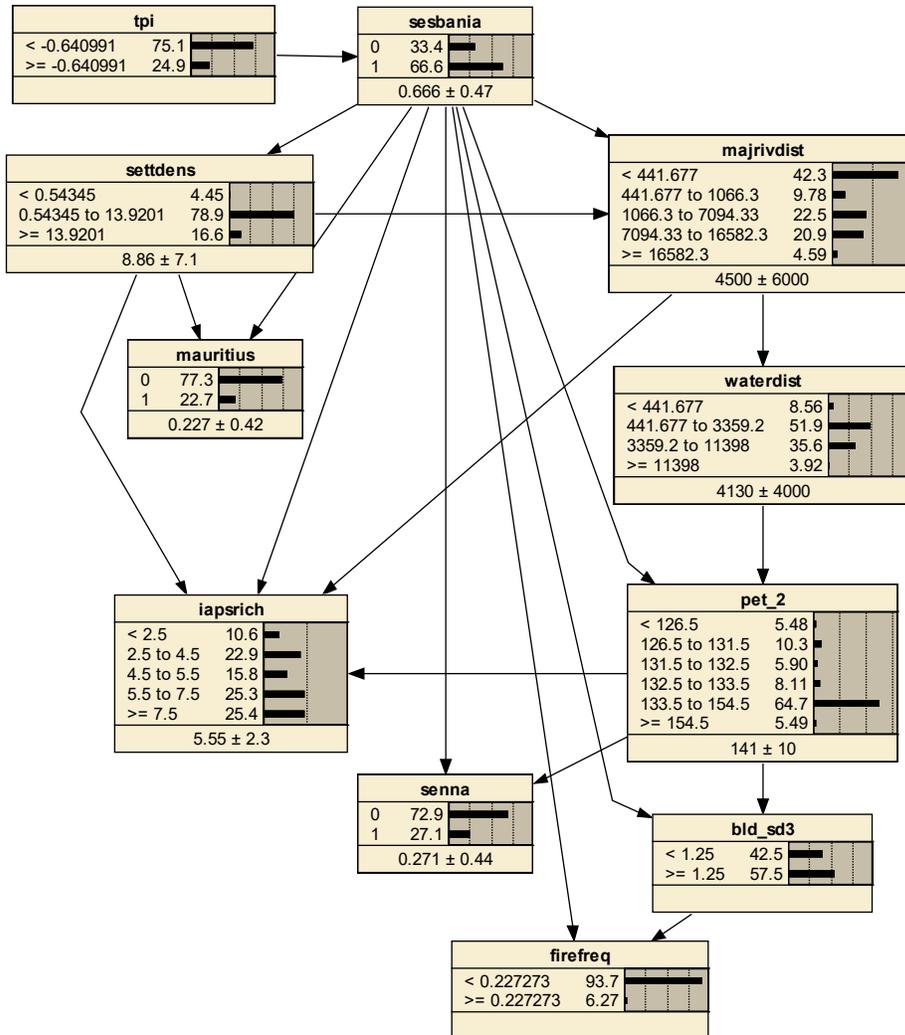


Figure 4.51: A learned Bayesian network for *Sesbania punicea* distribution.

The mutual information values indicate that *S. punicea* is strongly determined by proximity to major rivers and topographic position (Table 4.15). Fire frequency was comparatively of lower influence. Resource availability, enhanced by water transportation and human activity, is the primary determinant of *S. punicea* invasion. Biotic interactions too seem to play a role in this species' invasion process.

Table 4.15: Mutual information for selected *Sesbania punicea* predictor variables.

Variable	Mutual information
majrivdist	0.30586
tpi	0.13464
iapsrich	0.0926
pet_2	0.08981
bld_sd3	0.07736
waterdist	0.07213
settdens	0.0708
senna	0.07611
mauritus	0.07875
firefreq	0.02479

The dominant influence of major rivers is evidenced by the pattern in Figure 4.52 wherein areas close to watercourses and within river valleys have high posterior probabilities. This pattern confirms the findings in Table 4.15. Although areas where there was a good spatial correlation between species occurrence and posterior probabilities, those areas where *S. punicea* had marginal probability of occurrence had lower PPCI values (Figure 4.53).

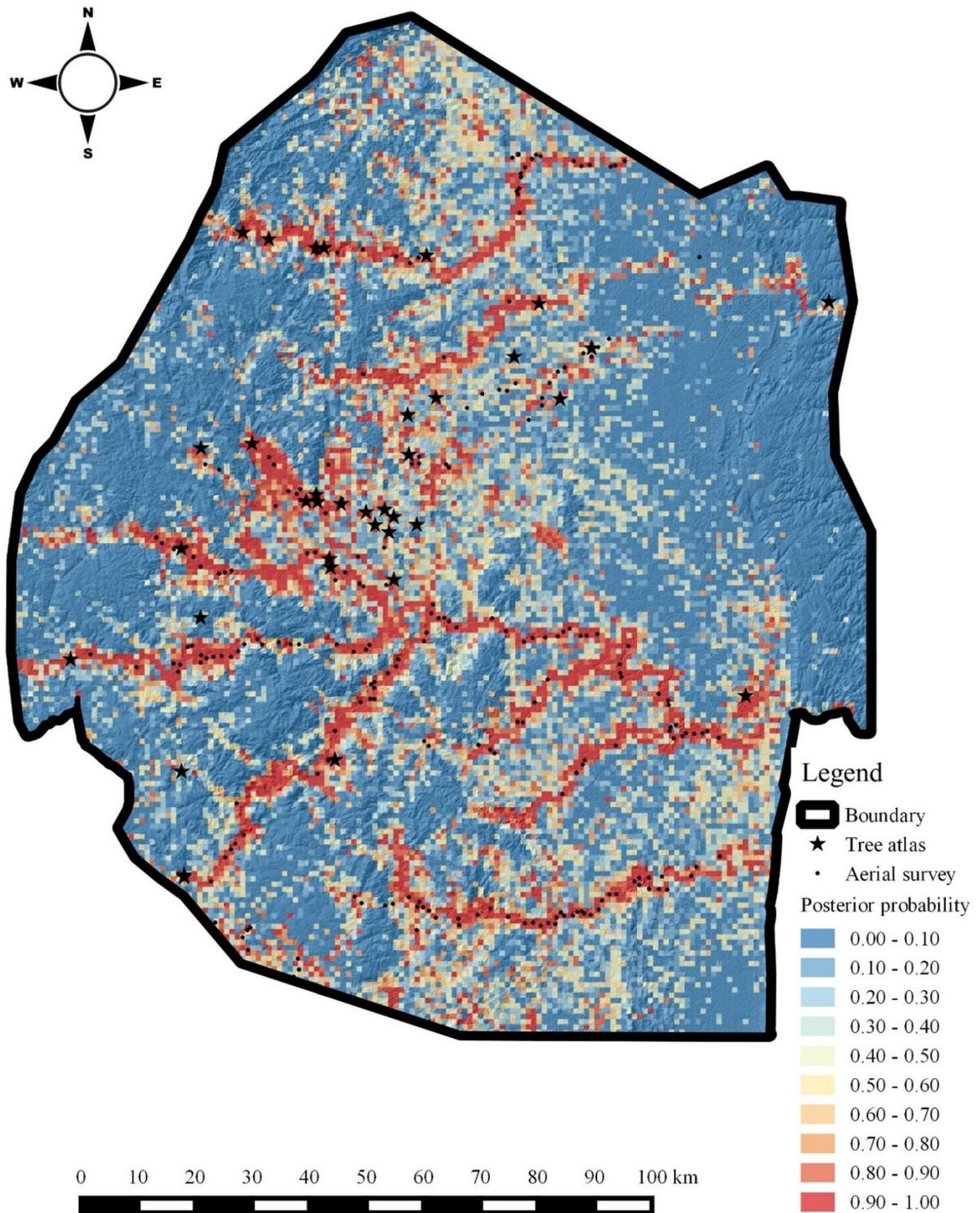


Figure 4.52: Posterior probability of occurrence for *S. punicea* in Swaziland (derived from the BN in Figure 4.51).

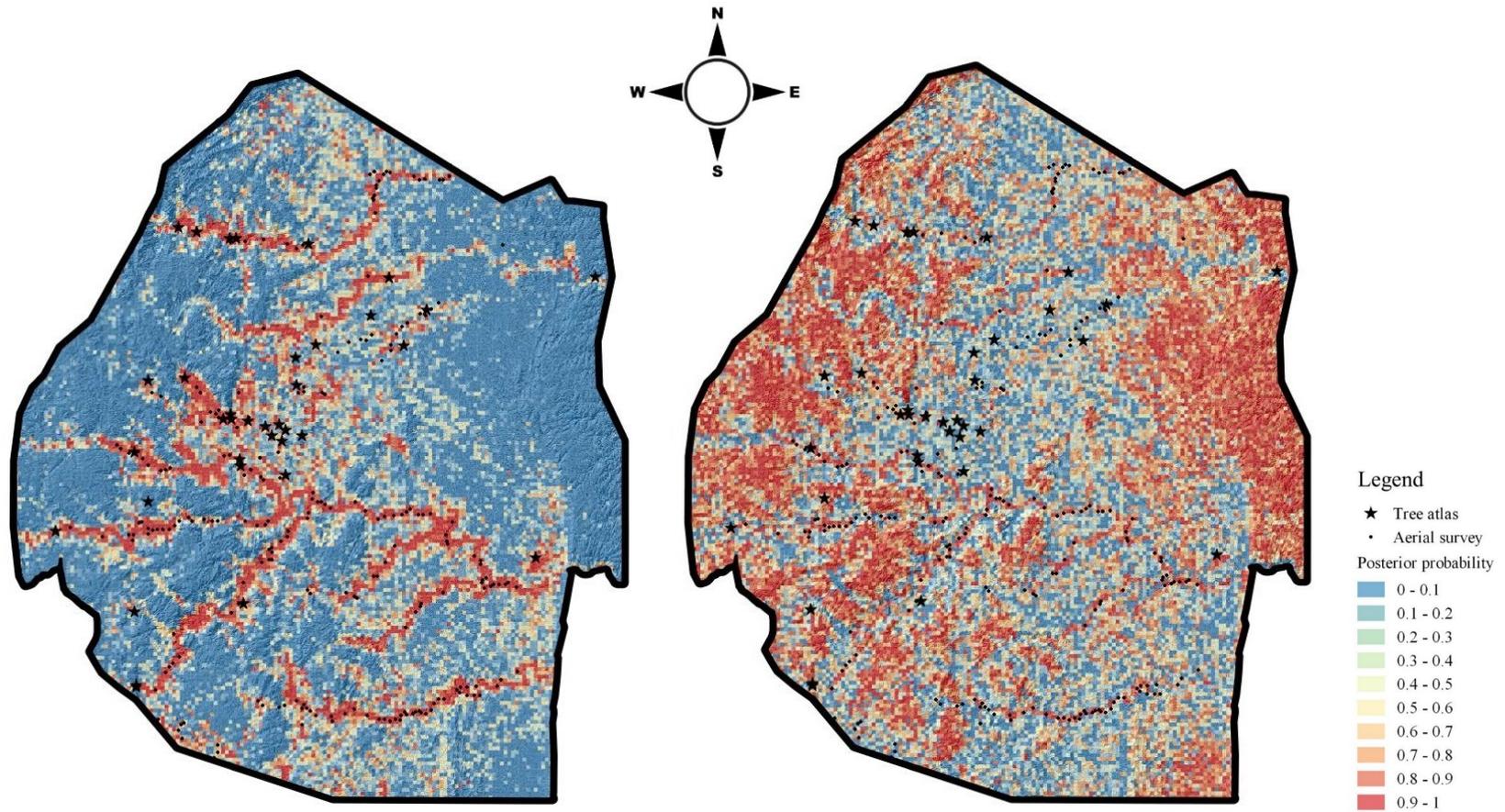


Figure 4.53: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *S. punicea* in Swaziland.

4.3.16 *Solanum mauritianum*

Land cover fragmentation, poverty rate, proximity to rivers, November actual evapotranspiration, and the presence of *S. didymobotrya*, *Rubus* species, *Eucalyptus* species, *A. mearnsii*, *Pinus* species and *C. jamararu* were found to be the primary determinants of *S. mauritianum* distribution in Swaziland. The ICS algorithm was again the relatively better performing algorithm for this species resulting in the causal BN shown in Figure 4.54.

Sites within 440m from rivers or streams and those with high land cover fragmentation (Shannon index of 1.07 to 1.61) were found to be prone to *S. mauritianum* invasion. Areas with moderate to high (>50% and < 83%) poverty levels and November AET values higher than 68.5mm were similarly found to provide suitable habitat.

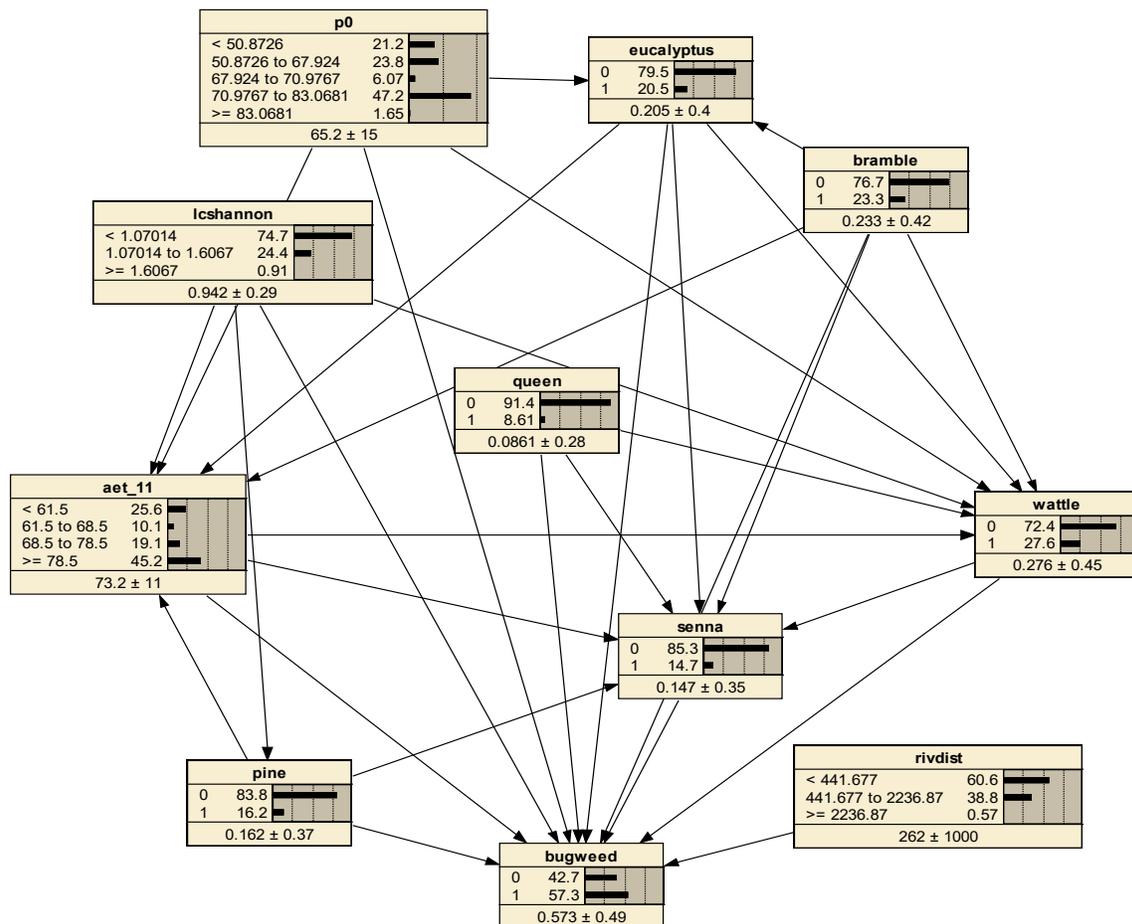


Figure 4.54: A learned Bayesian network for *Solanum mauritianum* distribution.

The November actual evapotranspiration was the strongest predictor of *S. mauritianum* followed by the occurrence of *A. mearnsii* and *Rubus* species (Table 4.16). The occurrence of *C. jamacaru* and land cover fragmentation had the lowest mutual information with *S. mauritianum* occurrence. Hence, resource availability, biotic interactions and human activity are the key drivers of *S. mauritianum* invasion in Swaziland.

Table 4.16: Mutual information for selected *Solanum mauritianum* predictor variables.

Variable	Mutual information
aet_11	0.14275
bramble	0.03595
wattle	0.03356
eucalyptus	0.02944
p0	0.02082
pine	0.02017
senna	0.00691
rivdist	0.00141
lcshannon	0.00045
queen	0.00039

The posterior probabilities in Figure 4.55 affirm the strong co-occurrence with *Rubus* species, *A. mearnsii*, and *Pinus* and *Eucalyptus* species as shown in Figure 4.54. These are the same species that are found in the cooler western part of the country as affirmed in Figure 4.56. Similar to the trend with the other species, the PPCI values were low in suitable areas bordering those where the species occurs was either not observed or least frequently detected (Figure 4.56).

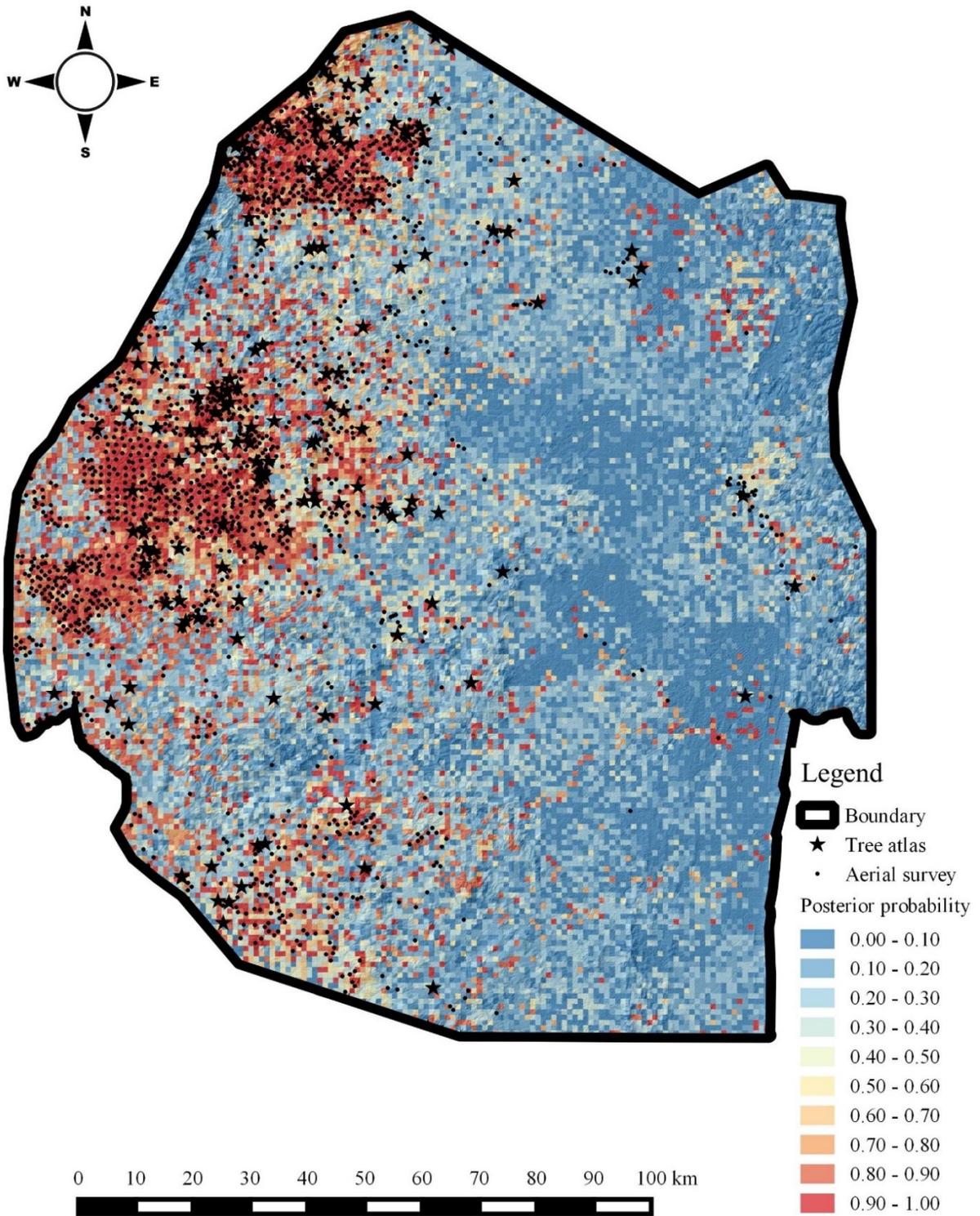


Figure 4.55: Posterior probability of occurrence for *S. mauritianum* in Swaziland (derived from the BN in Figure 4.54).

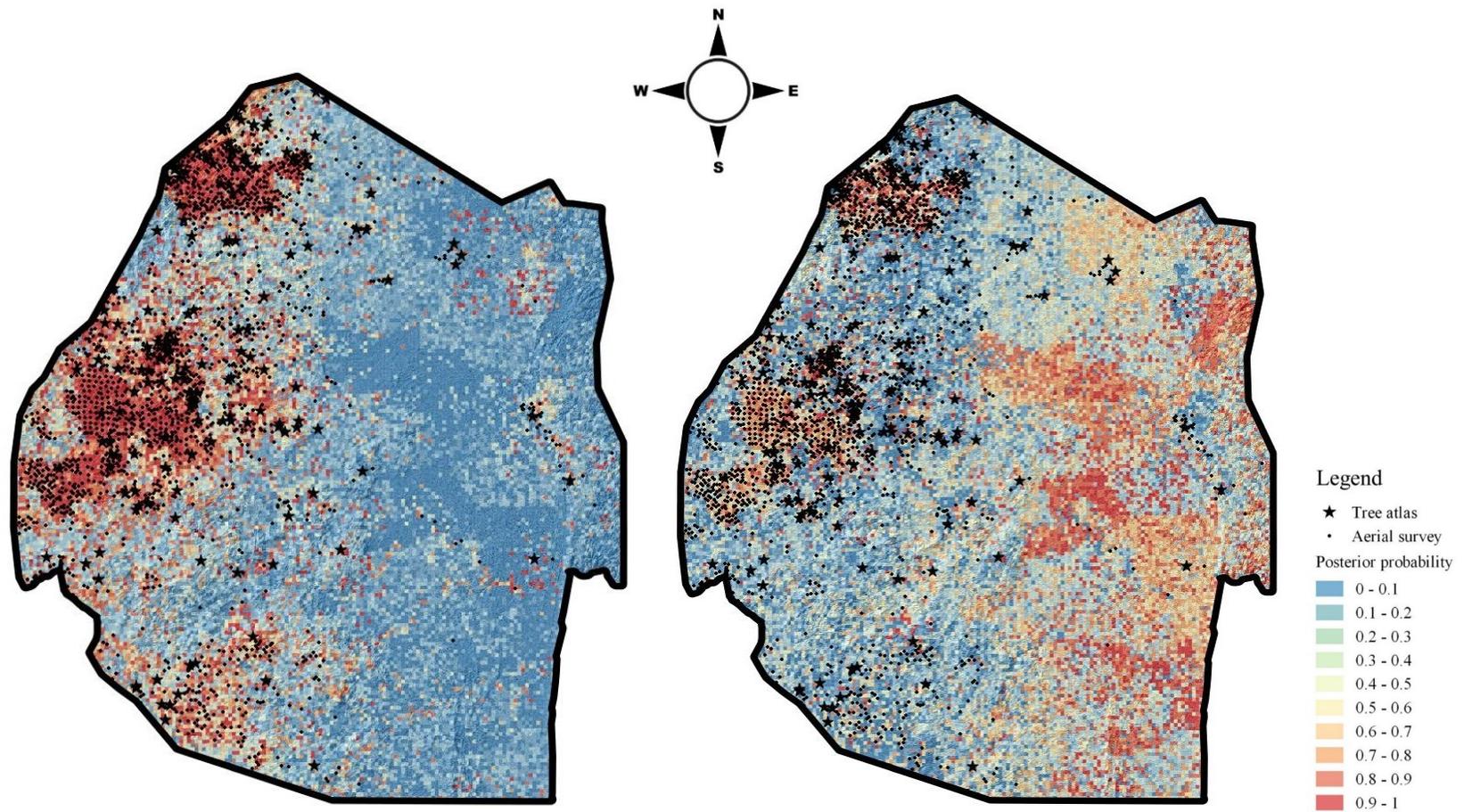


Figure 4.56: Posterior probability distribution maps derived from the ensemble of all the algorithms (left) and the PPCI (right) for *S. mauritanium* in Swaziland.

4.4 GENERAL FINDINGS ON LEARNED BAYESIAN NETWORK MODELS

4.4.1 Learned Bayesian network structures

To understand those variables that were most relevant in determining each species' distribution, the frequency at which they form part of the target species' Markov blankets are shown in Appendix 4. The variables are also classified into four categories for ease of interpretation: biotic, anthropogenic, climatic and topo-edaphic. It is evident that the number of key influential variables selected for the final models are much less than the 170 provided. In all, 68 variables were used for all the species, implying that 102 predictor variables were redundant and least relevant, i.e. they were not strong determinants of any of the species' spatial distribution. The number of variables selected per species ranged from a minimum of five for *A. mearnsii* to a maximum of 12 for *P. x canescens* with a median of eight variables.

Appendix 4 indicates that all the categories of variables were almost equally selected. *J. mimosifolia*, *S. didymobotrya*, and *S. mauritianum* were the foremost species found to be co-occurring with other species. The minimum temperature of the coldest month, precipitation seasonality and frost frequency were the most important bioclimatic variables determining the distribution of most of the plants. Land cover fragmentation, proximity to major roads and human population/settlement density were the frequently selected anthropogenic variables. Regarding topo-edaphic factors, proximity to rivers and slope aspect were the most relevant. Furthermore, Appendix 5 provides additional photographic evidence for some of the species and their occurrence with the country.

Although most of the variables were selected for at least one species' BN model, their relative influence varied with species as presented in the preceding section. In general, four types of BNs were learned: the NB, TAN, BAN, GBN and causal BN, the latter four of which represented the structures that best represented the species-environment relations considering their performance. For instance, based on the log loss, BAN structures were learned for *J. mimosifolia*, *S. didymobotrya* and *M. azedarach* whilst GBNs were the best for *C. odorata*, *Opuntia* species, *C. jamacaru* and *S. punicea*. The TAN structure was best for *L. camara*, *Pinus* species, *P. x. canescens* and *A. mearnsii* whereas the causal BN structure performed better for the rest of the species. Despite dissimilarities in the BN structures, multiple runs of the same

algorithm produced very similar BNs. This indicates that despite some of the algorithms being stuck in local maxima, the quality of the results of each algorithm did not differ significantly. This also implies that the quality of the learned BNs were least affected by random effects of the learning algorithms. The learned BNs graphically reveal the complexity of the species-environment relationships that form the basis of each BN model. The direct arcs to the class variable intuitively indicate those factors that have a large informative and/or causal effect on the distribution of each species. Hence, the BNs provide insight into the dependence relations of the variables and further reveals the number and strength of the links connecting the variables that determine each species' distribution in the country. Notably, the BN learning algorithms generated structures that better describe the domain understanding of each species ecology in addition to providing new insights on variable linkages.

What is also interesting were the final discretized states for all the models, which were parsimonious considering the accuracy of the resultant models and the interpretability of the discretization ranges. The discretization technique efficiently captured the complex between-variable gradients in relation to the target variable. The discretization provides useful information not just for explaining the environmental requirements of each species but also for describing the processes driving the invasion processes as well.

4.4.2 Species distribution maps

The collated species distribution from the aerial survey and tree atlas were produced from the collated and cleaned data (Appendix 2) resulting in a comprehensive dataset of alien invasive plant species for Swaziland. A visual comparison of the maps reveals similar geographic distributions between the tree atlas and aerial survey data for all the species albeit differences in sampling intensity. However, the aerial survey data reveal subtler gradients, which may be hidden in the coarse-resolution tree atlas dataset. There are also notable differences in the prevalence of the species considering their geographical coverage. For example, species such as *C. odorata* and *L. camara* are widely distributed whilst *Opuntia* spp. and *J. mimosifolia* are sparsely distributed. This is not only important for understanding the geographic spread of their impacts but also for appreciating the effects on model performance.

A visual analysis of the tree atlas data likely indicates the broad climate-driven distribution patterns of each species while the aerial survey data reveal the effects of more localized

landscape-scale factors such as anthropogenic activities, biotic interactions and the environmental variables specific to each species. The BN models probabilistically reproduced these spatial patterns albeit with some minor differences arising from the BN structure learning algorithms used. The spatial distribution of the probabilities reveals the geographic occurrence patterns of each species as constrained by the selected factors for each species. The posterior probability distributions for predicted species occurrence were generated via each BN model and subsequently the outputs from the best performing algorithm and the ensemble (mean) of all the algorithms for each species were mapped. Unlike the typical habitat suitability (or similar) indices from conventional classifiers and species distribution modelling approaches, the posterior probabilities of BNs learned from data express the likelihood of occurrence of each species conditioned on (uncertainties in) the selected explanatory data layers. A visual comparison of the prediction maps corroborates the high accuracy of the BN models considering the close match between the higher posterior probabilities and the field data in Appendix 2.

The key role of climate is evident in most those species for which climatic variables were selected and part of the Markov blanket. For instance, *A. mearnsii*, *Eucalyptus*, *Pinus spp.*, *Populus spp.*, *Rubus spp.* and *S. mauritianum* are more restricted to the colder high elevation areas to the west of the country characterized by a grassland ecosystem. The rest of the species are associated with the warmer eastern parts of the country. Species such as *C. odorata* and *L. camara* are widely distributed although they have limited incursions into the cold western part of the country. Such incursions occur along different variable gradients such as watercourses or drainage lines and human population and associated disturbance regimes. *P. guajava* and *C. decapetala* have similar optimal habitats in the more temperate conditions of the central part of the country albeit with varying sizes of niches largely regulated by anthropogenic factors, primarily the presence of human disturbance. Although large areas of the central part of the country offers substantial niches for *C. jamacaru*, *L. camara*, *J. mimosifolia* and *S. didymobotrya*, their distribution also extends to the warmer eastern parts of the country.

The patchy spatial distribution of species such as *P. x canescens* and *J. mimosifolia* indicates the predominant influence of both localized and anthropogenic activities, in addition to climatic conditions. The importance of watercourses is accentuated for species such as *S. didymobotrya* and *S. punicea*. The spatial patterning in the posterior probabilities reveals that all the species,

in particular *A. mearnsii*, *C. odorata* and *L. camara*, have large potential niches or proportion of the total land area where the species are likely to occur than where they currently do. Figure 4.57 shows the box plots of the posterior probability values for all the species. The differences in the box plots are indicative of the differences in the spatial coverage of probabilities within the study area. The box plots affirm that species such as *L. camara* and *C. odorata* have larger areas having relatively higher probabilities (>0.5), hence larger potential ranges. Species such as *J. mimosifolia* and *Opuntia* species have smaller potential ranges concentrated in few and localized high probability areas. The prediction maps, together with the box plots, could be simultaneously used for prioritization of control activities and management interventions.

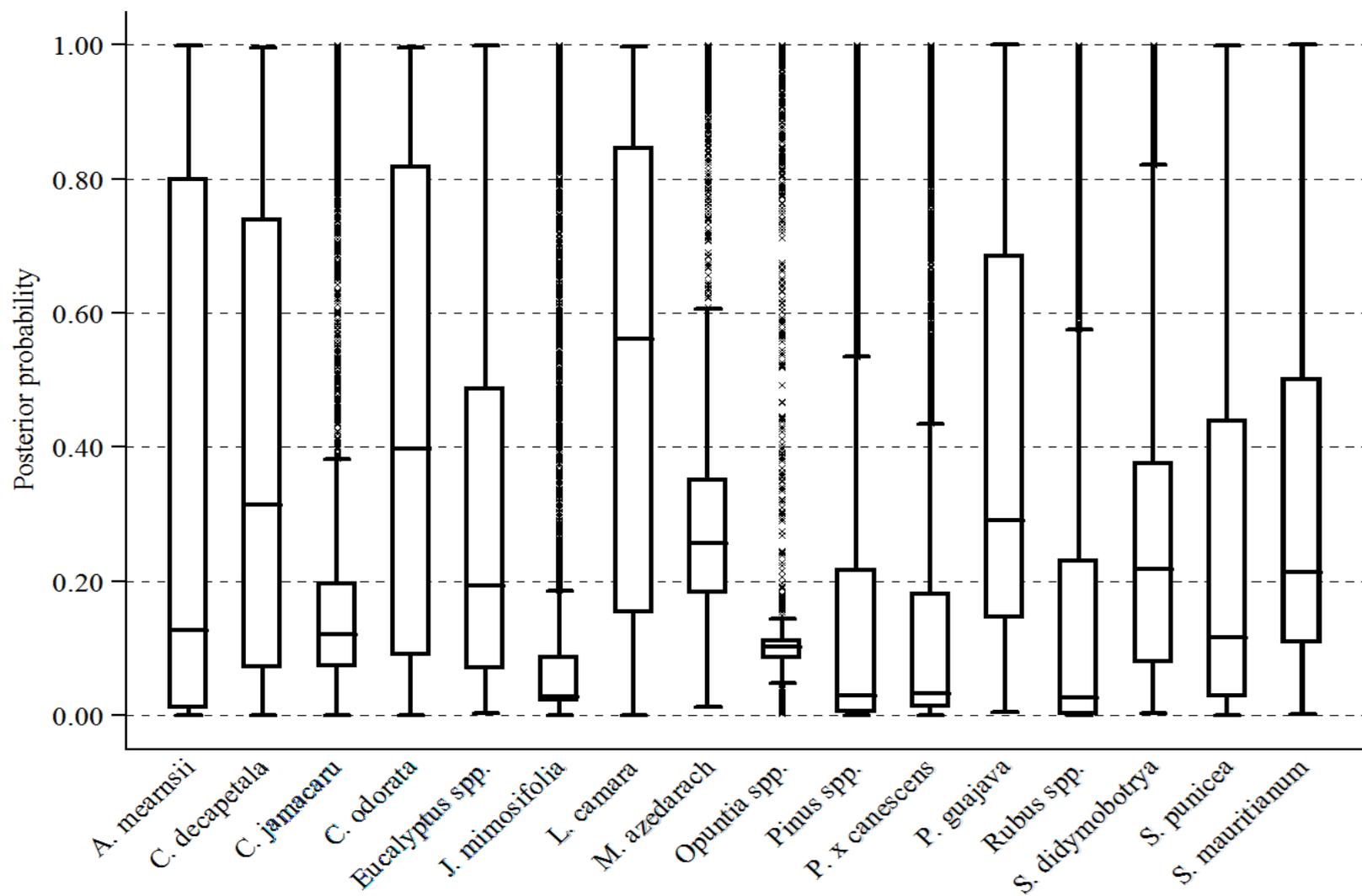


Figure 4.57: Box plots of the mean posterior probabilities from all the algorithms implemented for each species (source: own).

4.4.3 Species distribution uncertainty

The grid cell level uncertainty for two bounded continuous variables are mapped on the [0; 1] interval. Although continuous in nature, the maps essentially divide the uncertainty into three classes of model predictions:

- areas where a species is most certain to be either present or absent ($0.11 \leq \text{PPCI} \leq 0.89$),
- areas where a species is likely to be present but is not represented or there is insufficient information in the data ($0.11 > \text{PPCI} < 0.89$ excluding 0.5), and
- areas where any probabilistic statement about the species presence is uncertain ($\text{PPCI} = 0.5$).

The mapped degree of certainty reveals noticeable spatial patterning for all species. The PPCI values were high (>0.5) in areas where there was a good agreement between the actual species distribution and the model posterior probabilities. This indicates that there was a greater loading of posterior probabilities into either species presence or absence. Hence, higher PPCI values signify greater certainty in species distribution predictions. These include areas where high posterior probabilities coincide with species occurrence. Other high PPCI areas were low posterior probabilities areas coinciding with species absences from field data.

On the contrary, visualization inspection of the predictions shows low PPCI values (<0.5) or high uncertainty in areas where there were high posterior probabilities at the edges of most species' actual or observed distribution. These are areas with a suitable habitat for a species but are presently uninhabited or least habitat by that species. These marginal areas showed more inconsistencies amongst models resulting in low PPCI values. Interpretation of the SDM outputs, therefore, should simultaneously take into account the PPCI values.

Figure 4.58 is a generalized plot of the relationship between the posterior probabilities and the PPCI values for all the species. This relationship was expected considering the PPCI estimation formulae in equations 3.4 to 3.7. Since there are two states (presence/absence) for this study, a PPCI of zero is obtained for a model with a posterior probability prediction of 0.5. A mean PPCI value of 0.5 corresponds to two solutions: an upper posterior probability of approximately 0.89 and a lower value of 0.11. It follows that most of the species were predicted with high certainty, which means that the occurrence probabilities for most species were generally far away from 0.5 (the most uncertain condition).

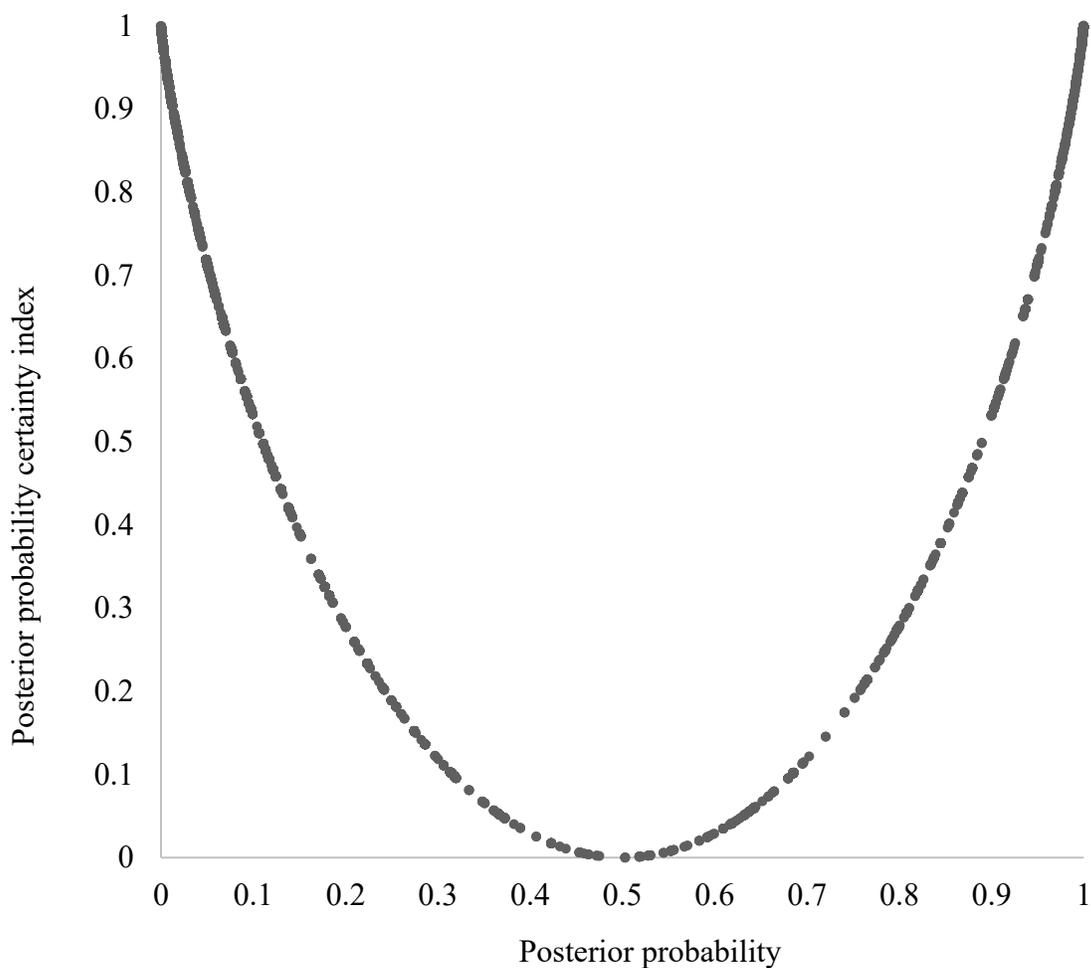


Figure 4.58: A plot of posterior probability against the posterior probability certainty index (source: own).

However, there were variations in the PPCI values amongst the species where species such as *J. mimosifolia*, *Pinus* species, *P. x canescens* and *Rubus* species were largely predicted with high certainty (Figure 4.59). On the contrary, species such as *M. azedarach* and *P. guajava* were predicted with relatively low certainty due to the nature of their spatial distribution. Such species are characterized by localized high PPCI values mainly concentrated around the grid cells where the species occur.

Notably, there was a weak negative relationship between species prevalence and mean PPCI values (Figure 4.60) implying that more prevalent species were relatively more difficult to model and predict with certainty than those that are less prevalent. On the contrary, there was no significant relationship between prediction certainty and model accuracy as measured using the logarithmic loss (Figure 4.60). This chapter presented in detail the learned BNs together with their accompanying parameters and maps. These BNs and the plotted outputs are discussed in Chapter 5 taking into account domain knowledge, the inherent uncertainties and each species's ecology.

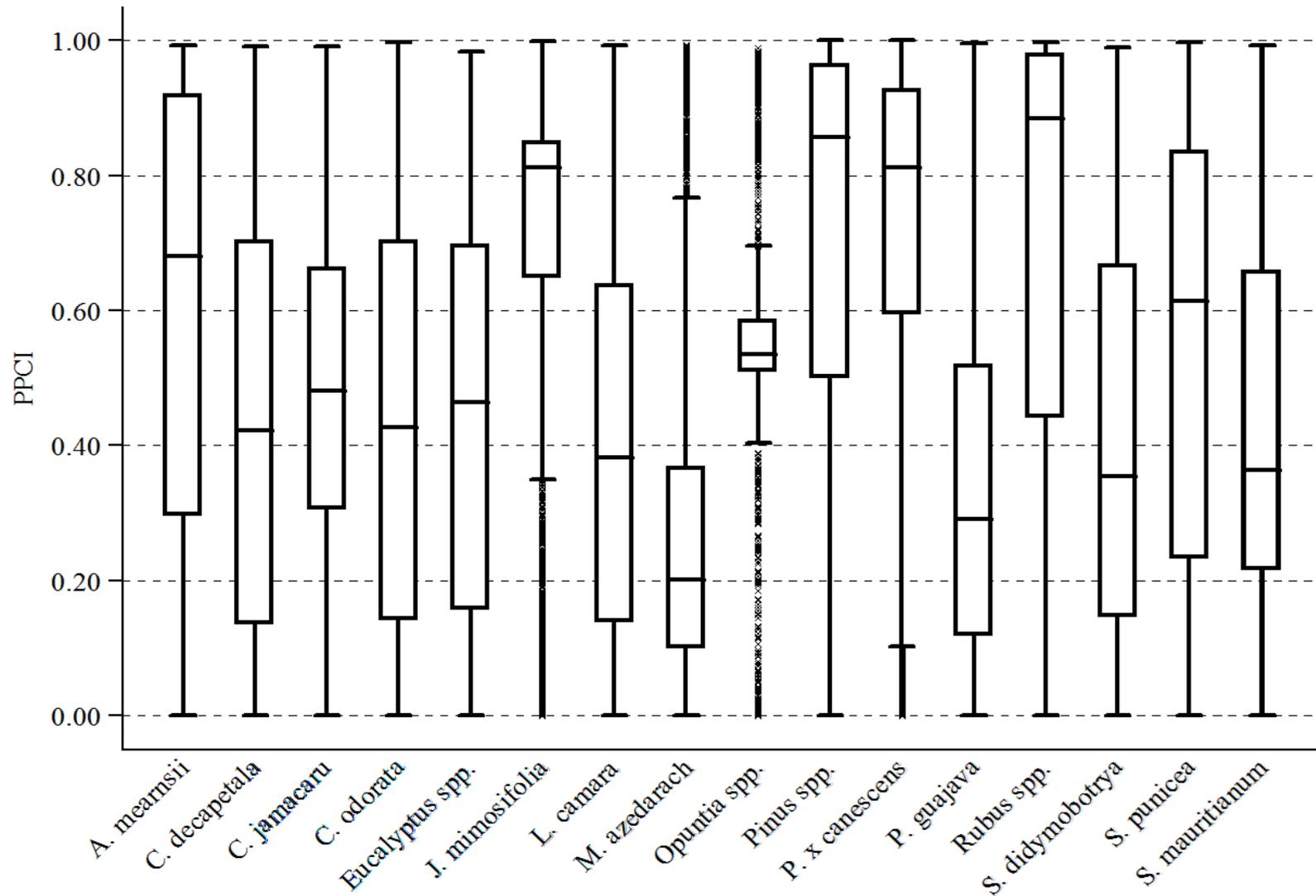


Figure 4.59: Box plots of the posterior probability certainty indices for all the species (source: own).

© University of South Africa 2016

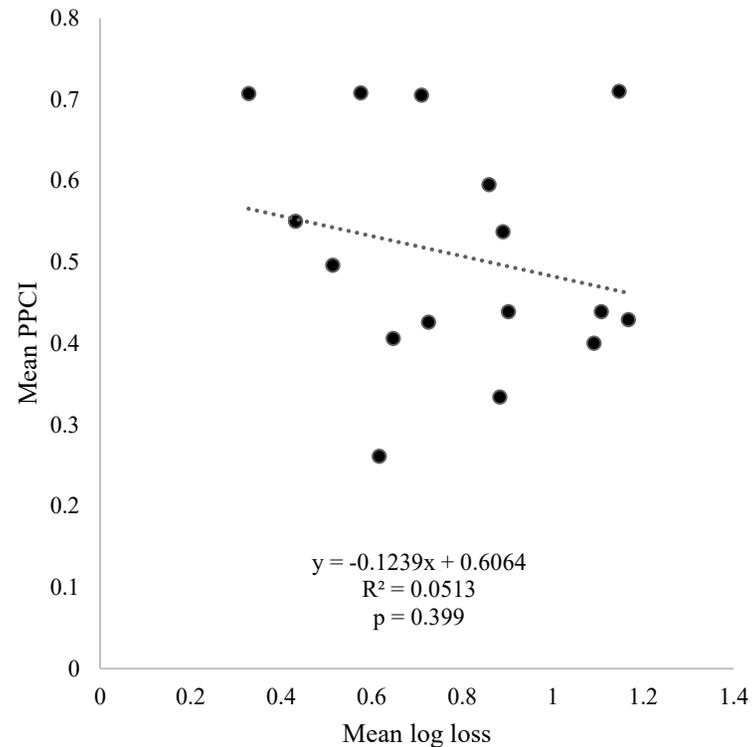
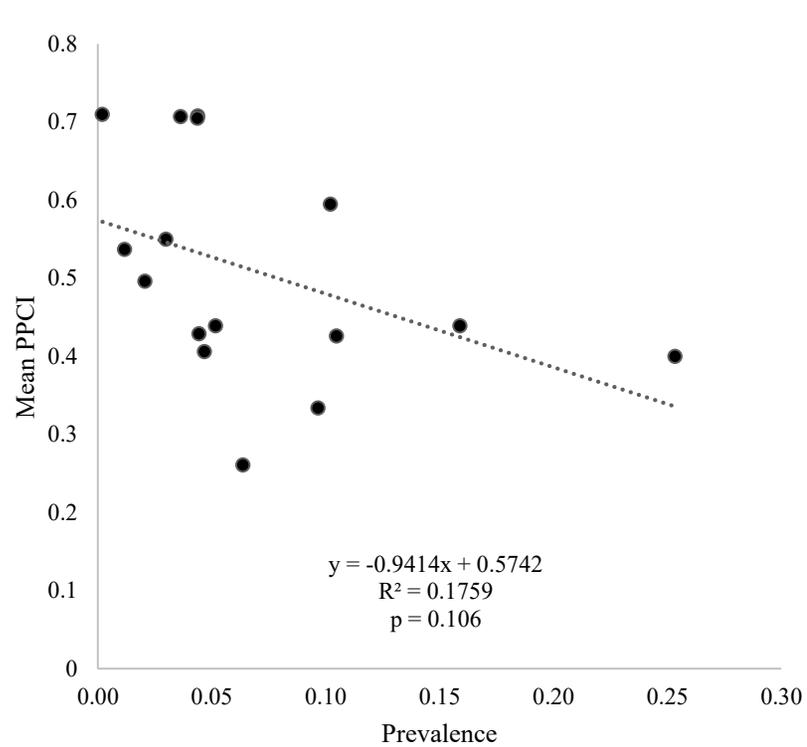


Figure 4.60: Scatter plots PPCI against prevalence and mean logarithmic loss (source: own).

CHAPTER 5 : DISCUSSION

5.1 INTRODUCTION

The recent accumulation of datasets on selected invasive species together with those on Swaziland's social, economic and environmental conditions has created a good foundation for developing models to investigate invasion patterns and processes. This study used data-driven BN models to identify the interactions of multiple environmental and socio-economic factors and their relative importance in shaping alien plant invasion patterns and processes in the country. Besides the predictive capacity of the different models, the BNs have additional advantages in terms of interpretability and inferential capacity. This chapter discusses the findings from the previous chapter with a focus on the observed spatial patterns and invasion processes. General observations on the BN learning algorithms used and the derived models are first discussed followed by species-specific findings. The performance of the BN models as well as the usefulness of the data mining or machine learning-based BN modelling framework for (invasive) species distribution modelling are discussed in the light of the findings.

5.2 BAYESIAN NETWORK MODEL DEVELOPMENT

Existing SDMs typically consider fewer variables and the use of a large number of variables is often considered disadvantageous since it is rare to have complete observations for all relevant variables. This requirement for complete observations does not hold for BNs, which can handle incomplete or missing data. Based on a BN model, the prediction of each species' distribution depends only on the variables of its Markov blanket. If the observation of the Markov blanket variables is incomplete, i.e. not all the variables are observed at inference time, information from outside the Markov blanket flows into the prediction by indirectly marginalizing (summing) missing variables out (Vogel *et al.*, 2014). This means that including many variables is desirable and provides additional knowledge, which is one of the advantages of BNs. Markov blankets have been shown to produce effective probabilistic models (Aliferis *et al.*, 2010). Moreover, the capability of BNs to predict from incomplete observations enable predictions to be made at an early stage of an event or with fewer

observations, employing only the information that is present at any given time such as early invasion detections. The predictions can subsequently be updated as new information or invasion data becomes available.

This study utilized discrete BNs, which use discrete or discretized variables. Some of the datasets used in this study were continuous in nature, thereby requiring discretization for BN modelling. The discretization method used (Section 3.5.4) was rigorous and enhanced the feature selection process by providing ecologically meaningful intervals and capturing the non-linear multivariate relationships based on each species' occurrence. The discretized attributes are easier to understand and interpret from a practical perspective such as when describing the climatic envelope of a species. The supervised entropy-based Kononenko discretization method effectively mapped the high-dimensional data sets into lower intrinsic dimensional spaces while keeping the intrinsic correlation structure of the original data. García *et al.* (2013) view the discretization process as a data reduction method in practice since it maps data from a large spectrum of numeric values to a greatly reduced subset of discrete values. The discretization process, therefore, facilitates the interpretation of the obtained BNs and maps and improves the accuracy of the classification task. The effectiveness of the method used may be attributed to the fact that entropy-based methods perform well on high-dimensional data regarding both the discretization intervals and classification accuracy (Sang *et al.*, 2014).

For the first time in species distribution modelling, the Kononenko MDL discretization method (Kononenko, 1995) was used together with the mRmR-based re-ranking feature subset selection technique (Bermejo *et al.*, 2012). The mRmR-based re-ranking feature selection technique was able to take advantage of the species–environment relationships or associations from the discretized and categorical variables to select those variables that were both predictively important and statistically relevant for each species. Both the number of variables and bins resulting from the feature selection and discretization processes were reduced thereby ensuring fewer ambiguous regions and a decrease in the number of uncertain links in the model structure. Too many discretization states or bins can make it hard to ascertain conditional (in)dependence and to determine the direction of the

relationships linking variables because of having too few instances or cases in each of the defined bins (Alameddine *et al.*, 2011).

This approach showed effectiveness in identifying a parsimonious set of uncorrelated but highly predictive and associated variables whilst minimizing noise and improving model performance. Furthermore, the reduced attribute space was appropriate for discrete BN-based SDM development and helped to reduce overfitting. Despite the possible loss of some information through the discretization process, the reduction and simplification of the input data made the BN learning process faster, yielding more accurate and compact results whilst reducing possible noise in the data. Discretization, therefore, balanced the complexity of the learned BN structure with efficiently modelling the training data. This ensured that variable discretization introduced an optimal number of intervals that captured interactions amongst adjacent variables in the BN.

The discretization problem remains one of the challenges associated with developing BNs with continuous variables and is one of the active areas of research. However, there are approaches to develop BNs that can work with continuous values and these include Conditional Gaussian models and Mixtures of Truncated Exponentials (Moral *et al.*, 2001; Aguilera *et al.*, 2010). Whilst the comparison of discrete BNs to continuous BNs by Aguilera *et al.* (2010) indicated a relatively better performance by the latter, the differences were likely artefacts of the discretization and parameter learning methods used. Nevertheless, there is need for investigation of the effectiveness of other contemporary discretization algorithms particularly those that can efficiently handle high-dimension and imbalance datasets.

The application of the Markov blanket, coupled with the automatic discretization and feature selection processes resulted in the learning of BN structures that had more arcs connecting the variables to each other and direct connections from some of the explanatory variables to the target (species occurrence) node. The direct arcs to the target variable intuitively indicate those factors that have a large informative effect are thus assumed to have an important influence on the distribution of each species. Most importantly, complex non-linear relationships between the target and predictor variables were uncovered. Moreover, the DAGs show which variables are the most relevant for the prediction of each species'

distribution and graphically decipher the probabilistic dependence and conditional relationships between those variables. It is important to emphasize that links or arcs between variables in a BN indicate that one variable is useful for predicting the other variable and vice versa. However, spatial scales can similarly influence the links between variables (Milns *et al.*, 2010), the mismatch of which can conceal the relationships between those variables.

For the naïve Bayes networks, the arcs which intuitively seem to be in the opposite direction indicate an important aspect of BNs. While in certain cases it is possible to infer causality between variables from the arcs in a BN, in most situations these arcs refer to predictive probabilistic relationships between variables rather a causal one (Smith, 2010). Similarly, common species co-occurrence patterns could be deciphered through the BNs. For example, whilst the presence of one species could (directly or indirectly) or could not influence the presence of another species, it was possible to correctly predict the possible occurrence of one species from the observation of the other species, e.g. *C. odorata* and *L. camara*. The presence of two or more arcs to a variable result in combinatoric, non-additive relationships among parent variables as was often the case with most of the learned BNs. This implies that only with the knowledge of the parameters or the CPT can it be ascertained whether multiple parent variables in a BN act independently or not. In most of the species studied, there was evidence of the fact that the variables acted together, i.e. the minimal set of dependencies in the resultant probability distributions require knowledge of two or more parent variables in order to predict the likelihood occurrence of a species. This is supported by the observation by Pearl (2000) that establishing a cause-effect relationship is always conditional on the “universe” one defines.

The interactions of different factors can be traced through the analysis of the dependence among variables. Uncovering these dependence relationships helped to reveal the underlying invasion dynamics and was useful for generating invasion hypotheses. The dependence relationships were dominated by indications of possible interactions which vary in their type, strength and symmetry. The mechanisms vary from species to species, ranging from mere co-location to association (positive/negative), competition/exclusion and facilitation through trophic interactions. Morales-Castilla *et al.* (2015) observe that whilst a

link between two species could be uncovered, it is often difficult to ascertain the expected type of interaction involved (e.g., antagonistic, mutualistic, facilitative, direct, or indirect) and their prevalence. However, the findings indicate that the uncovered relationships are potentially associative, mutualistic or facilitative considering that the feature selection process considered variables or factors that improved the BN models of the target species in addition to maximum relevance. Such interactions, which can be determined using geographical data, have been reviewed and observed elsewhere (Heikkinen *et al.*, 2007; Ovaskainen *et al.*, 2010; Araújo and Rozenfeld, 2014; Morales-Castilla *et al.*, 2015). It is also important that some of the selected variables may be proxy variables that are representative of latent (unobserved or hidden) variables either not directly measurable or not present in the data used in this study (Korb and Nicholson, 2011). A correlation or statistical association between two variables may also indicate the presence of hidden variables that are common causes for both the observed variables (Koski and Noble, 2012). The nature of such variables could be elucidated through in-depth research.

Perhaps the most interesting aspect of the learned BNs is that the predictors that have traditionally been used to explain species distributions were duly represented in the Markov blankets. This indicates that BN learning from data, coupled with rigorous feature selection techniques, can approximate expert and/or domain knowledge. However, some of the models generated by the algorithms had several arcs that were causally misdirected. This is due to the stochastic nature in which arcs are introduced by the rules in each algorithm. The BN algorithms revealed few locally known patterns of functional relationship and provided novel insights into the observed spatial structuring of the alien invasive plants in Swaziland. From the learned DAGs of direct relationships, BNs are evidently most useful in revealing the sets of functional relationships that influence alien plant distribution. The fact that a BN is a compact representation of probabilistic dependence (relevance) and independence (irrelevance) statements can be acknowledged from the learned BN structures. This confirms the unique ability of graphical models such as BNs to perform inter-causal and informative prediction. The learned models could be further used for what-if or scenario analyses in terms of both the environmental variables and each invasive species' distribution.

5.3 BAYESIAN NETWORK MODEL PERFORMANCE

The AUC, logarithmic loss, MCC and TSS values in Appendix 3: Performance of the Bayesian learning algorithms suggested good to excellent model performance, which was indicative of reliable predictions that have fewer cross-classification errors. Nevertheless, the empirical analyses showed that all the other accuracy metrics, save for the MCC and logarithmic loss, had strong negative or positive responses to prevalence. This is indicative of a known effect of prevalence (or ecological characteristics associated with prevalence) on predictive accuracy. The performance of SDMs is now widely acknowledged to be influenced by species traits (Hanspach *et al.*, 2010). Hanspach *et al.* (2010) and Tessarolo *et al.* (2014) found that TSS showed a strong response as opposed to being independent of prevalence, contrary to the findings of Allouche *et al.* (2006). The relationship with logarithmic loss and MCC, for instance, indicated that model predictions were not affected by species prevalence, contrary to general observations that specialist species SDMs are more accurate than generalist ones (Franklin *et al.*, 2009; Grenouillet *et al.*, 2011). However, when using the other metrics, the opposite was true whereby less prevalent species were predicted with a higher degree of accuracy (see Tessarolo *et al.*, 2014).

The fact that the AUC, which is known to be independent of prevalence, was sensitive to prevalence supports this interpretation. This can be explained by the fact that prevalent species such as *C. odorata* and *L. camara* often occupy wide niches (Allouche *et al.*, 2006; Soininen and Luoto, 2014) and hence data on such species lack strong ecological contrast to allow for meaningful species distribution modelling (Hanspach *et al.*, 2010; Tessarolo *et al.*, 2014). The area of predicted presence for such species is therefore much larger than that of scarce (low prevalence) species. On the contrary, Franklin *et al.* (2009) found that prevalence did not have an effect but that models that are more accurate can be developed for habitat-specialist species than habitat-generalist species. Nevertheless, as shown by the differences in the different metrics used, the choice of the commonly used evaluation metrics may create this artefact (Lobo *et al.*, 2008). The relative better performance of the globally scored BNs when considering the AUC and to a certain extent, the MCC, may be a result of the nature of the metrics themselves.

The data used in this study had very little bias considering the sampling method used to collect the aerial survey data. After all, SDM performance depends on the evaluation criteria (Aguirre-Gutiérrez *et al.*, 2013). Hence, the suggestions by Jarnevich *et al.* (2015) to use multiple metrics for model evaluation are affirmed in this study. The MCC was considered by Ding (2011) to be the best metric for imbalanced data learning. However, in this study the logarithmic loss was observed to be insensitive to data imbalance and hence a reliable metric. The logarithmic loss is only determined by the probability of the species presence or absence. The logarithmic loss' threshold independence and its suitability for evaluating probabilistic outputs made it more attractive for this study. The low logarithmic loss values indicate that the beliefs or posterior probabilities from the individual BN models were in better agreement with the data and more so for the ICS algorithm. Overall, all the classifiers achieved the best possible performance on the datasets and the findings suggest that even when the available data is highly skewed, the BNs perform well.

The findings generally indicate that GBN and BAN structures perform no better than that of the NB, CI and TAN when the same parameter estimation procedure is used. This is in agreement with the classical findings of Friedman *et al.* (1997) which showed that GBNs and BANs performed no better than NB and CI. This implies that the inclusion of dependencies among the features improved the prediction accuracy for species, especially *L. camara*, *Pinus* species, *P. x. canescens*, *A. mearnsii*, *J. mimosifolia*, *S. didymobotrya* and *M. azedarach*. On the other hand, since the target variables were treated the same way as the predictor variables in GBNs, the resultant dependencies could not be captured, suggesting that treating the target variable differently is useful for some species, e.g. *C. odorata*, *Opuntia* species, *C. jamacaru* and *S. punicea*. Similarly, when considering interpretability, the GBNs, BANs and the ICS algorithm performance was superior to that of NB and the CI as similarly observed by Madden (2009). This also suggests that the JPDs for a species could be represented with equal validity by an equivalence class, the collection of BN structures representing the same JPD with differences only in the direction of (some of) their arcs. However, the more flexible structure of GBNs and BANs endows them with the capacity to express approximate cause and effect relationships among not only the target variable and the explanatory biotic and abiotic variables, but also amongst the variables explanatory themselves (Milns *et al.*, 2010; Lee and Cho, 2012).

On the other hand, the performance of the hill-climbing algorithms informs us that better results can be obtained by choosing more intelligent moves through a sequence of moves (i.e. hill climbing) instead of a single move in each BN learning step. Correspondingly, the hill climbing and TAN structures included more arcs between variables in direct contrast to the conditional independence assumption of the NB and CI algorithms. This is an important observation from an ecological and species distribution modelling perspective because of the uncovered species-environment relationships. The generalizations of the NB and CI algorithm were better relaxed by recognizing and accounting for some of the dependencies in the training data. Therefore, the GBNs, BANs and TAN models were found to have the edge especially because some dependencies are necessary for indicating each species' relationship with its environment.

These findings corroborate the observations of Madden (2009) and Lee and Cho (2012) who found out GBNs and BANs generally perform relatively better than NB classifiers. Aguilera *et al.* (2010) and Lorena *et al.* (2011) noted that the limitations and relative poor performance of the NB might be attributed to its independence assumptions, which is not practical for ecological applications where interactions between variables are important. It is also unrealistic to assume conditional independence amongst the predictors with respect to target species' occurrences because:

- (a) a particular invasive plant species can partially respond to two or more environmental variables,
- (b) a particular ecological or landscape process or variable can be partially responsible for two or more invasive plant species, or
- (c) two or more ecological processes or variables can be related.

In addition, the response of an invasive plant species to one variable may be conditioned by the response of another species to a different variable. This indicates that BNs may be very helpful in species distribution mapping where species prevalence, robustness and causality and the interactions between variables are important considerations. The BN approach offers even a much better option because it can integrate more parameters that interact, either directly or indirectly, resulting in fast, robust, scalable and interpretable models that are simpler to understand.

The ICS algorithm, which is relatively computationally complex, attempted to discover the causal structure of the species-environment relationships resulting in BNs with probability distributions that closely matched the true distribution. The increasing development of novel constraint-based BN learning techniques, particularly hybrid approaches, presents a promising way to further explore the usefulness of this approach to BN learning as an attempt to approximate and even improve on expert knowledge. The advantage of constraint-based algorithms is the possibility of incorporating domain knowledge as additional constraints, resulting in better classification or prediction accuracy. New open-source packages that are mainly based on the R and Python platforms are of keen interest and promising in this regard. This is evidenced by the proliferation of more robust BN learning algorithms such as those reviewed by Nagarajan *et al.* (2013).

The performance of the scoring-based learning methods may be attributed to the objectives of the two approaches. The global scoring methods, for instance, use exact search algorithms that attempt to create an optimal DAG or arcs between the environmental variables (nodes) and the target variable from the available data. Furthermore, global score metrics measure the performance of a BN on a given data set by predicting its future performance through estimating the classification accuracy (Witten *et al.*, 2011). On the other hand, the local score metrics maximize the quality metric (e.g. Bayesian metric) of the given network structure, given the training data. Hence, the globally scored BNs produced DAGs that generally made better predictions of species occurrence than those obtained through local scoring. Nonetheless, whilst the local scoring-based BN models were mainly better at reproducing a given species distribution pattern, they were less so in predicting future patterns given that mechanisms driving distribution are not explicitly accounted for.

However, as noted in Chapter 2, finding the best BN structure is an NP-hard problem (Chickering *et al.*, 2003), hence the large training times for the global scoring models as opposed to locally score models. It would be prudent to suggest, therefore, that when computational cost is an important consideration, globally scored hill-climbing and simulated annealing algorithms be avoided. Clearly, contrary to the findings by Cutler *et al.* (2007), the number of covariates did not directly influence the computational cost of the BN models, as well as the complexity of the functional response obtained. It is possible, though,

that model-specific parameters in the BN learning algorithms could potentially alter the computational complexity responses and the posterior probabilities. García-Callejas and Araújo (2016) observe that model complexity is unrelated to predictive capacity for transferability if the estimated relationships between species distributions and environmental covariates are indirect. The geometrical characteristics of the data used in this study were not related to model performance and the choice and/or number of predictors did not affect the complexity of the models. Therefore, the topological complexity of BNs had no significant effect on their predictive ability and that the properties of species distributions data and their relationship with the environment, as determined by the CPTs, are strong determinants of model performance.

The findings imply that both BN structure and parameters are important in the prediction process. The general advantage of both scoring approaches is that the scoring functions balance goodness-of-fit to the data with a penalty term for model complexity, thereby avoiding overfitting. The relative better performance of the ICS algorithm when considering the logarithmic loss bears testimony to the importance of a causal structure that can result in better learning of species distribution parameters from available data. The NB also performed relatively well and was computationally cheaper than the other algorithms, hence its continued use for habitat modelling (e.g. Altartouri and Jolma, 2013). However, due to the strong conditional independence assumption, it is not an optimal method for probabilistic prediction especially in situations where there is imperfect and imprecise data from which the probabilities cannot be effectively characterized (Tütüncü and Kayaalp, 2015). This explains the relatively higher logarithmic loss values compared to the other BN learning algorithms.

BNs, combined with feature-selection approaches, may be more useful in reducing the number of direct input variables for prediction and in reducing the risk of mismatching the highest invasion risk areas with the lower invasion risk areas and vice versa through accounting for causality, dependencies and uncertainty. However, it is important that further studies investigate the efficacy of other robust feature-selection techniques especially hybrid methods which combine filter and wrapper models. The datasets used in this study provided the benefit of including more potentially relevant variables balanced against the detrimental

effect of including too many irrelevant variables in the classification node's Markov blanket. This implies that the Markov blanket used was an essential knowledge needed to predict the occurrence of all the species, and is probably another reason the NB was outperformed by the other BN classifiers.

Visual comparison of the prediction maps from the aerial survey data and the tree atlas data reveals that for some species, e.g. *C. odorata* and *L. camara*, some of the false presences were actually true. This is likely due to species detectability thus omission errors especially at the early stages of invasion. Hence, the models and the subsequent prediction maps invariably incorporated species detection uncertainty, which is a crucial component of species distribution modelling, although it is not always accounted for (Lahoz-Monfort *et al.*, 2014; Rota *et al.*, 2011).

5.4 SPECIES DISTRIBUTION PATTERNS AND INVASION PROCESSES

5.4.1 *Acacia mearnsii*

A. mearnsii is one of the widespread invasive species in the country. The poles derived from the plant were used as mine props in local mines when the species was introduced in the early 1900s (Menne and Carrere, 2007). Moreover, Doveton (1937) provides an earlier account of *A. mearnsii* being used for constructing traditional houses and being planted for economic uses by European settlers, particularly in the southwestern part of the country. It is important to note that *A. mearnsii* is often grown in the forestry industry for tannin production, and building material. Loffler and Loffler (2005) noted that this species was initially cultivated in woodlots for their barks as well as for fuelwood and building purposes. However, due to poor management of the woodlots, this species has spread uncontrolled (Bleys *et al.*, 1982). This study reveals that *A. mearnsii* occurs in human-disturbed areas as evidenced by the high occurrence probabilities in populated areas (densities greater than 5 people/km²) and the association with *J. mimosifolia*. The influence of human population density points to the assertion that human beings are primarily responsible for the seed dispersal and hence the spread of the species through its use as a hedge/windbreaker and its use for firewood through woodlots in the colder grassland ecosystem.

The distribution of the species is restricted to the western part of the country with smaller remnants in the eastern mountain ranges as determined primarily by its preference of mean temperature of coldest quarter less than 15.2°C as similarly observed by Duke (1983). Doran and Turnbull (1997) found that in the native range, minimum temperature of the coolest month (-3-7°C) is one of the key climatic constraints. In addition, the BN-derived models show strong association with *Rubus* spp. and *P. x canescens* probably due to their strong association with *Eucalyptus* species. *A. mearnsii* most often forms part of *Eucalyptus* species understorey (Weber, 2003; Loffler and Loffler, 2005) and their niche requirements generally overlap. *A. mearnsii* is a known invader along river corridors, forest and grassland including pine plantations (Henderson, 2001; Weber, 2003; Loffler and Loffler, 2005). The BN models captured all these variables including the temperature ranges and thresholds when taking into account the discretization states. These factors, coupled with the ability to transform ecosystems, make this species highly invasive as similarly observed by Richardson *et al.* (2011).

The management and control of this species, therefore, should take into account the strong influence of human activity particularly the deliberate planting for the identified uses. Control activities should focus on the areas of greatest impacts focusing on riparian areas and important grassland habitats. Control of *A. mearnsii* is also likely to help minimize infestations by other alien invasive plant species with which it co-occurs.

5.4.2 *Caesalpinia decapetala*

The BN models indicated that the optimum minimum temperature for *C. decapetala* ranges between 5 and 8.5°C affirming the observation that temperatures outside this range restrict this species from expanding westwards and to the eastern Lowveld as this species requires warm and moist conditions (De Beer, 1987). This study additionally finds that the species is linked to human population density thereby highlighting the role of humans in its introduction for ornamental purposes and as security barrier/hedge against people or animals (Henderson, 1986; 2001). This practice can be seen in many parts of the invaded range where fences are lined up with this species to form impenetrable barriers around homesteads and fields (pers. obs.). As a result, one area nearing the city of Manzini was aptly named

Lugaganeni, a local vernacular name for the species. In this and other invaded areas, *C. decapetala* can be observed around homesteads and other built infrastructure.

The relationship with land surface curvature indicates that *C. decapetala* prefers bushy hillsides, uplands, and along streams. This supports the observations by Loffler and Loffler (2005) that this species forms impenetrable thickets along the fringes of riverine habitats. Henderson (2007) also found strong invasion of this species along watercourses. Water currents are known to carry the large seeds downstream to form new infestations (De Beer, 1987; Henderson, 1989). The importance of *S. didymobotrya* and *S. punicea* in the *C. decapetala* models indicate a likely associative relationship derived from the fact that the former two species invade watercourses or riparian habitats (Weber, 2003) which are also the commonly invaded localities of *C. decapetala*.

The consociation with *L. camara* is also a noteworthy finding. A visual analysis of the prediction maps for both species indicates significant overlaps that may be attributed to shared niche requirements save for the larger range of *L. camara*. This concurrence is indicative of a likely biotic interaction that can be explained by *C. decapetala*'s role as a facilitator and transformer species through its ability to climb over and outcompete indigenous vegetation in particular. This it does through forming dense thickets mainly along watercourses and valleys thus suppressing native vegetation and changing the vegetation composition. By so doing, *C. decapetala* likely reduces the competitive effects of native species thereby allowing *L. camara* to establish itself. However, in-depth studies are needed to determine the exact nature of this interaction between the two species.

The strong association of *C. decapetala* with human activities suggests that control of this species should pay attention to disturbed areas particularly along riparian habitats in order to minimise its spread. This should include discouraging planting this species for the observed purposes. Moreover, the co-occurrence with species such as *L. camara* points to the need for a multi-species approach to the control of invasive plants to optimise costs.

5.4.3 *Cereus jamacaru*

Succulent collectors probably brought *C. jamacaru*, which originates from South America, to the southern Africa region (De Beer, 1987). The BN models indicate associations and co-occurrence with other species namely *J. mimosifolia*, *S. didymobotrya*, *M. azedarach* and *Opuntia* species whose distributions are largely human determined. A notable finding is that the only key abiotic predictors of *C. jamacaru* occurrence were human-related: proximity to tourism and human-disturbed areas. This illustrates that human-driven disturbances are the key processes driving the invasion of this species. These findings confirm the observations of De Beer (1987), Loffler and Loffler (2005) and Novoa *et al.* (2015) who observed that this species is predominantly horticultural and mainly cultivated for ornamental purposes in many gardens and sometimes planted as security hedges. Human disturbance also highlights the possible effect of human conversion of the landscape (e.g. through cultivation and gardening) which may facilitate the creation of suitable niches for the species whilst suppressing or removing native vegetation. Such areas, which are scattered in various parts of the country, are the potential sources of infestation resulting in the observed patchy distribution throughout the country showing no discernible pattern.

Additional to the biotic and human factors is the importance of bioclimatic variables specifically precipitation seasonality and precipitation of the driest quarter which highlight the importance of moisture as a limiting or controlling factor. The derived models point to preference for areas with low precipitation seasonality as well as areas with drier summer conditions, which is expected for the physiological functioning of a succulent cactus species. In its native range, *C. jamacaru* demonstrates annual flowering and fruiting related to seasonal rainfall (Gomes *et al.*, 2014). However, the prediction and accompanying uncertainty maps of *C. jamacaru* indicate that this species has a very large potential range in the country as similarly observed by Novoa *et al.* (2015) for cactus invaders at the global level.

Therefore, efforts to control this species should focus on minimizing its use for ornamental purposes focusing on the invaded areas, which are potential propagule sources. This will ensure the species does not spread to the wide areas where it has suitable niches.

5.4.4 *Chromolaena odorata*

One of the most aggressive invaders is *C. odorata*, which has invaded large tracts of land in the country. The species seems less tolerant of frost or very low temperatures as evidenced by low temperatures of the coldest month restricting its westward expansion to high altitude grasslands. However, incursions along warmer river valleys can be observed. A minimum temperature of 7.2°C is found to be the threshold for *C. odorata* distribution. Binggeli (1999) also observed this intolerance to frost as well as drought. The intolerance to drought perhaps explains low probabilities in the south-eastern parts of the country, an area characterized by very low rainfall and frequent drought conditions (Goodall and Erasmus, 1996). Goodall *et al.* (1994) and Goodall and Erasmus (1996) provided the first detailed account of the species invasion in South Africa and Swaziland wherein they highlighted that the species may have first invaded Swaziland in the mid- to late 1980s through the warmer south-eastern parts of the country. As predicted by Goodall *et al.* (1994), the species has since spread to most of the areas that were uninvaded two decades ago.

The importance of land surface curvature and surface form highlights the influence topography on the establishment of the plant. Although the species establishes itself in most terrain types, hillsides and drainage lines are particularly vulnerable to invasion. Another notable factor is the percentage of people with access to electricity. This compound variable is a measure of human welfare whilst at the same time it is an indirect indicator of the percentage of people who may be using other alternative sources of energy especially firewood, and thereby cutting natural vegetation, for heating. This highlights the potential effects of human welfare on patterns of spatial development and hence alien plant invasion on the landscape. Socio-economic variables are known to be indicative drivers of land use and land cover change including forest fragmentation through the creating of gaps that can facilitate invasion (Allen *et al.*, 2013; Mandal and Joshi, 2014).

Another important finding was the frequent co-occurrence of *C. odorata* with *L. camara* and *C. decapetala*. Ramaswani and Sukumar (2013) observed this co-occurrence with *L. camara* whereby there was a general increase in the abundance of *C. odorata* in areas where *L. camara* occurred. However, this increase was truncated by a further increase in the density of *L. camara*. *C. odorata* populations have also been observed to increase soil

fertility through enhancing the nutrient and organic matter content and reducing soil bulk density, thereby facilitating the invasion of species such as *L. camara* (Mandal and Joshi, 2014) and probably *C. decapetala*. This affirms the classification of this species as a transformer species (Henderson, 2001) through allelopathy, amongst other mechanisms (Sahid and Sugau, 1993; Zachariades and Goodall, 2002).

This species is vigorous in its invasion and as a result, it is invading many disturbed areas in many parts of the country. This highlights the need to focus on it as a priority species particularly in the high probability areas. Its control should also minimize new incursions into new areas primarily those that are disturbed by human activities.

5.4.5 *Eucalyptus* species

Eucalyptus species are largely planted in the forest plantations in the wetter and cooler western part of the country. In the 1930s, these were planted around many of the small towns of the Swaziland Highveld (Doveton, 1937). Currently, the main cultivated Eucalyptus species are *E. saligna* and *E. grandis* (Menne and Carrere, 2007). The confinement of *Eucalyptus* species to the wetter parts of the country is affirmed by the selection of the aridity index as the only climatic variable influencing its distribution. The relatively high aridity indices where this species is found indicates the species' affinity to dry-sub-humid to humid conditions thereby highlighting the importance of moisture to this species and its non-tolerance of arid conditions.

The importance of proximity to rivers and streams further affirms the affinity of this species to moisture or water availability. Some *Eucalyptus* such as *E. grandis*, have been observed to rely heavily on water abstraction from the lower soil profile in order to withstand dry seasons while some are deliberately planted along water courses (Forsyth *et al.*, 2004; Loffler and Loffler, 2005). Through a deep root system, they can exploit available groundwater particularly in areas where the water table is relatively high (Loffler and Loffler, 2005; Le Maitre *et al.*, 2015). In addition, rivers have been observed to spread *Eucalyptus* seeds (Forsyth *et al.*, 2004). The influence of slope aspect points to the common knowledge that the *Eucalyptus* species grow on hill slopes and lower slopes of valleys (Orwa *et al.*, 2009). Furthermore, slope has an influence on the microclimatic regime of a given

area in addition to other site-specific factors that form the niche of the species (du Plessis and Kotze, 2011).

The rest of the factors point to the strong influence of land use and the importance of areas close to major roads, which explain landscape patterns driven by human development (Allen *et al.*, 2013) particularly the importance of transportation routes. The models correctly identified that the species is most likely to be found in areas under plantation forestry where it is cultivated commercially for pulp, timber and firewood in addition to uses for donga rehabilitation around Swaziland (Loffler and Loffler, 2005). The association with *S. mauritianum* and *A. mearnsii* is a result of these two species' shade tolerance and ability to thrive under *Eucalyptus* plantations (Weber, 2003). This, therefore, is largely an associative and facilitative relationship that primarily highlights the other species dependence on *Eucalyptus* for part of their survival most likely through outcompeting non-shade tolerance indigenous species.

Similar to the other species, this species should be controlled in such a way that invasion of riparian habitats is minimized. The cultivation of *Eucalyptus* species should also be controlled such that it is contained within plantations to avoid impacts of other unintended areas such as riparian ecosystems.

5.4.6 *Jacaranda mimosifolia*

The dominant human influence on alien plant invasion is similarly evident in the *J. mimosifolia* models wherein human settlement density was selected as one of the key variables. *J. mimosifolia* distribution in the country is largely associated with human settlements wherein the species is planted for ornamental purposes along roads and as a hedge, mainly in urban and other areas where human settlement densities are above 13 settlements/km². Doveton (1937) provides an earlier recognition of the proliferation of the species in the urban areas particularly Mbabane. Loffler and Loffler (2005) also observes that this species is frequently observed in urban areas, where it is planted in gardens as an ornamental.

The species was positively associated with *P. guajava*, *Eucalyptus* spp., *S. mauritianum*, *Opuntia* spp., *M. azedarach* and *A. mearnsii*. There are no known biotic interactions of these

species with *J. mimosifolia*. Hence, the consociations are possibly indicative of shared niche requirements and characteristics such as the affinity to human disturbed areas. However, additional research on these interactions could shed more light.

Temperature annual range was the only bioclimatic variable found to be a determinant of this species' distribution with an upper threshold of approximately 22°C. This is an indication that that the species prefers sub-humid areas with lower intra-annual variation in temperature. The environmental conditions in those areas are similar to those in its native north-eastern Argentina, where it occurs mainly on riverbanks under warmer-temperate, sub-humid conditions (Poynton, 1973). Similarly, Henderson (2007) observed *J. mimosifolia* to be invasive in the moister parts of the savanna and forest biomes of South Africa.

The findings indicate that areas where this species needs to be controlled are primarily watercourses and other disturbed areas. Whilst complete eradication may not be necessary or possible, it is important that the species be controlled to avoid spread through riverine and other sensitive ecosystems.

5.4.7 *Lantana camara*

L. camara is the most widespread species when considering both the field data and prediction maps. This species has been in the country for a long time and has now naturalized. Magagula (2010) observes that there are now at least three biotypes found in the country, a pointer to the extent of adaptation of this species to the various eco-climatic conditions in the country. According to the BN model outputs, *L. camara* can be found under a wide range of climatic conditions in the country. Nevertheless, precipitation regimes (seasonality and precipitation of the wettest month) are the key climatic determinants of its distribution. The derived optimal conditions based on precipitation of the wettest month and precipitation seasonality are characteristic of warm to temperate conditions that are neither too wet nor too dry. Henderson (1989) also observed the wider climatic tolerance of this species. *L. camara* density has also been observed to increase with increasing rainfall in the absence of fire (Ramaswani and Sukumar, 2013).

The study finds that *L. camara* often occurs with a variety of other invasive species, mainly *C. decapetala* and *C. odorata*. The ability of *C. decapetala* to form thickets that tend to shade and climb over other (indigenous) species, thereby outcompeting them, may be playing a facilitatory role for *L. camara* invasion. There are also apparent mutualistic relationships between *L. camara* and *C. odorata* whereby the former acts a facilitator species for the latter's invasion. This is in line with Ramaswani and Sukumar's (2013) observation of a mutualistic relationship between the two species, which switches to be competitive at higher densities where *C. odorata* often outcompetes *L. camara*. This biotic interaction is supported by Mandal and Joshi (2014) who observed that *L. camara* invasion is enhanced by the improvement of edaphic and soil nutrient conditions through the transformative effects of *C. odorata*. Additionally, the two species share similar niche requirements. Le Maitre *et al.* (2002), who made similar observations in neighbouring South Africa, support the co-occurrence with *C. decapetala*.

The relationship with slope aspect points to a wide variety of land surface habitats ranging from north to west facing slopes and flat terrain albeit with fewer northwest facing slopes. Humans (population density), hence human activity and disturbance, is an important dispersal pathway in accordance with Henderson (2007). Densities as low as 2 homesteads/km² are enough to facilitate an incursion. Loffler and Loffler (2005) affirm the influence of human activity on *L. camara* invasion in the understorey of industrial timber plantations, urban areas, degraded land, and roadsides. Such disturbances, which are commonly through land conversion and other infrastructure developments, facilitate invasion through removal of indigenous plant competitors, increasing light availability, enhancing seed dispersal and improving soil nutrients (August-Schmidt *et al.*, 2015). Although the edaphic relationships are not straightforward, they are generally indicative of this species' requirement for soils with a good water retention capacity and aeration for its deep root system. It would seem that *L. camara* is able to effectively utilize resources distributed in edaphic space to its advantage, thus taking advantage of a variety of soil conditions.

L. camara is another priority species that requires urgent attention considering the prediction maps. The findings of this study indicate that this species will spread into most parts of the

country, predominantly those that are in close proximity to human settlements and those that are disturbed. Hence, control of this species should focus on those areas especially high propagule pressure areas.

5.4.8 *Melia azedarach*

The BN models and the maps indicate that *M. azedarach* is highly adaptable and tolerates a wide range of environmental conditions as similarly noted by Loffler and Loffler (2005). However, precipitation seasonality and August water content were selected as important climatic variables that determine the species' spatial distribution in Swaziland. Bisht and Toky (1993) observed that *M. azedarach* completes most of its growth during the initial dry part of the growing season, indicating that it uses reserves from the preceding year for growth. This could explain the importance of precipitation seasonality and August (winter) soil water content as learned by the models. The temperature coefficient of variation values where this species is found provides further evidence of the species' tolerance for seasonally warm and dry conditions. Moreover, August coincides with the fruit production season for this species (Orwa *et al.*, 2009). *M. azedarach*'s shallow root system is confined to the top layers of the soil and allocates most of its photosynthate to aboveground shoots, hence the soil water content requirement.

The importance of direct radiation duration indicates that *M. azedarach* requires reasonable sunlight, possibly the open sun, and is not shade tolerant. A notable finding is the positive association of the species with other invasive plants namely *S. didymobotrya*, *Opuntia* spp., *C. jamaecaru* and *J. mimosifolia*. Strong associations with *J. mimosifolia*, for instance, can be seen in densely populated areas. For instance, *M. azedarach* is cultivated for shade and ornamental purposes in gardens as well as along roads as similarly observed for *J. mimosifolia* (Henderson, 2001; Loffler and Loffler, 2005). The associations with these species points to the possibility of overlapping niche requirements characterized by human-disturbed areas and riparian ecosystems. As a transformer species (Henderson, 2001), *M. azedarach* likely enhances the invasion by the co-occurring species.

Riparian areas and disturbed areas, therefore, should be the focus of control activities. The wide niche of *M. azedarach* indicates that this species requires urgent attention to minimise its spread in the country.

5.4.9 *Opuntia* species

The natural range of the two recognized invasive species of *Opuntia* (*O. ficus-indica* and *O. stricta*) spans a large area with varying climates and habitats in the country. These species are found scattered throughout the country with no visually discernible pattern in the geographic distribution. This is similar to the observations by Kavirindi *et al.* (2010) who found that most populations in Namibia occurred either as isolated plants or as scattered satellite infestations. In neighbouring South Africa, the species have been observed scattered in many habitats and climates (Henderson, 2001; Smith *et al.*, 2011). The genus *Opuntia* is generally regarded as highly invasive and has a significantly higher proportion of invasive species (Novoa *et al.*, 2015). This is attributed to characteristics such as prolific fruiting coupled with strong vegetative reproduction and effective dispersal mechanisms (Walters *et al.* 2011).

The models in this study indicates that *Opuntia* species thrive best in conditions where winter (May) actual evapotranspiration is lower than 69.5mm. This suggests that the species may be slightly sensitive to dry season water availability. Although least important of all the selected variables, actual evapotranspiration is a measure of the amount of water that is removed from the soil due to both evaporation and transpiration processes. Cactus species such as *O. stricta* are well adapted to survive extreme drought through the Crassulacean Acid Metabolism (CAM) photosynthetic pathway.

The findings indicate that *Opuntia* invasions are enhanced by human disturbances. Major roads are a possible pathway for infestations as evidenced by higher probabilities in areas within 11km of major roads. Roads, which are often associated with human settlements and other developments, are disturbed areas and may facilitate long distance dispersal via seeds that are carried by animals, including humans that feed on the sweet fruit. Roads are often lined with telephone and electricity poles and fences that may provide good nesting sites for *Opuntia* seed dispersers such as birds (crows). Dean and Milton (2000) observed this association in South Africa. *O. ficus-indica*, in particular, is the most utilized introduced species for horticulture, one of the earliest reasons being its use as a drought-tolerant crop and for hedging (Walters *et al.*, 2011). In Swaziland, the species is often propagated for hedges, fodder, fruit, and donga stabilisation (Loffler and Loffler, 2005). Foxcroft *et al.*

(2004) also observed that propagule pressure was an important determinant of invasion of *O. stricta* invasion in South Africa's Kruger National Park.

The importance of proximity to rivers points to rivers and floodwaters as pathways of infestation along riverbanks as similarly observed by Lotter and Hoffman (1998). *Opuntia* seeds are possibly carried long distances by water after heavy rains or flooding. This could also indicate general preference for well-drained areas.

There is an observed niche overlap or association between *Opuntia* species and other species namely *J. mimosifolia*, *P. guajava*, *C. jamaicaru* and *M. azedarach*. It is not known if the transformer status of this species (Henderson, 2001) could be playing a facilitative role for the invasion of these other species. These consociation patterns are most likely to be a result of having similar niche requirements. Notably, *J. mimosifolia* is associated with human settlements where *Opuntia* species are sometimes used as barrier plants (Henderson, 1986). However, in-depth research is required to determine the exact nature and direction of these associations.

The control of this species should focus on the high propagule pressure areas particularly those nearing human settlements. The importance of rivers in *Opuntia* species distribution also indicates the need to focus on riparian areas. This could be done together with the control of most of the other species that seem to occur in these habitats.

5.4.10 *Pinus* species

Pinus species, predominantly *P. patula*, *P. radiata* and *P. taeda*, are grown in the country's plantation forest industry, which is important for economic development (Menne and Carrere, 2007). Established in the 1940s (Evans, 1974), these plantations grow species such as *P. elliottii* and *P. taeda*. Plantation forestry is practiced in specific land parcels in the cooler western part of the country. As such, the BN models uncovered land use and land cover as key determinants of *Pinus* species distribution. Human-disturbed areas also have relatively higher probabilities due to the species being used as an ornamental and wind break around homesteads. Hence, the BN models produced higher probabilities in those land areas. Eight decades ago, Doveton (1937) noted that *Pinus* species were common in the city of Mbabane. Interestingly, the introduction effort of *Pinus* species is related to its use as

ornamental species and is known to be highly invasive (Richardson and Rejmánek, 2004; Richardson *et al.*, 2011).

The optimum mean temperatures of the wettest quarter where the species occurs were below 21.5°C. This indicates that suitable sites must have lower temperatures to minimize evapotranspiration during the warm and wet season. These factors explain the wide distribution of the species in the cold western Highveld of the country. This relationship with temperature is not surprising as these species are montane species that prefer moist conditions coupled with low temperature conditions. Evans (1974), for instance, observed an optimum annual mean temperature range of between 15.5 and 19.5 °C, vales that in agreement with the threshold obtained in the models considering that the 21.5°C threshold corresponds to the wettest (normally warmest) quarter.

The co-occurrence of pine species with *S. mauritianum*, which forms the understorey, was uncovered. This can often be observed in the field throughout the entire species range. The frequent consociation with other invasive plants is also emphasized whereby areas with 4 to 5 other invasive species, in particular *A. mearnsii*, *Eucalyptus species* and *Rubus species*, had higher *Pinus* species occurrence probability. These patterns of co-occurrence can be ascertained from a visual analysis of the prediction maps. Worth noting is that the model indicates that this species tends to invade moderate to high tree species rich areas. This may be an indicator of the threat of *Pinus* species in biodiversity hotspots.

West-facing slopes were preferred together with those areas near perennial rivers, high stream/river density as well as areas with high cattle density. This indicates that watercourses are vulnerable to invasion by this species. Evans (1974) showed that the growth of *P. patula* in the Swaziland Highveld was strongly influenced by topography and soil. Furthermore, *Pinus* species are deep-rooted; a trait that they use to exploit deep soil for accessible moisture or groundwater particularly near water courses (Le Maitre *et al.*, 2015). The relationship with cattle density may indicate that highly overgrazed areas are vulnerable to invasion.

Pine species require control and management approaches that are focused primarily on containing these species within plantations and areas where they are planted for commercial

purposes. Efforts should be made to ensure that cultivation of pines is prohibited in riverine habitats and that such invaded areas are rid of this species.

5.4.11 *Populus x canescens*

P. x canescens is a plant that is normally introduced intentionally as an ornamental plant due to its attractive leaves of contrasting colour (Remaley and Swearingen, 1998). The models reveal that this species is tolerant to frost as it grows in areas with more than six frost days per year.

P. x canescens is also known to occur spontaneously along river valleys in its expansive native range around the Mediterranean and in central Asia (Modir-Rahmati, 1997; Stobrawa, 2014). In the country, the same pattern is observed (Loffler and Loffler, 2005). Hence, proximity to rivers was selected as a key determinant of *P. x canescens* distribution in the country in particular areas within 1km from rivers or streams. In these habitats, *Populus* species are observed to be dominant and pioneer species due to their tolerance for even flooding (Stobrawa, 2014) thereby forming dense stands in river valleys and near water sources (Loffler and Loffler, 2005). The species also establishes itself in areas of high late summer/March potential evapotranspiration indicating its preference to sunlight and adequate moisture. Although the species' distribution is currently restricted to the wetter western, particularly the southwestern part of the country, it has been observed to be thermophilic and light demanding whilst also tolerant of dry summers (Sekawin, 1975).

The high occurrence probabilities in highly fragmented land cover conditions and areas with high road density explains the influence of human disturbance in *Populus* species distribution. The construction of roads, settlements and other infrastructure fragments the landscape thereby creating important introduction pathways. This finding conforms to Remaley and Swearingen's (1998) observation that the species grows best in full sun habitats such as fields, forest edges, and wetland fringes. Additionally, Loffler and Loffler (2005) stated that *Populus* species are initially propagated in woodlots for matchwood purposes in the country.

The affinity to soils with moderate bulk density, coarse fragments and soils with silt content more than 12.5g/kg indicates the specific edaphic requirements of the species. *P. x*

canescens is known to tolerate dry, saline and calcareous soils, but prefers neutral, well-textured soil and good water availability (Os'kina and Bepalov, 1992; Rédei, 1998; Stobrawa, 2014). This is important for the species because its local spread is primarily by vegetative means via root suckers (Remaley and Swearingen, 1998).

The southwestern part of the country is, according to the maps, a priority area for control of this species. Other focal areas should include high propagule pressure areas such as those that are disturbed by human activities including areas within proximity to human settlements and other infrastructure developments.

5.4.12 *Psidium guajava*

The findings indicate that *P. guajava* is associated with *J. mimosifolia*, *S. didymobotrya* and *Opuntia* species. A visual comparison in the distribution patterns of *P. guajava* and these species shows some similarities in the distribution pattern. This may suggest that these species generally occupy similar habitats. Furthermore, the positive effect of *L. camara* on the probability of *S. didymobotrya* presence indicates a positive interaction, possibly a mutualistic or facilitative relationship. Indeed, these species were frequently observed in similar sample sites/plots from both the aerial and tree atlas surveys. *P. guajava* was also observed to form dense thickets that displace native vegetation. This transformative effect most likely facilitates the invasion of other species such as *S. didymobotrya*.

Areas with August (winter) evapotranspiration values between 29.5 and 53.5 mm offer optimal conditions for this species, which highlight the species' preference for sub-humid to humid climates. These conditions are prevalent in the central parts of the country (Middleveld) to parts of the Highveld. *P. guajava* is also found to be intolerant of minimum temperatures less than 8.4°C, which is very close to the lower limit of 9°C observed by Yadava (1996). These are the same localities identified by Loffler and Loffler (2005) to have an increase in *P. guajava*.

The importance of electric power lines indicates that the construction of these, through disturbance mechanisms, may play a key role in its distribution. The construction of electricity pylons and access roads creates gaps and ecotones, which facilitate invasion and possibly reduce competition from native vegetation. Furthermore, a high level of

fragmentation on the land surface cover, which is an indicator of disturbance and intense human land modification, seems to create edges and other suitable habitats for the species to establish itself (Weber, 2003; Loffler and Loffler, 2005). Henderson (2001) observed these patterns in neighbouring South Africa. *P. guajava* fruits are also edible and are consumed by a majority of the population in Swaziland (Ogle and Grivetti, 1985; Loffler and Loffler, 2005). During the fruiting season, vendors can be seen on the roadsides selling buckets of the fruit along the major roads to passing motorists and passers-by. This is most probably the main mechanism through which the *P. guajava* seeds are dispersed in the country. Frost conditions of the western part of the country and the drier and warmer conditions to the east constrain its distribution into those areas.

5.4.13 *Rubus* species

The *Rubus* species appear to be frost tolerant whilst minimum temperatures of the coldest month higher than 7.5°C present unsuitable conditions. Dean *et al.* (1986) first attributed this to a seed dormancy mechanism that is terminated by cold winter temperatures. This provides an explanation to the high posterior probabilities in the cooler western part of the country. The requirement for relatively high precipitation seasonality shows that the species is not drought tolerant and prefers areas with distinct summers and winter climates.

The species is found to co-occur with *A. mearnsii* and *Eucalyptus* species under which it forms the understorey (Henderson, 1989; 2001; Loffler and Loffler, 2005). This suggests some kind of aggregation of the *A. mearnsii* and *Eucalyptus* trees and *Rubus* shrubs, which is not surprising because this phenomenon has been observed many times in many different plant communities (Rejmánek, 2015). One such facilitation mechanism that may be responsible for this interaction is the nurse effects of trees on shrubs via microclimate or soil modifications. There are also associative relationships with *P. x canescens* and *S. mauritianum*, the latter species observed by Le Maitre *et al.* (2002) to co-occur with *Rubus* species under forest plantations.

Similarly, the high posterior probabilities in high river/stream density areas indicates preference for moisture, riparian habitats and gullies. Erasmus (1984) and Loffler and Loffler (2005) noted that *Rubus* species invaded gullies, valleys and watercourses which may be a pointer to the invasion pathway for this species whereby rivers transport the seeds.

Likewise, the species was frequently found in areas that are within 9km of major roads as well as within 25km from major tourism routes. These factors, which are indicative of the effects of anthropogenic activities and human beings being important dispersal agents. Human activities facilitate the dispersal of the species through ornamental and consumptive uses (Erasmus, 1984; Loffler and Loffler, 2005). Road clearings could create openings thereby allowing the species to invade whilst roads provide routes through which the species is spread.

Moderate to high tree species richness likely correlates with bird and other mammal species richness. Birds and other mammals are known to be very important dispersal agents for this plant (Buddenhagen and Jewell, 2006; Loffler and Loffler, 2005). It is, therefore, likely that the species shares seed dispersers with the native vegetation and its other associate invaders. Possible impacts on biodiversity (tree species rich areas) is correspondingly of concern considering that this species is a transformer species (Henderson, 2001).

Control activities for *Rubus* species should focus of those areas where it is currently located especially near human disturbed areas. Plantation forests are likewise prone to invasion and as are riparian habitats, thereby indicating that efforts should be put to those areas to minimize *Rubus* species spread.

5.4.14 *Senna didymobotrya*

The results suggest that *S. didymobotrya* and *S. punicea* distribution patterns are more similar and that the two species normally occupy similar habitats. Indeed, the two species were frequently observed in similar sites particularly along watercourses. The models highlighted proximity to water sources as an important factor confirming the observation by Orwa *et al.* (2009) that, in its natural habitat, *S. didymobotrya* is often ruderal in areas with regular water supply such as riparian areas and wetlands. Henderson (2001) and Loffler and Loffler (2005) observed this spatial patterning along riverbanks and within riverine vegetation. It is likely that rivers play a role in the dispersal of seeds.

Therefore, the positive effect of *S. punicea* on the probability of *S. didymobotrya* presence may not only indicate positive interactions between species but habitat overlapping. The

models indicate that the species was often found within close proximity or together with five other species mainly *P. guajava* and *M. azedarach*, as well as *S. punicea*.

The species is found to prefer well-drained soils with low clay content although this factor was relatively less influential in the model. The high posterior probabilities in areas with high land cover fragmentation indicates that the species finds refuge in disturbed areas. Human disturbance may also facilitate invasion through creating ruderal or sage habitats. Loffler and Loffler (2005) made the same observation where disturbed areas, roadsides and wastelands were frequently invaded by *S. didymobotrya*.

Interestingly this species can tolerate frost exposure up to 25 frost days per year thereby restricting this species to the temperate and relatively warmer parts of the country with minor incursions into the west along low-lying valleys. Orwa *et al.* (2009) observed that the species tolerates light frost.

Hence, the control of this species should focus on removal from disturbed areas with a focus on riparian areas where it co-occurs with other species.

5.4.15 *Sesbania punicea*

S. punicea is adapted for wetland and riparian zones, the buoyant seed pods of which are capable of being dispersed over long distances by water currents (Henderson, 1989; Hunter and Platenkamp, 2003). In the same way, the species requires sufficient moisture for the survival of seedlings during dry summers (Hoffmann and Moran, 1991). Wetlands are also frequently invaded by this species (Henderson, 2007) as are rivers and stream banks (Loffler and Loffler, 2005). This explains the importance of proximity to major (perennial) rivers and water sources in the BN models. The affinity to areas with low topographic index further affirms that relatively flat areas such as drainage lines provide suitable habitat. Foxcroft *et al.* (2008) found that *S. punicea* invasion in Kruger National Park was facilitated by the conditions provided by riverine channels through frequent inundation or disturbance.

Other suitable habits coincided with areas with high soil bulk density as learned by the BNs. Foxcroft *et al.* (2008) observed that *S. punicea* responded mainly to geomorphic units representing differences in the bedrock morphology and texture of the underlying channel sediment. These suitable areas are characterized by low fire frequencies where fires burn at

most once in five years. Frequent fires likely kill the plant and may suppress its growth and spread.

As indicated in the *S. didymobotrya* and *C. decapetala* model, the consociation of *S. punicea* with these two species is confirmed, pointing to possible niche partitioning and biotic interactions between these species. Henderson (2001) classifies this species as a transformer species, which would imply that it further facilitates the invasion of other species through transforming the natural ecosystems. The biotic interactions and niche partitioning are affirmed by the strong relationship with invasive plant species richness where *S. punicea* occurrence probabilities were high in areas with at least five other invasive alien plant species.

Since this species is largely used for ornamental purposes because of its attractiveness especially when flowering (Hoffmann and Moran, 1991), it is found near human settlements especially where densities exceed 13 homesteads/km². Hence, human habitation creates patchiness or gaps, which provide for increased *S. punicea* invasion ability in the landscape (Foxcroft *et al.*, 2008). Furthermore, this highlights the role of humans in the dispersal of the species. Hence, control of this species should focus on areas near rivers where the species is proliferating including areas that are disturbed such as those near human settlements.

5.4.16 *Solanum mauritianum*

The strong co-occurrence of this species with *A. mearnsii*, *Eucalyptus*, *Pinus*, and *Rubus* species is further affirmed in this species' model. *S. mauritianum* was observed by Henderson (2001) to be shade and frost-tolerant and highly invasive in forest plantations (where *Pinus*, *A. mearnsii* and *Eucalyptus* are the key species), riparian zones and watercourses, roadsides and other disturbed areas (e.g. disturbed natural vegetation, urban open spaces, along the course of electricity pylons etc.). Similarly, Loffler and Loffler (2005) made the same observation of *S. mauritianum* in timber plantations. In South Africa, *S. mauritianum* is a major problem in commercial forestry plantations where it competes with *Pinus* and *Eucalyptus* species seedlings thereby inhibiting their growth (Hinze, 1985). These biotic interactions were uncovered in the individual models for those species. However, the association with *C. jamaecaru* was relatively weak.

The influence of human disturbance was confirmed through the selection of land cover fragmentation as a key variable. Areas with moderate land cover fragmentation (Shannon index of 1.07 to 1.61) were found to create optimal conditions for *S. mauritianum* invasion. Loffler and Loffler (2005) also observed *S. mauritianum* infestations in disturbed areas. Land use intensification could be creating suitable habitats for frugivores that are responsible for seed dispersal rates and hence the spread of the species. However, Schor *et al.* (2015) observed that land use intensification decreased frugivore abundance, which translated into decreased fruit removal rates, and hence reduced spread, of *S. mauritianum*. Although bird species richness was part of the dataset used in this study, there was no specific information on frugivores to ascertain these relationships. The poverty index further provides an indicator of anthropogenic influence whereby areas with low poverty, and thereby high probable levels of land conversion and human mobility, were more vulnerable to invasion.

The proximity to rivers was an important variable with high *S. mauritianum* occurrence probabilities predicted in areas within half a kilometre from a stream or river. The species was mostly abundant in humid areas as evidenced by the high mutual information value with November actual evapotranspiration. Areas with November AET higher than 68.5mm were specifically found suitable pointing to the species' requirement for high moisture levels.

Considering the importance of the identified factors, *S. mauritianum* control should focus on plantation forests and human disturbed areas. Areas that need to be specifically targeted include riparian areas and ruderal areas.

5.5 GENERAL DISCUSSION ON INVASION PATTERNS AND PROCESSES

5.5.1 Invasion patterns and processes

Most of the invasive alien plant species studied arrived in the southern African region and the country during the 16th and 17th centuries (Doveton, 1937; Henderson and Wells 1986; Foxcroft *et al.*, 2010). This would imply that most of them have naturalized in the country and hence have become invasive. Since the alien plant invasion patterns and processes for each of the species have not been studied in detail, there has been a need to determine the drivers of invasion. The high-resolution aerial survey dataset used in this study helped to uncover important landscape level environmental gradients and relationships. There is no indication from literature that a survey has ever been undertaken at this scale and magnitude (nationwide) elsewhere in the world. Such a detailed dataset, therefore, presented an opportunity to test the ability of BN-based data mining techniques to uncover the finer scale spatial patterning of alien plant invasion and possible underlying processes.

Predicting the potential distribution of the selected alien plants is pivotal to planning their effective control and management more so because such species are rarely at equilibrium with their environment due to the progression of invasion coupled with the ongoing biotic and environmental changes in the invaded landscape. Through the BN-based machine learning approach, together with the feature-selection and other data pre-processing techniques used, it was possible to identify and select the environmental variables that have a greater bearing on the distribution pattern of each species. Accordingly, differences in BN structure appear to account for some of the disparities in model performance between all the species, pointing to dissimilar interrelationships among the variables. This is a result of disparate influences of the various explanatory variables on the spatial distribution of the species. The learned BN structures graphically reveal the complexity of the species-environment conditional dependencies that determine the occurrence of each species. This complexity may be attributed to the country's high climatic and topographic heterogeneity.

The BN models performed the automated searches for hidden spatial patterns in the large and multi-dimensional species-environment data. Hence, the distribution of most of the studied plants was similar to the general descriptions by Loffler and Loffler (2005). It is interesting to note that for all the species there is an observed interplay of key bioclimatic,

topo-edaphic and disturbance and propagule pressure (primarily anthropogenic) factors along with co-occurrences (or possible biotic interactions). These explain the processes that govern the observed distribution patterns. This is in line with the observations that the distribution of most plants is strongly affected by the nonlinear dependence on, as well as the influence of, the dominant environmental factors (Santika, 2011). Expectedly, the generally positive relationships among the invasive plant species and anthropogenic factors suggests that human disturbance promotes alien plant invasion through a combination of disturbance itself providing resource opportunities, and propagule pressure of ruderal alien plants from adjacent developed or disturbed areas.

The significant influence of human activities on the distribution of invasive alien plants is evidenced by the high posterior probabilities in populated and disturbed areas for all the species. Most of the species are located closer to the points of introduction (whether intentional or accidental) and human activities. The concentration of these plants around areas of human activity reflects human-mediated dispersal and disturbance through fragmentation of the landscape and creation of suitable habitats. These observations corroborate several studies which reveal the role of human activity in alien plant invasion in the region (e.g. Taylor and Irwin, 2004; Thuiller *et al.*, 2006; Fuentes *et al.*, 2015). The frequent selection of transportation (road) infrastructure, human settlements/population and land use into the models affirms the propagule pressure hypothesis (Lockwood *et al.*, 2005) which influences the spatial patterning of invasion in the country. As a result, the distribution of alien invasive plant species in Swaziland strongly follows the pattern of human alteration of ecosystems through agriculture, urbanisation and settlements, road and infrastructure construction and other land uses as similarly observed by Rouget *et al.* (2015).

In certain instances, the influence of anthropogenic activities overshadows the limiting effects of both biotic and climatic variables, resulting in variations in the species-environment relationships. For instance, the degree and nature of this clustering around human activities varies with species and is not necessarily influenced by residence time as suggested by Hui *et al.* (2013) but is constrained by biotic, bioclimatic and topo-edaphic factors. For instance, *C. odorata* is a recent invader that has spread to cover a larger area than those species that have been in the country for much longer (e.g. *Pinus* species).

Nevertheless, the findings indicate that most major invaders spread from forestry plantations and human settlements into new areas through human-mediated propagation as hedges, windbreaks, ornamental, silvicultural and agricultural plants.

The utilization of alien plants is also an important aspect of the human-invasive plant interaction as this determines the levels and patterns of invasion thereby shaping invasion pathways (Wilson *et al.*, 2009; Pyšek *et al.*, 2010; Bigirimana *et al.*, 2012). Ornamental species such as *C. jamaicaru*, *J. mimosifolia* and *Opuntia* species are primarily disseminated through horticultural practices often close to a wide range of potential habitats (Richardson and Rejmánek, 2011) and thus more widespread in their distribution. The gardens of tourism hotspots such as hotels and lodges are often decorated with various plant species some of which are alien (pers. obs.). *Pinus* and *Eucalyptus* species are commercial species mainly used in plantation forestry because of their fast growth rate, among other traits that are typically associated with adaptations for rapid colonization and the inherent invasiveness (Grotkopp *et al.*, 2010).

In Swaziland, these species are widely grown in large plantations in the western part (Highveld) of the country, allowing for the accumulation of large amounts of propagule banks. *A. mearnsii* is a woody plant most widely used in agroforestry because of its tolerance of a wide range of conditions, rapid growth and frequently precocious and prolific fruiting and/or seed production (Richardson and Rejmánek, 2011). This species is grown in the cooler western part of the country and is found in highly disturbed areas mainly near human settlements, which defines the introduction and invasion pathways. These factors, and the role of cultivation methods in mediating invasiveness, are fundamental filters that have resulted in the observed patterns of occurrence shown in Figure 4.20 and Appendix 2.

Human-mediated dispersal of alien plants is of direct relevance to invasion ecology particularly in understanding and managing dispersal pathways (Pyšek *et al.*, 2010; Auffret *et al.*, 2014). This is important to note because present-day patterns of alien species invasions have been observed to be a function of human disturbance history, particularly land use, going back to decades or even centuries (Lim *et al.*, 2014; Beauséjour *et al.*, 2015; Oswalt *et al.*, 2015). Human-mediated dispersal through vectors such as livestock and transportation mechanisms and their relationships with the physical environment are

moreover valuable for understanding and managing pathways of dispersal including invasion (Auffret *et al.*, 2014). This relationship between anthropogenic parameters and invasive alien plants reflect the intensity of human activities, which results in more such plants being transported and introduced, increasing the risk of their invasion. Hence, an increasing human population, coupled with increasing developments through agricultural development and road and infrastructure construction, is expected to result in more invasions in the country within the near future.

The importance of socio-economic factors is underscored in the learned BN models and helps to explain the patterns of landscape fragmentation that relate to plant invasion. Gaviera-Pizarro *et al.* (2010) and Allen *et al.* (2013) revealed that socioeconomic factors explain development, forest fragmentation and landscape patterns driven by human development and are linked to increased woody plant invasions. These variables provide a link on how human welfare influences landscape patterns of development and anthropogenic utilization of the landscape. However, socio-economic variables, although important, have received very little attention in species distribution modelling perhaps due to the difficulties in explaining their direct influence on species distribution and ecology.

The study demonstrates the general capability of BN models to uncover species co-location (co-occurrence) and association patterns where species are frequently located close to each other. It is common knowledge that species do not only interact in pairs, but can do so in complex networks as well (Bascompte, 2009). These higher-order interactions may lead to non-additive effects (Dormann and Roxburgh, 2005) which were better represented by the BN models. For example, if two interacting species, *C. odorata* and *L. camara*, are considered, a third component species *C. decapetala*, can be added which may affect the way *C. odorata* interacts with *L. camara* under changing environmental conditions or as mediated by other abiotic variables. All the models accounted for such possible biotic interactions that were graphically revealed by the BN topological features. Hence, if the distribution of one species is highly dependent on the distribution of another species, the BN models indicated the direction and strength of the relationship. This is not only indicative of possible biotic interactions between the species but also closeness in terms of niche requirements. Hence, invasive plant species richness was an important factor for some

of the species. These are important findings considering that the spatial patterns of alien species assemblages or co-occurrences across landscapes, though very important, have received much less attention (Hui *et al.*, 2013; Rouget *et al.*, 2015).

Globally, alien woody species are more likely to have positive associations or co-occurrences (Kuebbing and Nuñez, 2015). Coincidentally, with the exception of *C. jamaicaru* and *Opuntia* species, all the species studied were woody plants. According to Ovaskainen *et al.* (2010) and Kissling *et al.* (2012), such positive co-occurrences are often attributed to species having similar environmental, dispersal or biotic requirements, or to some direct or indirect positive interaction between the species. Most of the inter-alien plant relationships identified in this study were positive, such as the strong predictive power provided by the presence of *C. decapetala* on *L. camara* and vice versa. These positive relationships may reflect mutually beneficial relationships such as direct facilitation (e.g., shade tolerant *S. mauritianum* may benefit through competitive exclusion of shade-intolerant species by *Eucalyptus* or *Pinus* spp.), use of other species as indicators of habitat suitability, or mediation by unmeasured variables.

Positive co-occurrences can also be used to infer facilitative interactions (Ovaskainen *et al.*, 2010) while the likelihood of geographic co-occurrences is high for mutualistic interactions (Traveset and Richardson, 2014; Araújo and Rozenfeld, 2014; Morales-Castilla *et al.*, 2015). This is most likely for woody plants that often act as nurse plants through creating novel and favourable microenvironments, thereby promoting the establishment of other invasive plants that may otherwise not invade an ecosystem (Kuebbing and Nuñez, 2015). At the microhabitat scale, differences in the resource use can lead to the differential use of an ecosystem resulting in coexistence between species and the overlap in habitat use as shown by the models. However, it is important to state that competition can limit the population size of another species without completely excluding other species from habitats and further microhabitat analysis may provide valuable additional information. This was evident in the case of *L. camara* and *C. odorata*.

The frequent co-occurrences or associations suggest that the country, or parts thereof, is undergoing an ‘invasional meltdown’ (Simberloff and Von Holle, 1999) whereby the presence of other invasive species are facilitating the invasion by additional species, thereby

increasing their likelihood of survival (Traveset and Richardson, 2014). The transformative effects of most of the species may be exacerbating their spread through the ecosystem engineering mechanism (Jones *et al.*, 1994; Cuddington and Hastings, 2004). The ‘ecosystem engineers’ hypothesis asserts that some species have the inherent ability to alter their invaded environment through non-trophic interactions so as to promote themselves and to decrease neighbours. The known allelopathic traits of some of the species such as *L. camara* and *C. odorata* (Sahid and Sugau, 1993) may be supportive of the ‘novel weapons hypothesis’ (Callaway and Aschehoug, 2000; Callaway and Ridenour, 2004) which explains the competitive advantage of invasive species against native species thereby affecting them negatively and reducing competition. Furthermore, Henderson (2001) classifies a large majority of the species studied as ‘transformer species’, i.e. species that change the ecosystem characteristics and attributes over a substantial area relative to the extent of that ecosystem (Richardson *et al.*, 2000). Invasion hypotheses that consider invader-ecosystem interactions are better supported by empirical evidence (Jeschke *et al.*, 2012). This, therefore, suggests that invader-ecosystem interactions are working in a hierarchical way to produce the observed spatial patterns of invasion in the country.

The key task in the inclusion of biotic interactions in the modelling of species distribution is to identify the species with strong positive interactions that are capable of affecting distributions and coexistence across scales (Araujo and Rozenfeld, 2014). It is important to note that most modelling approaches designed to account for biotic interactions have limitations in inferring causation from spatial data. The modelling approach used in this study, coupled with the feature selection process, presented a novel way to graphically represent such possible interactions. However, if the distribution of one species is shown to be highly dependent on another species it can be difficult to differentiate if this is due to a real biotic interaction between the two species or is better explained by one or more missing or overlooked environmental factors not accounted for in the model (Wiszniewski *et al.*, 2013). While the precise nature of these relationships may not be fully known, they are reflective of underlying interactions within the vegetation community. This is important because of the rising interest in joint distributions models (JSDM) or methods that use data from multiple species simultaneously to study species co-occurrences (Kissling *et al.*, 2012; Clark *et al.*, 2013; Golding *et al.*, 2015; Thorson *et al.*, 2015). However, in-depth studies are

required to ascertain and differentiate if these relationships are due to real biotic interactions between species or are better explained by one or more overlooked environmental factors not accounted for in the models (see Wisz *et al.*, 2013).

The importance of climatic variables was expected, in particular the role of temperature. Temperature is considered the most important factor determining a species' distribution due to its effect on biochemical and cellular processes that affect an organism's performance (Kelley, 2014). Of all the bioclimatic variables used, the minimum temperature of the coldest month and frost occurrence appear to have the greatest bearing on the distribution of some of the invasive alien plant species in the country. These low temperature extremes have direct physiological roles in limiting the ability of plants to survive and grow, while some species have a chilling requirement for processes such as bud break and seed germination. These factors are important because many invasive alien species are ectothermic (i.e. poikilothermic) and temperature directly affects developmental, reproductive and survival rates (Venette, 2015). Low temperatures are, therefore, important factors in limiting the westward and altitudinal distribution of the invasive plants in the country. Precipitation seasonality, on the other hand, determines temporal variation in moisture availability.

The fact that the plants under investigation vary greatly among themselves in their resistance to cold is evidenced by the lack of a coincidence in the western limit of distribution. The line that marks the limit of frost is the most important climatic boundary in restricting the westward extension of mostly perennial species such as *C. decapetala*, *C. odorata*, *L. camara* and *P. guajava*. In any consideration of the geographical importance of the operation of these factors, therefore, it is apparent that the low temperatures serve as a check on the possible movements by these species. A study by Henderson (2006) showed that the current distributions of invasive plants in southern Africa are similar to the climatic zones of their native range. This has implications when considering the potential impact of climate variability and change on invasion processes whereby expected higher minimum temperatures and shorter frost duration may enhance the spread these species further west.

While large-scale geographic distribution of the invasive plants is limited by climatic, biotic and other anthropogenic variables, small-scale distribution is often limited by topo-edaphic

variables (Diekmann *et al.*, 2015). Visual analysis of the predictions reveals that rivers and riparian habitats are important dispersal routes for most of the alien plants. Riparian areas are highly vulnerable to invasion because they are often disturbed by flooding events and the extensive presence of ecotones, which are the preferred habitat of many invaders (Le Maitre *et al.*, 2002; Richardson *et al.*, 2007). Furthermore, in these areas water is freely available for plant growth and seed dispersal (Le Maitre *et al.*, 2002; Foxcroft *et al.*, 2008). For instance, the floods in the year 2000 are linked to the upsurge in invasion in the Kruger National Park (Foxcroft *et al.*, 2008). This would imply that the 1984 Cyclone Domoina, which was the largest in the country in recent times (Kovács *et al.*, 1985), had a significant impact on the present-day distribution of some of the invasive alien plants in the country.

Riparian zones are characterized by anthropogenic activity and disturbance, which facilitates the proliferation of invasive species (pers. obs.). The observed and potential distribution of species such as *M. azedarach*, *S. didymobotrya* and *S. punicea* is significant along rivers across the climatic and edaphic gradients. The constraint of edaphic factors is related to the thresholds beyond which species are not able to survive. The importance of edaphic variables for predicting plant distributions has been highlighted before (Dubuis *et al.*, 2013; Thuiller, 2013; Beauregard and de Blois, 2014). These thresholds differ even between species considered to have very similar ecological requirements (Diekmann *et al.*, 2015). Similarly, Wamelink *et al.* (2014) found that species with limited geographic distribution had narrower habitat preferences in terms of soil parameters than common species. In this study, this was particularly important for low prevalence species such as *S. didymobotrya*, *S. punicea* and *P. x canescens*.

Generally, the central part of the country is intensely invaded by most of the species largely due to a near-optimal combination of the key driving factors: moderate temperatures and precipitation, less frost, high human population and disturbance and suitable topo-edaphic conditions. However, changes in land-use patterns, decreasing isolation of mountain areas (e.g. by increased tourism and human settlement expansion), introduction of ornamental and commercial species directly to the western (mountainous) parts of the country and climate change are expected to increase invasion in the very near future. The relatively broad bioclimatic niche of some of the species such as *A. mearnsii*, *L. camara*, *C. odorata*, *C.*

decapetala and *P. guajava* most likely reflects their significant plasticity whilst the present unfilling of the identified niche suggests that the potential for near-future spread of invasive species is very high.

Certainly, the enemy release hypothesis (Maron and Vilà, 2001) is conceivably the overarching explanation for the recent invasion increases in the country where most of the species do not have enemies. This is in addition to the possible effect of a general lack of shared evolutionary histories with the components of a country's ecosystems (Cox, 2004). The dominance of anthropogenic activities highlights the need for control activities to focus on key invasion pathways, particularly high propagule pressure areas such as human disturbed areas. Highly populated areas such as the Middleveld require urgent attention as shown in the various maps. Regulating the movement and trade in these species is a necessary intervention that should be integrated into the country's biosecurity policies and strategies. Riparian ecosystems appear to be highly vulnerable to invasion by most of the species over and above rivers being invasion pathways. Therefore, a catchment approach is recommended to ensure that all the species are controlled at once thereby minimizing control costs whilst maximizing benefits including the restoration of these ecosystems. The prediction maps provide useful decision-making tools for prioritization of interventions and formulation of control strategies.

5.5.2 Species distribution uncertainty

To prevent the over-interpretation of outputs it is important to recognize that uncertainty is a part of species distribution modelling and as a result, the data and model outputs may not provide all of the answers. Although studies which have used multiple statistical models to predict species distribution have identified areas of consistency or divergence by comparing distribution maps (Elith *et al.*, 2006; Beale and Lennon, 2012; Tessarolo *et al.*, 2014; Watling *et al.*, 2015), uncertainty maps are rarely provided (Gould *et al.*, 2014). One of the key advantages of BN representations is in simplifying and modelling conditional independencies, making decisions under uncertainty and explaining the outcome of stochastic processes (Milns *et al.*, 2010). Through the BNs, posterior probabilities were calculated for each species' occurrence given evidence from the predictor variables. Thus, when the uncertainties of predictive indicators were propagated across the network, the

models determined the distribution of the expected species occurrence with respect to these uncertainties. The BN approach is appropriate in that it provides for the specification of uncertainty in model components, as well as the predictions, and allows for the incorporation of heterogeneous datasets collected with varying degrees of accuracy, spatial and temporal scales (Uusitalo, 2007; Low Choy *et al.*, 2009; Aguilera *et al.*, 2011; Korb and Nicholson, 2011).

Even though the posterior probability maps are, by themselves, a representation of uncertainty, the PPCI maps further highlighted areas that were predicted to have suitable habitat even though the species has not yet been observed. The uncertainty maps are useful in situations where there is need to be conservative and careful in inference and decision-making (Keil, 2014). They also serve as a basis for development of new probability thresholding techniques such as deriving binary presence-absence maps based on the probabilistic maps (e.g. using Figure 4.58). Compared with the field data, both the posterior probability and certainty maps revealed more details at the spatial level and were indicative of the level of uncertainties and confidence regarding each species' predicted presence.

Model uncertainty was spatially structured for all 16 species where most low PPCI values were found at the edges of the recorded occurrences in agreement with the observations of Hanspach *et al.* (2011) and Watling *et al.* (2015). The high uncertainty areas are also those located near the boundaries of currently observed predictions areas. Such low PPCI areas represent parts of the country where the models could not predict with certainty that a species would occur with posterior probabilities closer to 0.5. This supports the assertion that data-driven models may show high uncertainties since many alien species have not reached equilibrium and the relationships between environmental conditions and the occurrence of the species are uncertain (Boets *et al.*, 2015). Hence, this uncertainty is caused by the fact that the model finds suitable environmental conditions for a species to occur but is absent in the data or there is insufficient information in the data due to under-representation of those species-environment associations. Watling *et al.* (2015) also found that the choice of the modelling algorithm was the greatest source of uncertainty, with some additional variation in performance attributed to the comprehensiveness of the species presences used for modelling.

Areas predicted with low certainty could also be attributed to low species detectability, which results in low frequency of observation (Hanspach *et al.*, 2011; Gould *et al.*, 2014). For species that may have low detectability or widely distributed species such as *M. azedarach* and *Opuntia* species, such uncertainties could highlight areas that require additional sampling effort or possibly the inclusion of additional explanatory variables. However, the outputs from this study showed that the degree of certainty was least dependent on species prevalence *per se*. The spatial patterning of uncertainty has been shown to indicate possible effects of spatial autocorrelation, especially at the boundaries of observed species distribution and in areas of relatively higher spatial heterogeneity where determining factor (such as climatic) gradients are steep (Naimi *et al.*, 2014).

Nevertheless, uncertainty was generally low in areas where species were actually observed suggesting that the sources of uncertainty may be largely attributed to species-specific niche requirements and the (in)ability of the models to depict that niche using the currently available data. However, the uncertainty maps may not be interpreted the same way as the traditional uncertainty ‘maps of ignorance’ (Rocchini *et al.*, 2011) which highlight areas where knowledge is lacking. In this case, the BNs effectively highlighted areas where each of the species has the potential to invade. The low certainty areas require effective monitoring to detect new incursions by the respective invading species.

Since the input datasets used in this study were derived from different sources and at different scales and resolutions, the uncertainty and accuracy in the predictions of the BN models is likely an indication of each BN model’s ability to represent the inherent uncertainty in the spatial relationships (Laskey *et al.*, 2010; Dlamini, 2011; Gould *et al.*, 2014). In BNs, the prediction errors are a result of accumulated epistemic uncertainty caused by data, model parameter, structural (including discretization) and technical uncertainty (Ascough *et al.*, 2008; Laskey *et al.*, 2010; Uusitalo *et al.*, 2015). Uncertainty in species distribution modelling is also associated with species characteristics, sampling design, and the completeness, accuracy, and resolution of the predictors used (Stohlgren *et al.*, 2010; Gould *et al.*, 2014; Watling *et al.*, 2015). Additionally, such uncertainty is attributed to input data, model misspecification, parameter uncertainty, equifinality, and model stochasticity (Dormann *et al.*, 2012; Gould *et al.*, 2014).

Therefore, when considering the PPCI values, the whole country is vulnerable to invasion by at least one of the species studied. As a rule of thumb, it is suggested that the areas of general low certainty ($PPCI < 0.5$) are considered as areas of potential establishment in the near future under current environmental conditions and should also be considered in risk management or planning possible interventions.

5.6 APPLICABILITY OF BAYESIAN NETWORK-BASED DATA MINING TO SPECIES DISTRIBUTION MODELLING PROBLEMS

The subject of biology is characterized by complex systems formed by networks of interacting variables, the identities of which are both unknown and of great interest to discover (Smith, 2010). Similarly, invasion science requires an understanding of complex ecological systems that have multi-faceted dynamics involving multiple drivers (Kueffer *et al.*, 2013). The objective of data mining is to extract knowledge and analyse large and complex data sets to find associations, extract structures, patterns and regularities (Lausch *et al.*, 2015). Modelling of spatial data to reveal intrinsic spatio-temporal patterns is of essence in ecology and is of particular theoretical significance for ecological knowledge discovery. The ever-increasing amount of environmental and ecological data from various sensors and sources presents an opportunity for ecology because such data has the potential to reveal previously unknown ecological knowledge on species distribution. This is very important in invasion science because most invading species are entering into novel environments with possibly new sets of interacting environmental forcings. Such data necessitates dynamic multidisciplinary research, which requires tools such as data mining and machine learning to help gain insight into associations and relationships between features on the Earth's surface. Unfortunately, such methods are least applied in the ecological sciences whilst they are maturing in other fields such as computer science and engineering (Lausch *et al.*, 2015).

Whilst data-driven BN application to molecular biology has advanced in recent times, applications in ecology are still at its infancy particularly in species distribution modelling. This is despite the fact that the ability of BNs to represent complex systems of relationships in a visually insightful and intuitive way has been recognised in many fields (Sierra and

Stephens, 2012). Dormann *et al.* (2012) identify the causality of detected correlations as a critical issue for the use of SDMs, where the input variables are often correlated among themselves. The BNs, when used in a data mining or machine learning framework, were able to detect and represent causal relationships and patterns making them useful in predicting the occurrence of invasive species using conditional probability distributions derived from field data. Besides elucidating species-environment relationships, the data-driven BNs graphically revealed species association and co-occurrence by showing dependency relationships between attributes in a database. Hence, the study was able to develop what might be termed “joint invasive species distribution models” (JiSDM) in line with the iSDM framework proposed by Uden *et al.* (2015).

Despite known limitations (see reviews by Uusitalo, 2007; Newton, 2009; Aguilera *et al.*, 2011; Johnson *et al.*, 2012a), the findings of this study highlight the efficacy of BNs as SDMs. The inherent non-linear and transient nature of heterogeneous and complex species-environment characteristics did not deter the BN models from graphically and probabilistically representing them. The BN models were able to discover the spatial patterns and probabilistic dependence relationships between the environmental data and the selected invasive alien plant species. Such interactions were represented in a qualitative manner, by means of the directed acyclic graphs, and in a quantitative manner through the conditional probability distributions for every variable represented in each BN. Most importantly, the study finds that BNs are a novel tool that can be used to understand and analyse the potential interactions between species as similarly suggested by Faisal *et al.* (2010), Milns *et al.* (2010) and Sierra and Stephens, (2012). The learned BNs in this study were able to reveal both unknown and known or explicable species-species and species-environment interactions, providing confidence for BN applications in ecology and in particular species distribution modelling. This is an important attribute that makes BNs a promising technique for joint distributions modelling wherein multiple species data are used to study species co-occurrences and biotic interactions (Kissling *et al.*, 2012; Clark *et al.*, 2013; Golding *et al.*, 2015; Thorson *et al.*, 2015).

Likewise, the study demonstrated that BNs are suited for integrative analysis of heterogeneous data, as they not only provide the means to model relations between

variables, but also to model relations between such data (Thomas and Sael, 2015). The ability to handle data from disparate sources and of different types, handling missing values, robustness to outliers, the ability to deal with irrelevant inputs (through the Markov blanket), the ability to extract non-linear combinations of features, as well as their predictive power and interpretability are the what makes BNs appealing over other species distribution modelling techniques. Missing data is often encountered in problems where there are limitations in the data gathering process such as with possible missing detection of species as well as in cases where there are hidden (latent) or unobserved variables in a system (Korb and Nicholson, 2011).

The findings show that BNs can achieve very good prediction accuracies whilst the posterior probabilities from the BN models are also an indicator of the uncertainty in making hard class (species presence or absence) allocations. This study has reaffirmed and demonstrated the distinguishing properties of BNs in being able to reduce the JPD of the model into a set of conditional probabilities and their capability to express model uncertainties, propagate information quickly and to represent complex topologies. These are all important areas of research in species distribution modelling. The BN structures developed in this study were observed to embody a good trade-off between the quality of the approximation of correlations and predictive power among attributes and the computational complexity in structure and parameter learning. The BN-based data mining techniques are a good means for estimating alien plant invasion risk since the probabilistic outputs could be used with cost coefficients of the possible impacts to obtain vulnerability indices. As new invaded areas are discovered, the BN models can be easily updated and validated with the new data to continually refine control strategies.

The use of properly learned and parameterized BNs within a spatially explicit environment also offers a more robust and intuitive solution for species distribution modelling and reasoning whilst explicitly and graphically revealing species-environment relationships and the associated uncertainty. Freedman and Humphreys (1999) concluded that the output produced by structure learning algorithms provides invaluable information about associations and the possibility of causal relations. However, it is important to mention that it may still be necessary, especially under data deficient situations, for researchers to use

expert knowledge to develop BN structures, examine the data and to utilize expert knowledge, taking into account circumstances and contexts additional to the raw data, to reach conclusions (Koski and Noble, 2012). Data driven and combined expert- and data-driven models have been found to perform better than purely expert or knowledge-based models (Alameddine *et al.*, 2011; Hamilton *et al.*, 2015; Meineri *et al.*, 2015; Boets *et al.*, 2015).

Marcot *et al.* (2006), Newton (2009), Uusitalo (2007), Aguilera *et al.* (2011) and Johnson *et al.* (2012a) provide reviews on the limitations of BNs but the main one is the requirement for acyclicity, hence disallowing feedback loops that would sometimes be beneficial in species distribution modelling. Building accurate models through optimal and efficient BN structure learning from data is often difficult due to the super-exponential number of possible graphs along with the requirement of acyclicity. Scalability issues and modelling of prior probability distributions for random variables are some of the other major challenges for the use of BNs (Thomas and Sael, 2015).

CHAPTER 6 : CONCLUSIONS AND RECOMMENDATIONS

6.1 CONCLUSIONS

The impacts of invasive organisms on natural ecosystems and human welfare are increasingly being documented worldwide. Their spread in Swaziland culminated in their declaration as a national disaster thereby requiring an urgent response. Control and management of these organisms requires an in-depth understanding of the underlying factors and processes governing their spatial distribution and spread. Species distribution models (SDMs) are used for this purpose albeit with differing degrees of usability and predictive performance. Most of the traditional SDMs have documented shortcomings including the inability to explicitly deal with uncertain and high dimensional data and their black box nature. Through the novel application of BN-based data mining of 170 spatial datasets and data on alien invasive plant distribution in Swaziland, this study has demonstrated that some of the limitations associated with conventional SDMs can be overcome to produce more robust species occurrence probability estimates through data mining of vast datasets. Through various BN learning algorithms in conjunction with robust data pre-processing and feature selection techniques, the models revealed the complexity of explanatory factors and their role in determining meso-scale invasive alien plant species distribution in Swaziland.

Although BN learning from data is currently least applied in the species distribution modelling domain, this study has demonstrated that this approach can achieve excellent performance as well as probabilistic and graphical explaining power even on unbalanced and multi-dimensional data from disparate sources. Such BN-based SDMs offer a common conceptual architecture where spatial and species distribution data can be expressed with a common and intuitive graphical formalism. The ability to graphically and geographically represent the patterns from field data, as well as the exploratory character of BNs, makes it easier to indicate the relationships between the explanatory variables. This not only improves the comprehensibility of the observed geographic patterns but also the identification of the practical usefulness and ecological relevance of these patterns. BN-based data mining or machine learning techniques have the potential to reveal new and

unexpected insights into species distribution by highlighting unusual links between predictor variables that other multivariate models may not show as clearly and explicitly. Hence, the learned BNs and the accompanying maps constitute hypotheses about each species' invasion dynamics.

The BN models produced testable and potentially useful predictions of the distributions of all the species, conditioned on the causal factors. In general, all the BNs performed well by producing parameters that matched better with field data although the TAN, GBNs, BANs and ICS algorithm had relatively higher predictive accuracy. This highlights the importance of including dependencies amongst the variables when modelling species distributions. When probabilistically evaluated, the ICS algorithm, which attempts to create a causal structure, frequently performed better. Hence, constraint-based algorithms are relatively more promising and require further investigation more so because of their approximation of human intuition of causality. At times, the naïve Bayes algorithm performed well and was the better performer overall in terms of computational complexity. The posterior probability and the associated uncertainty maps for the 16 species studied reveal that their potential niches or range extents are larger than their recently recorded distribution, although most recent invasions would occur near the recorded occurrence localities following key driving factor gradients. This indicates that most species are at the invasive stage of the invasion process owing to the longer residence times and other ecophysiological traits.

The data-driven BN models provided further insight into the factors that generate the observed spatial patterns and further elucidated the underlying ecological processes. A combination of high climatic and topographic heterogeneity coupled with intense anthropogenic activity results in complex species-environment relationships, which the BNs were able to uncover. It is interesting to note that for all the studied species, the BN models uncovered non-linear relationships of varying complexity showing an interplay of bioclimatic, topo-edaphic and anthropogenic factors along with co-occurrence patterns that indicate biotic interactions and shared niche requirements. The minimum temperature of the coldest month, temperature seasonality and number of frost days were found to be the most bioclimatic important determinants of alien plant invasion in the country. Land cover fragmentation, proximity to major roads, human population and settlement density and land

use were the frequently selected anthropogenic variables. The most important topo-edaphic variables were proximity to and density of rivers/streams, slope aspect and surface curvature.

The strong association of the invasive species plants with human-disturbed areas is apparent, as is the effect of propagule pressure. This implies that increasing human development and human activities such as agriculture, travel (including tourism), transportation and land use changes will increase propagule pressure and facilitate further invasions. Furthermore, rivers are an invasion pathway making riverine habitats more vulnerable to invasion. The findings indicate that each species' distribution is comprised of different sets and types of predictors that are highly interacting, and that a missed link between two variables can often be compensated for by other interactions. The elucidated biotic interactions are an important indicator of the usefulness of BNs as joint species distribution models. Hence, the term joint invasive species distribution models (JiSDM) is suggested for the derived models. The derived BNs provide very important ecological knowledge on these species, which have not been studied before in the country and at this level of spatial detail or scale. Most importantly, the learned BNs reveal both known (domain) and new knowledge on each species invasion ecology.

The combination of the driving factors results in the whole country being highly vulnerable to invasion by most of the species. The importance of the selected variables has implications on the invasion processes considering an increasing human population and a changing climate. This information is crucial for risk assessment and for providing guidance in early detection monitoring and control activities. Decision and policy-making to control the spread of invasive plants in Swaziland should take into account the findings of this study including the produced distribution and uncertainty maps following the recommendations provided in the following section.

6.2 RECOMMENDATIONS

Although this study focused on learning of BNs from data through identifying plausible factors influencing the occurrence of each species, the learning of the arcs still requires improvement. This points to the possible need to compare the machine learning approach to structure learning with expert defined constraints and arcs in order to generate more plausible structures and minimize the sensitivity of the learning algorithms. Similarly, a rigorous comparison with continuous Bayesian networks may be useful.

The process of undertaking this study uncovered the inadequacies of biodiversity monitoring in the country, specifically in relation to information needs. The data and information used was in heterogeneous forms and formats, and most seriously locked up in institutional and individual cupboards under the misconceptions of cost recovery and intellectual property. Although the 2009 national baseline data is available for reporting on the identity and distribution of invasive alien plants, the observations need to be repeated to produce trend information. There is, therefore, a need to continuously monitor the invasion of the species studied and other emerging species. This would require that the citizenry and volunteers are involved and data sharing and access is enhanced.

Hence, there is a need to conceive and establish National Biodiversity Information Infrastructure. Such an infrastructure should include museum or herbarium databases for taxonomic data and link to nationally supported institutional repositories (e.g. the proposed National Spatial Data Infrastructure – NSDI), as well as consortia of university and government and non-governmental institutions that can share operational costs. The increasing availability of massive high-resolution earth surface and other geospatial data (big data) presents an opportunity to explore complex geographical phenomena. Subsequently, the appropriate infrastructure is required to store, manage and analyse this ever-increasing amount of observational data. The voluminous data creates the daunting challenge of building a new generation of models and algorithms that can effectively accommodate these data-rich environments. Bringing together many novel ideas and techniques from ecological informatics and spatial data science is an absolute necessity.

Therefore, it is imperative that species distribution modelling takes advantage of developments in the computational sciences such as data science and/or machine learning. Furthermore, there is a proliferation of BN learning techniques that need to be explored especially those that are hybrid in nature, taking advantage of the strengths of new algorithms and multiple approaches. As such, the BN modelling approach needs to be further developed to take advantage of open-source frameworks such as R, Java and Python as well as big data frameworks that would facilitate parallel processing for analyses that are more complex and data intensive.

More patch scale studies are clearly needed to investigate the nature of biotic interactions including the role of frugivores and other seed dispersers, soils (especially because some of the species change soil properties) and population dynamics. This should include investigations on the role of species traits in determining the invasion success of some of the species. Some of these analyses could be preliminarily explored starting with the results and data from this study.

REFERENCES

Adriaenssens V, Goethals PLM, Charles J & De Pauw N 2004: Application of Bayesian belief networks for the prediction of macroinvertebrate taxa in rivers. *Annales de Limnologie - International Journal of Limnology*, 40(3), 181-191.

Afrogeo 2014: *Afrogeobase Version 3 (CD-ROM)*. Mbabane: Afrogeo/GeoSystems.

Aguilera PA, Fernández A, Reche F & Rumí R 2010: Hybrid Bayesian network classifiers: application to species distribution models. *Environmental Modeling & Software*, 25, 1630-1639.

Aguilera PA, Fernández A, Fernández R, Rumí R & Salmerón A 2011: Bayesian networks in environmental modelling. *Environmental Modelling & Software*, 26, 1376-1388.

Aguirre-Gutiérrez J, Carvalheiro LG, Polce C, van Loon EE, Raes N, Reemer M & Biesmeijer JC 2013: Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria – Dutch Hoverflies as a Case Study. *PLoS ONE*, 8(5), e63708, doi: <http://dx.doi.org/10.1371/journal.pone.0063708>.

Ahmed SE, McNerny G, O'Hara K, Harper R, Salido L, Emmott S & Joppa LN 2015: Scientists and software – surveying the species distribution modelling community. *Diversity and Distributions*, 21, 258-267.

Aitkenhead MJ & Aalders IH 2009: Predicting land cover using GIS, Bayesian and evolutionary algorithm methods. *Journal of Environmental Management*, 90, 236–250.

Alameddine I, Cha Y & Reckhow KH 2011: An evaluation of automated structure learning with Bayesian networks: An application to estuarine chlorophyll dynamics. *Environmental Modelling & Software*, 26, 163-172.

Aliferis CF, Statnikov A, Tsamardinos I, Mani S & Koutsoukos XD 2010: Local causal and Markov blanket induction for causal discovery and feature selection for classification -- Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11, 171-234.

Allan D, Yuan LL, Black P, Stockton T, Davies PE, Magierowski RH & Read SM 2012: Investigating the relationships between environmental stressors and stream condition using Bayesian belief networks. *Freshwater Biology*, 57, 58-73.

Allen JM, Leininger TJ, Hurd JD, Civco DL, Gelfand AE & Silander JA 2013: Socioeconomics drive woody invasive plant richness in New England, USA through forest fragmentation. *Landscape Ecology*, 28, 1671-1686.

Allouche O, Tsoar A & Kadmon R 2006: Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223-1232.

Altartouri A & Jolma A 2013: A Naive Bayes classifier for modeling distributions of the common reed in Southern Finland. In Chan F, Marinova D & Anderssen RS (eds.): *MODSIM2013, Proceedings of the 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1- 6 December 2013*, 1645-1651.

Ames DP & Anselmo A 2008: Bayesian Network Integration with GIS. In Shekhar S & Xiong H (eds.): *Encyclopedia of GIS*. Springer: New York, 39-45.

Amstrup SC, Marcot BG & Douglas DC 2008: A Bayesian network modeling approach to forecasting the 21st century world-wide status of polar bears. In DeWeaver ET, Bitz CM & Tremblay L-B (eds.): *Arctic Sea Ice Decline: Observations, Projections, Mechanisms, and Implications*. Geophysical Monograph Series, Vol. 180, American Geophysical Union, Washington, DC, USA, 213-268.

Amstrup SC, Deweaver E, Douglas DC, Marcot BC, Durner GM, Bitz CM & Bailey DA 2010: Greenhouse gas mitigation can reduce sea-ice loss and increase polar bear persistence. *Nature*, 468, 955-958.

Anderson RP 2012: Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences*, 1260, 66-80.

Aps R, Fetissof M, Herkül K, Kotta J, Leiger R, Mander Ü & Suursaar Ü 2009: Bayesian inference for predicting potential oil spill related ecological risk. In Guarascio M, Brebbia C & Garzia F (eds.): *Safety and Security Engineering III - WIT Transactions on the Built Environment (Vol. 108)*, Southampton: WIT Press, 149-159.

Araújo MB & Guisan A 2006: Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33, 1677-1688.

Araújo MB & New M 2007: Ensemble forecasting of species distributions. *Trends in Ecology and Evolution*, 22, 42-47.

Araújo MB & Peterson AT 2012: Uses and misuses of bioclimatic envelope modelling. *Ecology*, 93, 1527-1539.

Araújo MB & Rozenfeld A 2014: The geographic scaling of biotic interactions. *Ecography*, 37, 406-415.

Ascough II JC, Maier HR, Ravalico JK & Strudley MW 2008: Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling*, 219, 383-399.

Aspinall R 1992: An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. *International Journal of Geographical Information Systems*, 6(2), 105-121.

Aspinall R & Veitch N 1993: Habitat mapping from satellite imagery and wildlife survey data using a Bayesian modelling procedure in GIS. *Photogrammetric Engineering and Remote Sensing*, 59(4), 537-543.

Atzmanstorfer K, Oberthür T, Läderach P, O'Brien R, Collet L & Quiñonez G 2007: Probability modelling to reduce decision uncertainty in environmental niche identification and driving factor analysis: CaNaSTA case studies. In Zeil P & Kienberger S (eds.): *GeoInformation for development: Bridging the divide through partnerships*. Heidelberg: Wichmann, 33-43.

Auffret AG, Berg J & Cousins SAO 2014: The geography of human-mediated dispersal. *Diversity and Distributions*, 20(12), 1450-1456.

August-Schmidt EM, Haro G, Bontrager A & D'Antonio CM 2015: Preferential associations of invasive *Lantana camara* (Verbenaceae) in a seasonally dry Hawaiian woodland. *Pacific Science*, 69(3), 385-397.

Ayre KK, Caldwell CA, Stinson J & Landis WG 2014: Analysis of regional scale risk of whirling disease in populations of Colorado and Rio Grande cutthroat trout using a Bayesian belief network model. *Risk Analysis*, 34(9), 1589-1605.

Baldi P, Brunak S, Chauvin Y, Andersen CAF & Nielsen H 2000: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16, 412-424.

Ban SS, Pressey RL & Graham NAJ 2015: Assessing the Effectiveness of Local Management of Coral Reefs Using Expert Opinion and Spatial Bayesian Modeling. *PLoS ONE*, 10(8), e0135465. doi: <http://dx.doi.org/10.1371/journal.pone.0135465>.

Barbet-Massin M & Jetz W 2014: A 40-year, continent-wide, multi-species assessment of relevant climate predictors for species distribution modeling. *Diversity & Distributions*, 20, 1285-1295.

Barton DN, Kuikka S, Varis O, Uusitalo L, Henriksen HJ & Borsuk M, de la Hera A, Farmani R, Johnson S & Linnell JDC 2012: Bayesian Networks in Environmental and Resource Management. *Integrated Environmental Assessment and Management*, 8, 418-429.

Bascompte J 2009: Mutualistic networks. *Frontiers in Ecology and the Environment*, 7, 429–436.

Bashari H & Hemami M-R 2013: A predictive diagnostic model for wild sheep (*Ovis orientalis*) habitat suitability in Iran. *Journal for Nature Conservation*, 21, 319-325.

Bayes T 1764: An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45, 296-315.

Beale CM & Lennon JJ 2012: Incorporating uncertainty in predictive species distribution modelling. *Philosophical transactions of the Royal Society of London*, 367, 247-258.

Bean WT, Stafford R & Brashares JS 2012: The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, 35, 250–258.

Beauséjour R, Handa IT, Lechowicz MJ, Gilbert B & Vellend M 2015: Historical anthropogenic disturbances influence patterns of non-native earthworm and plant invasions in a temperate primary forest. *Biological Invasions*, 17, 1267-1281.

Beauregard F & de Blois S 2014: Beyond a climate-centric view of plant distribution: edaphic variables add value to distribution models. *PLoS ONE*, 9(3), e92642, doi: <http://dx.doi.org/10.1371/journal.pone.0092642>.

Bermejo P, de la Ossa L, Gámez JA & Puerta JM 2012. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems*, 25(1), 35-44.

Bhattacharya M 2013: Machine Learning for Bioclimatic Modelling. *International Journal of Advanced Computer Science and Applications*, 4, 1-8.

Bielza C & Larrañaga P 2014: Discrete Bayesian network classifiers: A survey. *ACM Computing Surveys*, 47(1), Article 60 (April 2014), doi: <http://dx.doi.org/10.1145/2576868>.

Bigirimana J, Bogaert J, De Cannière C, Bigendako M-J & Parmentiere I 2012: Domestic garden plant diversity in Bujumbura, Burundi: Role of the socio-economical status of the neighborhood and alien species invasion risk. *Landscape and Urban Planning*, 107, 118-126.

Binggeli P 1999: *Chromolaena odorata* (L.) King & Robinson (Asteraceae). (Online) Available from: (<http://www.fs.fed.us/global/iitf/pdf/shrubs/Chromolaena%odoratum.pdf>) (Accessed 7 April 2015).

Bisht RP & Toky OP 1993: Growth pattern and architectural analysis of nine important multipurpose trees in an arid region of India. *Canadian Journal of Forest Research*, 23(4), 722-730.

Bleys JA, Mazibuko WKM & Allen JA 1982: The Silviculture of Indigenous and Exotic Trees other than Pines and Eucalypts in Swaziland. *South African Forestry Journal*, 121(1), 24-27.

- Boets P, Landuyt D, Everaert G, Broekx S & Goethals PLM 2015: Evaluation and comparison of data-driven and knowledge-supported Bayesian Belief Networks to assess the habitat suitability for alien macroinvertebrates. *Environmental Modelling & Software*, 74, 92-103.
- Bolstad P 2012: *GIS Fundamentals* (3rd Ed.) Atlas Books.
- Borsuk ME, Reichert P, Peter A, Schager E & Burkhardt-Holm P 2006: Assessing the decline of brown trout (*Salmo trutta*) in Swiss rivers using a Bayesian probability network. *Ecological Modelling*, 192(1-2), 224-244.
- Borsuk ME, Stow CA & Reckhow KH 2004: A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. *Ecological Modelling*, 173, 219-239.
- Bouckaert RR 1995: *Bayesian Belief Networks: From Construction to Inference*. PhD Dissertation. Universiteit Utrecht.
- Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A & Scuse D 2014: *WEKA Manual for Version 3-7-12*. Hamilton: University of Waikato.
- Boulangeat I, Gravel D & Thuiller W 2012: Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, 15, 584-593.
- Braun KP & Dlamini SD 2005: *A database and publication of an atlas for plant alien invasive species (AIS) in Swaziland*. Douglas Consulting (Pty) Ltd., Mbabane.
- Buang N, Liu N, Caelli T, Lesslie R & Hill MJ 2006: Discover Knowledge From Distribution Maps Using Bayesian Networks. In Peter C, Kennedy PJ, Li J, Simoff SJ & Williams GJ (eds.): *Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006)*, Sydney, Australia. CRPIT, 69-74.
- Buddenhagen C & Jewell KJ 2006: Invasive plant seed viability after processing by some endemic Galapagos birds. *Ornitologica Neotropical*, 17, 73-80.
- Buntine WL 1996: A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8, 195-210.

Burkhardt-Holm P 2008: Decline of brown trout (*Salmo trutta*) in Switzerland - How to assess potential causes in a multi-factorial cause-effect relationship? *Marine Environmental Research*, 66, 181–182.

Butchart SH, Walpole M, Collen B, van Strien A, Scharlemann JP, Almond RE, Baillie JE, Bomhard B, Brown C, Bruno J, Carpenter KE, Carr GM, Chanson J, Chenery AM, Csirke J, Davidson NC, Dentener F, Foster M, Galli A, Galloway JN, Genovesi P, Gregory RD, Hockings M, Kapos V, Lamarque JF, Leverington F, Loh J, McGeoch MA, McRae L, Minasyan A, Hernández Morcillo M, Oldfield TE, Pauly D, Quader S, Revenga C, Sauer JR, Skolnik B, Spear D, Stanwell-Smith D, Stuart SN, Symes A, Tierney M, Tyrrell TD, Vié JC & Watson R 2010: Global biodiversity: indicators of recent declines. *Science*, 328, 1164–1168.

Callaway RM & Aschehoug ET 2000: Invasive plants versus their new and old neighbors: a mechanism for exotic invasion. *Science*, 290, 521-523.

Callaway RM & Ridenour WM 2004: Novel weapons: invasive success and the evolution of increased competitive ability. *Frontiers in Ecology and the Environment*, 2, 436-443.

Castets M, Degenne P, Poncelet P & Lo Seen D 2014: Integrating raster and vector spatial representations with interaction graphs for multi-scale environmental simulations. In Ames DP, Quinn NWT & Rizzoli AE (eds.): *Proceedings of the 7th International Congress on Environmental Modelling and Software*, June 15-19, San Diego, California, USA.

Central Statistics Office 2011: *Poverty in a Decade of Slow Economic Growth: Swaziland in the 2000s*. Mbabane: Central Statistical Office.

Chapman AD 2005: *Uses of primary species-occurrence data, version 1.0*. Copenhagen: Global Biodiversity Information Facility.

Chan T, Hart B, Kennard M, Pusey B, Shenton W, Douglas M, Valentine E & Patel S 2012: Bayesian network models for environmental flow decision making in the Daly River, Northern Territory, Australia. *River Research and Applications*, 28(3), 1-19.

Chen SH & Pollino CA 2012: Good practice in Bayesian network modelling. *Environmental Modelling & Software*, 37, 134-145.

Cheng J, Bell DA & Liu W 1997. An algorithm for Bayesian belief network construction from data. In Smyth P & Madigan D (eds.): *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, January 4-7, 1997, Fort Lauderdale, USA, 83-90.

Chickering D 1996: Learning Bayesian Networks is NP Complete. In Fisher D & Lenz HJ (eds.): *Learning from Data: Artificial Intelligence and Statistics V.*, Springer-Verlag: New York, 121-130.

Chickering D Meek C & Heckerman D. 2003: Large-sample learning of Bayesian networks is NP-hard. In *Proceedings of the 19th annual conference on uncertainty in artificial intelligence (UAI-03)*, San Francisco, CA: Morgan Kaufmann Publishers, 124-133.

Chow CK & Liu CN 1968: Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14, 426-467.

Chytrý M, Pyšek P, Wild J, Pino J, Maskell LC & Vilà M 2009: European map of alien plant invasions based on the quantitative assessment across habitats. *Diversity and Distributions*, 15, 98-107.

Clark JS, LaDeau S & Ibanez I 2004: Fecundity of trees and the colonization–competition hypothesis. *Ecological Monographs*, 74, 415-442.

Clark JS, Gelfand AE, Woodall CW & Zhu K 2013: More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24, 990-999.

Cooper GF & Herskovits E 1992: A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 309–347.

Copps SL, Yoklavich MM, Parkes GB, Wakefield WW, Bailey A, Greene KG, Goldfinger C & Burn RW 2007: Applying marine habitat data to fishery management on the U.S. west coast: Initiating a policy-science feedback loop, In Todd BJ & Greene KG (eds.): *Mapping the Seafloor for Habitat Characterization: Geological Association of Canada*, Special Paper 47, 451-462.

Cowell RG, Dawid AP & Spiegelhalter DJ 1993: Sequential model criticism in probabilistic expert systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 209-219.

Cowell RG, Dawid AP, Lauritzen SL & Spiegelhalter DJ 2007: *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer: New York.

Cox GW 2004: *Alien Species and Evolution: The Evolutionary Ecology of Exotic Plants, Animals, Microbes, and Interacting Native Species*. Covello, California: Island Press.

Crall AW, Jarnevich CS, Young NE, Panke BJ, Renz M & Stohlgren TJ 2015: Citizen science contributes to our knowledge of invasive plant species distributions. *Biological Invasions*, 17, 2415-2427.

Crowl TA, Crist TO, Parmenter RR, Belovsky G & Lugo AE 2008: The spread of invasive species and infectious diseases as drivers of ecosystem change. *Frontiers in Ecology and the Environment*, 6(5), 238-246.

Cuddington K & Hastings A 2004: Invasive engineers. *Ecological Modelling*, 178, 335-347.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J & Lawler JJ 2007: Random forests for classification in ecology. *Ecology*, 88, 2783-2792.

Daly R, Shen Q & Aitken S 2011: Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review*, 26, 99-157.

Danielsen F, Burgess ND & Balmford A 2005: Monitoring matters: examining the potential of locally-based approaches. *Biodiversity and Conservation*, 14, 2507-2542.

Darwiche A 2002: A logical approach to factoring belief networks. In Fensel D, Giunchiglia F, McGuinness DL & Williams M-A (eds.): *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR-02)*, Morgan Kaufmann: San Francisco, 409-420.

Darwiche A 2009: *Modeling and Reasoning with Bayesian Networks*. New York: Cambridge University Press.

Davis J & Goadrich M 2006: The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, 18-22 September 2006, Pittsburgh, PA, 233-240.

De Beer H 1987: Mauritius Thorn. *Farming in South Africa*, Weeds/A.22.

de Laplace PS 1820: *A Philosophical Essay on Probabilities*. New York: Dover Publications.

De O.Silva AC, Mello MP & Fonseca LMG 2014: *Enhancements to the Bayesian Network for Raster Data* (BayNeRD). In Davis Jr CA & Ferreira KR (eds.): *Proceedings of the 15th Brazilian Symposium on Geoinformatics (GeoInfo 2014)*. 30 November - 3 December 2014, Campos do Jordão, Sao Paulo, Brazil.

Dean SJ, Holmes PM & Weiss PW 1986: Seed biology of invasive alien plants in South Africa and South West Africa/Namibia. In MacDonald IAW, Kruger FJ & Ferrar AA (eds.): *The ecology and management of biological invasions in southern Africa*. Cape Town: Oxford University Press.

Dean WRJ & Milton SJ 2000: Directed dispersal of *Opuntia* species in the Karoo South Africa: are crows the responsible agents? *Journal of Arid Environments*, 45, 305-314.

Dickinson J, Zuckerberg B & Bonter DN 2010: Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41,149-172.

Diekmann M, Michaelis J & Pannek A 2015: Know your limits - the need for better data on species responses to soil variables. *Basic and Applied Ecology*, 16(7), 563-572.

Ding Z 2011: *Diversified Ensemble Classifier for Highly imbalanced Data Learning and their application in Bioinformatics*. Ph. D Thesis. Georgia State University: Department of Computer Science.

Dlamini WM 2011: A data mining approach to predictive vegetation mapping using probabilistic graphical models. *Ecological Informatics*, 6, 111–124.

Dlamini WM 2015: *Vulnerability and Adaptation Assessment: Biodiversity and ecosystems*. Mbabane: Ministry of Tourism and Environmental Affairs.

Dobson L & Lotter M 2004. Vegetation Map of Swaziland. In: Mucina L & Rutherford MC (eds.): *Vegetation Map of South Africa, Lesotho and Swaziland*: Shapefiles of basic mapping units. Beta version 4.0, February 2004, National Botanical Institute, Cape Town.

Domisch S, Kuemmelen M, Jähnig SC & Haase P 2013: Choice of study area and predictors affect habitat suitability projections, but not the performance of species distribution models of stream biota. *Ecological Modelling*, 257, 1-10.

Domingos P 1999: Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 155-164, ACM Press.

Dormann CF & Roxburgh SH 2005. Experimental evidence rejects pairwise modelling approach to coexistence in plant communities. *Proceedings of the Royal Society of London Series B*, 272, 1279-1285.

Dormann CF, Schymanski SJ, Cabral J, Chuine I, Graham C, Hartig F, Kearney M, Morin X, Römermann C, Schröder B & Singer A 2012: Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, 39, 2119-2131.

Douglas SJ & Newton AC 2014: Evaluation of Bayesian networks for modelling habitat suitability and management of a protected area. *Journal for Nature Conservation*, 22, 235-246.

Doran JC & Turnbull JW 1997. *Australian trees and shrubs: species for land rehabilitation and farm planting in the tropics*. ACIAR Monograph No. 24, Australian Centre for International Agricultural Research, Canberra.

Dormann CF, Schymanski SJ, Cabral J, Chuine I, Graham CH, Hartig F, Kearney M, Morin X, Römermann C, Schröder B & Singer A 2012: Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, 39, 2119-2131.

Doveton DM 1937: The Human Geography of Swaziland. *Transactions (Institute of British Geographers)*, 7/8, xi-xvi+1-110.

Drummond C & Holte RC 2003: 4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. *Workshop on Learning from Imbalanced Data sets II*, ICML, Washington DC, 2003, 1-8.

du Plessis M & Kotze H 2011: Growth and yield models for *Eucalyptus grandis* grown in Swaziland. *Southern Forests: a Journal of Forest Science*, 73(2), 81-89.

- Dubuis A, Giovanettina S, Pellissier L, Pottier J, Vittoz P & Guisan A 2013: Improving the prediction of plant species distribution and community composition by adding edaphic to topoclimatic variables. *Journal of Vegetation Science*, 24, 593–606.
- Duda R & Hart P 1973: *Pattern classification and scene analysis*. New York: Wiley.
- Duke JA 1983: *Handbook of Energy Crops*. Unpublished.
- Eberhart R & Kennedy J 1995: A New Optimizer Using Particle Swarm Theory. In: *Proceedings of the 6th International Symposium on Micro Machine and Human Science*, Nagoya, Japan, October 1995, IEEE Press: Piscataway, 39-43.
- Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JMcC, Peterson AT, Phillips SJ, Richardson KS, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS & Zimmermann NE 2006: Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129-151.
- Elith J & Leathwick JR 2009: Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 677-697.
- Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE & Yates CJ 2011: A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17, 43-57.
- Elton CS 1958: *The Ecology of Invasions by Animals and Plants*. Chicago: University of Chicago Press.
- Environmental Systems Research Institute 2011: *ArcGIS Desktop: Release 10.2*. Redlands, CA: Environmental Systems Research Institute.
- Erasmus DJ 1984: Bramble. *Farming in South Africa*, Weeds/A.3.
- Evans J 1974: Some aspects of the growth of *Pinus patula* in Swaziland. *The Commonwealth Forestry Review*, 53(1), 57-62.

Evans J & Cushman S 2009: Gradient modeling of conifer species using random forests. *Landscape Ecology*, 24, 673-683.

Faisal A, Dondelinger F, Husmeier D & Beale CM 2010: Inferring species interaction networks from species abundance data: a comparative evaluation of various statistical and machine learning methods. *Ecological Informatics*, 5, 451-464.

Falke JA, Flitcroft RL, Dunham JB, McNyset KM, Hessburg PF & Reeves GH 2015: Climate change and vulnerability of bull trout (*Salvelinus confluentus*) in a fire-prone landscape. *Canadian Journal of Fisheries and Aquatic Sciences*, 72(2), 304-318.

Fayyad U & Irani K 1993: Multiple-interval discretization of continuous-valued attributes for classification learning. In: *Thirteenth International Joint Conference on Artificial Intelligence*. Kaufmann: San Mateo, 1022-1027.

Fayyad U, Piatetsky-Shapiro G & Smyth P 1996: From data mining to knowledge discovery in databases. *AI Magazine*, Fall 1996, 37-54.

Fernandes JA, Irigoien X, Goikoetxea N, Lozano JA, Inza I, Pérez A & Bode A 2010: Fish recruitment prediction, using robust supervised classification methods. *Ecological Modelling*, 221, 338–352.

Fienen MN & Plant NG 2015: A cross-validation package driving Netica with Python. *Environmental Modelling & Software*, 63, 14-23.

Flores J, Gámez JA & Martínez AM 2012: Supervised classification with Bayesian networks: A review on models and applications. In *Intelligent Data Analysis for Real World Applications. Theory and Practice*. IGI Global, 72–102.

Forsyth GG, Richardson DM, Brown PJ, & van Wilgen BW 2004: A rapid assessment of the invasive status of Eucalyptus species in two South African provinces. *South African Journal of Science*, 100, 75-77.

Foxcroft LC, Rouget M, Richardson DM & MacFadyen S 2004: Reconstructing 50 years of *Opuntia stricta* invasion in the Kruger National Park, South Africa: environmental determinants and propagule pressure. *Diversity and Distributions*, 10, 427-437.

- Foxcroft LC, Parsons M, McLoughlin CA & Richardson DM 2008: Patterns of alien plant distribution in a river landscape following an extreme flood. *South African Journal of Botany*, 74, 463-475.
- Foxcroft LC, Richardson DM, Rejmánek M & Pysěk P 2010: Alien plant invasions in tropical and sub-tropical savannas: patterns, processes and prospects. *Biological Invasions*, 12, 3913-3933.
- Franklin J, Wejnert KE, Hathaway SA, Rochester CJ & Fisher RN 2009: Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. *Diversity and Distribution*, 15, 167-177.
- Franklin J 2010: *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge: Cambridge University Press.
- Freedman D & Humphreys P 1999: Are there algorithms that discover causal structure? *Synthese*, 121, 29-54.
- Freeman EA, Moisen GG & Frescino TS 2012: Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in Random Forest models of tree species distributions in Nevada. *Ecological Modelling*, 233, 1-10.
- Friedman N, Geiger D, Goldszmidt M 1997: Bayesian network classifiers. *Machine Learning*, 29, 131-163.
- Friedman N, Linial M, Nachman I, Pe'er D 2000: Using Bayesian Networks to Analyse Expression Data. *Journal of Computational Biology*, 7, 601 – 620.
- Fu B, Pollino CA, Cuddy SM & Andrews F 2015: Assessing climate change impacts on wetlands in a flow regulated catchment: A case study in the Macquarie Marshes, Australia. *Journal of Environmental Management*, 157, 127-138.
- Fuentes N, Pauchard A, Sánchez P, Esquivel J & Marticorena A 2013: A new comprehensive database of alien plant species in Chile based on herbarium records. *Biological Invasions*, 15(4), 847-858.

Fuentes N, Saldaña A, Kühn I & Klotz S 2015: Climatic and socioeconomic factors determine the level of invasion by alien plants in Chile. *Plant Ecology & Diversity*, 8(3), 371-377.

Gallien L, Münkemüller T, Albert CH, Boulangeat I & Thuiller W 2010: Predicting potential distributions of invasive species: where to go from here? *Diversity and Distributions*, 16, 331-342.

Gallien L, Douzet R, Pratte S, Zimmermann N & Thuiller W 2012: Invasive species distribution model - How violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography*, 21, 1126-1136.

García S, Luengo J, Sáez JA, López V & Herrera F 2013. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 734-750.

García-Callejas D & Araújo MB 2016: The effects of model and data complexity on predictions from species distributions models. *Ecological Modelling*, 326, 4-12.

Gavier-Pizarro GI, Stewart SI, Huebner C, Keuler NS & Radeloff VC 2010: Housing is positively associated with invasive exotic plant species richness in New England, USA. *Ecological Applications*, 20(7), 1913–1925.

Gawne B, Price A, Koehn JD, King AJ, Nielsen DL, Meredith S, Beesley L & Vilizzi L 2012: A Bayesian Belief Network Decision Support Tool for Watering Wetlands to Maximise Native Fish Outcomes. *Wetlands*, 32, 277-287.

Gelfand AE, Silander JA, Wu S, Latimer A, Lewis PO, Rebelo AG & Holder M 2003: Explaining Species Distribution Patterns through Hierarchical Modeling. *Bayesian Analysis*, 1, 1-35.

Gibbs MT 2007: Assessing the risk of an aquaculture development on shorebirds using a Bayesian belief model. *Human and Ecological Risk Assessment: An International Journal*, 13(1), 156-179.

Gibert K, Spate J, Sánchez-Marré M, Athanasiadis I & Comas J 2008: Data Mining for Environmental Systems. In Jakeman A, Voinov A, Rizzoli AE & Chen S (eds.): *Environmental*

Modelling, Software and Decision Support: State of the art and new perspective, 205-228, Elsevier.

Gieder KD, Karpantya SM, Fraser JD, Catlin DH, Gutierrez BT, Plant NG, Turecek AM & Thieler ER 2014: A Bayesian network approach to predicting nest presence of the federally-threatened piping plover (*Charadrius melodus*) using barrier island features. *Ecological Modelling*, 276, 38-50.

Giglio L, Loboda T, Roy DP, Quayle B & Justice CO 2009: An active-fire based burned area mapping algorithm for the MODIS sensor. *Remote Sensing of Environment*, 113, 408-420.

Giretti A, Carbonari A & Naticchia B 2012: A Spatio-temporal Bayesian Network for Adaptive Risk Management in Territorial Emergency Response Operations. In: Premchaiswadi W (ed.). *Bayesian Networks*. InTech, 49-70, Available from <http://www.intechopen.com/books/bayesian-networks/a-spatio-temporal-bayesian-network-for-adaptive-risk-management-in-territorial-emergency-response-ops>. (Accessed 17 June 2014).

Glover F 1989: Tabu search - part i. *ORSA Journal on Computing*, 1, 190-192.

Golding N, Nunn MA & Purse BV 2015: Identifying biotic interactions which drive the spatial distribution of a mosquito community. *Parasites & Vectors*, 8,367, doi: <http://dx.doi.org/10.1186/s13071-015-0915-1>.

Gomes, VGN, Quirino, ZGM & Araujo, HFP 2014: Frugivory and seed dispersal by birds in *Cereus jamacaru* DC. ssp. *jamacaru* (Cactaceae) in the Caatinga of Northeastern Brazil. *Brazilian Journal of Biology*, 74(1), 32-40.

González-Salazar C, Stephens CR & Marquet PA 2013: Comparing the relative contributions of biotic and abiotic factors as mediators of species' distributions. *Ecological Modelling*, 248, 57-70.

Goodall JM, Zimmermann HG & Zeller D 1994: *The distribution of Chromolaena odorata in Swaziland and implications for further spread*. Pretoria: Plant Protection Research Institute (Agricultural Research Council).

- Goodall JM & Erasmus DJ 1996: Review of the status and integrated control of the invasive alien weed, *Chromolaena odorata*, in South Africa. *Agriculture, Ecosystems and Environment*, 56, 151-164.
- Goodchild MF, Yuan M & Cova TJ 2007: Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21, 239-260.
- Goudarzi F, Hemami MR, Bashari H & Johnson S 2015: Assessing translocation success of the endangered Persian fallow deer using a Bayesian Belief Network. *Ecosphere*, 6(10), 1-14.
- Gould SF, Beeton NJ, Harris RMB, Hutchinson MF, Lechner AM, Porfirio LL & Mackey BG 2014: A tool for simulating and communicating uncertainty when modelling species distributions under future climates. *Ecology and Evolution*, 4(24), 4798-4811.
- Grech A & Coles RG 2010: An ecosystem-scale predictive model of coastal seagrass distribution. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 20, 437-444.
- Grenouillet G, Buisson L, Casajus N & Lek S 2011: Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, 34, 9-17.
- Grotkopp E, Erskine-Ogden J & Rejmánek M 2010: Assessing potential invasiveness of woody horticultural plant species using seedling growth rate traits. *Journal of Applied Ecology*, 47, 1320-1328.
- Guisan A & Thuiller W 2005: Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8, 993-1009.
- Guyon I & Elisseeff A 2003: An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
- Haas TC 1991: A Bayesian Belief Network advisory system for aspen regeneration. *Forest Science*, 37(2), 627-654.
- Hall MA 1998: *Correlation-based Feature Subset Selection for Machine Learning*. PhD Thesis, Waikato: University of Waikato.

- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P & Witten IH 2009: The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 10-18.
- Hamilton GS, Fielding F, Chiffings AW, Hart BT, Johnstone RW & Mengersen KR 2007: Investigating the Use of a Bayesian Network to Model the Risk of *Lyngbya majuscula* Bloom Initiation in Deception Bay, Queensland, Australia. *Human and Ecological Risk Assessment*, 13(6), 1271-1287.
- Hamilton G, McVinish R & Mengersen K 2009: Bayesian model averaging for harmful algal bloom prediction. *Ecological Applications*, 19, 1805–1814.
- Hamilton SH, Pollino CA & Jakeman AJ 2015: Habitat suitability modelling of rare species using Bayesian networks: Model evaluation under limited data. *Ecological Modelling*, 299, 64-78.
- Han J & Kamber M 2006: *Data Mining: Concepts and Techniques* (2nd Ed). San Francisco: Morgan Kaufmann.
- Hanberry BB, He HS & Dey DC 2012: Sample sizes and model comparison metrics for species distribution models. *Ecological Modelling*, 227, 29-33.
- Hand D 1997: *Construction and Assessment of Classification Rules*. New York: Wiley.
- Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A, Egorov A, Chini L, Justice CO & Townshend JRG 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160), 850-853.
- Hanspach J, Kühn I, Pompe S & Klotz S 2010: Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics*, 12, 219–225.
- Hanspach J, Kühn I, Schweiger O, Pompe S & Klotz S 2011: Geographical patterns in prediction errors of species distribution models. *Global Ecology and Biogeography*, 20, 779-788.

- Harrison S 1997: How natural habitat patchiness affects the distribution of diversity in Californian serpentine chaparral. *Ecology*, 78, 1898-1906.
- Hastie T, Tibshirani R & Friedman J 2009: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics)* (2nd Ed). Berlin: Springer.
- Heckerman D, Geiger D & Chickering DM 1995: Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20, 197-243.
- Heckerman D 1997: Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1, 79-119.
- Heckerman, D. 2007. A Bayesian approach to learning causal networks. In Edwards W & Miles RF Jr (eds.): *Advances in Decision Analysis: from Foundations to Applications*, Cambridge University Press, 202-220.
- Hegel TM, Cushman SA, Evans J & Huettmann F 2010: Current state of the art for statistical modelling of species distributions. In Cushman S & Huettmann F (eds.): *Spatial Complexity, Informatics, and Wildlife Conservation*. Springer: Tokyo, 273-311.
- Heikkinen RK, Luoto M, Virkkala R, Pearson RG & Körber J-H 2007: Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, 16, 754-763.
- Helle I, Lecklin T, Jolma A & Kuikka S 2011: Modeling the effectiveness of oil combating from an ecological perspective – A Bayesian network for the Gulf of Finland; the Baltic Sea. *Journal of Hazardous Materials*, 185, 182-192.
- Henderson L 1986: Barrier plants in South Africa. *Bothalia*, 14(3&4), 635-639.
- Henderson L 1989: Invasive alien woody plants of Natal and the north-eastern Orange Free State. *Bothalia*, 19(2), 237-261.
- Henderson L & Wells MJ 1986: Alien plant invasions in the grassland and savanna biomes. In: Macdonald IAW, Kruger FJ & Ferrar AA (eds.): *The ecology and management of biological invasions in southern Africa*. Cape Town: Oxford University Press, 109-117.

Henderson L 2001: *Alien Weeds and Invasive Plants*. Plant Protection Research Institute Handbook No. 12. Cape Town: Paarl Printers.

Henderson L 2006: Comparisons of invasive plants in southern Africa originating from southern temperate, northern temperate and tropical regions. *Bothalia*, 36, 201-222.

Henderson L 2007: Invasive, naturalized and casual alien plants in southern Africa: a summary based on the Southern African Plant Invaders Atlas (SAPIA). *Bothalia*, 37, 215-248.

Henriksen HJ & Barlebo HC 2008: Reflections on the use of Bayesian belief networks for adaptive management. *Journal of Environmental Management*, 88, 1025-1036.

Higgins SI, O'Hara RB & Röermann C 2012: A niche for biology in species distribution models. *Journal of Biogeography*, 39, 2091-2095.

Hijmans RJ, Cameron SE, Parra JL, Jones PG & Jarvis A 2005: Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965-1978.

Hill MO 1973: Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54, 427-432.

Hinze WHF 1985: *Solanum mauritianum Scop. Bugweed, luisboom*. Plant Invaders, Pamphlet 365/1. Directorate Forest Management, South Africa.

Hirzel AH & Le Lay G 2008: Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, 45, 1372-1381.

Hochachka WM, Caruana R, Munson A, Riedewald M, Sorokina D & Kelling S 2007: Data-Mining Discovery of Pattern and Process in Ecological Systems. *Journal of Wildlife Management*, 71, 2427-2437.

Hochachka WM, Fink D, Hutchinson RA, Sheldon D, Wong W-K & Kelling S 2012: Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, 27, 130-137.

- Hoffmann JH & Moran VC, 1991: Biological control of *Sesbania punicea* (Fabaceae) in South Africa. *Agriculture, Ecosystems and the Environment*, 37, 157-173.
- Holland A, Fathi M, Abramovici M & Neubach M 2008: Competing Fusion for Bayesian Applications. In *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2008)*, June 2008, Malaga, Spain, 378-385.
- Hortal J, Lobo JM & Jiménez-Valverde A 2012: Basic questions in biogeography and the (lack of) simplicity of species distributions: Putting species distribution models in the right place. *Natureza & Conservação*, 10, 108-118.
- Howes AL, Maron M & McAlpine CA 2010: Bayesian networks and adaptive management of wildlife habitat. *Conservation Biology*, 24(4), 974-983.
- Huang J & Yuan Y 2007: Construction and Application of Bayesian Network Model for Spatial Data Mining. In *Proceedings of IEEE International Conference on Control and Automation, Guangzhou, China, 30 May 30 to 1 June 2007*, 2802-2805.
- Huang M, Guo L, Gong J & Yang W 2013. Bayesian Network and Factor Analysis for Modeling Pine Wilt Disease Prevalence. *Journal of Software Engineering and Applications*, 6(3B), 13-17.
- Hugin 2014: *Hugin software*. Alborg, Denmark: Hugin Expert A/S.
- Hui C, Richardson DM, Pyšek P, Le Roux JJ, Kučera T & Jarošík V 2013: Increasing functional modularity with residence time in co-distribution of native and introduced vascular plants. *Nature Communications*, 4, 2454. <http://dx.doi.org/10.1038/ncomms3454>.
- Hunter J & Platenkamp GAJ 2003: The hunt for red sesbania: biology, spread and prospects for control. *California Exotic Pest Plant Council Newsletter*, 11(2), 4-6.
- ISRIC – World Soil Information 2013: Soil property maps of Africa at 1km. (Online) Available from <http://www.isric.org> (Accessed 17 March 2014).

- Japkowicz N & Stephen S 2002: The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6, 429-449.
- Jarnevich CS, Stohlgren TJ, Kumar S, Morisette JT & Holcombe TR 2015: Caveats for correlative species distribution modeling. *Ecological Informatics*, 29(1), 6-15.
- Jarvis A, Reuter HI, Nelson A & Guevara E 2008. *Hole-filled SRTM for the globe Version 4 (online)*. Available from the CGIAR-CSI SRTM 90m Database. Available from: <http://srtm.csi.cgiar.org> (Accessed on: 7 February 2014).
- Jay CV, Marcot BG & Douglas DC 2011: Projected status of the Pacific walrus (*Odobenus rosmarus divergens*) in the twenty-first century. *Polar Biology*, 34, 1065-1084.
- Jellinek S, Rumpff L, Driscoll DA, Parris KM & Wintle BA 2014: Modelling the benefits of habitat restoration in socio-ecological systems. *Biological Conservation*, 169, 60-67.
- Jensen FV & Nielsen TD 2007: *Bayesian networks and decision graphs. Information Science and Statistics* (2nd Ed.). New York: Springer.
- Jensen FV, Olesen KG & Andersen SK 1990: An algebra of Bayesian belief universes for knowledge based systems. *Networks*, 20, 637-659.
- Jeschke JM 2014: General hypotheses in invasion ecology. *Diversity and Distributions*, 20, 1229-1234.
- Jeschke JM, Gómez Aparicio L, Haider S, Heger T, Lortie CJ, Pyšek P & Strayer DL 2012: Support for major hypotheses in invasion biology is uneven and declining. *NeoBiota*, 14, 1-20.
- Jiménez-Valverde A, Peterson AT, Soberón J, Overton J, Aragón P & Lobo J 2011: Use of niche models in invasive species risk assessments. *Biological Invasions*, 13, 2785-2797.
- Johnson RA, Chawla NV & Hellmann JJ 2012a: Species distribution modeling and prediction: A class imbalance problem. In *Proceedings of Conference on Intelligent Data Understanding (CIDU), 24-26 October 2012*, Boulder, Colorado, 9-16.

Johnson S, Low-Choy S & Mengersen K 2012b: Integrating Bayesian Networks and Geographic Information Systems: Good Practice Examples. *Integrated Environmental Assessment and Management*, 8, 473-479.

Johnson S, Mengersen K, de Waal A, Marnewickc K, Cilliers D, Houser AM & Boast L 2010a. Modelling cheetah relocation success in southern Africa using an Iterative Bayesian Network Development Cycle. *Ecological Modelling*, 221, 641–651.

Johnson S, Fielding F, Hamilton G & Mengersen K 2010b: An integrated Bayesian network approach to *Lyngbya majuscula* bloom initiation. *Marine Environmental Research*, 69(1), 27-37.

Jones CG, Lawton JH & Shachak M 1994: Organisms as ecosystem engineers. *Oikos*, 69, 373-386.

Joost S, Colli L, Baret PV, Garcia JF, Boettcher PJ, Tixier-Boichard M, Ajmone-Marsan P & The GLOBALDIV Consortium 2010: Integrating geo-referenced multiscale and multidisciplinary data for the management of biodiversity in livestock genetic resources. *Animal Genetics*, 41(1), 47-63.

Joppa LN, McInerney G, Harper R, Salido L, Takeda K, O'Hara K, Gavaghan D & Emmott S 2013: Troubling Trends in Scientific Software Use. *Science*, 340, 814-815.

Kavirindi IU, Du Preez PJ & Brown LR 2010: Distribution and potential invasion of *Opuntia* spp. on selected Namibian sites. *Proceedings of the Second RUFORUM Biennial Meeting*, 20-24 September 2010, Entebbe, Uganda, 339-343.

Kaya E, Findik O, Babaoğlu İ & Arslan A 2011: Effect of discretization method on the diagnosis of Parkinson's disease. *International Journal of Innovative Computing, Information and Control*, 7, 4669-4678.

Keil P 2014: Limits of uncertainty about estimates of probability of ecological events. *PeerJ PrePrints*, 2:e446v1, doi: <http://dx.doi.org/10.7287/peerj.preprints.446v1>.

Kelley AL 2014: The role thermal physiology plays in species invasion. *Conservation Physiology*, 2(1), doi: <http://dx.doi.org/10.1093/conphys/cou045>.

Kéry M, Gardner B & Monnerat C 2010: Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, 37, 1851-1862.

Kirkpatrick SC, Gelatt D & Vecchi MP 1983: Optimization by Simulated Annealing. *Science*, 220, 671-680.

Kissling WD, Dormann CF, Groeneveld J, Hickler T, Kühn I, McNerny GJ, Montoya JM, Römermann C, Schiffers K, Schurr FM, Singer A, Svenning J-C, Zimmermann NE & O'Hara RB 2012: Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, 39(12), 2163-2178.

Kjærulff UB & Madsen AL 2013: Bayesian networks and influence diagrams: a guide to construction and analysis (2nd Ed). *Information Science and Statistics*. New York: Springer.

Kononenko I 1995: On biases in estimating multi-valued attributes. In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, Morgan Kaufmann: San Francisco, 1034-1040.

Korb KB & Nicholson AE 2011: *Bayesian Artificial Intelligence* (2nd Ed.). London: Chapman & Hall/CRC Press.

Koski T & Noble J 2009: *Bayesian Networks: An Introduction*, Wiley Series in Probability and Statistics. Chichester:Wiley.

Koski T & Noble J 2012: A Review of Bayesian Networks and Structure Learning. *Mathematica Applicanda*, 40, 53-103.

Kotzé I, Sibandze P, Beukes H, van den Berg E, Weepener H & Newby T 2010a: *Surveying and Mapping the Distribution and Intensity of Infestation of Selected Category 1 Invasive Alien Plant Species in Swaziland*, Report Number: GW/A/2010/28. Pretoria: Agricultural Research Council - Institute for Soil, Climate and Water.

Kotzé I, Sibandze P, Beukes H, van den Berg E, Weepener H & Newby T 2010b: *Surveying and Mapping the Distribution and Intensity of Infestation of Selected Category 1 Invasive Alien Plant Species in Swaziland – DATA SET (CD-ROM)*. Pretoria: Agricultural Research Council: Institute for Soil, Climate and Water.

- Kovács ZP, Du Plessis DB, Bracher PR, Dunn P & Mallory GCL 1985: *Documentation of the 1984 Domoina Floods*. Pretoria: Department of Water Affairs (South Africa).
- Kozlov, AV & Koller D 1997: Nonuniform dynamic discretization in hybrid networks. In: Geiger D & Shenoy PP (eds.): *Thirteenth Conference on Uncertainty in Artificial Intelligence*. Providence: Morgan Kaufmann, 314-325.
- Kraak MJ & Ormeling FJ 2011: *Cartography, Visualization of Spatial Data*. New York: Guildford Press.
- Krishnapuram B, Hartemink AJ, Carin L & Figueiredo MAT 2004: A Bayesian Approach to Joint Feature Selection and Classifier Design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 1105-1111.
- Kuebbing SE & Nuñez MA 2015: Negative, neutral, and positive interactions among nonnative plants: patterns, processes, and management implications. *Global Change Biology*, 21, 926-934.
- Kuikka S, Hilden M, Gislason H, Hansson S, Sparholt H & Varis O 1999: Modeling environmentally driven uncertainties in Baltic cod (*Gadus morhua*) management by Bayesian influence diagrams. *Canadian Journal of Fisheries and Aquatic Sciences*, 56, 629-641.
- Kueffer C, Pyšek P & Richardson DM 2013: Integrative invasion science: model systems, multi-site studies, focused meta-analysis and invasion syndromes. *New Phytologist*, 200, 615-633.
- Lahoz-Monfort JJ, Guillera-Arroita G & Wintle BA 2014: Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, 23, 504-515.
- Landuyt D, Broekx S, D'hondt R, Engelen G, Aertsens J & Goethals PLM 2013: A review of Bayesian belief networks in ecosystem service modeling. *Environmental Modelling & Software*, 46, 1-11.
- Landuyt D, Van der Biest K, Broekx S, Staes J, Meire P & Goethals PLM 2015: A GIS plug-in for Bayesian belief networks: Towards a transparent software framework to assess and visualise uncertainties in ecosystem service mapping. *Environmental Modelling & Software*, 71, 30-38.

- Langseth H, Nielsen TD, Rumí R & Salmerón A 2012. Mixtures of truncated basis functions. *International Journal of Approximate Reasoning*, 53(2), 212-227.
- Larjo A, Shmulevich I & Lähdesmäki H 2013: Structure learning for Bayesian networks as models of biological networks. *Methods in Molecular Biology*, 939, 35-45.
- Laskey KB, Wright EJ & Costa P 2010: Envisioning Uncertainty in Geospatial Information. *International Journal of Approximate Reasoning*, 51(2), 209-223.
- Lauritzen SL 1992: Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87, 1098-1108.
- Lauritzen SL 1995: The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19, 191-201.
- Lauritzen SL & Jensen F 2001: Stable local computation with conditional Gaussian distributions. *Statistics and Computing*, 11, 191-203.
- Lauritzen SL & Spiegelhalter DJ 1988: Local Computations of Probabilities on Graphical Structures and their Applications to Expert Systems. *Journal of the Royal Statistical Society B (Methodological)*, 50, 157-224.
- Lausch A, Schmidt A & Tischendorf L 2015: Data mining and linked open data – New perspectives for data analysis in environmental research. *Ecological Modelling*, 295, 5-17.
- Lavoie C, Shah MA, Bergeron A & Villeneuve P 2013: Explaining invasiveness from the extent of native range: New insights from plant atlases and herbarium specimens. *Diversity and Distributions*, 19, 98-105.
- Laws RJ & Kesler DC 2012: A Bayesian network approach for selecting translocation sites for endangered island birds. *Biological Conservation*, 155, 178-185.
- Lawton JH 1999: Are there general laws in ecology? *Oikos*, 84, 177-192, cited in McMahon SM 2005: Quantifying the community: using Bayesian Learning Networks to find structure and conduct inference in invasions biology. *Biological Invasions*, 7, 833–844.

- Lecklin T, Ryömä R & Kuikka S 2011: A Bayesian network for analyzing biological acute and long-term impacts of an oil spill in the Gulf of Finland. *Marine Pollution Bulletin*, 62, 2822-2835.
- Lee DC 2000: Assessing land-use impacts on bull trout using Bayesian belief networks. In Ferson S (ed.): *Quantitative methods for conservation biology*, New York: Springer, 127-147.
- Lee DC & Irwin LL 2005: Assessing risks to spotted owls from forest thinning in fire-adapted forest of the western United States. *Forest Ecology and Management*, 211, 191-209.
- Lee KC & Cho H 2012: Integration of General Bayesian Network and Ubiquitous Decision Support to Provide Context Prediction Capability. *Expert Systems with Applications*, 39(5), 6116-6121.
- Lehmkuhl JF, Kie JG, Bender LC, Servheen G & Nyberg H 2001: Evaluating the effects of ecosystem management alternatives on elk, mule deer, and white-tailed deer in the interior Columbia River basin, USA. *Forest Ecology and Management*, 153(1-3), 89-104.
- Le Maitre DC, van Wilgen BW, Gelderblom CM, Bailey C, Chapman RA & Nel JA 2002: Invasive alien trees and water resources in South Africa: case studies of the costs and benefits of management. *Forest Ecology and Management*, 160, 143-159.
- Le Maitre DC, Gush MB & Dzikiti S 2015: Impacts of invading alien plant species on water flows at stand and catchment scales. *AoB PLANTS*, 7, plv043; doi: <http://dx.doi.org/10.1093/aobpla/plv043>.
- Lim J, Crawley MJ, De Vere N, Rich T & Savolainen V 2014: A phylogenetic analysis of the British flora sheds light on the evolutionary and ecological factors driving plant invasions. *Ecology and Evolution*, 4(22), 4258-4269.
- Liu KF-R, Yeh K, Chen C-W, Liang H-H & Shen Y-S 2013: Using Bayesian Belief Networks and Fuzzy Logic to Evaluate Aquatic Ecological Risk. *International Journal of Environmental Science and Development*, 4(4), 419-424.

- Liu KF-R, Kuo J-Y, Yeh K, Chen C-W, Liang H-H & Sun Y-H 2015: Using fuzzy logic to generate conditional probabilities in Bayesian belief networks: a case study of ecological assessment. *International Journal of Environmental Science and Technology*, 12, 871-884.
- Liu X-Y & Zhou Z-H 2006: The influence of class imbalance on cost-sensitive learning: An empirical study. In *Proceedings of the 6th IEEE International Conference on Data Mining*, 18 – 22 December, Hong Kong, 970-974.
- Lobo JM, Jimenez-Valverde A & Real R 2008: AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17, 145-151.
- Lobo JM, Jiménez-Valverde A & Hortal J 2010: The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33:103-114.
- Lockwood JL, Cassey P & Blackburn T 2005: The role of propagule pressure in explaining species invasions. *Trends in Ecology and Evolution*, 20, 223-228.
- Loffler L & Loffler P 2005: *Swaziland Tree Atlas - including selected shrubs and climbers*. Southern African Botanical Diversity Network (SABONET) Report No. 38, Pretoria.
- Loffler L 2013: Personal Communication. Private Ecologist/Botanist, Mbabane.
- Lorena AC, Jacintho LFO, Siqueira MF, De Giovanni R, Lohmann LG, de Carvalho ACPLF & Yamamoto M 2011: Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, 38, 5268-5275.
- Lotter WD & Hoffman JH 1998: An integrated management plan for the control of *Opuntia stricta* (Cactaceae) in the Kruger National Park, South Africa. *Koedoe*, 41(1), 63-68.
- Low Choy SJ, O'leary RA & Mengersen K 2009: Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology*, 90, 265-277.
- Lucena-Moya P, Brawata R, Kath J, Harrison E, ElSawah S & Dyer F 2015: Discretization of continuous predictor variables in Bayesian networks: An ecological threshold approach. *Environmental Modelling & Software*, 66, 36-45.

MacCracken JG, Garlich-Miller J, Snyder J & Meehan R 2013: Bayesian belief network models for species assessments: an example with the Pacific walrus. *Wildlife Society Bulletin*, 37(1), 226-235.

Madden MG 2009: On the Classification Performance of TAN and General Bayesian Networks. *Knowledge-Based Systems*, 22(7), 489-495.

Magagula CN 2010: Determination of *Lantana camara* L. (Verbanaceae) varieties and evaluation of its biological control agents in selected areas of Swaziland. *Journal of Entomological Research*, 34(2), 103-109.

Maldonado AD, Aguilera PA & Salmerón in press: Continuous Bayesian networks for probabilistic environmental risk mapping. *Stochastic Environmental Research and Risk Assessment*.

Mandal G & Joshi SP 2014: Invasion establishment and habitat suitability of *Chromolaena odorata* (L.) King and Robinson over time and space in the western Himalayan forests of India. *Journal of Asia-Pacific Biodiversity*, 7(4), 391-400.

Mantyka-Pringle CS, Martin TG, Moffatt DB, Linke S & Rhodes JR 2014: Understanding and predicting the combined effects of climate change and land-use change on freshwater macroinvertebrates and fish. *Journal of Applied Ecology*, 51, 572-581.

Marceau DJ 1999: The scale issue in the social and natural sciences. *Canadian Journal of Remote Sensing*, 25, 347-356.

Marcer A, Pino J, Pons X & Brotons L 2012: Modelling invasive alien species distributions from digital biodiversity atlases – model upscaling as a means of reconciling data at different scales. *Diversity and Distributions*, 18, 1177-1189.

Marcot BG, Holthausen RS, Raphael MG, Rowland MM & Wisdom MJ 2001: Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management*, 153 (1-3), 29-42.

Marcot BG 2006: Characterizing species at risk I: modeling rare species under the Northwest Forest Plan. *Ecology and Society*, 11(2), 10. Available from: <http://www.ecologyandsociety.org/vol11/iss2/art10/> (Accessed on: 18 February 2015).

Marcot BG, Steventon JD, Sutherland GD & McCann RK 2006: Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, 36, 3063-3074.

Marcot BG 2012: Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling*, 230, 50-62.

Marcot BG, Allen CS, Morey S, Shively D & White R 2012: An Expert Panel Approach to Assessing Potential Effects of Bull Trout Reintroduction on Federally Listed Salmonids in the Clackamas River, Oregon. *North American Journal of Fisheries Management*, 32(3), 450-465.

Markowitz F & Spang R 2007: Inferring Cellular Networks - A Review. *BMC Bioinformatics*, 8, S5.

Marmion M, Parvainien M, Luoto M, Heikkinen RK & Thuiller W 2009: Evaluation of consensus methods in predictive species distribution modeling. *Diversity and Distributions*, 15, 59-69.

Maron JL & Vilà M 2001: When do herbivores affect plant invasion? Evidence for the natural enemies and biotic resistance hypotheses. *Oikos*, 95, 361-373.

Martin TG, Murphy H, Liedloff A, Thomas C, Chadès I, Cook G, Fensham R, McIvor J & van Klinken RD 2015: Buffel grass and climate change: a framework for projecting invasive species distributions when data are scarce. *Biological Invasions*, 17(11), 3197-3210.

McCann RK, Marcot BG & Ellis R 2006: Bayesian belief networks: Applications in ecology and natural resource management. *Canadian Journal of Forestry Research*, 36, 3053-3062.

McCarthy K, Zaber B & Weiss G 2005: Does cost-sensitive learning beat sampling for classifying rare classes? UBDM '05. In: *Proceedings of the 1st International Workshop on Utility-Based Data Mining*, Chicago, Illinois, 69-77.

- McMahon SM 2005: Quantifying the community: using Bayesian Learning Networks to find structure and conduct inference in invasions biology. *Biological Invasions*, 7, 833–844.
- McNay RS, Marcot BG, Brumovsky V & Ellis R 2006: A Bayesian approach to evaluating habitat for woodland caribou in north-central British Columbia. *Canadian Journal of Forest Research*, 36, 3117-3133.
- McNay RS, Sutherland G & Morgan DG 2011: Standardized occupancy maps for selected wildlife in Central British Columbia. *BC Journal of Ecosystems and Management*, 12(1), 118-135.
- McPherson JM, Jetz W & Rogers DJ 2004: The effects of species range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, 41, 811-823.
- Meier ES, Kienast F, Pearman PB, Svenning J-C, Thuiller W, Araújo MB, Guisan A & Zimmermann NE 2010: Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography*, 33, 1038-1048.
- Meineri E, Dahlberg CJ & Hylander K 2015. Using Gaussian Bayesian Networks to disentangle direct and indirect associations between landscape physiography, environmental variables and species distribution. *Ecological Modelling*, 313, 127-136.
- Mello MP, Risso J, Atzberger C, Aplin P, Pebesma E, Vieira CAO & Rudorff BFT 2013: Bayesian Networks for Raster Data (BayNeRD): Plausible Reasoning from Observations. *Remote Sensing*, 5, 5999-6025.
- Menne W & Carrere R 2007: *Swaziland: The myth of sustainable timber plantations*. World Rainforest Movement: Montevideo, Uruguay.
- Mennis J & Guo D 2009: Spatial data mining and geographic knowledge discovery - An introduction. *Computers, Environment and Urban Systems*, 33, 403-408.
- Merow C, Smith M, Edwards T, Guisan A, McMahon S, Normand S, Thuiller W, Wuest R, Zimmermann N & Elith J 2014: What do we gain from simplicity versus complexity in species distribution models? *Ecography*, 37, 1267-1281.

- Millsaugh JJ & Thompson III FR 2009: *Models for planning wildlife conservation in large landscapes*. Boston: Elsevier/Academic Press.
- Milns I, Beale CM & Smith VA 2010: Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, 91, 1892-1899.
- Modir-Rahmati A 1997: Cultivation of poplar in Iran. *Holzzucht*, 51(2), 41-43.
- Monadjem A, Boycott R, Parker V & Culverwell J 2003: *Threatened Vertebrates of Swaziland. Swaziland Red Data Book: Fishes, Amphibians, Reptiles, Birds and Mammals*. Mbabane: Ministry of Tourism, Environment and Communications.
- Moraglio A, Di Chio C & Poli R 2007: Geometric Particle Swarm Optimization. In *Proceedings of the 10th European Conference on Genetic Programming*, Berlin, Heidelberg, 125-136.
- Moral S, Rumí R & Salmerón A 2001: Mixtures of truncated exponentials in hybrid Bayesian networks. ECSQARU. *Lecture Notes in Artificial Intelligence*, 2143, 135-143.
- Morales-Castilla I, Matias MG, Gravel D & Araújo MB 2015: Inferring biotic interactions from proxies. *Trends in Ecology and Evolution*, 30(6), 347-356.
- Morgan JD, Hutchins MW, Fox J & Rogers KR 2012: A Methodological Framework focused on integrating GIS and BBN Data for Probabilistic Map Algebra Analysis. *Extended Abstracts Volume, 7th International Conference on Geographic Information Science, GIScience 2012, September 2012, Columbus, Ohio*.
- Morgan MG, Henrion M 1990: *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. New York: Cambridge University Press.
- Mouton AM, Baets BD & Goethals PLM 2010: Ecological relevance of performance criteria for species distribution models. *Ecological Modelling*, 221, 1995-2002.
- Mandal G & Joshi SP 2014: Invasion establishment and habitat suitability of *Chromolaena odorata* (L.) King and Robinson over time and space in the western Himalayan forests of India. *Journal of Asia-Pacific Biodiversity*, 7, 391-400.

- Murphy K 2014: Software packages for Graphical Models (Online). Available from: www.cs.bc.ca/~murphyk/Software/bnsoft.html (Accessed on: 17 December 2014).
- Murray JV, Stokes KE & van Klinken RD 2012: Predicting the potential distribution of a riparian invasive plant: the effects of changing climate, flood regimes and land-use patterns. *Global Change Biology*, 18, 1738-1753.
- Murray JV, Berman DM & van Klinken RD 2014: Predictive modelling to aid the regional-scale management of a vertebrate pest. *Biological Invasions*, 16, 2403-2425.
- Murty US, Rao MS & Arunachalam N 2009: Prediction of Japanese encephalitis vectors in Kurnool district of andhra pradesh, India by using Bayesian network. *Applied Artificial Intelligence*, 23(9), 828-834.
- Myllymaki P, Silander T, Tirri H & Uronen P 2002: B-course: a web-based tool for Bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools*, 11, 369-387.
- Nadeau C & Bengio Y 2003: Inference for the Generalization Error. *Machine Learning*, 52, 239-281.
- Nagarajan R, Scutari M & Lebre S 2013: *Bayesian Networks in R*. New York: Springer.
- Naimi B, Hamm NAS, Groen TA, Skidmore AK & Toxopeus AG 2014: Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37(2), 191-203.
- Neapolitan RE 2004: *Learning Bayesian Networks. Series in Artificial Intelligence*. Prentice Hall: New Jersey.
- Neapolitan RE 2009: *Probabilistic methods for bioinformatics: with an introduction to Bayesian networks*. Morgan Kaufmann: Burlington.
- Newbold T 2010: Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, 34, 3-22.

- Newton AC, Stewart GB, Diaz A, Golicher D & Pullin AS 2007: Bayesian Belief Networks as a tool for evidence-based conservation management. *Journal for Nature Conservation*, 15, 144-160.
- Newton AC 2009: Bayesian belief networks in environmental modelling: a review of recent progress. In: Findley PN (ed.): *Environmental Modelling: New Research*, New York: Nova Science Publishers, 13-50.
- Nielsen C, Hartvig P & Kollmann J 2008: Predicting the distribution of the invasive alien *Heracleum mantegazzianum* at two different spatial scales. *Diversity and Distributions*, 14, 307-317.
- Norsys Software Corporation 2014a: *Netica Application User's Guide*. Vancouver: Norsys Software Corporation, Norsys Software Corporation.
- Norsys Software Corporation 2014b: *Geo-Netica*. Norsys Software Corporation, Norsys Software Corporation, Vancouver, BC, Canada. (Online) Available from: <http://www.norsys.com> (Accessed 7 September 2014).
- Novoa A, Le Roux JJ, Robertson MP, Wilson JRU & Richardson DM 2015: Introduced and invasive cactus species: a global review. *AoB PLANTS*, 7, plu078; doi: <http://dx.doi.org/10.1093/aobpla/plu078>.
- Nyberg JB, Marcot BG & Sulyma R 2006: Using Bayesian belief networks in adaptive management. *Canadian Journal of Forestry Research*, 36, 3104-3116.
- O'Brien R 2006: Spatial Driving Factor Analysis for Specialty Crops. In *Interactions and Processes: Proceedings of the 18th Annual Colloquium of the Spatial Information Research Centre, University of Otago, Dunedin, New Zealand, 6-7 November 2006*, SIRC, 6-7.
- Ogle BM & Grivetti LE 1985: Legacy of the chameleon: Edible wild plants in the kingdom of Swaziland, Southern Africa. A cultural, ecological, nutritional study. Part I - Introduction, objectives, methods, Swazi culture, landscape and diet. *Ecology of Food and Nutrition*, 16(3), 193-208

- Orwa C, Mutua A, Kindt R, Jamnadass R, Anthony S 2009: Agroforestry Database: a tree reference and selection guide version 4.0 (Online) Available from: (<http://www.worldagroforestry.org/sites/treedbs/treedatabases.asp>) (Accessed 7 March 2015).
- Os'kina NV & Bepalov VP 1992: Biological productivity of white poplar stands in the floodplain of the river Ural. *Lesovedenie*, 6, 39-47.
- Oswalt CM, Fei S, Guo Q, Iannone III BV, Oswalt SN, Pijanowski BC & Potter KM 2015: A subcontinental view of forest plant invasions. *NeoBiota*, 24, 49-54.
- Ovaskainen O, Hottola J & Siitonen J 2010: Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91(9), 2514-2521.
- Parker V 1994. *The Swaziland Bird Atlas 1985-1991*. Mbabane: Webster's.
- Pauchard A & Shea K 2006: Integrating the study of non-native plant invasions across spatial scales. *Biological Invasions*, 8, 399-413.
- Pearl J 1982: Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. In *Proceedings of the National Conference on Artificial Intelligence, Pittsburgh, Pasadena, Menlo Park*: AAAI Press, 133-136.
- Pearl J 1988: *Probabilistic Reasoning in Intelligent Systems*. San Mateo: Morgan Kaufmann.
- Pearl J 1995: Causal diagrams for empirical research. *Biometrika*, 82, 669-709.
- Pearl J 1997: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (2nd Ed). San Mateo: Morgan Kaufmann.
- Pearl J 2000: *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
- Pearl J & Verma TS 1991: A theory of inferred causation. In Allen JF, Fikes R & Sandewall E (eds.): *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann: San Mateo, 441-452.

- Pejchar L & Mooney HA 2009: Invasive species, ecosystem services and human well-being. *Trends in Ecology & Evolution*, 24, 497-504.
- Pellikka J, Kuikka S, Lindén H & Varis O 2005: The role of game management in wildlife populations: uncertainty analysis of expert knowledge. *European Journal of Wildlife Research*, 51, 48-59.
- Peng H, Long F & Ding C 2005: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1226-1238.
- Peterson DP, Rieman BE, Dunham JB, Faush KD & Young MK 2008: Analysis of trade-offs between threats of invasion by modelling brook trout (*Salvelinus fontinalis*) and intentional isolation for native westslope cutthroat trout (*Oncorhynchus lewisi*). *Canadian Journal of Fisheries and Aquatic Sciences*, 65, 557-573.
- Peterson AT, Soberón J & Sanchez-Cordero VV 1999: Conservatism of ecological niches in evolutionary time. *Science*, 285(5431), 1265-1267.
- Peterson AT 2006: Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, 3, 59-72.
- Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M & Araújo MB 2011: *Ecological Niches and Geographic Distributions*. Princeton: Princeton University Press.
- Peterson AT & Soberón J 2012: Species Distribution Modeling and Ecological Niche Modeling: Getting the concepts right. *Natureza & Conservação*, 10, 102-107.
- Peterson DP, Wenger SJ, Rieman BE & Isaak DJ 2013. Linking Climate Change and Fish Conservation Efforts Using Spatially Explicit Decision Support Tools. *Fisheries*, 38(3), 112-127.
- Phillips SJ & Elith J 2013: On estimating probability of presence from use-availability or presence-background data. *Ecology*, 94(6), 1409-1419.

Pilz J & Spöck G 2007: Why Do We Need and How Should We Implement Bayesian Kriging Methods. *Stochastic Environmental Research and Risk Assessment*, 22, 621-632.

Polar Bear Specialist Group 2014. *Minutes of the 17th Meeting of the IUCN/SSC Polar Bear Specialist Group, Fort Collins, Colorado, USA 9-14 June 2014*. (Online) Available from: https://polarbearsociety.files.wordpress.com/2014/07/pbsg17-minutes-final_posted-online-july-18-after-the-rest-jun-26.pdf (Accessed 27 February 2015).

Pollino CA, White AK & Hart BT 2007: Examination of conflicts and improved strategies for the management of an endangered Eucalypt species using Bayesian networks. *Ecological Modelling*, 201, 37-59.

Pollino CA & Hart BT 2008: Developing Bayesian network models within a risk assessment framework. In Sanchez M (ed.): *Proceedings of the International Environmental Modelling and Simulation Society*, Barcelona, Spain. (Online) Available from: <https://hdl.handle.net/1885/53304> (Accessed 21 February 2015).

Pourret O, Nam P, Naïm P, Marcot B 2008: *Bayesian Networks: A Practical Guide to Applications*. Wiley: New York.

Poynton RJ 1973: Trees in South Africa - two hundred selected indigenous and exotic species: how to recognise and grow them. In Immelman WFE, Wicht CL & Ackerman DP (eds.): *Our green heritage*. Cape Town: Tafelberg.

Press WH, Flannery BP, Teukolsky SA & Vetterling WT 1988: *Numerical Recipes in C*. Cambridge: Cambridge University Press.

Pressey RL 2004: Conservation planning and biodiversity: assembling the best data for the job. *Conservation Biology*, 18, 1677-1681.

Pullar DV & Phan TH 2007: Using a Bayesian Network in a GIS to Model Relationships and Threats to Koala Populations Close to Urban Environments. In Oxley L & Kulasiri D (eds.): *MODSIM 2007: Land, Water and Environmental Management: Integrated Systems for Sustainability. International Congress on Modelling and Simulation (December 2007)*, University of Canterbury, Christchurch, New Zealand, 1370-1375.

Pyšek P, Jarošík V, Hulme PE, Kühn I, Wild J, Arianoutsou M, Bacher S, Chiron F, Didžiulis V, Essl F, Genovesi P, Gherardi F, Hejda M, Kark S, Lambdon PW, Desprez-Loustau M-L, Nentwig W, Pergl J, Pobljšaj K, Rabitsch W, Roques A, Roy DB, Shirley S, Solarz, W, Vilá M & Winter M 2010: Disentangling the role of environmental and human pressures on biological invasions across Europe. *Proceedings of the National Academy of Sciences*, 107, 12157-12162.

QGIS Development Team 2012: QGIS Geographic Information System. Open Source Geospatial Foundation Project. Technical report. (Online) Available from: <http://qgis.osgeo.org> (Accessed 7 September 2014).

Qian SS & Miltner RJ 2015: A continuous variable Bayesian networks model for water quality modeling: A case study of setting nitrogen criterion for small rivers and streams in Ohio, USA. *Environmental Modelling & Software*, 69, 14-22.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabus B, Eineder M, Roth A & Bamler R 2003: The shuttle radar topography mission- a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57, 241-262.

Ramaswami G & Sukumar R 2013: Long-Term Environmental Correlates of Invasion by *Lantana camara* (Verbenaceae) in a Seasonally Dry Tropical Forest. *PLoS ONE*, 8(10): e76995. doi: <http://dx.doi.org/10.1371/journal.pone.0076995>.

Rangel TF & Loyola RD 2012: Labeling Ecological Niche Models. *Natureza & Conservação*, 10, 119-126.

Raphael MG, Wisdom MJ, Rowland MM, Holthausen RS, Wales BC, Marcot BG & Rich TD 2001: Status and trends of habitats of terrestrial vertebrates in relation to land management in the interior Columbia River Basin. *Forest Ecology and Management*, 153, 63-88.

Rédei K 1998: Early evaluation of promising white poplar (*Populus alba* L.) clones in sandy ridges between the rivers Danube and Tisza in Hungary. *Silva Lusitana*, 6(1), 63-71.

Rehr AP, Williams GD, Tolimieri N & Levin PS 2014: Impacts of Terrestrial and Shoreline Stressors on Eelgrass in Puget Sound: An Expert Elicitation. *Coastal Management*, 42, 246-262.

Rejmánek M 2015: Invasion of *Rubus praecox* (Rosaceae) is promoted by the native tree *Aristotelia chilensis* (Elaeocarpaceae) due to seed dispersal facilitation. *Gayana Botanica*, 72(1), 27-33.

Remaley T & Swearingen JM 1998: *White poplar, Populus alba fact sheet. Alien plant invaders of natural areas – trees.* (Online) Available from: <http://www.nps.gov/plants/alien/fact/poal1.htm> (Accessed 10 November 2014).

Rommelzwaal A & Dlamini WS 1994: *Present Land Use Map of Swaziland, Scale 1:250,000.* FAO/UNDP/GOS Land Use Planning for Rational Utilization of Land and Water Resources Project SWA/89/001. Field Doc. 9, Mbabane.

Rommelzwaal A & Vilakati JD 1994: *Land Tenure Map of Swaziland, Scale 1:250,000.* FAO/UNDP/GOS Land Use Planning for Rational Utilization of Land and Water Resources Project SWA/89/001. Field Doc. 10, Mbabane.

Renken H & Mumby PJ 2009: Modelling the dynamics of coral reef macroalgae using a Bayesian belief network approach. *Ecological Modelling*, 220(9–10), 1305-1314.

Richardson DM, Pysěk P, Rejmánek M, Barbour MG, Panetta FD & West CJ 2000: Naturalization and invasion of alien plants: concepts and definitions. *Diversity and Distributions*, 6, 93-107.

Richardson DM, Rejmánek M. 2004. Conifers as invasive aliens?: a global survey and predictive framework. *Diversity and Distributions*, 10, 321-331.

Richardson DM, Holmes PM, Esler KJ, Galatowitsch SM, Stromberg JC, Kirkman SP, Pysěk P & Hobbs RJ 2007: Riparian vegetation: degradation, alien plant invasions, and restoration prospects. *Diversity and Distributions*, 13, 126-139.

- Richardson DM, Carruthers J, Hui C, Impson FAC, Miller JT, Robertson MP, Rouget M, Le Roux JJ & Wilson JRU 2011: Human-mediated introductions of Australian acacias – a global experiment in biogeography. *Diversity and Distributions*, 17, 771-787.
- Richardson DM & Rejmánek M 2011: Trees and shrubs as invasive alien species – a global review. *Diversity and Distributions*, 17, 788-809.
- Richardson DM & Whittaker RJ 2010: Conservation biogeography – foundations, concepts and challenges. *Diversity and Distributions*, 16, 313-320.
- Rieman B, Peterson JT, Clayton J, Howell P, Thurow R, Thompson W & Lee D 2001. Evaluation of potential effects of federal land management alternatives on trends of salmonids and their habitats in the interior Columbia River basin. *Forest Ecology and Management*, 153, 43-62.
- Robinson RW 1977: Counting unlabeled acyclic digraphs. In Little CHC (ed.): *Combinatorial Mathematics V*. Vol. 622, Berlin: Springer, 28-43.
- Rocchini D, Lobo JM, Jime A, Bacaro G & Chiarucci A 2011: Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, 35, 211-226.
- Ropero RF, Aguilera PA, Fernández A & Rumí R 2014: Redes bayesianas: una herramienta probabilística en los modelos de distribución de especies. *Ecosistemas*, 23(1), 54-60.
- Roques KG 2002: *A Preliminary Field Assessment of Protection-Worthy Areas of Swaziland*. Mbabane: Swaziland Environment Authority, Ministry of Tourism, Environment and Communications.
- Rota CT, Fletcher RJ, Evans JM & Hutto RL 2011: Does accounting for imperfect detection improve species distribution models? *Ecography*, 34, 659-670.
- Rouget M, Hui C, Renteria J, Richardson DM & Wilson JRU 2015: Plant invasions as a biogeographical assay: Vegetation biomes constrain the distribution of invasive alien species assemblages. *South African Journal of Botany*, 101, 24-31.

- Rowland MM, Wisdom MJ, Johnson DH, Wales BC, Copeland JP & Edelman FB 2003: Evaluation of landscape models for wolverines in the interior northwest, United States of America. *Journal of Mammalogy*, 84(1), 92-105.
- Royle JA, Link WA & Sauer JR 2002. Statistical mapping of count survey data. In Scott JM, Heglund PJ, Morrison ML, Haufler JB, Raphael MG, Wall WA & Samson FB (eds.): *Predicting species occurrences: issues of scale and accuracy*, Covello, California: Island Press, 625-638.
- Rüger N, Schlüter M, Matthies M 2005: A fuzzy habitat suitability index for *Populus euphratica* in the Northern Amudarya delta (Uzbekistan). *Ecological Modelling*, 184, 313-328.
- Sahid IB & Sugau, JB 1993: Allelopathic effects of lantana (*Lantana camara*) and Siam weed (*Chromolaena odorata*) on selected crops. *Weed Science*, 41, 303-308.
- SANBI 2014: *The Southern African Bird Atlas Project 2 (Dataset)*. Pretoria: South African National Biodiversity Institute.
- Sang Y, Qi H, Li K, Jin Y, Yan D & Gao S 2014: An effective discretization method for disposing high-dimensional data. *Information Sciences*, 270, 73–91.
- Santika T 2011: Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, 20, 181-192.
- Schor J, Farwig N & Berens DG 2015: Intensive land-use and high native fruit availability reduce fruit removal of the invasive *Solanum mauritianum* in South Africa. *South African Journal of Botany*, 96, 6-12.
- Schulze RE 1997: *South African atlas of agrohydrology and climatology*. Water Research Commission, Pretoria, Report TT82/96.
- Sekawin M 1975: Genetics of *Populus alba*. *Annales Forestales*, 6(6), 159-189.
- Shachter RD 1986: Evaluating influence diagrams. *Operations Research*, 34, 871–882.
- Shenoy PP & West JC 2011: Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning*, 52(5), 641-657.

Shenton W, Hart BT & Chan T 2011: Bayesian network models for environmental flow decision-making: 1. Latrobe River Australia. *River Research and Applications*, 27(3), 283-296.

Shenton W, Hart BT & Chan TU 2014: A Bayesian network approach to support environmental flow restoration decisions in the Yarra River, Australia. *Stochastic Environmental Research and Risk Assessment*, 28(1), 57-65.

Sierra R & Stephens CR 2012: Exploratory analysis of the interrelations between co-located boolean spatial features using network graphs. *International Journal of Geographical Information Science*, 26, 441-468.

Sillero N 2011: What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling*, 222, 1343-1346.

Simberloff D & Von Holle B 1999: Positive interactions of nonindigenous species: invasional meltdown? *Biological Invasions*, 1, 21-32.

Silverman BW 1986: *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.

Skandrani Z, Lepetz S & Prévot-Julliard A 2014: Nuisance species: beyond the ecological perspective. *Ecological Processes*, 3:3, doi: <http://dx.doi.org/10.1186/2192-1709-3-3>.

Smith AV 2010: Revealing Structure of Complex Biological Systems Using Bayesian Networks. In Estrada E, Fox M, Higham DJ & Oppo G-L (eds.): *Network Science: Complexity in Nature and Technology*. London: Springer, 185-204.

Smith CS, Howes AL, Price B, McAlpine CA 2007: Using a Bayesian belief network to predict suitable habitat of an endangered mammal - The Julia Creek Dunnart (*Sminthopsis douglasi*). *Biological Conservation*, 139, 333-347.

Smith C, van Klinken RD, Seabrook L, McAlpine C 2012: Estimating the influence of land management change on weed invasion potential using expert knowledge. *Diversity and Distributions*, 18, 818-831.

- Smith GF, Figueiredo E, Boatwright JS & Crouch NR 2011: South Africa's ongoing *Opuntia* Mill. (Cactaceae) problem: the case of *Opuntia microdasys* (Lehm.) Pfeiff. *Bradleya*, 29, 73-78.
- Soininen J & Luoto M 2014: Predictability in species distributions: a global analysis across organisms and ecosystems. *Global Ecology and Biogeography*, 23, 1264-1274.
- Spiegelhalter DJ & Lauritzen SL 1990: Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20, 579-605.
- Spirtes P, Glymour C & Scheines R 1993: *Causation, Prediction and Search. Lecture Notes in Statistics No. 81*. New York: Springer-Verlag.
- Spirtes P, Glymour C & Scheines R 2000: *Causation, Prediction, and Search*. Cambridge: The MIT Press.
- Stafford R, Williams RL & Herbert RJH 2015: Simple, policy friendly, ecological interaction models from uncertain data and expert opinion. *Ocean and Coastal Management*, 118, 88-96.
- Stassopoulou A, Petrou M & Kittler J 1998: Application of a Bayesian network in a GIS based decision making system. *International Journal of Geographic Information Science*, 12, 23-46.
- Steventon JD & Daust DK 2009: Management strategies for a large-scale mountain pine beetle outbreak: modelling impacts on American martens. *Forest Ecology and Management*, 257(9), 1976-1985.
- Stobrawa K 2014: Poplars (*Populus* spp.): Ecological role, applications and scientific perspectives in the 21st century. *Baltic Forestry*, 20(1), 204-213.
- Stockwell DRB 2006: Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecological Modelling*, 192, 188-196.
- Stohlgren TJ, Ma P, Kumar S, Rocca M, Morisette JT, Jarnevich CS & Benson N 2010: Ensemble habitat mapping of invasive plant species. *Risk Analysis*, 30, 224-235.
- Strayer DL 2012: Eight questions about invasions and ecosystem functioning. *Ecology Letters*, 15, 1199-1210.

Su C, Andrew A, Karagas MR & Borsuk ME 2013: Using Bayesian networks to discover relations between genes, environment, and disease. *BioData Mining*, 6(1), 6, doi: <http://dx.doi.org/10.1186/1756-0381-6-6>.

Suring LH, Gaines WL, Wales BC, Mellen-McLean K, Begley JS & Mohoric S 2011: Maintaining populations of terrestrial wildlife through land management planning: a case study. *The Journal of Wildlife Management*, 75(4), 945-958.

Syphard AD & Franklin J 2009: Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, 32(6), 907-918.

Syphard AD & Franklin J 2010: Species traits affect the performance of species distribution models for plants in southern California. *Journal of Vegetation Science*, 21, 177-189.

Tantipisanuh N, Gale GA & Pollino C 2014: Bayesian networks for habitat suitability modeling: a potential tool for conservation planning with scarce resources. *Ecological Applications*, 24, 1705-1718.

Tatem AJ, Noor AM, von Hagen C, Di Gregorio A & Hay SI 2007. High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa. *PLoS ONE*, 2(12): e1298. doi: <http://dx.doi.org/10.1371/journal.pone.0001298>.

Tattari S, Schultz T & Kuussaari M 2003: Use of belief network modelling to assess the impact of buffer zones on water protection and biodiversity. *Agriculture, Ecosystems & Environment*, 96, (1-3), 119-132.

Taylor BW & Irwin RE 2004: Linking economic activities to the distribution of exotic plants. *Proceedings of the National Academy of Science (USA)*, 101, 17725–17730.

Tessarolo G, Rangel TF, Araújo & Hortal J 2014: Uncertainty associated with survey design in Species Distribution Models. *Diversity and Distributions*, 20, 1258-1269.

Thomas J & Sael L 2015: Overview of Integrative Analysis Methods for Heterogeneous Data. In *Proceedings of the 2015 International Conference on Big Data and Smart Computing (BigComp 2015)*, 9-11 February 2015, Jeju, South Korea, 266-270.

Thorson JT, Scheuerell MD, Shelton AO, See KE, Skaug HJ & Kristensen K 2015: Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, 6, 627-637.

Thuiller, W 2013: On the importance of edaphic variables to predict plant species distributions – limits and prospects. *Journal of Vegetation Science*, 24, 591–592.

Thuiller W, Richardson DM, Rouget M, Prochesx S & Wilson JRU 2006: Interactions between environment, species traits, and human uses describe patterns of plant invasions. *Ecology*, 87(7), 1755–1769.

Ticehurst JL, Newham LTH, Rissik D, Letcher RA & Jakeman AJ 2007: A Bayesian network approach for assessing the sustainability of coastal lakes in New South Wales, Australia. *Environmental Modelling & Software*, 22(8), 1129-1139.

Tobler WR 1970: A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240.

Trabucco A & Zomer RJ 2010: *Global Soil Water Balance Geospatial Database*. (Online) CGIAR Consortium for Spatial Information. Available from <http://www.cgiar-csi.org> (Accessed 7 September 2013).

Traveset DM & Richardson DM 2014: Mutualistic Interactions and Biological Invasions. *Annual Review of Ecology, Evolution, and Systematics*, 45, 89-113.

Trifonova N, Kenny A, Maxwell D, Duplisea D, Fernandes J & Tucker A 2015: Spatio-temporal Bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. *Ecological Informatics*, 30, 142-158.

Tütüncü GY & Kayaalp N 2015: An Aggregated Fuzzy Naive Bayes Data Classifier. *Journal of Computational and Applied Mathematics*, 286, 17-27.

Uden DR, Allen CR, Angeler DG, Corral L & Fricke KA 2015: Adaptive invasive species distribution models: a framework for modelling recipient invasions. *Biological Invasions*, 17(10), 2831-2850.

- Uusitalo L 2007: Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, 203, 312-318.
- Uusitalo L, Kuikka S, Kauppila P, Söderkuntalahti P & Bäck S 2011: Assessing the Roles of Environmental Factors in Coastal Fish Production in the Northern Baltic Sea: A Bayesian Network Application. *Integrated Environmental Assessment and Management*, 8(3), 445–455.
- Uusitalo L, Lehtikoinen A, Helle I & Myrberg K 2015: An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*, 63, 24-31.
- Václavík T & Meentemeyer RK 2009: Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, 220, 3248–3258.
- van Klinken RD & Murray J 2011: Challenges, constraints and solutions for modeling regional-scale dispersal of invasive organisms: from practice to policy. In Chan F, Marinova D & Anderssen RS (eds.): *MODSIM2011, 19th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, 12–16 December 2011, Perth, Australia*, 2570-2577.
- van Klinken RD, Murray J & Smith C 2015: Process-based Pest Risk Mapping using Bayesian Networks and GIS. In Venette RC (ed.): *Pest Risk Modelling and Mapping for Invasive Alien Species*, CABI: Boston, 171-188.
- van Wilgen BW, Cowling RM, Marais C, Esler KJ, McConnachie M & Sharp D 2012: Challenges in invasive alien plant control in South Africa. *South African Journal of Science*, 108, 11-12.
- Varis O & Kuikka S 1999: Learning Bayesian decision analysis by doing: lessons from environmental and natural resources management. *Ecological Modelling*, 119, 177–195.
- Vegter JR 1995: *Geology map of South Africa with simplified lithostratigraphy for geohydrological use*. (Simplified lithostratigraphy digitised by A Havenga, Council for

Geosciences, 1994). Water Research Commission Report No. TT 74/95, Pretoria: Water Research Commission.

Velikova M, Lucas PJF, Samulski M & Karssemeijer N 2013: On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks. *Artificial Intelligence in Medicine*, 57, 73–86.

Venette RC, Kriticos DJ, Magarey RD, Koch FH, Baker RH, Worner SP, Gómez Raboteaux NN, McKenney DW, Dobesberger EJ, Yemshanov D, De Barro PJ, Hutchison WD, Fowler G, Kalaris TM & Pedlar J 2010: Pest risk maps for invasive alien species: a roadmap for improvement. *BioScience*, 60, 349–362.

Venette RC 2015: The Challenge of Modelling and Mapping the Future Distribution and Impact of Invasive Alien Species. In Venette RC (ed.): *Pest Risk Modelling and Mapping for Invasive Alien Species*, Boston: CABI, 1-17.

Verma T & Pearl J 1992: An algorithm for deciding if a set of observed independencies has a causal explanation. In Dubois D & Wellman MP (eds.): *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann: Stanford, 323-330.

Verweij P, Simoes M, Alves A, Ferraz R & Cormont A 2014: Linking Bayesian Belief Networks and GIS to assess the Ecosystem Integrity in the Brazilian Amazon. In Ames DP, Quinn NWT & Rizzoli AE (eds.): *International Environmental Modelling and Software Society (iEMSs) 7th International Congress on Environmental Modelling and Software*, San Diego, CA, USA.

Vilà M, Espina, JL, Hejda M, Hulme, PE, Jarošík V, Maron JL, Pergl J, Schaffner U, Sun Y & Pyšek P 2011: Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. *Ecology Letters*, 14: 702–708.

Vilizzi L, Price A, Beesley L, Gawne B, King AJ, Koehn JD, Meredith SN & Nielsen DL 2013: Model development of a Bayesian belief network for managing inundation events for wetland fish. *Environmental Modelling & Software*, 41, 1-14.

- Vogel K, Riggelsen C, Korup O & Scherbaum F 2014: Bayesian network learning for natural hazard analyses. *Natural Hazards and Earth System Science*, 14, 2605-2626.
- Voie OA 2003: Information indexing in a Bayesian network for risk assessment of land-use decisions on threatened species. In Frost DA (ed.): *MODSIM 2003: Integrative Modelling of Biophysical, Social, Economic Systems for Resource Management Solutions*. International Congress on Modelling and Simulation, 14-17 July 2003, Townsville, Australia, 1370-1375.
- Walker A, Pham B & Maeder A 2004: A Bayesian framework for automated dataset retrieval in Geographic Information Systems. In *Proceedings of the 10th International Multimedia Modeling Conference, 5-7 January 2004. Brisbane, Australia*, 138-144.
- Walters M, Figueiredo E, Crouch NR, Winter PJD, Smith GF, Zimmermann HG & Mashope BK 2011: *Naturalised and invasive succulents of southern Africa*. Belgium: ABC Taxa.
- Wamelink GWW, Goedhart PW & Frissel JY 2014: Why some plant species are rare. *Plos ONE*, 9, e102674.
- Warren DL 2012: In defense of ‘niche modeling’. *Trends in Ecology & Evolution*, 27, 497-500.
- Watling JI, Brandt LA, Bucklin DN, Fujisaki I, Mazzotti FJ, Romañach SS & Speroterra C 2015: Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. *Ecological Modelling*, 309-310, 48-59.
- Weber E 2003: *Invasive plant species of the world: A reference guide to environmental weeds*. Wallingford: CAB International.
- Wikle CK 2002: Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains. In: A. Lawson & D. Denison (eds): *Spatial Cluster Modelling*. Boca Raton: CRC Press: 199-209.
- Wikle CK 2003: Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84, 1382-1394.
- Wikle CK & Royle JA 2004: Spatial statistical modeling in biology. In *Encyclopedia of Life Support Systems (EOLSS)*. Oxford: Eolss Publishers.

Wilhere GF 2012: Using Bayesian networks to incorporate uncertainty in habitat suitability index models. *The Journal of Wildlife Management*, 76, 1298–1309.

Wilson AC 1982: *Geological Map of Swaziland, 1:250,000 (including Geological Summary)*. Mbabane: Geological Surveys and Mines Department.

Wilson JRU, Dormontt EE, Prentis PJ, Lowe AJ & Richardson, DM 2009. Something in the way you move: dispersal pathways affect invasion success. *Trends in Ecology and Evolution*, 24, 136–144.

Wilson DS, Stoddard MA & Puettmann KJ 2008: Monitoring amphibian populations with incomplete survey information using a Bayesian probabilistic model. *Ecological Modelling*, 214, 210–218.

Wisz MS, Pottier J, Kissling WD, Pellissier L, Lenoir J, Damgaard CF, Dormann CF, Forchhammer MC, Grytnes JA, Guisan A, Heikkinen RK, Høye TT, Kühn I, Luoto M, Maiorano L, Nilsson MC, Normand S, Öckinger E, Schmidt NM, Termansen M, Timmermann A, Wardle DA, Aastrup P, Svenning JC 2013: The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, 88, 15-30.

Witten IH, Frank E & Hall MA 2011: *Data Mining: Practical Machine Learning Tools and Techniques* (3rd Ed.). San Francisco: Morgan Kaufmann.

Yadava UL 1996: Guava production in Georgia under cold protection structure. In Janick J (ed.), *Progress in new crops*, Arlington: ASHS Press, 451-457.

Yu J, Wong W-K & Hutchinson RA 2010: Modeling Experts and Novices in Citizen Science Data for Species Distribution Modeling. In *Proceedings of the IEEE International Conference on Data Mining*, 1157-1162.

Yu L & Liu H 2003: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *Proceedings of the 20th International Conference on Machine Learning*, Washington, D.C., August 21-24, 2003, 20 (2), 856-863.

Yun C, Shin D, Jo H, Yang J & Kim S 2007. An Experimental Study on Feature Subset Selection Methods. *Seventh International Conference on Computer and Information Technology*, 77-82.

Zachariades C & Goodall JM 2002: Distribution, impact and management of *Chromolaena odorata* in southern Africa. In Zachariades C, Muniappan R & Strathie LW (eds.): *Proceedings of the 5th International Workshop on the Biological Control and Management of Chromolaena Odorata*, Durban, South Africa, Pretoria, South Africa: ARC-PPRI, 34-39.

Zimmermann NE, Edwards TC, Graham CG, Pearman PB & Svenning J-C 2010: New trends in species distribution modelling. *Ecography*, 33, 985-989.

Zomer RJ, Trabucco A, Bossio DA, van Straaten O & Verchot LV 2008: Climate Change Mitigation: A Spatial Analysis of Global Land Suitability for Clean Development Mechanism Afforestation and Reforestation. *Agriculture, Ecosystems & Environment*, 126, 67-80.

APPENDIX 1: LIST OF VARIABLES (DATASETS) USED IN THE STUDY

Variable (Node)	Code	Unit	Type ³	Scale/ resolution	Source
Actual Evapotranspiration -October	aet_1	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - January	aet_10	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - November	aet_11	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - December	aet_12	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - February	aet_2	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - March	aet_3	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - April	aet_4	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - May	aet_5	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - June	aet_6	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - July	aet_7	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - August	aet_8	mm	C	880m	Trabucco and Zomer (2010)
Actual Evapotranspiration - September	aet_9	mm	C	880m	Trabucco and Zomer (2010)
Aridity index	ai_yr	dimensionless	C	880m	Zomer <i>et al.</i> (2008)
Proximity to human-disturbed areas	anthrodist	km	C	880m	Derived from land cover data
Slope aspect	aspect	degrees	C	880m	Derived from digital elevation model
Annual Mean Temperature	bio1	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Mean Temperature of Warmest Quarter	bio10	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Mean Temperature of Coldest Quarter	bio11	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Annual Precipitation	bio12	mm	C	880m	Hijmans <i>et al.</i> (2005)
Precipitation of Wettest Month	bio13	mm	C	880m	Hijmans <i>et al.</i> (2005)
Precipitation of Driest Month	bio14	mm	C	880m	Hijmans <i>et al.</i> (2005)
Precipitation Seasonality	bio15	fraction	C	880m	Hijmans <i>et al.</i> (2005)
Precipitation of Wettest Quarter	bio16	mm	C	880m	Hijmans <i>et al.</i> (2005)

³ C- Continuous, D - Discrete

Variable (Node)	Code	Unit	Type ³	Scale/ resolution	Source
Precipitation of Driest Quarter	bio17	mm	C	880m	Hijmans <i>et al.</i> (2005)
Precipitation of Warmest Quarter	bio18	mm	C	880m	Hijmans <i>et al.</i> (2005)
Precipitation of Coldest Quarter	bio19	mm	C	880m	Hijmans <i>et al.</i> (2005)
Mean Diurnal Range	bio2	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Isothermality	bio3	dimensionless (x100)	C	880m	Hijmans <i>et al.</i> (2005)
Temperature Seasonality	bio4	°C (x100)	C	880m	Hijmans <i>et al.</i> (2005)
Max Temperature of Warmest Month	bio5	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Min Temperature of Coldest Month	bio6	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Temperature Annual Range	bio7	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Mean Temperature of Wettest Quarter	bio8	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Mean Temperature of Driest Quarter	bio9	°C (x10)	C	880m	Hijmans <i>et al.</i> (2005)
Bird species richness	birdrich	count	C	8th degree square	Parker (1994); SANBI (2014)
Bulk density (0-5cm depth)	bld_sd1	kg/m ³	C	880m	ISRIC (2013)
Bulk density (5-15cm depth)	bld_sd2	kg/m ³	C	880m	ISRIC (2013)
Bulk density (15-30cm depth)	bld_sd3	kg/m ³	C	880m	ISRIC (2013)
Bulk density (30-60cm depth)	bld_sd4	kg/m ³	C	880m	ISRIC (2013)
Bulk density (60-100cm depth)	bld_sd5	kg/m ³	C	880m	ISRIC (2013)
Proximity to entry points (border posts)	borddist	km	C	880m	Derived from Afrogeo (2014) data
<i>Rubus spp.</i>	bramble	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
<i>Solanum mauritianum</i>	bugweed	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Cattle density	cattdens	cattle/100 km ²	C	1:50,000	Ministry of Agriculture
Cation exchange capacity (0-5cm depth)	cec_sd1	cmol+/kg	C	880m	ISRIC (2013)
Cation exchange capacity (5-15cm depth)	cec_sd2	cmol+/kg	C	880m	ISRIC (2013)
Cation exchange capacity (15-30cm depth)	cec_sd3	cmol+/kg	C	880m	ISRIC (2013)
Cation exchange capacity (30-60cm depth)	cec_sd4	cmol+/kg	C	880m	ISRIC (2013)
Cation exchange capacity (60-100cm depth)	cec_sd5	cmol+/kg	C	880m	ISRIC (2013)

Variable (Node)	Code	Unit	Type ³	Scale/ resolution	Source
<i>Chromolaena odorata</i>	chromolaen	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Clay fraction (0-5cm depth)	clyppt_sd1	g/kg	C	880m	ISRIC (2013)
Clay fraction (5-15cm depth)	clyppt_sd2	g/kg	C	880m	ISRIC (2013)
Clay fraction (15-30cm depth)	clyppt_sd3	g/kg	C	880m	ISRIC (2013)
Clay fraction (30-60cm depth)	clyppt_sd4	g/kg	C	880m	ISRIC (2013)
Clay fraction (60-100cm depth)	clyppt_sd5	g/kg	C	880m	ISRIC (2013)
Clay fraction (100-200cm depth)	clyppt_sd6	g/kg	C	880m	ISRIC (2013)
Coarse fragments (0-5cm depth)	crfvol_sd1	cm ³ /cm ³	C	880m	ISRIC (2013)
Coarse fragments (5-15cm depth)	crfvol_sd2	cm ³ /cm ³	C	880m	ISRIC (2013)
Coarse fragments (15-30cm depth)	crfvol_sd3	cm ³ /cm ³	C	880m	ISRIC (2013)
Coarse fragments (30-60cm depth)	crfvol_sd4	cm ³ /cm ³	C	880m	ISRIC (2013)
Coarse fragments (60-100cm depth)	crfvol_sd5	cm ³ /cm ³	C	880m	ISRIC (2013)
Coarse fragments (100-200cm depth)	crfvol_sd6	cm ³ /cm ³	C	880m	ISRIC (2013)
Compound Topography Index	cti	dimensionless	C	30m	Derived from digital elevation model
Surface curvature	curvature	radians/m	C	880m	Derived from digital elevation model
Solar radiation duration	dirduratio	dimensionless	C	800m	Zomer <i>et al.</i> , 2008
Solar radiation total/annum	dirradiati	dimensionless	C	800m	Zomer <i>et al.</i> , 2008
Proximity to main electricity line	electdist	m	C	880m	Swaziland Electricity Company
Elevation	elevation	m	C	90m	SRTM (Jarvis <i>et al.</i> , 2008)
<i>Eucalyptus spp.</i>	eucalyptus	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Fire frequency	firefreq	count	C	880m	Derived from Giglio <i>et al.</i> (2009)
Proximity to deforested pixel	flossdist	m	C	800m	Derived from Hansen <i>et al.</i> (2013)
Forest loss patch density	forlosdens	patches/km ²	C	880m	Derived from Hansen <i>et al.</i> (2013)
Goat density	goatdens	goats/ha	C	1:50,000	Ministry of Agriculture
Coefficient of variation (%) of annual precipitation	gpcvapre	dimensionless	C	1km	Schulze (1997)
Annual total number of frost days	gpfrostd	dimensionless	C	1km	Schulze (1997)

Variable (Node)	Code	Unit	Type ³	Scale/ resolution	Source
<i>Psidium guajava</i>	guava	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Heat load index	heatload	dimensionless	C	90m	Derived from digital elevation model
Invasive species richness	iapsrich	dimensionless	C	880m	Derived from Kotze <i>et al.</i> (2010)
Percent of people who use woodfuel for cooking	icook	fraction	C	1:50,000	Central Statistics Office (2011)
Percent of people who have electricity	ielec	fraction	C	1:50,000	Central Statistics Office (2011)
<i>Jacaranda mimosifolia</i>	jacaranda	dimensionless	D	1:10,000	Kotzé <i>et al.</i> (2010)
Land cover	landcov	dimensionless	D	10m	Kotze <i>et al.</i> (2010)
Land use	landuse	dimensionless	D	1:250,000	Remmelzwaal and Dlamini (1994)
<i>Lantana camara</i>	lantana	dimensionless	D	1:10,000	Kotzé <i>et al.</i> (2010)
Land cover diversity	lcs Shannon	dimensionless	C	880m	Derived from land cover data
Geology (Lithostratigraphy)	lithostrat	dimensionless	D	1:250,000	Vegter (1995)
Livestock density	lsu	livestock units/ha	C	1:50,000	Ministry of Agriculture
Proximity to major (perennial) rivers	majrivdist	km	C	880m	Derived from hydrology data
Proximity to major roads	majroaddis	km	C	880m	Derived from roads data
<i>Ceasalpinia decapetala</i>	mauritus	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Disturbance (natural/non-natural)	natural	dimensionless	D	10m	Derived from land cover data
Proximity to natural (non-human disturbed) areas	naturdist	km	C	880m	Derived from land cover data
<i>Opuntia spp</i>	opuntia	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Soil organic carbon content (0-5cm depth)	orcdrc_sd1	g/kg	C	880m	ISRIC (2013)
Soil organic carbon content (5-15cm depth)	orcdrc_sd2	g/kg	C	880m	ISRIC (2013)
Soil organic carbon content (15-30cm depth)	orcdrc_sd3	g/kg	C	880m	ISRIC (2013)
Soil organic carbon content (30-60cm depth)	orcdrc_sd4	g/kg	C	880m	ISRIC (2013)
Soil organic carbon content (60-100cm depth)	orcdrc_sd5	g/kg	C	880m	ISRIC (2013)
Poverty (Headcount) rate	p0	%	C	1:50,000	Central Statistics Office (2011)

Variable (Node)	Code	Unit	Type ³	Scale/ resolution	Source
Protected areas status	pastatus	dimensionless	D	1:50,000	Roques (2002); updated by author
Potential EvapoTranspiration - October	pet_1	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - January	pet_10	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - November	pet_11	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - December	pet_12	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - February	pet_2	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - March	pet_3	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - April	pet_4	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - May	pet_5	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - June	pet_6	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - July	pet_7	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - August	pet_8	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential EvapoTranspiration - September	pet_9	mm	C	880m	Zomer <i>et al.</i> (2008)
Potential evapotranspiration	pet_yr	mm	C	880m	Zomer <i>et al.</i> , 2008
Soil pH in water (0-5cm depth)	phih05_sd1	dimensionless	C	880m	ISRIC (2013)
Soil pH in water (5-15cm depth)	phih05_sd2	dimensionless	C	880m	ISRIC (2013)
Soil pH in water (15-30cm depth)	phih05_sd3	dimensionless	C	880m	ISRIC (2013)
Soil pH in water (30-60cm depth)	phih05_sd4	dimensionless	C	880m	ISRIC (2013)
Soil pH in water (60-100cm depth)	phih05_sd5	dimensionless	C	880m	ISRIC (2013)
Soil pH in water (100-200cm depth)	phih05_sd6	dimensionless	C	880m	ISRIC (2013)
<i>Pinus spp.</i>	pine	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Human population density	popdens	people/km ²	C	100m	Tatem <i>et al.</i> (2013)
<i>Populus x canescens</i>	poplar	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
<i>Cereus jamacaru</i>	queen	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Proximity to rail line	raildist	km	C	880m	Derived from Afrogeo (2014) data
River/stream density	rivdens	km/km ²	C	880m	Derived from Afrogeo (2014) data
Proximity to rivers/streams	rivdist	km	C	880m	Derived from Afrogeo (2014) data

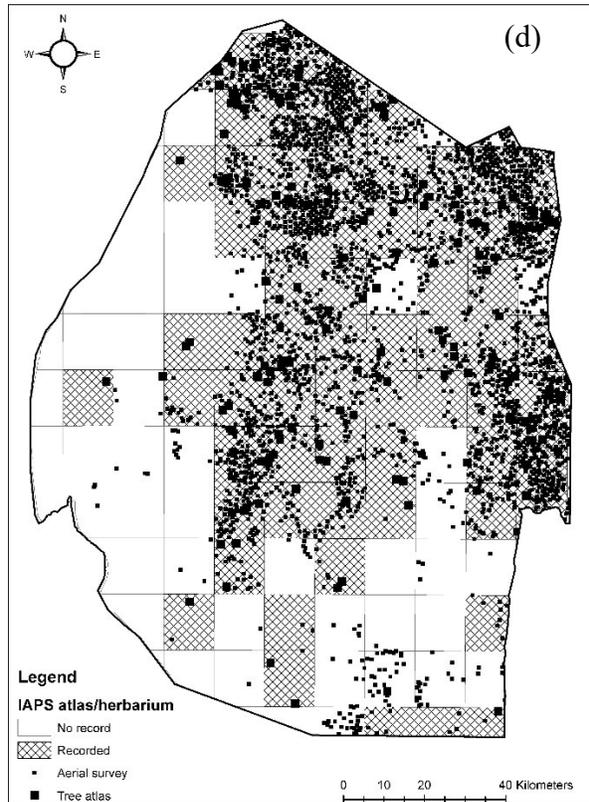
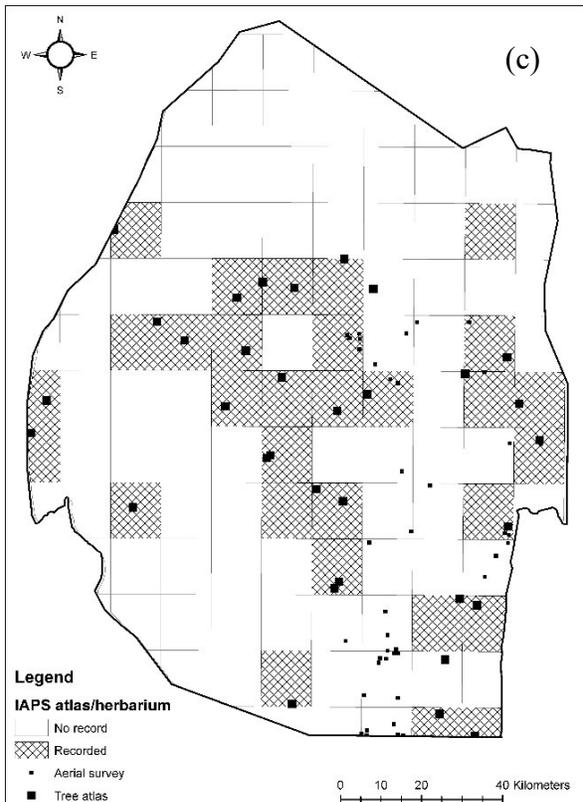
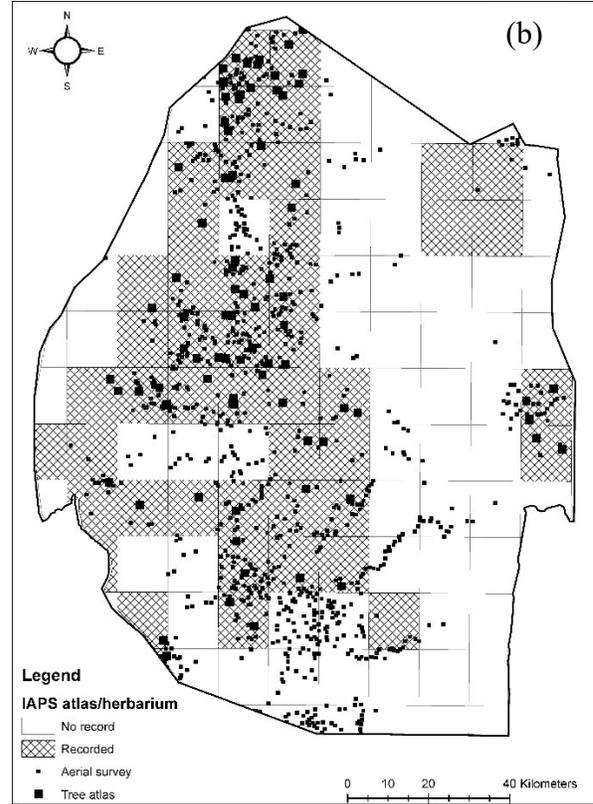
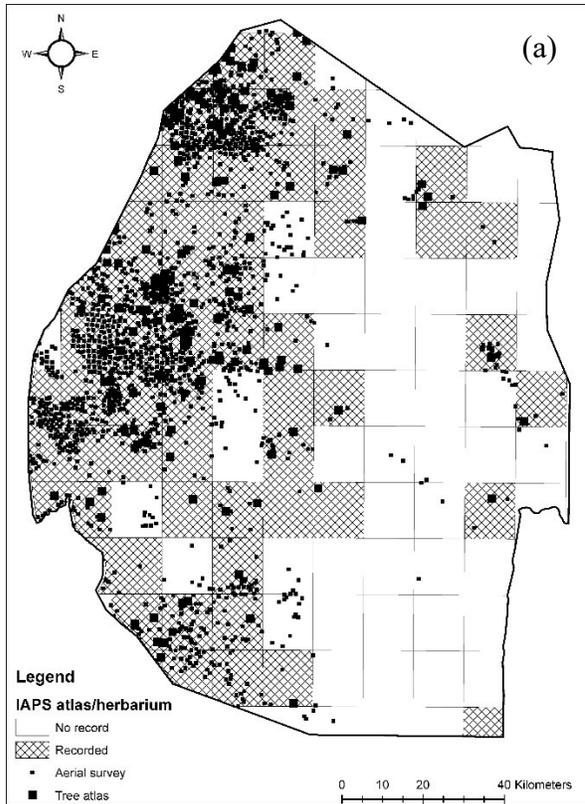
Variable (Node)	Code	Unit	Type ³	Scale/ resolution	Source
Road density	roaddens	km/km ²	C	880m	Derived from Afrogeo (2014) data
Proximity to roads	roaddist	km	C	880m	Derived from Afrogeo (2014) data
Surface roughness	roughness	dimensionless	C	800m	Derived from digital elevation model
<i>Senna didymobotrya</i>	senna	dimensionless	D	1:10000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
<i>Sesbania punicea</i>	sesbania	dimensionless	D	1:10000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Proximity to human settlements	settdist	km	C	880m	Afrogeo (2014)
Human settlement density	settdist	settlement/km ²	C	880m	Afrogeo (2014)
Slope	slope	°	C	30m	Derived from digital elevation model
Silt fraction (0-5cm depth)	sltppt_sd1	g/kg	C	880m	ISRIC (2013)
Silt fraction (5-15cm depth)	sltppt_sd2	g/kg	C	880m	ISRIC (2013)
Silt fraction (15-30cm depth)	sltppt_sd3	g/kg	C	880m	ISRIC (2013)
Silt fraction (30-60cm depth)	sltppt_sd4	g/kg	C	880m	ISRIC (2013)
Silt fraction (60-100cm depth)	sltppt_sd5	g/kg	C	880m	ISRIC (2013)
Silt fraction (100-200cm depth)	sltppt_sd6	g/kg	C	880m	ISRIC (2013)
Sand fraction (0-5cm depth)	sndppt_sd1	g/kg	C	880m	ISRIC (2013)
Sand fraction (5-15cm depth)	sndppt_sd2	g/kg	C	880m	ISRIC (2013)
Sand fraction (15-30cm depth)	sndppt_sd3	g/kg	C	880m	ISRIC (2013)
Sand fraction (30-60cm depth)	sndppt_sd4	g/kg	C	880m	ISRIC (2013)
Sand fraction (60-100cm depth)	sndppt_sd5	g/kg	C	880m	ISRIC (2013)
Sand fraction (100-200cm depth)	sndppt_sd6	g/kg	C	880m	ISRIC (2013)
Landform type	surfform	dimensionless	D	30m	Derived from digital elevation model
Soil Water Content - January	swc_fr_1	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - October	swc_fr_10	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - November	swc_fr_11	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - December	swc_fr_12	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - February	swc_fr_2	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - March	swc_fr_3	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - April	swc_fr_4	mm	C	880m	Trabucco and Zomer (2010)

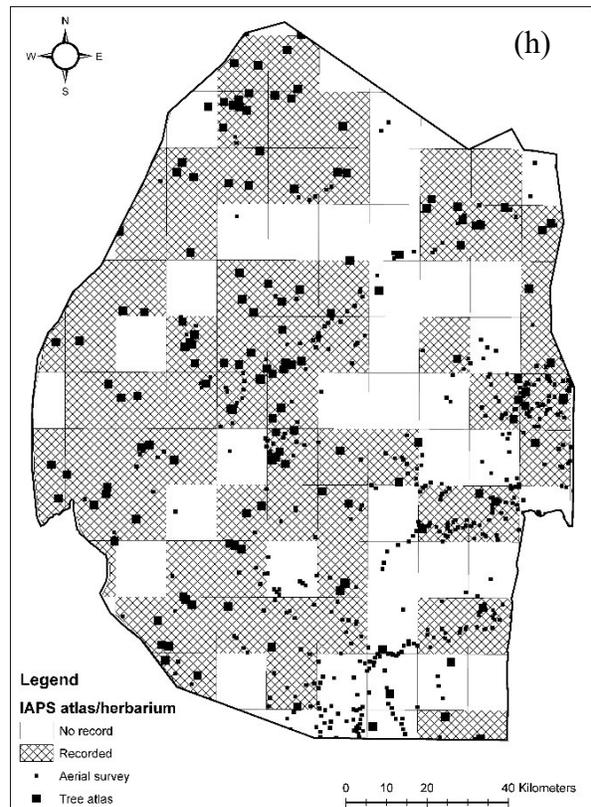
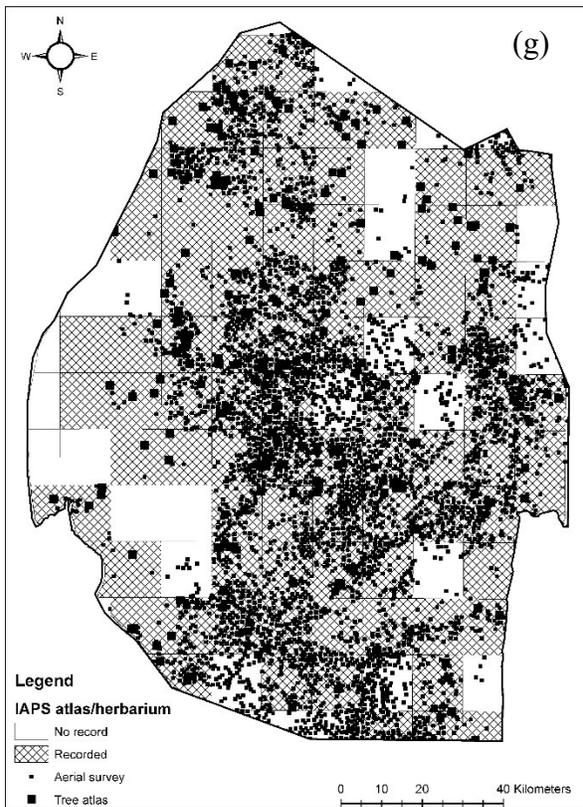
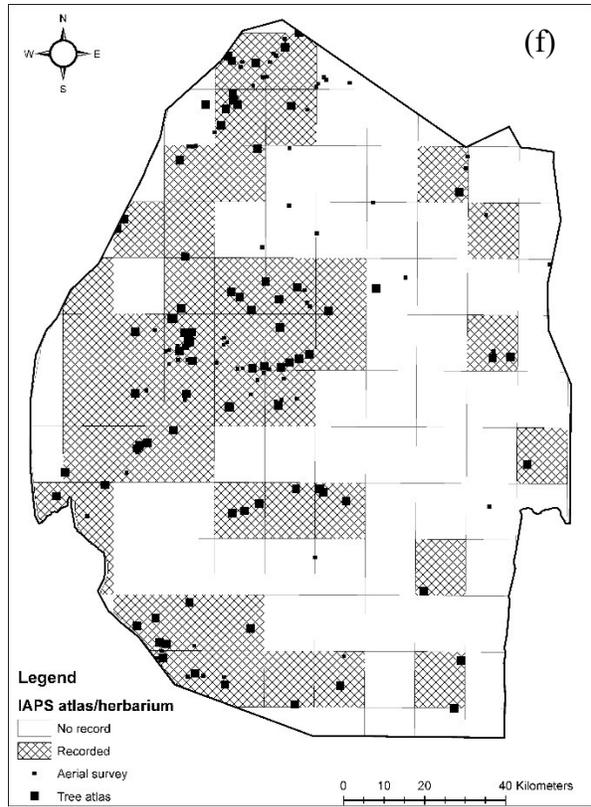
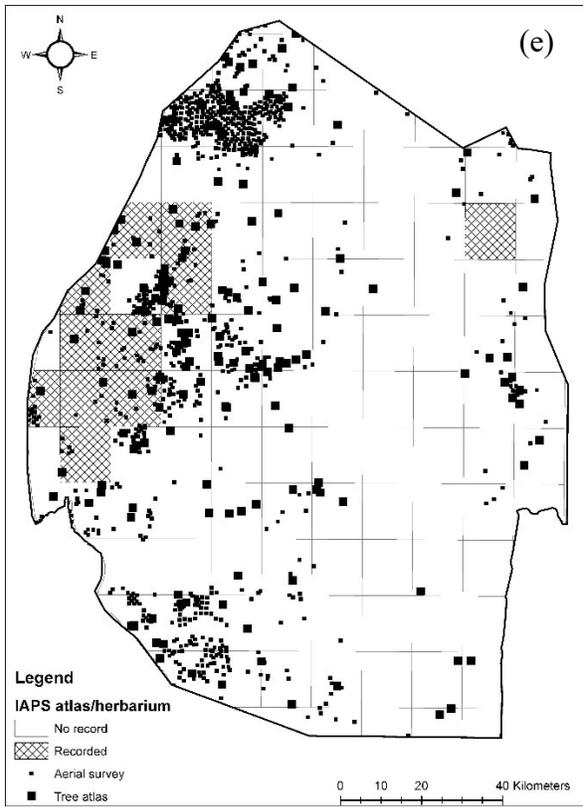
Variable (Node)	Code	Unit	Type ³	Scale/ resolution	Source
Soil Water Content - May	swc_fr_5	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - June	swc_fr_6	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - July	swc_fr_7	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - August	swc_fr_8	mm	C	880m	Trabucco and Zomer (2010)
Soil Water Content - September	swc_fr_9	mm	C	880m	Trabucco and Zomer (2010)
<i>Melia azedarach</i>	syringa	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);
Land tenure	tenure	dimensionless	D	1:250,000	Rommelzwaal and Vilakati (1994)
Proximity to tourism sites	tourdist	m	C	880m	Derived from Afrogeo (2014) data
Proximity to tourism routes	touroutedi	km	C	880m	Derived from Afrogeo (2014) data
Proximity to towns/cities	towndist	km	C	880m	Derived from Afrogeo (2014) data
Topographic position index	tpi	dimensionless	C	30m	Derived from digital elevation model
Tree species richness	treerich	count	C	880m	Dlamini (2015)
Vegetation type	vegetation	dimensionless	D	1:50,000	Dobson and Lotter (2004)
Percentage of people who are waged (employed)	wage	fraction	C	880m	Central Statistics Office (2011)
Proximity to wetlands	waterdist	km	C	880m	Derived from land cover data
<i>Acacia mearnsii</i>	wattle	dimensionless	D	1:10,000	Kotze <i>et al.</i> (2010); Loffler and Loffler (2005);

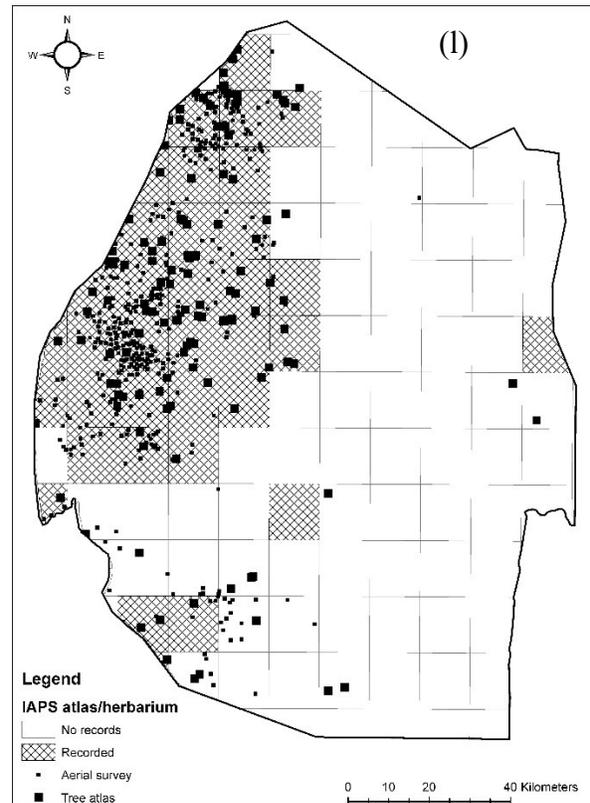
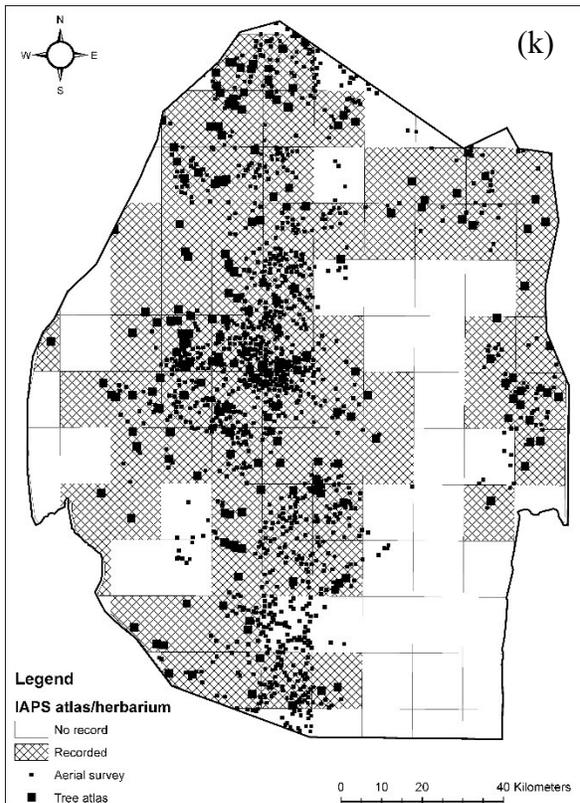
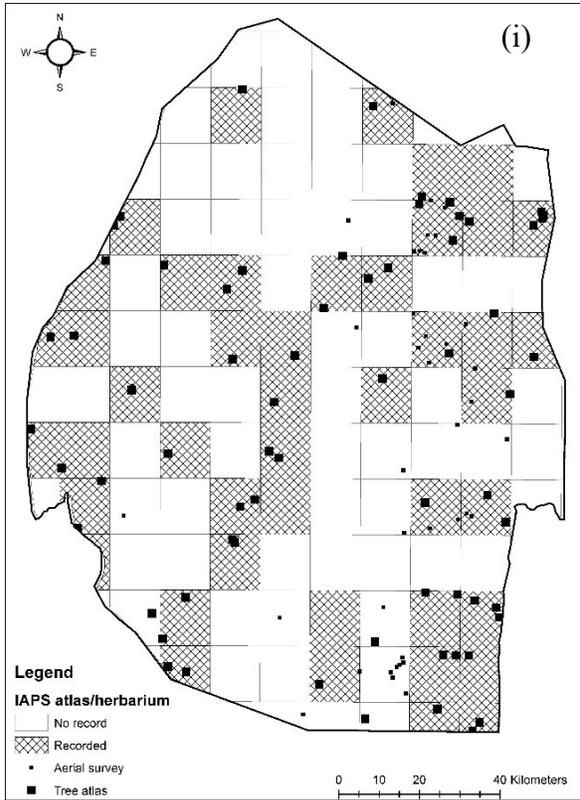
APPENDIX 2: OBSERVED DISTRIBUTION MAPS OF ALL THE SPECIES FROM THE AERIAL SURVEY AND TREE ATLAS DATASETS⁴

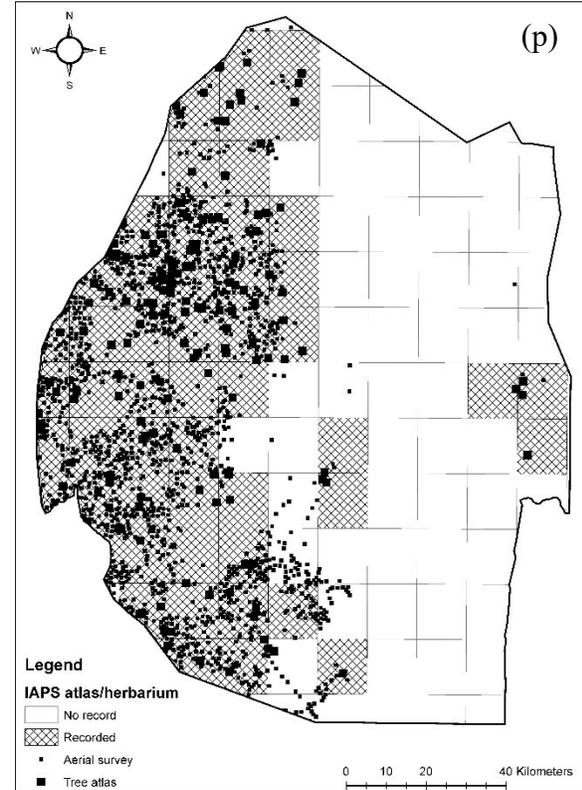
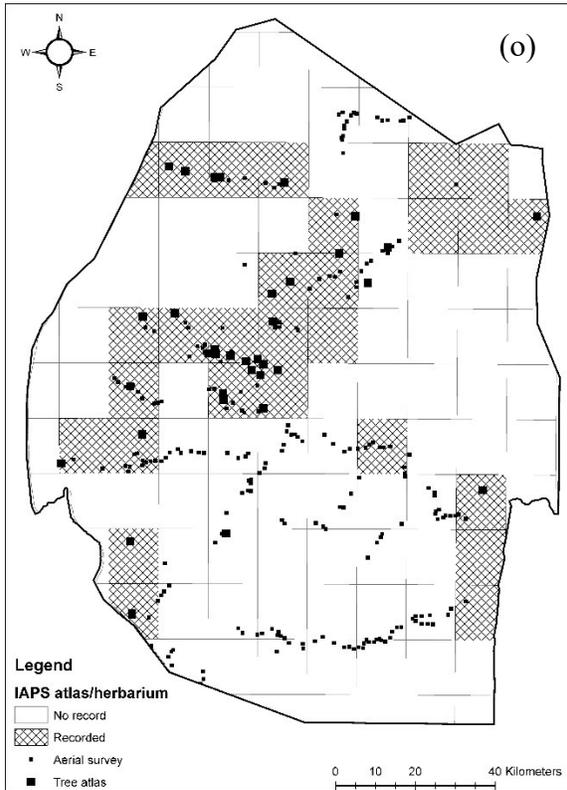
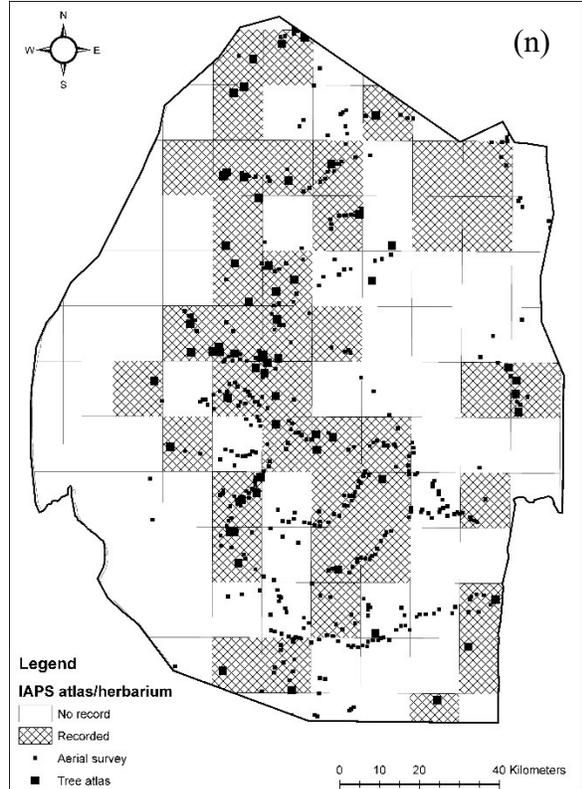
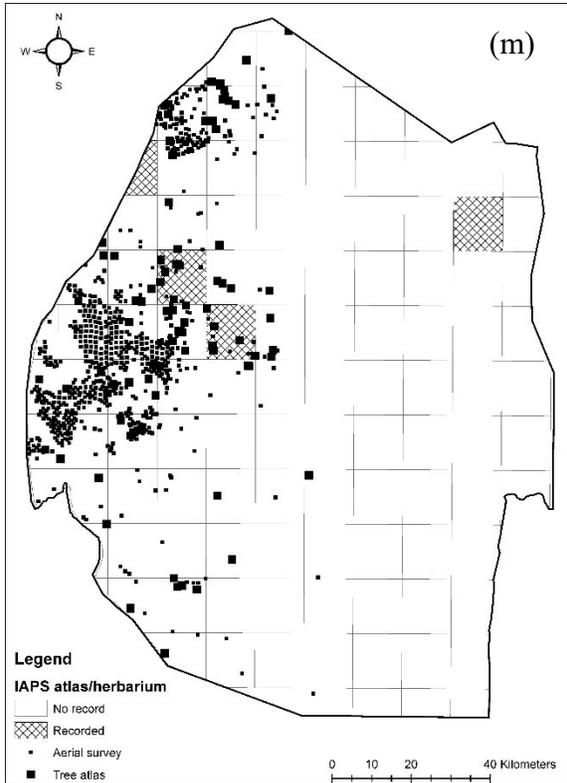
Maps are shown in the following order: (a) *A. mearnsii*, (b) *C. decapetala*, (c) *C. jamacaru*, (d) *C. odorata*, (e) *Eucalyptus spp.*, (f) *J. mimosifolia*, (g) *L. camara*, (h) *M. azedarach*, (i) *Opuntia spp.*, (j) *P. x canescens*, (k) *P. guajava*, (l) *Pinus spp.*, (m) *Rubus spp.*, (n) *S. didymobotrya*, (o) *S. punicea* and (p) *S. mauritianum*.

⁴ Mapped using data from Braun and Dlamini (2005), Kotzé *et al.* (2010b) and Loffler and Loffler (2005).









APPENDIX 3: PERFORMANCE OF THE BAYESIAN LEARNING ALGORITHMS⁵

Performance comparison using the logarithmic loss (Log loss) for all the BN learning algorithms.

	ci	global-hc	global-k2	global-rhc	global-sa	global-tan	global-ts	ics	local-hc	local-k2	local-lagd	local-rhc	local-sa	local-tan	local-ts	nb	Best
bramble	0.539	0.548	0.511	0.529	0.524	0.529	0.545	0.453	0.485	0.474	0.485	0.457	0.478	0.523	0.523	0.536	ics
bugweed	0.669	0.649	0.665	0.641	0.639	0.626	0.639	0.622	0.689	0.674	0.683	0.685	0.681	0.640	0.634	0.667	ics
chromolaena	1.119	1.053	1.115	-	1.008	1.080	1.097	0.966	0.920	1.004	0.931	0.923	0.935	0.999	1.002	1.121	local-hc
eucalyptus	0.846	0.813	0.786	0.840	0.798	0.802	0.793	0.779	0.846	0.821	0.846	0.835	0.838	0.821	0.838	0.844	ics
guava	0.798	0.852	0.781	0.877	0.786	0.793	0.773	0.764	0.824	0.868	0.821	0.807	0.858	0.873	0.815	0.797	ics
jacaranda	0.308	0.265	0.273	0.278	0.250	0.285	0.285	0.270	0.309	0.300	0.307	0.299	0.314	0.305	0.302	0.304	global-sa
lantana	1.085	-	1.166	-	-	1.112	1.102	0.963	0.964	0.995	0.972	0.952	1.007	0.996	0.974	1.085	local-rhc
mauritius	1.104	1.100	1.067	1.100	1.127	1.078	1.073	0.917	0.939	1.010	0.956	0.994	1.037	1.022	0.945	1.104	ics
opuntia	0.374	0.380	0.362	0.393	0.356	0.358	0.354	0.374	0.401	0.399	0.399	0.400	0.409	0.420	0.392	0.372	global-ts
pine	0.701	0.658	0.726	-	0.592	0.641	0.685	0.619	0.538	0.557	0.538	0.522	0.548	0.502	0.525	0.701	local-tan
poplar	0.721	0.803	0.729	0.860	0.840	0.753	0.744	0.795	0.677	0.651	0.765	0.677	0.664	0.610	0.705	0.728	local-tan
queen	0.410	0.383	0.374	0.393	0.370	0.377	0.374	0.463	0.484	0.483	0.484	0.484	0.481	0.480	0.502	0.408	global-sa
senna	0.593	0.560	0.548	0.561	0.564	0.566	0.560	0.610	0.632	0.648	0.643	0.648	0.659	0.675	0.667	0.592	global-k2
sesbania	0.803	0.729	0.792	0.787	0.796	0.782	0.741	0.750	0.647	0.704	0.677	0.698	0.709	0.733	0.713	0.796	local-hc
syringa	0.455	0.457	0.452	0.455	0.488	0.454	0.454	0.677	0.686	0.695	0.686	0.690	0.688	0.597	0.653	0.454	global-k2
wattle	0.733	0.781	0.699	0.781	0.763	0.691	0.778	0.716	0.720	0.718	0.720	0.703	0.717	0.727	0.739	0.734	global-tan

⁵ Shown are the values for the best performing run for each algorithm

Performance comparison using Matthew's correlation coefficient (MCC) for all the BN learning algorithms.

	ci	global-hc	global-k2	global-rhc	global-sa	global-tan	global-ts	ics	local-hc	local-k2	local-lagd	local-rhc	local-sa	local-tan	local-ts	nb	Best
bramble	0.488	0.495	0.478	0.496	0.463	0.490	0.495	0.484	0.502	0.496	0.502	0.474	0.496	0.499	0.498	0.493	local-hc
bugweed	0.507	0.510	0.508	0.508	0.507	0.513	0.505	0.505	0.503	0.516	0.508	0.506	0.513	0.517	0.506	0.507	local-tan
chromolaena	0.384	0.450	0.426	-	0.443	0.398	0.426	0.419	0.423	0.426	0.430	0.431	0.438	0.400	0.395	0.384	global-hc
eucalyptus	0.375	0.387	0.374	0.377	0.383	0.384	0.390	0.343	0.338	0.330	0.338	0.338	0.335	0.339	0.329	0.375	global-ts
guava	0.384	0.384	0.377	0.377	0.381	0.386	0.387	0.353	0.344	0.365	0.344	0.343	0.361	0.358	0.349	0.384	global-ts
jacaranda	0.588	0.602	0.619	0.607	0.598	0.611	0.603	0.588	0.594	0.591	0.583	0.594	0.576	0.583	0.615	0.588	global-k2
lantana	0.433	-	0.438	-	-	0.439	0.432	0.387	0.406	0.444	0.401	0.430	0.425	0.414	0.405	0.433	local-k2
mauritius	0.229	0.227	0.247	0.234	0.237	0.234	0.239	0.224	0.221	0.220	0.238	0.232	0.225	0.216	0.236	0.229	global-k2
opuntia	0.570	0.575	0.589	0.573	0.554	0.586	0.586	0.578	0.591	0.588	0.576	0.574	0.586	0.540	0.602	0.570	local-ts
pine	0.465	0.484	0.449	-	0.471	0.409	0.443	0.374	0.428	0.421	0.428	0.408	0.411	0.438	0.441	0.465	global-hc
poplar	0.094	0.100	0.099	0.102	0.101	0.097	0.096	0.120	0.105	0.107	0.103	0.098	0.095	0.099	0.099	0.094	ics
queen	0.454	0.459	0.455	0.459	0.459	0.459	0.452	0.468	0.475	0.486	0.474	0.470	0.470	0.455	0.464	0.454	local-k2
senna	0.472	0.480	0.484	0.480	0.483	0.478	0.481	0.455	0.472	0.474	0.472	0.475	0.478	0.478	0.470	0.472	global-k2
sesbania	0.203	0.214	0.205	0.201	0.204	0.196	0.220	0.204	0.194	0.202	0.194	0.198	0.186	0.192	0.206	0.202	global-ts
syringa	0.697	0.696	0.696	0.699	0.699	0.696	0.696	0.599	0.604	0.556	0.553	0.599	0.602	0.696	0.627	0.697	global-rhc
wattle	0.413	0.438	0.427	0.431	0.429	0.430	0.425	0.423	0.437	0.425	0.437	0.434	0.424	0.407	0.414	0.413	global-hc

Performance comparison using area true skill statistic (TSS) for all the BN learning algorithms.

	ci	global-hc	global-k2	global-rhc	global-sa	global-tan	global-ts	ics	local-hc	local-k2	local-lagd	local-rhc	local-sa	local-tan	local-ts	nb	Best
bramble	0.860	0.862	0.850	0.858	0.849	0.858	0.860	0.874	0.856	0.853	0.867	0.854	0.852	0.862	0.865	0.860	ics
bugweed	0.761	0.753	0.745	0.753	0.773	0.745	0.751	0.750	0.775	0.769	0.778	0.776	0.782	0.767	0.777	0.766	local-sa
chromolaena	0.596	0.617	0.608	-	0.637	0.596	0.579	0.631	0.594	0.598	0.584	0.601	0.631	0.573	0.551	0.595	global-sa
eucalyptus	0.734	0.750	0.723	0.743	0.747	0.728	0.741	0.693	0.688	0.675	0.703	0.691	0.707	0.687	0.691	0.734	global-hc
guava	0.659	0.642	0.626	0.631	0.631	0.620	0.623	0.625	0.584	0.640	0.611	0.596	0.617	0.620	0.588	0.659	ci
jacaranda	0.906	0.919	0.909	0.931	0.941	0.918	0.917	0.929	0.921	0.906	0.926	0.938	0.927	0.911	0.930	0.906	global-sa
lantana	0.507	-	0.517	-	-	0.511	0.523	0.480	0.529	0.527	0.526	0.511	0.519	0.503	0.484	0.507	local-hc
mauritius	0.559	0.589	0.578	0.587	0.602	0.562	0.571	0.626	0.573	0.584	0.596	0.561	0.550	0.567	0.589	0.560	ics
opuntia	0.917	0.931	0.933	0.930	0.920	0.933	0.933	0.925	0.934	0.927	0.925	0.927	0.933	0.908	0.933	0.917	local-hc
pine	0.854	0.869	0.843	-	0.874	0.867	0.840	0.849	0.810	0.824	0.810	0.819	0.815	0.849	0.826	0.854	global-sa
poplar	0.878	0.862	0.878	0.870	0.860	0.854	0.877	0.945	0.864	0.867	0.859	0.865	0.857	0.870	0.873	0.872	ics
queen	0.906	0.905	0.883	0.905	0.883	0.885	0.907	0.865	0.888	0.891	0.888	0.887	0.887	0.883	0.885	0.905	global-ts
senna	0.806	0.811	0.821	0.811	0.811	0.811	0.821	0.820	0.810	0.826	0.810	0.818	0.818	0.828	0.809	0.806	local-tan
sesbania	0.780	0.791	0.783	0.788	0.826	0.820	0.790	0.826	0.778	0.827	0.778	0.806	0.819	0.822	0.774	0.781	local-k2
syringa	0.783	0.804	0.804	0.783	0.825	0.804	0.804	0.822	0.822	0.809	0.808	0.821	0.820	0.819	0.804	0.783	global-sa
wattle	0.677	0.707	0.693	0.700	0.702	0.688	0.700	0.701	0.706	0.695	0.706	0.689	0.705	0.663	0.666	0.677	global-hc

Performance comparison using the area under the ROC curve (AUC) for all the BN learning algorithms.

	ci	global-hc	global-k2	global-rhc	global-sa	global-tan	global-ts	ics	local-hc	local-k2	local-lagd	local-rhc	local-sa	local-tan	local-ts	nb	Best
bramble	0.969	0.969	0.971	0.969	0.969	0.971	0.971	0.962	0.971	0.972	0.970	0.970	0.971	0.974	0.970	0.969	local-tan
bugweed	0.932	0.937	0.932	0.937	0.938	0.933	0.934	0.930	0.934	0.930	0.934	0.931	0.933	0.931	0.933	0.932	global-sa
chromolaena	0.832	0.885	0.863	-	0.881	0.852	0.854	0.846	0.863	0.863	0.865	0.869	0.867	0.856	0.849	0.832	global-hc
eucalyptus	0.914	0.922	0.917	0.922	0.919	0.917	0.916	0.892	0.915	0.913	0.915	0.915	0.922	0.911	0.907	0.914	global-hc
guava	0.885	0.889	0.890	0.894	0.892	0.887	0.890	0.882	0.884	0.891	0.884	0.885	0.883	0.890	0.891	0.885	global-rhc
jacaranda	0.992	0.994	0.992	0.993	0.994	0.992	0.992	0.991	0.991	0.992	0.992	0.993	0.992	0.993	0.992	0.992	global-sa
lantana	0.817	-	0.824	-	-	0.826	0.816	0.807	0.821	0.828	0.821	0.822	0.830	0.822	0.816	0.817	local-sa
mauritius	0.861	0.867	0.869	0.875	0.870	0.869	0.867	0.834	0.875	0.868	0.875	0.870	0.875	0.869	0.852	0.860	global-rhc
opuntia	0.970	0.976	0.975	0.975	0.975	0.975	0.974	0.974	0.971	0.975	0.975	0.975	0.970	0.976	0.971	0.970	local-tan
pine	0.971	0.974	0.971	-	0.975	0.966	0.971	0.950	0.969	0.968	0.969	0.970	0.966	0.971	0.972	0.971	global-sa
poplar	0.982	0.972	0.981	0.972	0.953	0.981	0.971	0.976	0.976	0.975	0.961	0.976	0.970	0.976	0.982	0.983	nb
queen	0.963	0.963	0.960	0.962	0.965	0.965	0.966	0.962	0.962	0.962	0.962	0.962	0.963	0.960	0.957	0.963	global-ts
senna	0.939	0.945	0.947	0.945	0.941	0.944	0.945	0.938	0.944	0.942	0.940	0.939	0.938	0.940	0.941	0.940	global-k2
sesbania	0.953	0.955	0.950	0.955	0.954	0.957	0.956	0.930	0.959	0.955	0.958	0.959	0.958	0.959	0.952	0.951	local-hc
syringa	0.918	0.919	0.919	0.919	0.919	0.918	0.917	0.915	0.917	0.918	0.917	0.918	0.919	0.920	0.921	0.918	local-ts
wattle	0.914	0.921	0.919	0.923	0.921	0.924	0.921	0.923	0.925	0.924	0.925	0.924	0.923	0.922	0.922	0.914	local-hc

Computation time (in seconds) for all the BN learning algorithms.

	ci	global-hc	global-k2	global-rhc	global-sa	global-tan	global-ts	ics	local-hc	local-k2	local-lagd	local-rhc	local-sa	local-tan	local-ts	nb	Best
bramble	4.266	44.156	6.891	345.859	433.156	11.547	24.422	4.609	3.469	3.359	4.281	5.516	9.719	3.344	3.453	3.344	local-tan
bugweed	10.344	223.328	21.938	1349.141	1136.797	42.297	79.531	12.578	11.516	9.766	11.984	15.516	22.813	9.234	9.656	10.406	local-tan
chromolaena	15.313	262.219	63.203	-	233.094	54.297	109.141	16.563	15.422	14.438	21.203	210.328	118.391	13.891	14.344	13.531	nb
eucalyptus	3.984	86.188	8.625	545.328	487.141	15.234	32.703	4.969	3.906	3.750	5.063	7.063	12.297	3.750	3.750	3.750	local-k2
guava	8.688	106.094	14.422	547.188	795.813	20.281	43.281	9.359	8.484	8.344	9.516	13.688	22.469	8.094	8.453	8.609	local-tan
jacaranda	2.359	44.188	5.875	352.250	359.516	11.313	23.141	3.563	2.578	2.469	3.688	4.734	8.891	2.500	2.578	2.266	nb
lantana	27.297	-	115.813	-	-	154.422	291.641	32.656	30.672	28.078	39.109	647.203	133.609	26.906	27.734	25.891	nb
mauritius	3.563	71.484	7.875	424.594	466.266	14.250	30.172	4.625	3.781	3.797	4.828	6.672	9.938	3.609	3.641	3.500	nb
opuntia	3.031	26.500	3.844	165.891	262.188	5.875	13.328	3.203	1.922	1.781	2.281	2.641	5.000	1.766	1.859	2.016	local-tan
pine	2.891	163.203	11.344	-	121.781	24.797	44.906	4.609	3.313	2.984	5.281	565.125	20.359	2.891	2.984	2.953	ci
poplar	0.375	3.563	0.578	30.344	23.672	1.016	1.813	1.563	0.344	0.344	0.641	1.656	1.906	0.297	0.359	0.328	local-tan
queen	2.156	21.156	2.594	127.281	174.797	4.297	9.688	2.156	1.266	1.156	1.625	1.875	3.531	1.172	1.141	1.156	local-ts
senna	3.375	19.000	5.375	125.516	361.203	7.828	16.469	3.703	4.016	3.797	4.297	4.813	9.266	3.563	3.625	3.172	nb
sesbania	0.844	31.859	2.469	212.391	160.547	5.172	10.344	2.063	0.938	0.797	1.328	4.688	3.922	0.938	1.000	0.828	local-k2
syringa	5.859	60.922	9.375	378.547	572.906	14.406	30.688	6.047	5.141	5.047	6.047	6.906	12.266	4.891	5.109	4.844	nb
wattle	8.719	104.563	15.250	569.766	886.078	22.813	48.969	9.813	9.813	8.781	9.938	12.438	21.594	8.891	8.922	8.672	nb

APPENDIX 4: VARIABLES SELECTED TO FORM THE MARKOV BLANKET OF ALL THE SPECIES' DISTRIBUTION MODELS⁶

Variable	Anthropogenic	Bioclimatic	Biotic	Topo-edaphic
bugweed			5	
jacaranda			5	
senna			5	
bio6		4		
iapsrich			4	
lclshannon	4			
opuntia			4	
rivdist				4
syringa			4	
wattle			4	
aspect				3
bio15		3		
eucalyptus			3	
guava			3	
majroaddis	3			
mauritus			3	
popdens	3			
queen			3	
sesbania			3	
settdens	3			
bramble			2	
curvature				2
gpfrostd		2		
landuse	2			
lantana			2	
majrivdist				2
poplar			2	
rivdens				2
tourdist	2			
treerich			2	
waterdist				2
aet_11		1		
aet_4		1		
aet_5		1		
aet_8		1		
ai_yr		1		

⁶ The selected variables were within the Markov blanket of all the species; the number of species for which the variable was selected is shown within the highlighted (grey) boxes.

Variable	Anthropogenic	Bioclimatic	Biotic	Topo-edaphic
anthrodist	1			
bio11		1		
bio13		1		
bio17		1		
bio4		1		
bio7		1		
bio8		1		
bld_sd3				1
bld_sd4				1
bld_sd5				1
cattdens	1			
chromolaen			1	
clyppt_sd2				1
crfvol_sd4				1
dirduratio				1
electdist	1			
firefreq	1			
ielec	1			
landcov	1			
p0	1			
pet_2		1		
pet_3		1		
pine			1	
roaddens	1			
sltppt_sd4				1
sltppt_sd5				1
sndppt_sd2				1
sndppt_sd4				1
surfform				1
swc_fr_8		1		
touroutedi	1			
tpi				1
Total number of variables	15	17	18	18
Total frequency of selection	26	23	56	27

APPENDIX 5: PHOTOGRAPHS SHOWING ALIEN PLANT INVASION IN SWAZILAND⁷



S. mauritianum, *Pinus* and *Eucalyptus* species co-occurring within an urban environment (left), and *Eucalyptus* along a road (right)

⁷ Ground photos taken by author; aerial photos are from the report by Kotze *et al.* (2010a).



A *M. azeradach* tree as shade in a homestead (left) and *M. azedarach* near human settlements



C. jamaecaru infestation around a human settlement (left) and a *S. mauritianum* and *L. camara* invasion (right).



A *C. odorata* invasion in disturbed areas



Opuntia near an abandoned human settlement (left) and an existing settlement (right)



A *P. canescens* infestation (left) and a *Pinus* species along a river (right)



A. *P. guajava* infestation (left) and a *S. didymobotrya* infestation near a water source (right)

GLOSSARY

Arc (or edge): a representation (in the form of a line) of a conditional statistical dependence between a pair of nodes in a Bayesian network.

Bayesian network (or Bayesian belief network or graphical model): a model visually representing the joint probability distribution of a set of random variables by means of a directed acyclic graph and conditional probability distributions for each node in the graph.

Conditional probability: the probability of an event happening given that some event has already occurred.

Data mining: the process of using analytical tools to discover non-obvious valuable patterns from a large collection of data.

Directed acyclic graph: a set of nodes and directed edges/arcs, which does not contain any cycle (i.e. it is not possible to get from one node back to itself, when following the directed edges/arcs).

Directed edge/arc: an arc or edge with specified direction, which represents causal relationship between two connected nodes.

Discretization: the process of dividing a continuous data geometry into finite elements or discrete categories.

Feature selection: a process by which the most useful subset of useful features are chosen from a large number of predictors in order to find a good predictive model for some phenomenon of interest.

Geographic information system (GIS): an organized collection of computer hardware, software, data, and personnel designed to efficiently capture, store, update, manipulate, analyze, and display all forms of geographically referenced information.

Knowledge discovery: a term often used interchangeably with data mining which denotes the derivation of rules, patterns, and decisions from models derived solely from data.

Likelihood function: a retrospective probability of the observed data.

Machine learning: a branch of artificial intelligence that deals with methods of data analysis that automate analytical model building and learning.

Markov blanket: a set of nodes consisting of a node's parents, its children, and any other parents of its children.

Minimum Description Length (MDL) Principle: the notion that the least complex predictive model (with acceptable accuracy) will be the one that best reflects the true underlying model and performs most accurately on new data.

Parameter learning: the process of learning the parameters (probabilities) of the conditional probability tables from data

Posterior distribution: a probability distribution of a random variable composed of the prior distribution and the likelihood function of the data.

Prior distribution: a probability distribution assigned to a random variable before the incorporation of data.

Sensitivity analysis: a process that determines the sensitivity of a predictive model to small fluctuations in predictor value.

Species distribution model (or ecological niche model): associative model relating occurrence or abundance data at known locations of individual species (distribution data) to information on the environmental characteristics of those locations.

Structure learning: a method of automatic construction of a Bayesian network from a database using an appropriate software.

Uncertainty: imperfect knowledge.

Visualization: graphical display of data and models which helps the user in understanding the structure and meaning of the information contained in them.