Comparison of Bagged Multivariate Adaptive Regression Splines (MARS) and Boosted Regression Trees (BRT) using spatial autocorrelation for modelling the habitat niche of four tree species in the Mount Lofty Ranges, South Australia

Yue Zhuo

Master Degree of Geospatial Information Science

School of the Environment

Faculty Science and Engineering

Flinders University of South Australia

Submitted on 22th June 2017

ABSTRACT

The habitat niche of four Australia native trees in the Mount Lofty Ranges of South Australia have been modelled: *Allocasuarina verticillata*, *Eucalyptus fasciculosa*, *Eucalyptus goniocalyx* and *Eucalyptus obliqua*. Two non-parametric modelling techniques have been compared: bagged Multivariate Adaptive Regression Splines (MARS - Friedman 1991) and Boosted Regression Trees (BRT - Freund & Schapire 1996) modelling. Each regression model was conducted using presence/absence data and a set of 34 environmental variables. Among these predictors, climate data, including rainfall and temperature variables, were found to be most contributed determinants of the distribution of selected trees.

In order to combat spatial autocorrelation, a method was used to re-sample the data, separating the distances between sample points. This was compared with an entirely different method where an index of spatial autocorrelation was explicitly incorporated in each model. Spatial weights set did not contribute to bagged MARS models while it slightly altered the structure of BRT models. However, spatial variable was not a crucial predictor and it cannot significantly affect the response (occurrence of trees).

The performance of each model was evaluated through the Area Under Curve (AUC) values of Receiver Operator Characteristic (ROC) analysis. Generally, all models in this study performed well. However, BRT models had better fit of data than bagged MARS algorithm and shown relatively stable prediction results.

As the sample size of data was limited, not enough data could be set aside for independent testing. Instead, final prediction surfaces of habitat niche were compared with an expert opinion approach undertaken by the Department of Environment, Water and Natural Resources, South Australia.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Signed.....

Date.....

ACKNOWLEDGEMENTS

This thesis research is a summary of my master's degree study at Flinders University. During the past two years, I paid a lot of hard work on graduate study. Master's degree research is difficult and full of challenges, but with the care, support and help from my teachers and classmates, the road to scientific research has gradually become smooth.

I would like to extend my most sincere thanks to all the teachers who taught me during my study these years; especially my supervisor, Dr. Mark Lethbridge, I would like to express my appreciation to him for all his help at my thesis work, he guided me patiently, give beneficial suggestion and comments. Furthermore, I want to say thanks to my friends who give me care and support over these years. I also am grateful to my relatives in China, for their unselfish supports and infinite love. Finally, I would like specially to say thank to my husband for his accompany with me in Australia, also for his contribution to our family!

CONTENTS

APPENDIX

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AUC	Areas Under the Curve
BRT	Boosted Regression Trees
CAD	Computer Aided Drafting
CITES	Convention on International Trade in Endangered Species
DEH	Department for Environment and Heritage of South Australia
DEWNR	Department of Environment, Water and Natural Resources
FLB	Flinders Lofty Block
GAMs	Generalized Additive Models
GIS	Geographical Information Systems
GLMs	Generalized Linear Models
MARS	Multivariate Adaptive Regression Splines
MLR	Mount Lofty Ranges
NND	Nearest Neighbour Distances
PA	Presence and Absence
PIRSA	Primary Industries and Resources of South Australia
ROC	Receiver Operating Characteristic
SA	South Australia
SD	Sample Distance
SDMs	Species Distribution Models

CHAPTER ONE: INTRODUCTION

1.1 Introduction

According to Wilson's (1999) analysis, the diversity of life, represented by the total number of different organisms in the world, is between 2 and 100 million. Unfortunately, by 2000, approximately one quarter of bird species would become extinct, and only 8% of terrestrial vegetation will remain (Wilson 1999). Moreover, the survival of 44% existing tree species is still threatened, tropical and temperate forests are likely to continue to reduce by 1% to 4% each year (Chen 2011; Wilson 1999). Thus, the protection of species' diversity and the sustainable use of biological resources are two key issues for the survival and future development of humans (Franco 2013; Vold & Buffett 2000). Trees are a significant component of species' diversity. They are of central importance to the survival of all animal species as they are the primary producers of food; trees also produce oxygen and define habitats (Elliot & Jones 1990; Szabo et al. 2011). Their conservation is thus central for the future of a functioning and sustainable environment. However, most of the forests in South Australia (SA) were cleared during the 19th and early 20th centuries (Szabo et al. 2011). As a result, little of the original vegetation which once covered SA still survives due to the extensive clearing for agricultural purposes (Bradshaw 2012; Laut et al. 1977). This is particularly true in the Mount Lofty Ranges (MLR), where only less than 10% of the original plants remain (Paton & O'Connor 2010). Therefore, understanding how trees will respond to the environmental variables is fundamental for further re-cultivation projects in this region (Bradshaw 2012). Trees' living preference can also be used to predict potential habitat niche. This is essential for improving the management of native trees in SA (Bradshaw 2012).

Species' habitat niche is fundamental to species' management and the conservation of biodiversity; thus, it is one of the most important concepts in ecology (De Oliveira et al. 2005; Elith et al. 2006; Folke et al. 1996). To this end, Species Distribution Models (SDMs) have been widely used in conservation biology, bio-geography and other disciplines (Ahmed et al. 2015; Guisan et al. 2013; Caswell 1987). The development of computers and Geographical Information Systems (GIS) has also laid the technical foundation for ecological studies over geographical space, and has been a framework for providing a variety of algorithms for developing SDMs (Guisan & Thuiller 2005; Hijmans & Elith 2013). An SDM is based on the relationship between species' occurrence data (presence/absence or abundance) and environmental data (Elith et al. 2006). SDMs are then used to explore trees' environmental requirements and predict environmental suitability or the realised habitat niche of the species (Guisan et al. 2013). There are several notable SDMs which include the Bio-climatic Models, Ecological Niche Models, Climatic Envelope Models and Habitat Models (Ahmed et al. 2015; Elith et al. 2006). At present, these models have been extensively studied and widely applied, including the assessment of species' distribution and abundance under disturbance, predicting the potential direction for invasion species, supporting the development of conservation and the selection of suitable reserve, exploring distribution in changing environment and so on (Elith et al. 2006; Elith & Leathwick 2009; Guisan & Thuiller 2005; Kong 2015).

1.2 Habitat models

Habitat models can estimate the relationship between species' locations and their environmental features such as soil types, elevation, rainfall level and so on (Elith et al. 2011; Franklin, 2009; Boulesteix et al. 2012). A suitable habitat model can help to identify species' environmental requirements and also be used to predict tree's distributions by projecting the model to an unknown geographical area (Elith & Leathwick 2009; Gutiérrez et al. 2009). A wide variety of habitat models' techniques have been developed to express the

typical characteristics of ecological data (Austin 2002). These contain parametric approaches including Generalized Linear Models (GLMs - Hilborn & Mangel 1997); and nonparametric approaches such as classification trees (Lim et al. 2000), regression trees (Breiman et al. 1984), Generalized Additive Models (GAMs - Austin, 2002) and Multivariate Adaptive Regression Splines (MARS - Friedman 1991).

At the beginning of the 1970s, Nelder and Wedderburn proposed the use of GLMs, which include linear regression and logistic regression in statistical learning (Hilborn & Mangel 1997; James et al. 2014; Ball & Brunner 2010). GLMs are flexible in practice as they were designed as means of assimilating the properties of other statistical models (Hilborn & Mangel 1997; McCallum 2000). Since then, GLMs have become useful analytical tools in the fields of social science, biology, medicine, and so on (Hilborn & Mangel 1997; Nakagawa & Cuthill 2007). GLMs provide a cohesive overview of linear normal, categorical and survival models (Miller & Franklin 2002; Vittinghoff et al. 2011). The general form of a GLM function is shown in **Equation 1**. Here $g(y)$ is a link function relating the combination of explanatory variables (x_1, x_2, … x_n) to the mean of the response. β_0, β_1, β_2, … β_n are the parameters being estimated and ε is the stochastic term.

$$g(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ ... + \beta_n x_n + \varepsilon$$

Equation 1

However, candidate predictors (explanatory variables) may have a more complex relationship with the response variable (Cutler et al. 2007). For example, despite the fact that the link function $g(y)$ in GLMs can deal with a non-normal error structure, GLMs are not able to deal with non-linear multi-modal relationships (Hilborn & Mangel 1997; McCallum 2000). This instead requires the use of non-parametric models (Smyth 1989; Dobbertin & Biging 1998).

Breiman et al. (1984) proposed classification trees and regression trees (Lim et al. 2000). These tree-based models apply recursive binary segmentation to divide sample data into two subsets (Skidmore et al. 1996). Formulated specifically for dependent variables with unordered values of a finite number, classification trees calculate prediction error based on misclassification cost (Lim et al. 2000; Rokach & Maimon 2014). Classification trees generally deal with categorical response variables and can be applied in a diverse range of fields and contexts (Lim et al. 2000; Rokach & Maimon 2014). Regression trees are designed for dependent variables with discrete values that are ordered or continuous. For regression trees, the squared difference between values that are observed and predicted is generally used to calculate prediction error (Breiman et al. 1984; Lemon et al. 2003; Loh 2014; Rokach & Maimon 2014). An example of a regression tree is shown in Figure 1. Regression tree recursive partitioning generates a set of terminal nodes (N), and each node applies a GLM equation to explain the remaining variation in the response data (Hastie et al. 2001; Moisen & Frescino 2002).

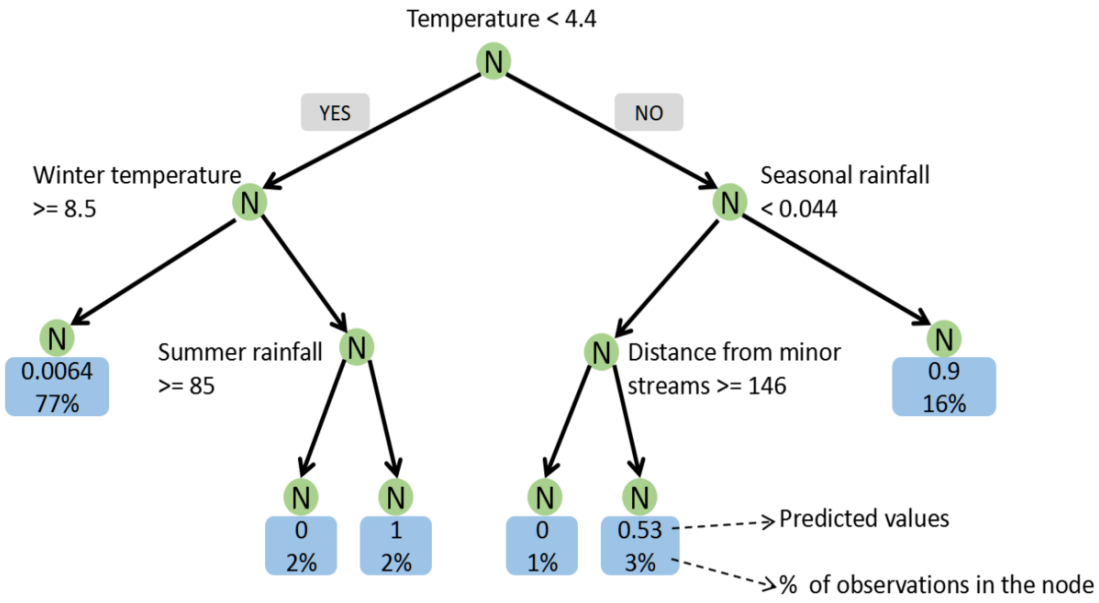


Figure 1: Regression tree example with a continuous response.

Regression trees have been widely used in business, particularly in data mining or in the field of finance for the purposes of credit scoring (Rokach & Maimon 2014; Loh 2011; Chu 2014; Sumathi & Sivanandam 2006). Regression trees can predict the likelihood of customers defaulting and the level of acceptable risk associated with credit (Breiman et al. 1984; Loh 2014). Given that regression trees employ GLMs, they suppose that the terminal node data have a linear relationship with the response (or the transformed response) variable. However, this is not always the case. In this situation, non-linear models are extended to Generalized Linear Models (GAMs - Austin 2002; Bolker et al. 2009).

GAMs are one of the valuable methods for the analysis of the potential geographical distribution of ecological species (Austin 2002; He et al. 2008; Yee & Mitchell 1991). As they allow each predictor variable to be related to the response variable using a nonlinear and non-parametric basis function, meanwhile GAMs also maintain the overall 'additivity' (Austin 2002; Guisan et al. 2002; Lehmann et al. 2002a; Nyström Sandman 2011). Hidden patterns in data can be isolated using GAMs, and predictor functions of GAMs can be regularized to prevent overfitting (Yee & Mitchell 1991). GAMs are often applied in instances when the statistician has no predefined reason for selecting a specific response function and aims to let the data speak for itself (Guisan et al. 2002). The general form of a GAM function is displayed in **Equation 2**, which has the same model structure as that of a GLM function (**Equation 1**). It should be noted that, the estimated parameters ($\beta_0,\ \beta_1,\ \beta_2,\ \dots\ \beta_n$) of GLMs are replaced by univariate spline functions ($f_1,\ f_2,\ \dots\ f_n$) corresponding to each explanatory variable ($x_1,\ x_2,\ \dots\ x_n$).

$$g(y) = \beta_0 + f_1(x_1) + f_2(x_2) + \ \dots + f_n(x_n) + \varepsilon$$

Equation 2

However, GAMs cannot deal with complex interactions other than additive effects (univariate splines) without the brute-force testing of predictor interactions using the multiplication of more than one predictor/response spline (Austin 2002).

Unlike GAMs, Multivariate Adaptive Regression Splines (MARS - Friedman 1991) algorithms test both multivariate and univariate splines up-front in the learning process (Austin 2002). MARS begin by classifying explanatory variables into several regions (Elith & Leathwick 2007; Friedman 1991). The second step involves fitting piecewise linear basis functions (called basis functions) to each region (Friedman 1991). The regions are separated by 'knots' which are then pruned to remove basis functions that make no contribution to the model fit (Gutiérrez et al. 2008). These processes are repeated until there is a parsimonious number of basis functions in the final MARS model (Elith & Leathwick 2007). The order of basis functions is evaluated to select the MARS model with the best predictive fit of data. For example, the **earth** (Milborrow 2013) package in R software (R Core Team 2014) is developed to build a regression model using MARS function (Steinberg et al. 1999). And then the **earth** (Milborrow 2013) package selects the most suitable model through Generalized Cross Validation (GCV - Craven & Wahba 1979; Friedman 1991). The general form of a MARS function is shown in **Equation 3**. Here $s_{ij}(x_i, x_j, \ldots)$ is a general form of a basis function. And $ij\ldots$ or $ijk\ldots$ subscripts denote two to n-dimensional interactions for explanatory variables x_i, x_j, etc., while a single subscript (e.g. $s_1(x_1)$) denotes univariate piecewise linear segment.

$$y = \beta_0 + s_1(x_1) + s_2(x_2) + s_{ij}(x_i, x_{j,\ldots}) \ \ldots + s_n(x_n) + \varepsilon$$

Equation 3

For high dimensional dataset, MARS have the ability to assess the contribution of high dimensional interactive effects to explanatory variables, whereas GLMs cannot achieve that (Archer & Kimes 2008; Friedman 1991; Hastie & Tibshirani 1990; Smyth 1989). MARS are now widely used in the field of data mining as they do not impose a specific class of relationship between predictor variables and dependent variables and are ideal for problems that contain two or more variables (Bellman 1961; Bishop 1995; Friedman 1991; Sumathi & Sivanandam 2006). To demonstrate this, Elith and Leathwick (2007) trained presence-only data using museum and herbarium records and MARS (Friedman 1991). They attempted to predict the distribution of over 226 species from six regions and found that MARS were able to model multiple responses to determine the dominant environmental drivers (Elith & Leathwick 2007; Friedman 1991). This also improved the stability of variable selection (Elith & Leathwick 2007).

1.3 Ensemble learning

In the early 1990s, ensemble learning began to appear, that is the simultaneous use of a combination of models (Xue 2016). The primary task of the integrated method is to create a group of models rather than only apply one individual model (Yu-Wei 2015; Xue 2016). The final prediction is determined by the form of the “major vote”. Each individual model from the set is a basic learner. Although one basic learner corresponds to one sample dataset, they usually have the same model form. The selection of sample data and the methods to combine multiple basic learners to achieve a reasonable “major vote” are two significant aspect of ensemble models (Kuhn & Johnson 2013). Bagging (Breiman 1996) and boosting (Freund & Schapire 1996) are two common ensemble learning techniques (Kuhn & Johnson 2013). They can be applied to classification or regression analysis to form an integrated combination (Xue 2016).

Bagging applies bootstrap (Breiman 1996) to randomly generate several training sets with replacement (Efron & Tibshirani 1986; James et al. 2013). These samples are then used to train basic learners (Kuhn & Johnson 2013). Bootstrap sample dataset has the same size with the original dataset; however, several samples may be repeatedly used many times, while others may not be used for the modelling (Kuhn & Johnson 2013). The final bagged models can generate a consensus prediction through voting all individual models. Integrated bagging is an effective approach to improve the prediction accuracy and robustness of individual models (Breiman 1996; Huang & Wang 2014; Xue 2016). For example, a bagged MARS begins with computing training sets through bootstrap sampling (Breiman 1996; Friedman 1991; Kuhn & Johnson 2013; James et al. 2013). These training sets are used to generate MARS models respectively with each of them having an individual classifier and independent prediction (Breiman 1996; Kuhn & Johnson 2013). Finally, combining a group of bootstrapped MARS models through majority vote to select an optimized MARS (Breiman 1996; Huang & Wang 2014).

Unlike bagging which uses a random selection algorithm, boosting begins with sequentially selecting variables for training basic learners (Breiman 1996; Freund & Schapire 1996; Huang & Wang 2014). It will evaluate the previous individual model in order to gradually shift emphasis on poor performance variables in the following basic learner (Elith et al. 2008; James et al. 2013). Boosted models can potentially reduce the loss of model performance through its interactive algorithm (Elith et al. 2008). For example, Boosted Regression Trees (BRT - Freund & Schapire 1996) are used by ecologists who require models that have the flexibility to express the principal correlations in a dataset. BRTs combine the benefits of regression trees and boosting to form an additive regression model that can fit complex nonlinear relationships and overcome the poor predictive performance issues associated with single tree models (Abeare 2009; Freund & Schapire 1996; Li et al. 2014; Jiao et al. 2015).

1.4 Spatial autocorrelation

Spatial analysis examines whether the data are geographically relevant in space (Deng et al. 2013). A spatial autocorrelation index describes the degree of independence between data at different positions. In the field of ecology or population studies, spatial autocorrelation can also be used to examine the dynamic properties of specific ecological datasets and determine the cause of spatial synchrony on a large scale (Deng et al. 2013). There are several approaches to remove spatial autocorrelation. One approach that is largely increasing the distance between points can potentially eliminate spatial effects (Brito et al. 1999). For example, Brito et al. (1999) calculated a critical distance, which was up to 10 kilometers; they separated out data with that distance to ensure data's independence. Unfortunately, this method may result in a small dataset, thereby failing to achieve an accurate modelling analysis. Alternatively, spatial autocorrelation can be accounted for and this allows us to understand spatial patterns (Nyström Sandman 2011). Rather than removing spatial autocorrelation, several techniques have been designed to measure and address it (Simard et al. 1992). For instance, Moran's Index (Cliff & Ord 1973) is used to examine the level of spatial autocorrelation. This spatial examination can study all data's spatial correlation or explore data at various distance bins (Anselin 1995).

1.5 Objective

The overall aim of this study is to demonstrate the usefullness of habitat models through examining the habitat niche of the target tree species within the Mount Lofty Ranges of South Australia and using two non-linear regression methods: bagged MARS and BRT. The complex interactions between the spatial location of the target tree species with the environment data will be modeled and explained.

This research dose not only compare two different regression algorithms, but also explores the influence of the spatial autocorrelation on the models. Spatial autocorrelation may bias the results by violating the assumption of independence in the sample data. In order to combat spatial autocorrelation, a method will be used to re-sample the data, separating the distances between sample points. This will be compared with a different method where an index of spatial autocorrelation is explicitly incorporated in each model.

As the sample size of data was limited, not enough data can be set aside for independent testing. Instead final prediction surfaces of habitat niche are compared with an expert opinion approach undertaken by the Department for Water and Natural Resources, South Australia.

The following key questions will be posed:

1. How do bagged MARS and BRT models compare in terms of their fit of the data?
2. What is the difference between the ranking of their important predictors and why?
3. How does spatial autocorrelation alter these models?
4. How do the important predictions compare with biological information from previous studies?
5. How do these predictions compare with the current distribution maps as drawn by expert opinion?

1.6 Thesis outline

This research is a combination of theories and empirical analysis. It is divided into four chapters. The remainder of the thesis is structured as follows:

The research methodology is covered in Chapter Two, introducing the overall data of the field of study. This is followed by general information of four

selected trees species. Next, the response (presence/absence data) and a set of 34 environmental predictors are introduced. Following that are the techniques and processes of testing data independence and spatial autocorrelation. The approaches to generating spatial weights are also described. This chapter also clarifies the modelling approaches of bagged MARS and BRT with and without spatial autocorrelation. This includes model training, evaluation and prediction. Finally, the chapter describes the thresholds used to transform continuous prediction into binary presence/absence.

Chapter Three displays the results of the previous chapter, showing the results of data preparation such as a summary of the PA dataset and the relationship of the sample distance with the percentage of spatial points. Next, it concludes the outcomes of modelling approaches, including model evaluation using AUC values of the ROC plot analysis, the relatively important predictors selected by bagged MARS and BRT models, trees' soil and geology requirements, spatial interaction and frequency histograms of model predictions. The end of this chapter develops the final predictive surface of models.

Chapter Four discusses the predictive models by comparing the environmental preference predicted by models with previous studies. Following that is the discussion of evaluating model performance and selecting appropriate threshold of prediction. Then this chapter discusses spatial autocorrelation. It also illustrates the comparison between the predictive surface and existing maps drawn with an expert's opinion. This chapter discusses the limitations of this study and concludes with the key findings and problems; and then gives suggestions for further relative studies.

CHAPTER TWO: RESEARCH METHODS

2.1 Study area

The Flinders Lofty Block (FLB) is located in the southeast of SA. It is one of the 89 Australian bio-regions developed by Interim Bio-geographic Regionalisation for Australia (IBRA - Guerin et al. 2016). FLB has a Mediterranean climate with warm to hot summers and cool, moist winters (Guerin et al. 2016). Within the FLB region, Laut et al. (1977) created several Environmental Associations (Kinnear et al. 2001). The study area (Figure 2) comprises a well-defined zone of uplands extending from the Barossa valley in the north through Fleuriue Peninsula in the south.

The study area has been chosen because of its varied terrain and diverse environment; it mainly comprises open forests and low open forests (Laut et al. 1977). Several peaks, such as Mt. Lofty, Mt. Torrens and Mt. Gawler stand above the general summit level and can represent hills. However, the landform in the Inman Valley is eroded deeply below the surface. Consequently, the area's elevation varies between -0.72m to 703 m. The hottest month is January which has a mean maximum temperature of 21.9 ^{o}C and a mean minimum temperature of 16.8 ^{o}C. The coolest month is July, when the mean maximum temperature is 11.8 ^{o}C and the mean minimum temperature is 8.2 ^{o}C. The relative humidity follows an annual cycle opposite to that for temperature (Laut et al. 1977). Due to the lower air temperatures, the humidity increases during winter in June and July whereas the humidity decreases in the hotter summer months of December and January. The study area contains the Mount Lofty Ranges which are located to the east of Adelaide. The watersheds from the Mount Lofty Ranges supply on average 60% of Adelaide's water demands (Wood 1986). The majority of the water derives from rainfall which is then captured by streams and reservoirs (Kuhnert et al. 2015). The study area's mean annual rainfall is between 306mm and 1138mm, where most of the rainfall occurs in winter; the mean winter rainfall is 533mm. The most common soil types are acidic sandy loam over clay and shallow soil.

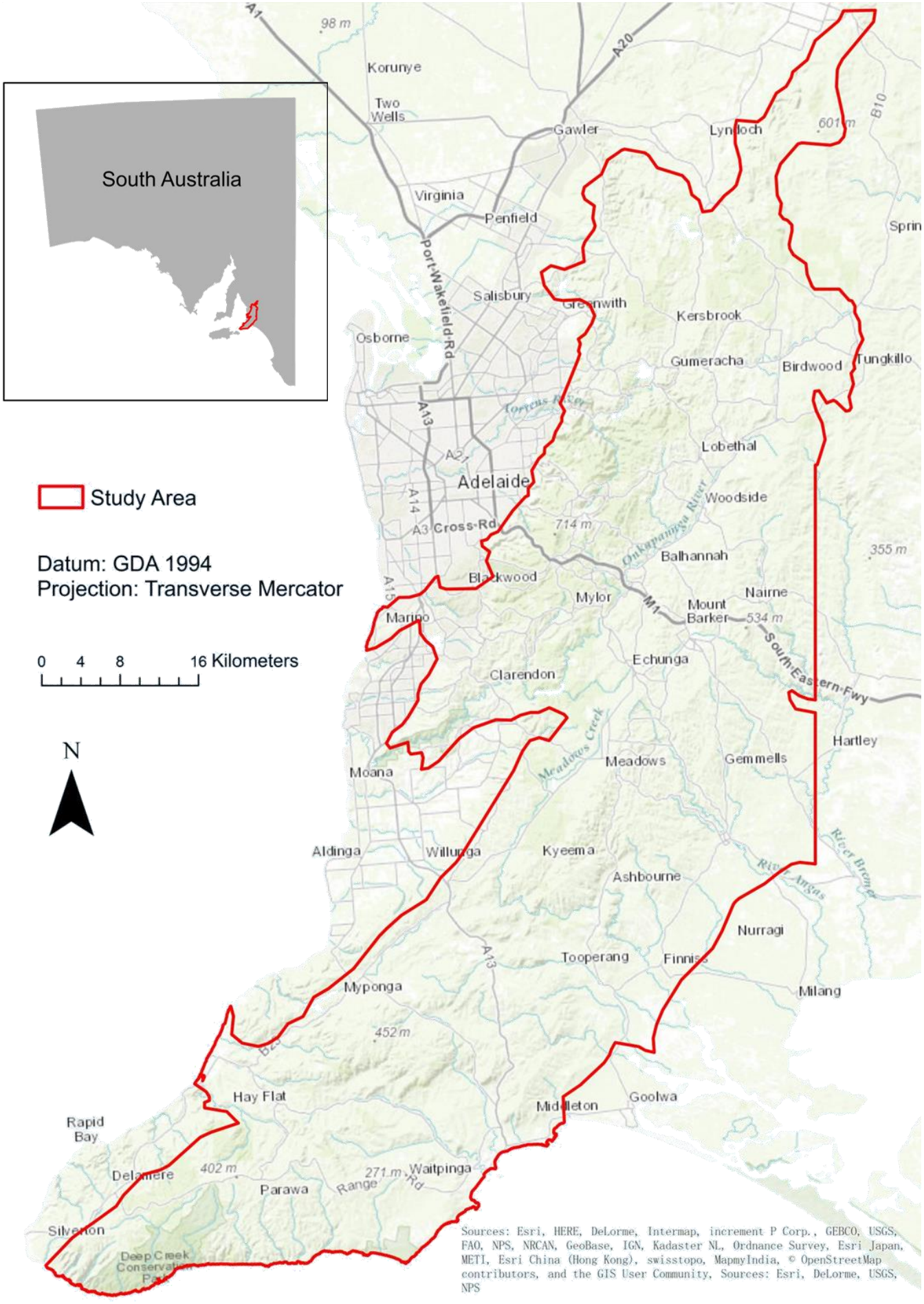


Figure 2: The study area covers approximately 3778.82 km2.

2.2 Species selection

There are approximately 1,500 native species found in this area, including gum trees, orchids, ferns, grasses, herbs, lilies, rushes, and so on. Historically, vegetation growing in the Mount Lofty Ranges was cleared for agriculture and residential development (Paton et al. 2000). As a result of that, there are only approximately 20% of the native plants still surviving in this region (Armstrong et al. 2003). Among these plants, tree was selected as the target species. The selection of native trees for regressions in this study should satisfy certain criteria. First of all, the target trees must have sufficient samples to achieve a meaningful statistical analysis and modelling. Secondly, the ecological dominance of the trees was considered because the distribution of a dominant or co-dominant species is more likely to have direct response to the specified environmental variables (Green 1994). The interaction between trees and environment was the key point of a regression model analysis. Thirdly, the conservation status of trees was another criterion, endangered or protected trees were the focus in this research. Finally, trees' significant features were also standard for the selection. Those trees with high economical values or with special functions were preferred.

2.3 Target tree species

With the above criteria, four Australian native trees (See Table 1) have been selected to model their habitat niche. This selection ensured a mix of Eucalyptus and non-Eucalyptus species. There were enough surveying samples of each target tree. And these species were all identified as either dominant or co-dominant tree within the Mount Lofty Range region (Armstrong et al. 2003). The target species were also considered representative of the overall biodiversity of the region (Lambeck 1997; Brooker 2002; Watson et al. 2001). Especially for *E. obliqua*, this tree was dominant to the formation of the sclerophyll open forest (Adamson & Osborn 1924). Although the target trees were not presently endangered, *A. verticillata* and *E. fasciculosa* were recorded

as vulnerable trees in SA (DEH 2001; Berkinshaw 2010), and each target tree has its own economic values and significant characteristics.

Table 1: Target trees' general information.

Species' Name	Common Name	Height	Family
Allocasuarina verticillata	Dropping Sheoak	5 - 9 m	Casuarinaceae
Eucalyptus goniocalyx	Long Leaved-box	15 m	Myrtaceae
Eucalyptus fasciculosa	Pink or Gill or Scrub Gum	15 m	Myrtaceae
Eucalyptus obliqua	Messmate Stringy-bark	30 m	Myrtaceae

Data source: Electronic Flora of SA species' Fact Sheet from Department of Environment, Water and Natural Resources (DEWNR)

Allocasuarina verticillata (Dropping she oak) has a rounded crown and usually lives for up to 80 years. It can potentially live in diverse habitats (Obertello et al. 2005). This species is mainly found in a wide range of soils from good loam to calcareous sands (Bonney 1997). Although *A. verticillata* is not identified as a threatened species, it still plays a crucial functional role (Williams & Cary 2002). For example, the tree's seed is the primary food source for Glossy Black Cockatoo (*Calyptorhynchus lathami*) which is endangered and protected by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES 2011; Joseph 1982; Pepper et al. 2000). Thus, *A. verticillata* is one of the local trees that assists the bird's survival, particularly within Kangaroo Island (Pepper et al. 2000). In terms of it values, it provides great materials for windbreaks, drought fodder and firewood (Broadhurst 2015). This tree is mainly used as an on-farm timber for fence posts and shelter purpose. It is also known as a good turning wood that can be used for craft items (Bonney 1997).

There are varieties of Eucalyptus species, it is a good economic tree species with a strong adaptability and has been widely used (Wang 2015). The majority of Eucalyptus trees are native to Australia. It is one of the most

common tree species in Australia (Boland et al. 2006). *Eucalyptus goniocalyx*'s (Long Leaved-box) height is approximately 15m with rough and fibrous bark covering the whole trunk (Baker & Smith 1902; Brooker & Slee 1996). The timer of long leaved box is hard, close-grained and pale-coloured; this can be used to differentiate it from other similar Eucalyptus species. Additionally, it is a very durable wood material (Menut et al. 1995). This species' habitats in coast range from the Blue Mountains, New South Wales into Victoria and SA (Guerin et al. 2016).

Eucalyptus fasciculosa (Pink or hill or scrub gum) is a small to medium tree that is endemic to Australia; its main habitat is in the states of Victoria and SA (Ward 2007; Whittington & Sinclair 1988). It is a vulnerable health woodland has ever-green foliage and an abundant quantity of flowers, ranging from white to cream (Slippers et al. 2004). This pink-brown timber with medium durability can also be used on farms such as fence and firewood (Elliot & Jones 1990; Bonney 1997). Furthermore, it plays an important role in shelter and habitat plantings (Bonney 1997).

Eucalyptus obliqua (Messmate Stringy-bark) is a very tall tree and has a hard wood (Bar-Ness et al. 2012). It is native to south-eastern Australia and largely confined to the sufficient rainfall areas in the Mount Lofty Range (Stead 2008). Messmate Stringy-bark has much reduced, but this tree is still relatively common in lower south east of SA (Bonney 2010). In its habitat areas, messmate stringy-bark may form a woody forest with other Eucalyptus species, such as *E. fastigata, E. delegatensis*, or *E. Viminalis* (Bassett & White 2001; Lutze 1999). The rough bark on the tree is thick and fibrous (Barry et al. 2015). They have several applications, for example, bark paintings and dishes after treated (Sinclair 1980). This tree is an economic species that can produce wood chips. It can also be used as construction materials, such as house frames, and internal flooring and furniture (Bonney 1997). The wood of messmate stringy-bark has moderate hardness and can be used for making tools and shelters, or else for pulp production (Facelli et al. 1999).

2.4 Presence and absence data

The trees' spatial location data were collected from Flora Species Observations dataset, provided by Department of Environment, Water and Natural Resources (DEWNR) and the Department for Environment and Heritage of South Australian (DEH). Biological species' surveying of the Southern Mount Lofty Ranges were conducted between 1997 and 2001. These field surveys totally recorded 54424 samples of thousands of tree's species. A further 350 samples were collected by Lethbridge in 2004 (Lethbridge *pers comm*) to supplement the sample dataset. A random sampled quadrat survey method was used in the field surveying. The quadrat size was 25m providing a total number of presences for each species in the study area in Table 2. For the remaining quadrats where a given target species was not found, these were deemed to be absences. These Presence and Absence (PA) data indicating existence as 1 and in-existence as 0 became the response variable which was used to train the regression models.

Table 2: Species' presence and absence amount.

Species [a]	**Presences** [b]	**Absences** [c]
A. verticillata	215	54209
E. goniocalyx	152	54272
E. fasciculosa	705	53719
E. obliqua	662	53762

Note:
a: there were thousands of trees species and four of them were selected.
b: presences were the sample's amount of the target tree species.
c: absences were sample's amount of all species except the target species.

2.5 Explanatory variables (Candidate predictors)

Regression modelling in this context is effectively testing the response of a species to environmental variables (Ahmed et al. 2015). Hence, the selection and testing of environmental variables (also called candidate predictors) can directly affect model performance (Guisan & Zimmermann 2000). In this study, 34 environmental predictors including 14 topographic variables, 10 soil predictors and another 10 climatic predictors were tested in regression models for their likely contribution to the presence and absence responses of the four target species. **Appendix 1** was a summary of predictor variables in terms of their units and data sources.

A raster grid of elevation data (called **elev**) was calculated by Lethbridge et al. (2006) in ArcGIS$_{TM}$ (ESRI 2017) from a combination of 2m, 5m and 10m vector contour data, depending on the coverage of the more precise of these data. The source data in an unedited form was provided by DEWNR in Computer Aided Drafting (CAD) format. The land system classes (**envas**) were also provided by DEWNR. Another 9 topographic variables were computed from elevation data by Lethbridge et al. (2006). These included the percentage of surface slope (**slope**), the number of cells in a raster grid that can contribute water to a given cell (**flowac**), the soil wetness which indicates soil water content (**wet**), the aspect of the terrain (in degrees), converted using **Equation 4** and **Equation 5** into the level of east-west aspect and north-south aspect (**ew** and **ns**) to avoid cyclic effects of using degrees in regression,

$$\mathrm{ew} = \sin(\mathrm{Aspect}) \qquad \textbf{Equation 4}$$

$$\mathrm{ns} = \cos(\mathrm{Aspect}) \qquad \textbf{Equation 5}$$

and 4 solar radiation data. Summer (**smrsrad**) and winter (**wntsrad**) solar radiation estimated the average monthly data over December to January and June to July, respectively. The annual solar radiation (**anlsrad**) was the

average value for the first day of each month of a year in hourly intervals. The seasonal data (**seasrad**) calculated the difference between the June and the annual solar radiation. The distance from major (4th to 7th order streams) and minor (1st to 3rd order streams) were also provided by DEWNR (**major** and **minor**).

The geological classification (**geol**) and all soil data were provided by the Department of Primary Industries and Resources of South Australia (PIRSA) in Land Information in Analysing Mapping Soil and Landscape Attribute Data of 2001. 10 soil data include the most common soil type (**comsoil**), the capacity of deep subsoil and the material immediately below the soil profile to allow excess water to move downwards into deep sediments or fractured rock (**drain**), the inherent fertility (**fertil**), the susceptibility to acidity (**acid**) and waterlogging (**waterlog**), the depth to water table (**wtdepth**), the alkalinity of the surface and subsoil (**alkali**), the water holding capacity (**awhc**), the depth of hard rock (**rkdepth**) and the overall amount surface stone and outcropping rock (**srock**).

All weather station data of climate were originally supplied by Australian Bureau of Meteorology. The rainfall and temperature data were converted by Lethbridge et al. (2006) into continuous surface using co-Kriging (Burrough & McDonnell 1998). Annual, summer (December to January) and winter (June to July) mean rainfall and average temperature data were converted into raster layers. The rainfall and temperature standard deviation (**sdrain1k** and **sdtmp1k**) were calculated in **Equation 6** and **Equation 7**. Here, x_i, y_i were the observed rainfall and temperature values, respectively, and $\bar{x}$, $\bar{y}$ were the mean monthly data.

$$\sigma_{Rainfall} = \sqrt{\frac{\sum_{i=1}^{n} = \left(x_i - \bar{x}\right)^2}{n-1}} \qquad \textbf{Equation 6}$$

$$\sigma_{Temperature} = \sqrt{\frac{\sum_{i=1}^{n} = \left(y_i - \bar{y}\right)^2}{n-1}} \qquad \textbf{Equation 7}$$

Seasonal rainfall index (**snrain1k**) was computed with **Equation 8**, it was the difference between the standard deviation of rainfall in the driest month (σ_i) and the standard deviation of annual rainfall (σ_R). Seasonal temperature index (**sntmp1k**) was calculated using **Equation 9**, the standard deviation of the monthly mean (σ_T) divided by the mean annual temperature ($\overline{x_t}$).

$$R_{index} = \sigma_i - \sigma_{Rainfall} \qquad \textbf{Equation 8}$$

$$T_{index} = \frac{\sigma_T}{\overline{x_t}} \qquad \textbf{Equation 9}$$

2.6 Testing for data independence and spatial autocorrelation

Geographic dataset may lose independent observations owing to the spatial interaction and diffusion (Brito et al. 1999; Burrough & McDonnell 1998; White 2000). However, all forms of regression require that the training data are independent (Tufféry 2011). Spatial autocorrelation violates this assumption and therefore can reduce the power of model (Anselin 1988). One way to ensure the sample data are independent is to spread the data out geographically according to a minimum distance threshold (Brito et al. 1999). The software “Sort PA” (Lethbridge 2005) was designed to resample all tree survey data into presence and absence data for the target species. The PA of the target specie were drawn from a much larger dataset (54772 samples). For the purpose of measuring spatial autocorrelation, in this study, the minimum spacing between quadrat centers was deliberately varied from 30 m (also the dimension of each quadrat) up to 2000m. More specifically, sampling was carried out at distances of 30m, 60m, 90m, 120m, 150m, 500m, 1000m, 1500m and 2000m. Absence data were randomly eliminated to ensure that absence data cannot dominant the modelling process. For those species that had fewer presence points (*A. verticillata* had 136 and *E. goniocalyx* had 148), the maximum absence points’ number was 500; for the rest tree that had sufficient presence data (*E. fasciculosa* had 631 and *E. obliqua* had 655), absence points equaled to twice of presence data.

Global Moran's Index is a well-known index of the strength of spatial autocorrelation (Cliff & Ord 1973). Local Moran's Index specifically identifies any localised spatial non-stationary autocorrelation changes in the data (Cliff & Ord 1973; Anselin 1995). In this study, the spatial statistical tool, Cluster and Outlier Analysis of ArcGIS™ (ESRI 2017) was used. The purpose of Cluster and Outlier Analysis is to determine an Anselin Local Moran's index value, together with a z-score, a pseudo p-value, as well as a code denoting the type of cluster associated with every point (Anselin 1995; Getis & Ord 1996). The statistical significance of the calculated index value is given by the z-score and the pseudo p-value, which represents standard deviation and the probability respectively (Levine 2004). The likelihood that an arbitrary process generated the observed spatial pattern is given by the p-value (Levine 2004; Zhang et al. 2008). In this study, a p-value <= 0.05 was used to select points with 95% confidence (a significant level) of spatial autocorrelation. The percentage of significant spatial related points was then calculated to indicate the effect that the spacing of the data had on spatial autocorrelation.

2.7 Spatial weights

The alternative method to the above is to explicitly model a spatial weights function in the regression. SpaceStat™ (BioMedware 2014) was used to create a set of a spatial weight file with no inflated spacing between sample data (i.e. quadrats theoretically could be as close as 30m apart). The spatial weights were an inverse distance rank for the nearest five sample points at each target point. For instance, the nearest neighbor was assigned a spatial weight 1, the second nearest neighbor had a spatial weight 1/2, and so on, up to a fifth neighbor. A new candidate regression predictor called **Spatial_A_influence** was added describing the level of spatial autocorrelation.

2.8 Bagged Multivariate Adaptive Regression Splines (MARS)

MARS is a sum of piecewise linear basis functions for single or more variables (Moisen & Frescino 2002; Friedman 1991). It can be tuned by determining the interaction depth (**degree**) and the number of basis functions (**nprune**) used in each model. In this study, single response and two-way interactions between the candidate predictors were tested. However, there was a variation in the number of terms for PA dataset. The optimal combination of **degree** and **nprune** for fitting MARS models can be determined by the **train** function from **caret** package (Kuhn 2013) of the statistical R software (R Core Team 2014) using resampling (Friedman 1991). Resampling techniques were employed to measure models' performance. One example of them was bootstrapping, which involved generating a series of altered datasets from the training samples (Efron & Tibshirani 1986; Kuhn 2013; Kuhn & Johnson 2013). An analogous set of hold-out samples was associated with every re-sample dataset. For every possible combination of tuning parameters, every sub-dataset was fitted with a model that was then employed to predict the equivalent held-out samples. The results of every hold-out sample set were merged to forecast the resampling performance, which in turn helped to determine the ideal tuning parameters (Kuhn 2013). MARS is a non-parametric regression approach, which has been increasingly applied in bioinformatics due to its flexibility (Friedman 1991; Lewis & Stevens 1991; Valavanis et al. 2008). However, assembling bagging can potentially improve the stability and reliability of an independent MARS (Breiman 1996; Borra & Di Ciaccio 2002; Friedman 1991; Kuhn & Johnson 2013). Bagged MARS is a stochastic modeling process, involving the records in the original sample data to be randomly selected with replacement and used to build several sub-models (Borra & Di Ciaccio 2002; Breiman 1996; Friedman 1991). The **caret** packages (Kuhn 2013) from the statistical R software (R Core Team 2014) provided **bagging** and **earth** functions to fit **bagEarth** (bagged MARS) models. In this study, the number of bootstrap samples was 500, the response variable was represented by the PA data in **Section 2.3**, while the predictor variables were the 34 environmental data from **Section 2.4**. Response and predictors were

together employed for purposes of training sub-MARS model for each bootstrap sample with the ideal values of tuning parameters (**degree** and **nprune**).

Bagged MARS has the ability to fit the complex shape of the response variable's relationship with explanatory variables including spatial term (Abraham & Steinberg 2001; Breiman 1996; Friedman 1991). Spatial relationship is commonly used in GIS studies to examine spatial autocorrelation (Nyström Sandman 2011). Bringing spatial terms into modeling analysis can access the spatial interaction between target variable and predictors. Repeated bagged MARS approaches in previous part provided that spatial weights set from **Section 2.6** was included as an extra predictor. Additionally, all the same environmental predictors were used here to maintain comparability and consistency. The spatial term was an inverse distance ranks computed from 30m PA dataset for four species. Spread out the data can reduce or eliminate spatial affect, but this will loss presence and/or absence samples. Any missing values could alter the spatial weights set due to the change of neighbour data. Hence, this study only studied spatial autocorrelation with 30m dataset.

2.9 Boosted Regression Trees (BRT)

BRT is a forward and stage-wise procedure (Elith et al. 2008; Freund & Schapire 1996; James et al. 2013). The final prediction of BRT is computed through averaging all generated regression trees (Freund & Schapire 1996; Schapire 2003). Its model fit depends on how boosting quantifies the low fitted observations and selects variables for the next iteration (Elith et al. 2008). Predictive models can be fitted by train function over a range of tuning parameters, and bootstrapping can be used to measure how the models perform, which in turn determines the most suitable combination(s) of tuning parameters. In the case of BRT models, tuning can be done over the tree complexity and learning rate by the **train** function from **caret** package (Kuhn 2013) of R software (R Core Team 2014; Freund & Schapire 1996; Kuhn 2013). As previously mentioned, only one- or two-interactions were considered in this study, meaning that tree complexity (**interaction.depth**) was either 1 or 2.

Secondly, when determining learning rate (**shrinkage**), which was the weight that every tree should be subjected to, it should be taken into account that the predictive value was enhanced. However, if there is a low learning rate, the number of trees necessary and the cost of computation will increase (Kuhn 2013). Assembling individual models is a new algorithm to improve accuracy. For those unstable models having a high variance like classification trees and regression trees, the bagged and/or boosted models are more stable (Breiman et al. 1984; Lim et al. 2000; Kuhn & Johnson 2013). For example, combining boosting with regression trees can generate an integrated regression model, called boosted regression trees (Breiman et al. 1984; Freund & Schapire 1996). The **gbm** package (Ridgeway 2006), and extension in the statistical R software (R Core Team 2014) can fit BRT models (Elith et al. 2008; Freund & Schapire 1996). The PA data were the response variables in **Section 2.3**, and the 34 environmental data from **Section 2.4** were the predictor variables, both of them were used to train BRTs with the optimal tuning parameters (**interaction.depth** and **shrinkage**). This required the use of Bernoulli character of the error structure and 0.5 proportion of observations (**bag.fraction**) employed in variable selection. To suggest the following tree in the expansion, selection of the fraction of training set observations was arbitrary. This afforded the model to fit randomness. Similar but not identical fits will be obtained by applying the same model twice if the **bag.fraction** is less than 1.

With the fast development of machine learning, novel integrated algorithms like BRT are able to overcome the weakness of single modes analysis (Breiman et al. 1984; Freund & Schapire 1996; Harrington 2012). Assembling boosting with regression trees can improve model accuracy and increase model performance through reducing the over-fittings, which are commonly occurring with single trees (Breiman et al. 1984; Freund & Schapire 1996). In the current part, spatial weights set computed in **Section 2.6** were introduced to study how spatial autocorrelation alter BRT models. Spatial autocorrelation can examine the spatial correlation of variables (Zhang et al. 1998). Taking spatial predictor into modelling, spatial autocorrelation can contribute to the fit and may alter models' structure (Anselin 1988). However, accounting for spatial terms is one way to ensure that a model is not biased by spatial autocorrelation. Regressed

the PA data and explanatory variables again using BRT method in that case **Spatial_A_influence** was one of the predictors. To keep the comparability and consistency, the same combination of environmental predictors was used to train BRT model with spatial autocorrelation. Moreover, only 30m PA dataset was trained to study spatial terms' influence due the missing values of spread out dataset.

2.10 Model evaluation

Model evaluation is a way to access the predictive power of the models (Guisan & Zimmermann 2000). It can be extrinsic, which involves examine model predictions with independent source of data (Anderson et al. 2003). For example, dividing all the sample data into training and testing subset for the purpose of fitting data and evaluating models' performance, this is an extrinsic evaluation. However, in this study, there were not enough sample data to be set aside and then perform such an independent testing. Instead, this study conducted an extrinsic comparison. That was, the final prediction maps of habitat niche applying to regression methods were compared with an existing distribution map as drawn by expert opinions undertaken by the DEWNR (Croft unpublished). In general, Tim Croft's distribution maps were qualitative analysis results combining expert opinions with field observations and visual interpretations from aerial photographs.

Model evaluation can also be intrinsic validation, which employs resampling techniques to decide an optimal model having a better performance; these include bootstrap sampling, cross validation and so on (Anderson et al. 2003; Craven & Wahba 1979; Efron & Tibshirani 1986). For example, bootstrapping from the **train** function in the **caret** package (Kuhn 2013) of R software (R Core Team 2014) was used in this study. This measures model's fits and then determines the optimal combination of tuning parameters for the regression.

Threshold-independent techniques like Receiver Operating Characteristic (ROC - Guisan & Zimmermann 2000) plots are another type of model evaluation method (Anderson et al. 2003). ROC plots take the true negative rate (the probability of predicting absence species' distribution as absence) as abscissa and the true positive rate (the probability of predicting actual species' distribution as presence) as ordinate in a curve (Fawcett 2006; Guisan & Zimmermann 2000). Areas Under the Curve (AUC) are calculated to evaluate the model predictive power. For example, the greater the AUC value is indicating the stronger the predictive ability of the model. AUC provides an evaluation of the estimated prediction probabilities without reliance on threshold (Guisan & Zimmermann 2000). Thus, this study not only carried out ROC analysis to evaluate each model's performance, but also used AUC values to compare two different regression algorithms. ROC curves were drawn using the **roc** function of **pROC** package (Robin et al 2011) in the statistical R software (R Core Team 2014). They compared the estimated PA values (the prediction) with the actual PA data (the response). The accuracy of model prediction can be explained by AUC values. For instance, Swets (1988) has developed an approximate guide to classify the accuracy measure, and these classifications (Table 3) were used in this study to evaluate regressions.

Table 3: Accuracy classification.

AUC values	Model performance
0.9 to 1.0	Excellent
0.8 to 0.9	Good
0.7 to 0.8	Fair
0.6 to 0.7	Poor
0.5 to 0.6	Fail

2.11 Model prediction

Regression models compute a 'formula' explaining the relationship between dependent and independent variables. Projecting this formula spatially allows one to forecast the distribution probability using the explanatory variables. As previously mentioned, this is called a Spatial Distribution Model (SDM). The predictive models used were bagged MARS models (**Section 2.7**) and BRT models (**Section 2.8**). One of the main purposes of an SDM is to create a probability or suitability map of species (Guisan et al. 2013). SDMs were created in the form of raster grids.

2.12 Calculate a threshold of distribution probabilities

The prediction results of SDMs of this study were on a 0 to 1 continuous scale corresponding to low to high probability of likely current or historical occurrence. This is because much of the landscape is either cleared or currently supports a land use. It is common practice to convert the probability of a continuous occurrence into a binary presence/absence map by selecting a threshold (Liu et al. 2005). That is, predicted values above the fixed threshold indicate species' appearance (presence) and those below that represent disappearance (absence). The determination of thresholds can be subjective and objective approaches (Liu et al. 2005). A representative fixed threshold of the first category is 0.5, has been widely used in ecology (Liu et al. 2005). This study examined 0.5 as one of thresholds. Two other statistical representations of the predictions were also recommended as thresholds; they were the mean and the median value of the predicted probabilities. The data correction rate for these thresholds were compared.

Only the model of each target tree species with the best performance was used to generate the final presence/absence map. Therefore, the predictive values of these models were compared with the corresponding actual PA data to calculate the positive (P) and negative (A) data correction rate. The optimal threshold is the one, which has a relatively high average correction and acceptable rate (> 50%) in both positive and negative.

CHAPTER THREE: RESULTS

3.1 Data preparation

As previously mentioned, the first regression method tested involved seeking independence in the sample data by spreading out the minimum distance between samples until the percentage of points showing significant Local Moran's indices was sufficiently low. This required the PA data of target trees to be re-sampled. The size of Sample Distance (SD) spacing in turn alters the size of a PA dataset. In Table 4 the Nearest Neighbour Distances (NND) are also shown. These are a measure of the average dispersion of the sample data after applying a minimum distance between samples threshold (SD). In this study, this ceased when presence or absence points were less than 100. *A. Verticillata* and *E. goniocalyx* only lose several (approximate 30) present points as SD increases from 30m to 500m. However, for *E. Fasciculosa* and *E. obliqua,* SD is extended further to 1500m and 2000m, respectively. For this reason, their total number of points decreases sharply. See PA dataset's details in Table 4.

Table 4: PA dataset summary. All distance is in meters.

A. verticillata					*E. goniocalyx*				
SD	**NND**	**P**	**A**	**Total**	**SD**	**NND**	**P**	**A**	**Total**
30	756	133	500	633	30	787	148	500	648
60	789	130	500	630	60	804	148	500	648
90	842	128	500	628	90	781	148	500	648
120	849	128	500	628	120	788	145	500	645
150	820	127	500	627	150	833	142	500	642
500	1095	104	500	604	500	1096	110	500	610

E. fasciculosa				
SD	**NND**	**P**	**A**	**Total**
30	460	605	864	1469
60	459	591	809	1400
90	516	579	767	1346
120	544	562	735	1297
150	571	543	708	1251
500	961	408	399	807
1000	1555	270	208	478
1500	2096	201	121	322

E. obliqua				
SD	**NND**	**P**	**A**	**Total**
30	128	623	1578	2201
60	116	605	1578	2183
90	109	596	1578	2174
120	127	588	1578	2166
150	121	573	1578	2151
500	81	470	1548	2018
1000	169	353	1016	1369
1500	324	289	616	905
2000	487	237	430	667

The percentage of significant points extracted by Anselin Local Morans Index that exhibited significant spatial autocorrelation, was shown in Figure 3. As SD spacing increased from 30m to 500m, percentage values for *A. Verticillata* and *E. oblique* slightly decreased to 14.74% and 6.05%, respectively. *E. obliqua*'s spatial points gradually climbed and peaked at 2000m to 12.59%. *A. verticillata* and *E. obliqua*'s spatial related points were less than 20% for all sample distance; that was to say, there was a limited or no spatial relationship for these two species. For *E. goniocalyx*, the percentage of spatial points had a fluctuation and peaked at 150m; that was 37.07%. If SD spacing was less or equaled to 500m, NND of *E. goniocalyx* was then close to or slightly over 1000m; furthermore, spatial related points accounted for in excess of 20%. Therefore, *E. goniocalyx* had spatial autocorrelation, but a strong relationship was not evident. For *E. fasciculosa*, the percentage of spatial-related points was basically unaltered, at approximately 45%. When the distance between sample points was over 500m, spatial points of *E. fasciculosa* dramatically decreased and concluded with 14.29% at 1500m. There was spatial autocorrelation for *E. fasciculosa*; indeed, it had the highest percentage of spatial related points. This species had a stronger spatial relationship than the other three. However, if the space distance was in advance of 1000m, *E. fasciculosa*'s spatial autocorrelation became weak.

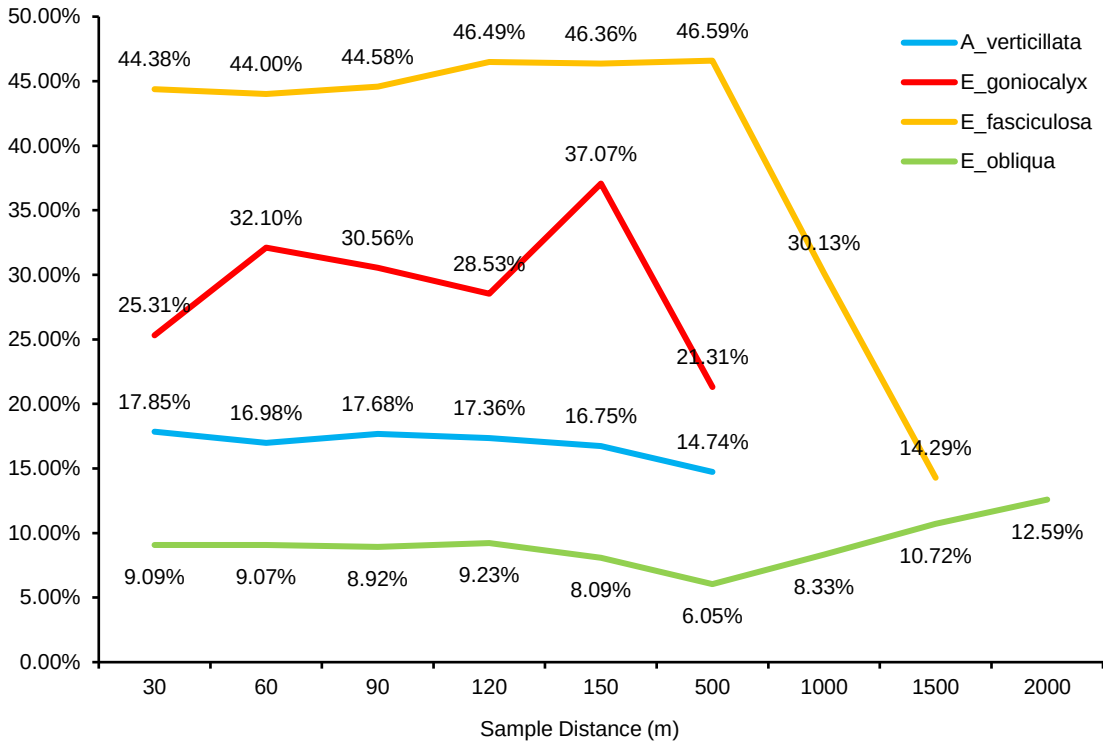


Figure 3: The relationship of sample distance and percentage of spatial points for *A. Verticillata* (Light blue), *E. goniocalyx* (Red), *E. Fasciculosa* (Orange) and *E. Obliqua* (Light green).

Specify models an index of spatial autocorrelation is one way to study interactions between response and spatial variable. In this study, spatial weights set (**Spatial_A_influence**) was introduced to two regressions for modelling all sample data (30m PA dataset) for all four target trees. Instead, an alternative method is to combat spatial autocorrelation. Spacing of the data reduced the percentage of significant spatial related points (see Figure 3), this can potentially minimise or eliminate spatial effects. This method was applied with those trees, *E. goniocalyx* and *E. fasciculosa*, which were had spatial relationship. Thus, these two tree species were also regressed with the spread-out PA dataset. The space distance was 500m and 1500m for *E. goniocalyx* and *E. fasciculosa,* respectively.

3.2 Intrinsic Receiver Operating Characteristic (ROC) tests

Different measures are designed to evaluate the quality of a prediction. The area under the ROC curve is considered as an acceptable measure of models' fit (Guisan & Zimmermann 2000). The discriminatory capability of a model is validated by a 0.5 score of AUC and invalidated by 1.0 score (Leathwick et al. 2006; Lethbridge et al. 2006; Fielding & Bell 1997). Models from the following three datasets were selected to compare their fit of data. They were 1) all the samples (30m PA dataset) without spatial weights; 2) the same dataset, but include spatial variable, and 3) the spread-out dataset that assumed the samples were all independent. According to the ROC analysis, the model of each tree with the highest AUC value was finally used to predict the potential habitat niche of each tree.

In general, all AUC values were significantly different from 0.7 indicating that bagged MARS and BRT models performed well for tree's PA dataset with the select explanatory predictors. The predictive ability of the *E. fasciculosa* models increased significantly when it was fitted using the spacing dataset, while the other trees had a consistent model performance for different dataset. In Table 5, the highest AUC value (0.995) indicating excellent performance represented the predictive ability of the model of *E. goniocalyx*, this method used all samples (30m PA dataset) excluding spatial weights; meanwhile the smallest AUC value (0.729) indicating fair performance was for the model fit of *E. obliqua,* this model used 30m PA dataset with spatial weights. Among all BRT models (Table 6), *E. goniocalyx* still had the highest AUC value (0.998) corresponding to excellent model performance, but it was for modelling of the spread-out PA dataset this time. And the smallest AUC value (0.748) corresponding to the worst fit of data was the mode of *E. obliqua* using 30m PA dataset without spatial weights.

Table 5: Intrinsic AUC values for four tree species using bagged MARS.

PA Dataset	*A. verticillata*	*E. goniocalyx*	*E. fasciculosa*	*E. obliqua*
	AUC	**AUC**	**AUC**	**AUC**
1	0.861	0.995	0.754	0.739
2	0.867	0.991	0.765	0.729
3		0.995 [a]	0.842 [b]	

a: SD spacing = 500m; b: SD spacing = 1500m.

Table 6: Intrinsic ROC values for four tree species using BRT.

PA Dataset	*A. verticillata*	*E. goniocalyx*	*E. fasciculosa*	*E. obliqua*
	AUC	**AUC**	**AUC**	**AUC**
1	0.939	0.996	0.896	0.748
2	0.946	0.996	0.893	0.765
3		0.998 [a]	0.917 [b]	

a: SD spacing = 500m; b: SD spacing = 1500m.

ROC plot is a threshold-independent model evaluation technique (Anderson et al. 2003; Guisan & Zimmermann 2000). It can not only access individual model's performance but also serve as a metric to compare the predictive ability of two or more different models. The better the model's fit is, the larger the AUC value is (Fielding & Bell 1997). The average AUC values of each regressions for four species were calculated and illustrated in Figure 4. The analysis of *E. goniocalyx* had the best model fit of data of any target species using both regressions, it shown excellent model performance (1> AUC > 0.9) (Swets 1988). However, modelling analysis of *E. obliqua* had a relatively poor performance with a fair fit of data; its AUC value was between 0.70 and 0.75. Regression models for the other two species had fair to good fit of data with their AUC values varied around 0.8. The comparisons also indicated that BRT models always had higher AUC values than bagged MARS models. Therefore, it had a better model performance in this study.

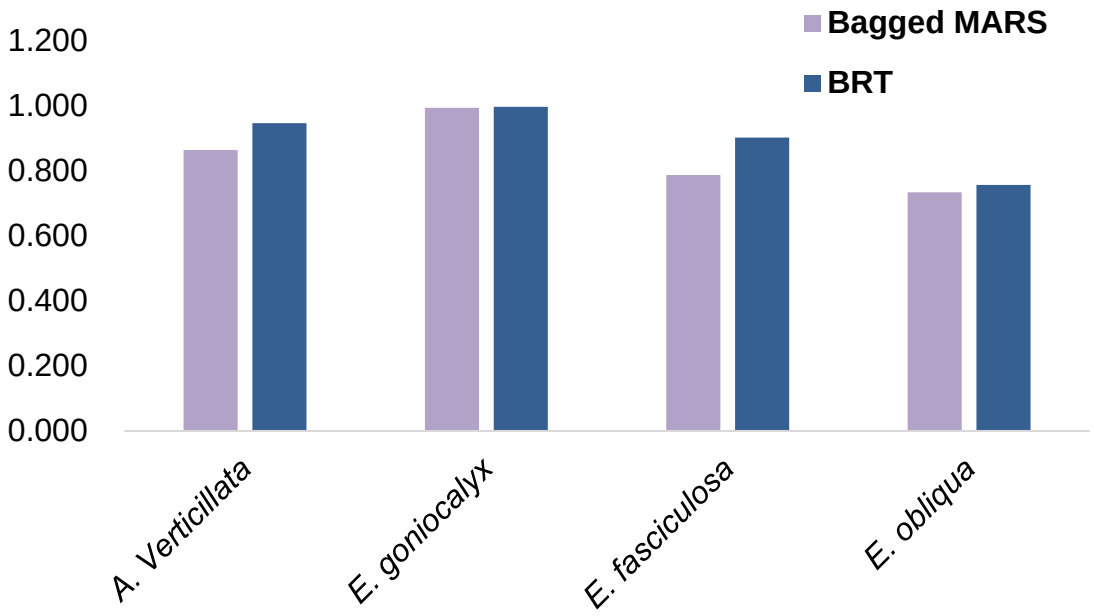


Figure 4: The average AUC values for four tree species using bagged MARS and BRT.

3.3 Candidate predictor's impact on the model

In ecology, an important aspect of the regression modeling is the ability to assess each candidate predictor's impact on the model. Some of these will be dropped. The highest performing predictors make a large contribution to the model. The function **VarImp** in **caret** package (Kuhn 2013) in statistical software R (R Core Team 2014) is a generic variable importance evaluation function for regression models. In this study, the rankings of candidate predictors were examined so as to explore the living preference of the selected trees. The relatively important predictors (top three variables) were considered to contribute environmental issues, which were selected by **bagEarth** (Kuhn 2013) and **gbm** (Ridgeway 2006) packages in the statistical software R (R Core Team 2014) corresponding to bagged MARS and BRT models, respectively. The ranks and explanation of these predictors for four species using two regression methods were summarised in Table 19 to 26 and displayed in **Appendix 2**.

As previously mentioned, two very different methods were used in this study to combat spatial autocorrelation. The first one was to specify regression models a spatial weights term. This was applied to all target trees. Secondly, SD was largely increased to minimise or eliminate spatial influence. Thus, *E. goniocalyx* and *E. fasciculosa*, which had been confirmed having spatial autocorrelation, were also regressed with 500m and 1500m spread out dataset, respectively. The predictor's impact on the regression models using the sample distance spacing method and the spatial weights method were also compared.

On the whole, the relatively important predictors obtained by the two regression algorithms for each tree were very similar. With regard to bagged MARS models, either the introduction of spatial autocorrelation or separating the distances between sample points resulted in slight changes in ranks of predictors. Comparatively speaking, BRT models seemed stable; their results were unchanged after including spatial weights.

For *A. verticillata*, the dominant predictors selected by bagged MARS varied greatly (Table 19) after accounting for spatial weights whereas variables' selection of BRT models (Table 20) did not change. Compared these two methods, they both have chosen temperature data (e.g. temperature standard deviation and the average summer temperature) as the most contributed variables for this species' distribution, but their rank was not the same. Using bagged MARS algorithm, the average winter temperature and environmental association were consistently found to exert the greatest influence over response, which was the distribution of *E. obliqua* (Table 21); while elevation, winter and summer rainfall were the most contributed determinants of the *E. obliqua*'s distribution in BRT models (Table 22). These two regression models generally indicated that elevation and rainfall variables may have a greater impact on *E. obliqua*. For the other two Eucalyptus species, the introduction of spatial weights altered the contributed predictors slightly in bagged MARS models whereas it did not influence BRT models at all. Indeed, the dominant variables selection was the same between bagged MARS models of all data including spatial weights and the spread-out dataset. As for *E. goniocalyx*, using different regressions or modelling with unlike dataset had little change on the

selection of relatively important predictors. For example, temperature standard deviation ranked the first place without change. According to the bagged MARS models (Table 23), temperature data were dominant environmental variables for *E. goniocalyx*, while BRT models (Table 24) indicated that, in addition to temperature variables, seasonal rainfall also contributed greatly to the distribution of this tree. In response to *E. fasciculosa*, bagged MARS and BRT both agreed that the average winter rainfall and the distance from the major water source were found to make greater contribution to the final model and might have a great influence on tree's occurrence.

3.4 Environment requirements

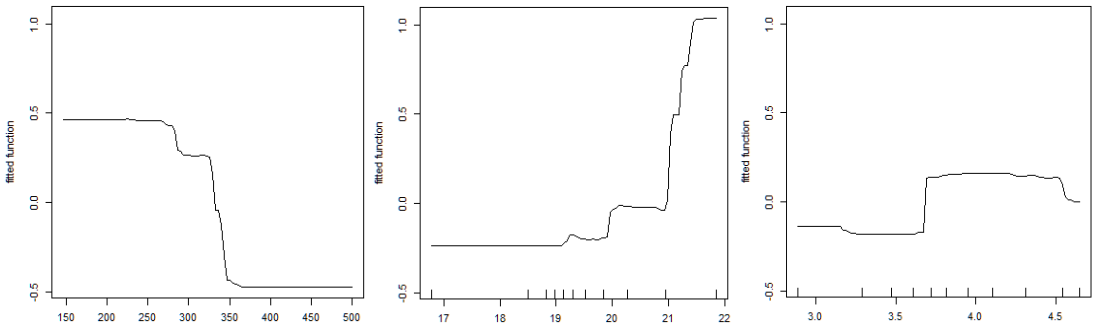
In this study, only the model with the best fit of data was used as the final regression to explore the environment requirements of the target trees. The dominant predictors (top three) and their importance of each tree were summarised in Table 7 to Table 10. Among the 34 environment data, the climate predictors were found to be more important than other predictor variables. 75% of the relatively important predictors (top three) was climate data while the other 25% was topographic variable. However, the spatial autocorrelation variable was not identified as relatively important predictors of any trees. The interactions between the response (PA) and those relatively important predictors were plotted and displayed in the response-predictors curve (see Figure 5 to Figure 8). These curves were basically showing the models fit of data. The abscissa of curves presented the values of environmental variable while the ordinate indicated the predicted probabilities of the existence of the tree (the distribution probabilities). Even though the ranges of predicted probabilities were from 0 to 1, the regression models suggested that certain ranges of predictor variables may have a negative contribution to the response.

In this study, the BRT model using the original PA dataset and spatial weights set had the best fit of data for *A. verticillata* (Dropping Sheoak). The modelling indicated that *A. verticillata* preferred to naturally grow in areas with

low level winter rainfall (from 150 to approximately 350mm) and high summer temperature (greater than 22 ^{o}C). The temperature standard deviation calculated from all mean monthly data was also important after the previous climate variables. More than that, the final regression suggested that excessive rainfall may have an obvious negative influence on *A. verticillata*. When the average winter rainfall increased to approximately 350mm or larger values, this variable maintained its negative correlation with response (distribution probabilities of Dropping Sheoak).

Table 7: Dominant predictors of *A. verticillata* (Dropping Sheoak).

Rank	Predictors	Importance
1	Winter rainfall	11.762
2	Summer temperature	11.045
3	Temperature standard deviation	7.378



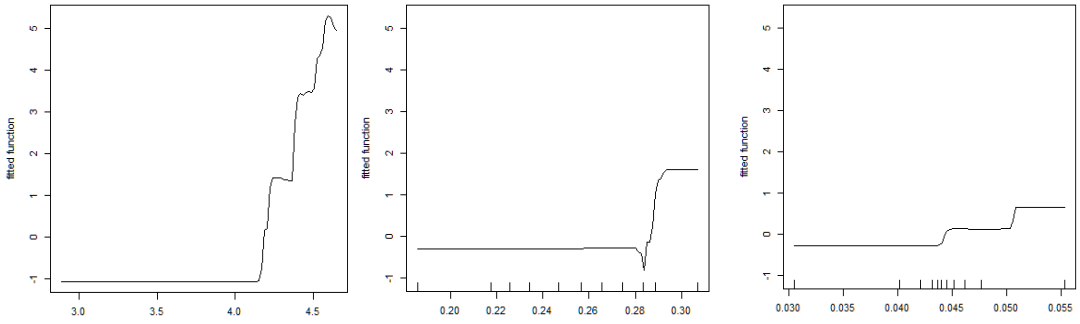
(a) Winter rainfall (b) Summer temperature (c) Temperature st deviation

Figure 5: Response-predictors curve of dominant factors of *A. verticillata* (Dropping Sheoak).

The final regression model of *E. goniocalyx* (Long Leaved-box) using BRT and the spacing method (SD spacing = 500) had the best performance. The modelling analysis found *E. goniocalyx*'s climate requirements. The standard deviation of the annual average temperature was a factor significantly affecting this tree's occurrence (importance = 65.78). If the standard deviation of the mean monthly temperature was less than 4.2, it was negatively correlated with the response. On the contrary, the greater the temperature dispersion was, the higher the predicted distribution probabilities of the tree were. According to the model, this tree might also prefer to occur in areas that had high seasonal temperature index (> 0.285) with greater difference between the standard deviation of rainfall in the driest month and the annual rainfall (> 0.05). The two seasonal indexes contributed relatively less than the temperature standard deviation, and these two indexes were almost of equal importance to the regression.

Table 8: Dominant predictors of *E. goniocalyx* (Long Leaved-box).

Rank	Predictors	Importance
1	Temperature standard deviation	65.777
2	Seasonal temperature	7.021
3	Seasonal rainfall	6.891



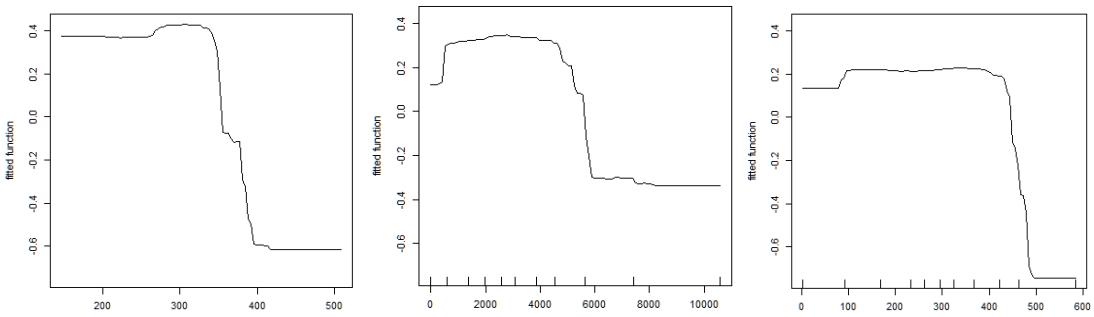
(a) Temperature st deviation (b) Seasonal temperature (c) Seasonal rainfall

Figure 6: Response-predictors curve of dominant factors of *E. goniocalyx* (Long Leaved-box).

The model evaluation results of *E. fasciculosa* (Pink Gum) indicating the best model performance was for the BRT model using the 1500m spread out dataset. According to the modelling analysis, *E. fasciculosa* was in favour of 100mm to 350mm average winter rainfall. Otherwise, excess rainfall may have negative influence on the response (presence/absence data). This tree also preferred to grow close to the water source; growing too far from the major streams (distance > 500 meters) may decrease the distribution probabilities of *E. fasciculosa*. Not only that, the regression modelling indicated that this tree species was also likely to be found in low or mid-high elevation areas. Areas above 450 meters were not suitable for the growth of the Pink Gum.

Table 9: Dominant predictors of *E. fasciculosa* (Pink Gum).

Rank	Predictors	Importance
1	Winter rainfall	11.914
2	Distance from major streams	9.231
3	Elevation	8.770



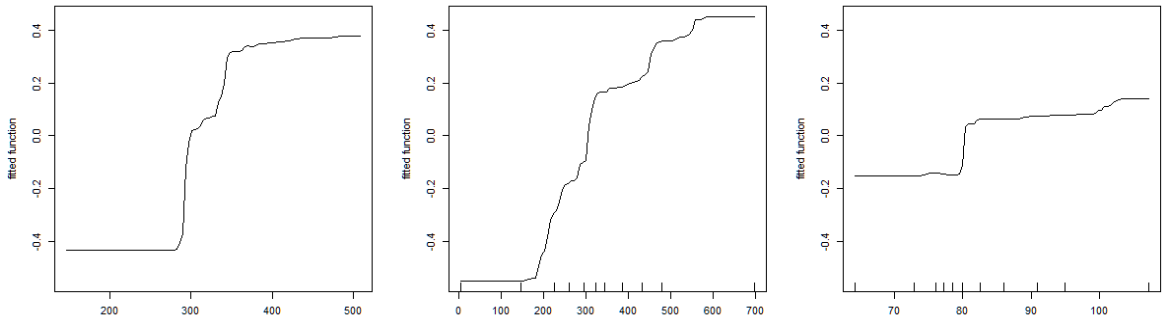
(a) Winter rainfall (b) Distance from major streams (c) Elevation

Figure 7: Response-predictors curve of dominant factors of *E. fasciculosa* (Pink Gum).

In this study, the regression model using all sample data and the spatial weights set had the best fit of data for *E. obliqua* (Messmate Stringy-bark). The regression model predicted that *E. obliqua* needed adequate rainfall amount (more than 300mm average winter rainfall and/or more than 80mm in summer). Provided its average rainfall requirement was satisfied, this tree also preferred to grow at high altitude regions (elevation greater than 300 meters). Comparatively speaking, the average winter rainfall and elevation made more contribution to the regression of *E. obliqua* rather than the summer rainfall.

Table 10: Dominant predictors of *E. obliqua* (Messmate Stringy-bark).

Rank	Predictors	Importance
1	Winter rainfall	22.561
2	Elevation	21.461
3	Summer rainfall	8.833



(a) Winter rainfall (b) Elevation (c) Summer rainfall

Figure 8: Response-predictors curve of dominant factors of *E. obliqua* (Messmate Stringy-bark).

The interactions of the most common soil type and geology with distribution probabilities of target trees were plotted (see Figure 13 and 14 in **Appendix 3**). The relevant codes of soils and geology data were listed in **Appendix 4** and **Appendix 5**. The target trees were found more likely to grow on a wide range of soils types including shallow to moderate deep acidic loams, shallow soils on rock, deep sands, saline soils and wet soils, while their geology requirements varied greatly. *A. verticillata* was more likely to grow on the following geology types: siltstone, limestone, dolomites, quartzite, mudstone or sandstone. The regression suggested that *E. goniocalyx* occurred more frequently at locations where the geology was one of the following categories, amphibolite, barite, basal quartzite, dolerite, pegmatite or pyrite. Notably, *E. fasciculosa* was more commonly found on granite, marble, limestone, mudstone, siltstone or shale. Most importantly, *E. obliqua* was likely to be found in the following geological categories: carbonaceous clay, dolomite, siltstone, limestone, shale, quartzite or sandstone.

3.5 Spatial weights method

An index of spatial autocorrelation (**Spatial_A_influence**) was introduced to each regression algorithm. These spatial weights method was one way to ensure a model was not biased by spatial autocorrelation. The ranks and importance of spatial weights set using regression models and all sample data were summarised in Table 11 to Table 14. In this study, importance of spatial variable using bagged MARS models was zero for all trees whereas for BRT models the ranks of the spatial variable increased up to the 17th. The highest importance was 1.794 for the regression of *E. fasciculosa*, followed by *E. obliqua* and *A. verticillata*. The smallest spatial importance was only 0.107 for the BRT model of *E. goniocalyx*. The interactions between the response (PA) and spatial weights set were plotted in R software (R Core Team 2014) showing the fitted values (predicted distribution probabilities) in ordinates and spatial variable values in abscissa. In Figure 9 (in **Appendix 6**), spatial interactions of each trees using BRT models were plotted to illustrate the influence of spatial

variable on response. In general, spatial variable of *A. verticillata* and *E. obliqua* had positive relation with the dependent variable, their corresponding maximum prediction probabilities were 0.07 and 0.03, respectively. At the same time, spatial weights set of *E. fasciculosa* was negatively related with response. In addition, its maximum fitted value was the highest of all trees modelled.

Table 11: Ranks and importance of spatial weights set recorded regressions for *A. verticillata*.

Model	Rank	Importance
Bagged MARS	35	0
BRT	23	1.301

Table 12: Ranks and importance of spatial weights set recorded regressions for *E. goniocalyx*.

Model	Rank	Importance
Bagged MARS	35	0
BRT	19	0.107

Table 13: Ranks and importance of spatial weights set recorded regressions for *E. fasciculosa*.

Model	Rank	Importance
Bagged MARS	35	0
BRT	17	1.794

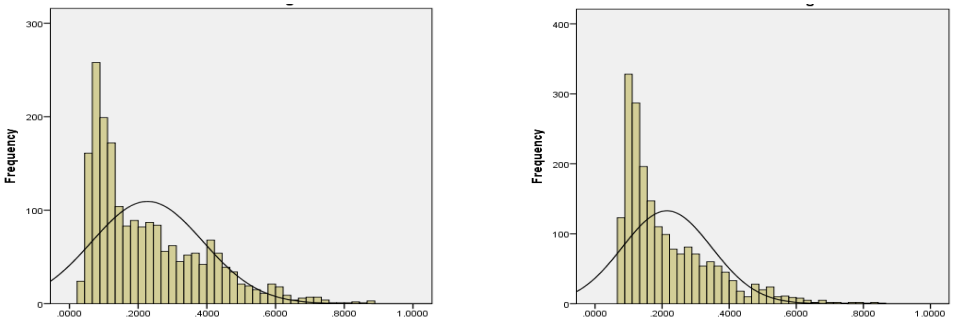
Table 14: Ranks and importance of spatial weights set recorded regression for *E. obliqua*.

Model	Rank	Importance
Bagged MARS	35	0
BRT	17	1.151

3.6 Model prediction

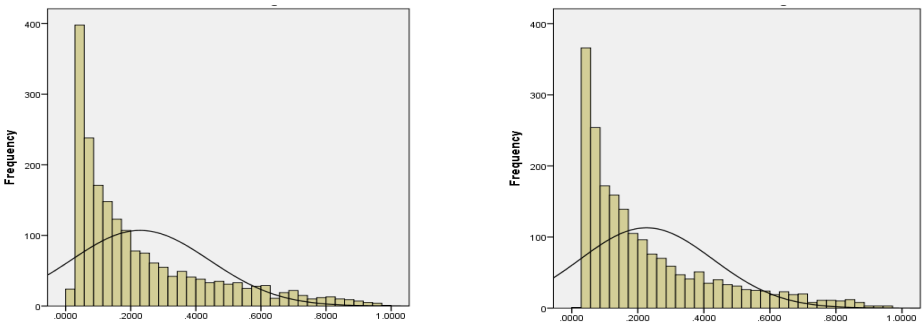
As previous mentioned, the models' fit, which indicates the interaction between response and predictors of input data can be projected to unknown area for the purpose of predicting (Guisan et al. 2013). Thus, the potential distribution probabilities of the study site of the four target species were predicted using the regression functions (bagged MARS and BRT in **Section 2.7** and **Section 2.8**) and the environmental variables (**Section 2.4**) and the spatial variable (**Section 2.6**). After the introduction of spatial weights, the central trend of models' prediction remained unchanged; their statistical data, including the mean value, variance, etc., fluctuated slightly. However, the distribution probabilities predicted by models using large spacing distance PA dataset had obvious changes, their mean values increased.

In Figure 9, the predicted probabilities of *A. Verticillata* using two regressions generally varied from 0.02 to 0.88, and the whole data were inclined to the left (low probabilities). Approximately 58% of prediction concentrated in the first interval (0 to 0.2).



(a) Exclude spatial weights (b) Include spatial weights

A: Bagged MARS

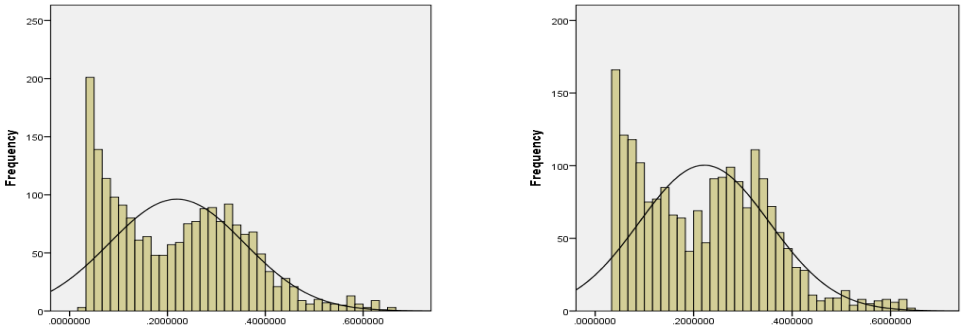


(c) Exclude spatial weights (d) Include spatial weights

B: BRT

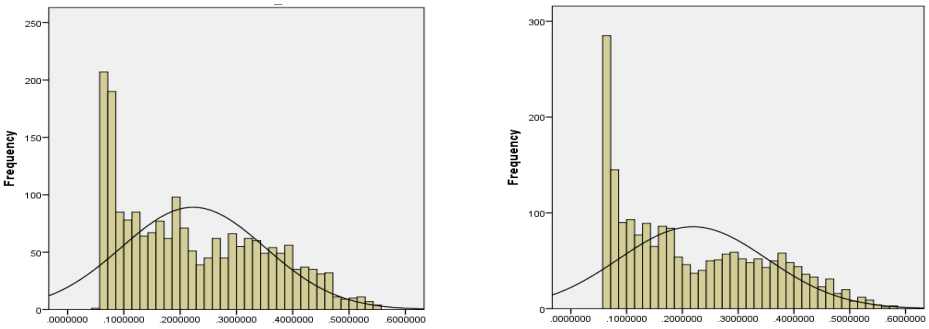
Figure 9: Frequency histogram of model prediction for *A. verticillata*.

For *E. obliqua* (Figure 10), prediction probabilities obtained by the two methods only varied from 0 to around 0.65 and they were not normal distribution showing double peaks. A large number of data (26%) clustered in 0.05 to 0.1 interval while a relatively small amount (15%) gathered around 0.3.



(a) Exclude spatial weights (b) Include spatial weights

A: Bagged MARS

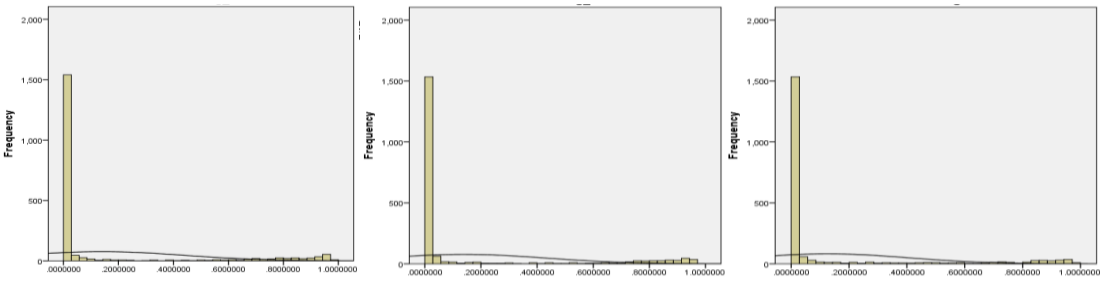


(c) Exclude spatial weights (d) Include spatial weights

B: BRT

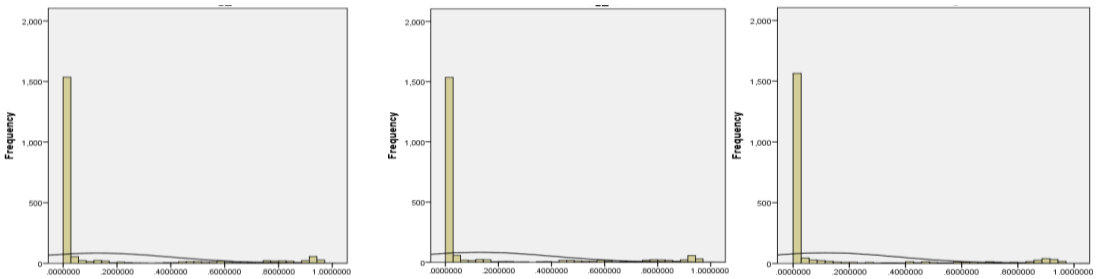
Figure 10: Frequency histogram of model prediction for *E. obliqua*.

The frequency histograms of *E. goniocalyx* (Figure 11) had no notice changes, prediction probabilities highly concentrated in the first column, which was the 0 to 0.029 range. Nearly 77.5% of the data were in this interval with a frequency about 1550.



(a) Exclude spatial weights (b) Include spatial weights (c) 500m PA dataset

A: Bagged MARS

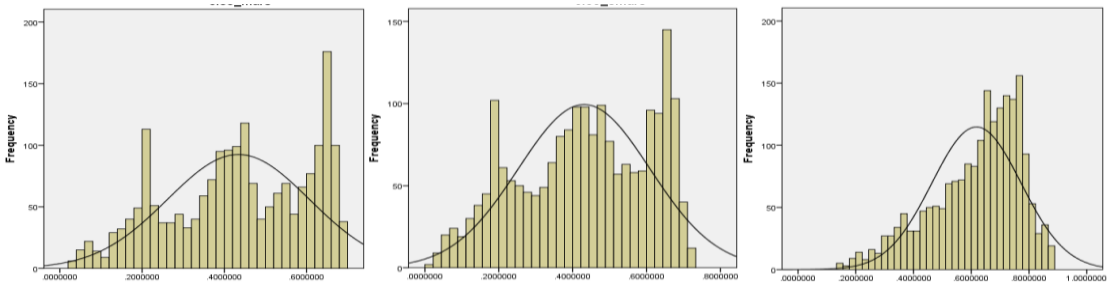


(d) Exclude spatial weights (e) Include spatial weights (f) 500m PA dataset

B: BRT

Figure 11: Frequency histogram of model prediction for E. goniocalyx.

For *E. fasciculosa*, Figure 12a and 12b were saw-tooth histograms showing model predictions' frequency using bagged MARS models with selected environmental variables. They had several peaks, with high frequency data were around 0.2, 0.4 and 0.7. Meanwhile, Figure 12d and 12e were flat topped histograms indicating that the overall predictions using BRT models of all samples was not very different. The frequency of data from 0.2 to 0.8 was relatively high and varied little. After increasing the sample distance to against spatial influence, the prediction using two algorithms were tilted to the right (high probabilities), with high frequency data gathered between 0.62 and 0.75 (Figure 12c and 12f).



(a) Exclude spatial weights (b) Include spatial weights (c) 1500m PA dataset

A: Bagged MARS

(d) Exclude spatial weights (e) Include spatial weights (f) 1500m PA dataset

B: BRT

Figure 12: Frequency histogram of model prediction for E. fasciculosa.

SDMs can also be used to create probability surface showing habitat niche (Guisan et al. 2013). In this study, IDW technique interpolated raster surfaces of previous regression models' prediction, and every grid cell was on a 0 to 1 continuous scale; 1 was the maximum probability of a certain species' occurrence. These raster surfaces of habitat niche were displayed by grouping cell values into ten evenly spaced classes. High distribution probabilities (e.g., 0.8 to 1) were assigned in red color series, low values (e.g., 0 to 0.2) were shown in blue-colored items while green and yellow colors were used for those in between (e.g., 0.2 to 0.8).

More than half of the model prediction of *A. verticillata* were less than 0.2, thus most areas of the study site had small or moderate distribution probabilities, and these areas located in the spine of the Mount Lofty Ranges (see **Appendix 7**). For *E. goniocalyx*, high distribution probabilities (> 0.8) concentrated in the north corner of the study site while the rest area only had little chance (< 0.2) for this tree to naturally live in there (see **Appendix 8**). With regarded to *E. fasciculosa*, high prediction values shown in the lower right edge areas and upper left corner. Spacing sample distance raised the overall distribution probabilities of this tree (see **Appendix 9**). The upper-middle area within the Mount Lofty Ranges had relatively high distribution probabilities (0.5 to 0.7) to grow *E. obliqua* naturally. And the low values area of this tree was predicted to be the lower right corner (see **Appendix 10**).

3.7 Threshold of distribution probabilities

As previously mentioned, transforming the prediction surfaces of habitat niche into binary presences/absences maps is often more practical for conservation projects and environmental management. Here, a specific threshold determines the information presented as species' occurrence. According to the intrinsic evaluation results from **Section 3.2**, for *A. Verticillata* and *E. oblique*, the best predictive ability was the BRT model using full sample data (30m PA dataset) and including spatial weights set; meanwhile, for *E. goniocalyx* and *E. fasciculosa,* the final BRT model using the spacing method

had the best performance. The data correction of these models was calculated with the specific threshold of distribution probabilities, see details in Table 15 to Table 18.

The positive (P) and negative (A) correction rate was the comparison between the predictions and the actual PA data. A larger correction rate in the following tables indicated a better model performance with the specific threshold; conversely, low values represented low model fit. For instance, 50% is an acceptable data correction rate, below this value, the model accuracy is too low to be adopted. The fixed threshold method using 0.5 performed worse for *E. obliqua* with a very low presence correction, only 11.72%. Rather than that, using the mean value as the threshold, both the true negative and positive rate were accepted and the overall correction of the model was the highest (see Table 18). For the other three trees, the presence and absence correct rate for each threshold was all acceptable (> 50%) and using mean values as threshold had the highest average correction.

Table 15: Correction rate for *A. verticillata*.

Type	Threshold	True Negative Rate (Absence correction)	True Positive Rate (Presence correction)	Average Correction
Fixed	0.500	97.40%	56.39%	76.90%
Mean	0.209	86.20%	91.73%	88.97%
Median	0.102	63.00%	98.50%	80.75%

Table 16: Correction rate for *E. goniocalyx*.

Type	Threshold	True Negative Rate (Absence correction)	True Positive Rate (Presence correction)	Average Correction
Fixed	0.500	98.20%	95.45%	96.83%
Mean	0.180	95.20%	99.09%	97.15%
Median	0.005	61.00%	100.00%	80.50%

Table 17: Correction rate for *E. fasciculosa*

Type	Threshold	True Negative Rate (Absence correction)	True Positive Rate (Presence correction)	Average Correction
Fixed	0.500	64.46%	94.03%	79.25%
Mean	0.625	82.64%	83.08%	82.86%
Median	0.690	89.26%	73.63%	81.45%

Table 18: Correction rate for *E. obliqua*

Type	Threshold	True Negative Rate (Absence correction)	True Positive Rate (Presence correction)	Average Correction
Fixed	0.500	86.82%	11.72%	49.27%
Mean	0.283	51.71%	78.17%	64.94%
Median	0.289	52.53%	76.89%	64.71%

3.8 Presence/Absence maps

Model predictions of BRTs including spatial weights term were transformed into presence/absence maps for *A. Verticillata* and *E. oblique*, while model predictions of BRTs using the spread out dataset were used to develop presence/absence maps for *E. goniocalyx* and *E. fasciculosa*. In this study, using the average values of prediction probabilities as threshold can provide acceptable presence/absence data correlation and the highest overall fit of data. Thus, the mean values were used to transform continuous model prediction into binary presence/absence map, which could directly show potential habitat niche areas. The final predictive maps indicated presence distribution in green and absence in grey.

Appendix 11 shows a large presence areas of *A. verticillata* to the central-east of the study area, with other scattered areas of high preference to the north, central-west and south. The total potential presence habitat niche area of *A. verticillata* is 1453.956 km^2. The area representing high preference of *E. goniocalyx* is the smallest of all target species modelled.

Appendix 12 shows that the presence area of *E. goniocalyx* concentrates in the north of the study area while the rest areas indicate absence distribution. The total potential presence habitat niche area of *E. goniocalyx* is 609.788 km^2. The area representing high preference of E. fasciculosa is the greatest of all target species modelled.

Appendix 13 shows the widespread distributions of *E. fasciculosa*. The total potential presence habitat niche area of *E. fasciculosa* is 1934.124 km^2.

Appendix 14 shows the presence area of *E. obliqua* in the middle-upper areas, with another two relatively small patches to the south. The total potential presence habitat niche area of *E. obliqua* is 1245.688 km^2.

3.9 Extrinsic comparison

As previously mentioned, due to the limit sample data size, the final predictive surfaces of habitat niche were compared with an existing map of expert opinions. It was shown with black border with the transparent color in **Appendix 11** to **Appendix 14**. The general locations of presence distribution were in good agreement with the predictive habitat niche. However, the predicted presence areas were larger than Croft's (unpublished) analysis results.

For *A. verticillate* (Dropping Sheoak), there were approximately 12.52% predictive niche areas in line with Croft's (unpublished) distribution map. In the middle, the two maps had similar presence locations, but models' predictive areas were much larger. In addition, in the north, according to Croft's (unpublished) map, the probability of the occurrence of the tree was little, while

the regression model predicted that the northern environment was also suitable for the natural growth of the tree.

For *E. goniocalyx* (Long Leaved-box), although only approximately 17.76% of the predictive presence distributions agreed with the areas observed by Croft (unpublished), they were all highly concentrated in the northern area.

For *E. fasciculosa* (Pink Gum), approximately 47.70% of the predictive presence distributions agreed with the areas observed by Croft (unpublished). In the northern of the study area, the modelling analysis largely extended the Croft's (unpublished) distribution areas from Barossa Goldfields and Para Wirra Recreation Park to Mount Torrens. Moreover, in the middle part of MLR, modelling outputs also revealed a preference of this tree for the Mount Barker and its surrounding areas.

For *E. obliqua* (Messmate Stringy-bark), approximately 65.51% of the predictive presence distributions agreed with the observation of Croft (unpublished). This was the highest degree of coincidence.

CHAPTER FOUR: DISCUSSION AND SUMMARY

4.1 Key findings

An understanding of the manner in which trees interact with the environment is fundamental for their cultivation and conservation (Bradshaw 2012). Environmental managers need robust species' predictions in order to appropriately capture the potential interactions within communities and ecosystems (Guisan et al. 2006). This study used two regression algorithms, bagged MARS and BRTs, to explore the complex relationships between each of four tree species and the environment. Generally, BRTs have a better model fit compared to the regressions using bagged MARS (see Figure 4). The modelling analysis tested the candidate predictors (35 in total) and accessed the environment preference of the target trees. The relatively important environmental variables (top three predictors) are thought to best describe each species' realised environmental requirements within the study site. The regressions found that among the 35 environmental predictors, climate data were the determinants to the distribution of target trees (see Table 7 to Table 10). This finding is supported by Adamson and Osborn (1924), Stead (2008) and Green (1994); they concluded that climate, particularly rainfall, significantly contributed to the tree occurrence within the Mount Lofty Ranges. The final regression models indicating the best of model fit of each tree were then used to predict the tree distribution probabilities.

The data independent test results (see Figure 3) explained that, the target trees had different strengths of spatial autocorrelation using 30m PA dataset. Spatial weights term method and spacing sample distance method were then used to against spatial autocorrelation. Comparing the impact on the models of these two methods, the spatial weights term method had less influence than the spacing sample method. For example, the candidate predictors' impact using BRTs were unchanged after the introduction of spatial weights term, while after increasing the sample distance, the modelling results slightly altered (see **Appendix 2**). The contribution of spatial weights term on the models was described in **Section 3.5**. Specifically, the spatial variable did not contribute to

any bagged MARS models, although it altered BRTs' structure. Indeed, although the spatial weights term contributed in BRTs, it was still not a crucial factor. Spatial autocorrelation cannot directly or significantly influence the response (plants occurrence of the target tree).

4.2 Predictive models

The algorithm of the model and the selection of variables will directly affect the prediction results (Guisan & Zimmermann 2000; Lehmann et al. 2002b). In addition, the distribution characteristics of the data itself are the main factors that affect the prediction results of the model. The Mount Lofty Ranges had a long history of vegetation clearance for accessing productive lands for agriculture, urban development, as well as for the collection of timbers for construction and fuel (Williams 1977; Szabo et al. 2011; Paton et al. 2000). For this reason, the majority of the remaining native vegetation occurred on unsuitable topography, soil types, and even climate (Lethbridge & Green unpublished; Stead 2008). Thus, the final regression model may be over fitted the plant's sensitivity to the environment. This bias reveals a common criticism of using regression models to predict the species' distribution; it relates to the difference between the fundamental and the realised niches of a species. The fundamental niche is the sum of all the abiotic conditions necessary to maintain the survival of the species. It reflects the physiological needs of a species, and is biologically significant (Hutchinson 1957). However, due to the interaction between different species, the fundamental niche of a species is usually not fully expressed or reflected in a specific geographical area (Hutchinson 1957). Therefore, Hutchinson (1957) used the realised habitat niche to represent the actual or observed range of a species within their geographical areas (Hutchinson 1957; Lethbridge et al. 2006). In this study, the regression models have been trained using extensive field survey information, thus the final models were more likely to describe the potential habitat preferences closer to the realised environmental niche at their geological locations rather than the fundamental niche of a target species (Guisan & Zimmermann 2000; Franklin

1995; Malanson et al. 1992). In other words, the environmental preference of the target tree habitat was driven by the input data; it explained the interactions between surveying PA data and provided environment data. The model predictions have regional characteristics, and they may not indicate the true environmental limits of the target tree species.

Based on the intrinsic model ROC tests in **Section 3.2**, the overall performance of all regression models was concluded to be good. Comparing the predictive ability between the bagged MARS and BRTs, it indicated that BRT models had a better fit of data and showed relatively stable prediction results (see Table 5, Table6 and Figure 4). Specifically, the BRT model of *A. verticillata*, *E. goniocalyx* and *E. fasciculosa* had excellent performance (0.9 < AUC < 1.0), while the predictive ability of the BRT model of *E. obliqua* was relatively low, corresponding to fair model performance (0.7 < AUC < 0.8) (see Figure 4). However, good model performance is of no value if the final model is not biologically relevant (Stead 2008). The biological relevance of each the predictive models were explained in **Section 3.3**, and the interactions between the response (PA) with the dominating predictors (top three) were further described in detail (see **Section 3.4**). These environmental requirements are compared with literature and expert opinions.

More specifically, the relatively important environmental predictor variables of *A. verticillata* (Dropping Sheoak) and *E. obliqua* (Messmate Stringy-bark) are in excellent agreement with previous researches. In this study, the regression models indicated that *A. verticillata* preferred to grow in areas with low level winter rainfall (< 350mm) or high summer temperature (> 22 ^{o}C) (see Table 7 and Figure 5). Furthermore, an excessive rainfall may have a negative influence on this tree. The rainfall requirements were in line with the observations of Armstrong et al. (2003). They found that, *A. verticillata* was likely to be found in the dry portion of the Mount Lofty Ranges (Armstrong et al. 2003). Similarity, Bonney (1997) found that *A. verticillata* preferred a hot dry climate with less annual rainfall (150mm to 250mm). Meanwhile, the temperature limits were also supported by Stead (2008), who found that a high

average summer temperature (> 21 ^{o}C) made great contribution to *A. verticillata*'s distribution.

E. obliqua (Messmate Stringy-bark) is a rainfall sensitive tree species (Specht & Perry 1948; Armstrong et al. 2003). Stead (2008) explained that this species preferred to occur in the high winter rainfall regions (greater than 252mm) within the Mount Lofty Ranges, and Bonney (1997) found that its average annual rainfall varied from 650mm to 1000mm. *E. obliqua* also prefers to live in hilly or mountainous areas with higher elevations (Kantvilas & Jarman 2004; Sinclair 1980; Stead 2008). These findings are also supported by Lethbridge et al. (2006), who found that, within the Mount Lofty Ranges, *E. obliqua* was likely to be found either in areas with sufficient rainfall, or in locations where elevation was greater than 345m. These views are consistent with the modelling analysis results obtained in this study. The regression models predicted that *E. obliqua* needed an adequate amount of rainfall, especially in winter (> 300mm), and also that this species preferred to grow at high altitudes (> 300m) (see Table 10 and Figure 8).

It should be noted that, for *E. goniocalyx* (Long Leaved-box), the regression prediction results indicated that the relatively important predictors to the response were not consistent with those of other studies. This study found *E. goniocalyx*'s climate requirements, and that the standard deviation of the annual average temperature significantly influenced this tree (Table 8). In addition, seasonal rainfall and temperature data can also affect this trees' distribution. Indeed, the modelling results of Lethbridge et al. (2006) were quite different from these findings. They found that topographic data such as elevation (> 235m), distance to major water source (< 2407m) and geology were dominant variables rather than climate data (Lethbridge et al. 2006). There are several possible reasons that may explain the difference between these two studies. Firstly, the modelling methods used are different. Their research applied MARS as regression method, while this study used BRTs. Secondly, in order to eliminate the spatial autocorrelation as much as possible, the final model of this tree species used a largely spaced sample dataset (SD = 500m).

On the whole, the predicted environment requirements of *E. fasciculosa* (Pink Gum) had similarities with the literatures. According to the modelling, *E. fasciculosa* was likely to be found in middle-level rainfall zone (100mm to 350mm average winter rainfall) (Table 9 and Figure 7). The rainfall preference here is supported by Bonney (1997). Additionally, Lethbridge et al. (2006) found that the distribution of *E. fasciculosa* was affected by the geology categories, given that the rainfall amount was met; at the same time, this tree also preferred to grow in regions with a certain surface slope, provided the soil types were suitable. These soil and geology preferences from Lethbridge et al. (2006) were in consistence with the modelling of this study; however, the BRT regression indicated that the effects of topography on the distribution of this species were not as great as either a close distance to major water source (< 500 m) or a moderate elevation (< 450 m). Again, the disagreements between the prediction results with the observations of Lethbridge et al. (2006) could be due to the different modelling algorithms used, or caused by the spacing sample distance method used for this tree in this study.

Not surprisingly, the species' distribution was generally closely related to the climate conditions. However, the necessary nutrients and the water for the trees' life are mainly absorbed from the soil through their roots (Sharma 2000). Soil and geology are also important ecological factors, which can directly affect the growth and yield of trees (Pierzynski et al. 2005). Prior studies found that soil types and geology can exert a great contribution over the distribution of aid tree species. For example, Stead (2008) explained that *A. verticillata* was more likely to occur on rocky sites. Baker and Smith (1902) found that *E. goniocalyx* preferred poor stony and well drained acidic to neutral soils (Guerin & Lowe 2013). They also reported that most of *E. goniocalyx* were likely to be found on basalt and granite ridges (Baker & Smith 1902). Armstrong et al. (2003) observed that *E. goniocalyx* was a species associated with dry sclerophyll forests, and it can survive on infertile soils within Mount Lofty Ranges. White (2015) discussed that *E. fasciculosa* is useful for streets and park planting, as it can perform well on sandy or rocky terrains, or even poor fertility soils (Dean & Ian 2013). *E. obliqua* is in favour of a wide variety of soils such as acidic loams, deep sands and shallow soils (Stead 2008). These views of soil and

geology requirements for the target trees had all been confirmed in this study, and the predictive range of soil and geology preferences were much wider. According to the model analysis, the target trees were found to be more likely to grow on a wide range of soils types with varied geology categories (see Figure 13 and Figure 14 in **Appendix 3**). Generally, the target trees had a strong ability to adapt to different soil types and varied geology categories, even soils with poor fertility or rocky sites with minimal soil (Sharma 2000; Pierzynski et al. 2005; Stead 2008). Apart from that, the significant vegetation clearance in the MLR region has resulted in that the remaining trees being distributed within unsuitable environment areas. Therefore, the regression models cannot explore the natural soil and geology requirements of the target trees; instead, they highlighted the regional soil and geological characteristics within the MLR. That is, the soil types and geology can affect the distribution of trees naturally; however, within the study area, they were not the decisive factors on the distribution of the trees, especially when compared with the dominant effects of climate variables.

4.3 Predictive surfaces

The realised niche of four native trees in the Mount Lofty Ranges of SA have been modelled. It is practical to display the habitat niche in binary presence/absence map rather than continuous predictive probabilities (Liu et al. 2005). The final PA surfaces were compared with prior knowledge and existing distribution maps.

The final PA surface did not always agree with the distribution map of Tim Croft (unpublished). Croft's maps were qualitative analysis results which partly depended on the subjective judgment. However, the predictive surfaces generated by the regression models provided quantitative analysis results. This study used a large amount of survey data and historical environment data records to provide an objective result. Due to the significant historical clearance of the Mount Lofty Ranges, Tim Croft may have subjectively underestimated the past distribution of species within the MLR region, therefore the quantitative

results of habitat models had a wide range of fit than Croft's map (unpublished). For example, the existing distribution map of *A. verticillata* and *E. goniocalyx* had an extremely low consistency (12.52% and 17.76%, respectively) with the predictive presence distribution (**Appendix 11** and **Appendix 12**). In general, the two distribution maps had similar results indicating the approximate presence locations of the trees; however, the areas of the PA surface were large, and the edges of the predictive surface were much smoother.

The modelling outputs indicated that *A. verticillata* demonstrated a preference to drier areas with high average summer temperature (see Table 7 and Figure 5). **Appendix 11** showed the areas that satisfied the above environmental requirements. Adamson and Osborn (1924), Specht & Perry (1948), and Boomsma and Lewis (1980) indicated that *A. verticillata* commonly occurred on cliffs and rocky outcrops in the Mount Lofty Ranges. More specifically, Stead (2008) predicted a large niche area of *A. verticillata* near the town of Strathalbyn, and other small patches in the eastern side of the Barossa Valley as well as the north-west areas of the Fleurieu Peninsula. These presence distributions of Stead (2008) and the observation data of Dashorst and Jessop (1990) were in agreement with the model prediction of this study. A large presence areas were found in the middle-right of the study area, extending through Strathalbyn to Nairne, with other scattered areas of high preference to Barossa valley in the north, and the Fleuriue Peninsula in the south (see **Appendix 11**).

Armstrong et al. (2003) reported that *E. goniocalyx* was largely confined to the Barossa regions, and was mainly distributed in the area near Mount Crawford and the River Torrens. Croft (unpublished) reported that this tree occurred as a woodland formation in the north of the Mount Lofty Ranges. Dashorst and Jessop (1990) found large amount of presence data of this tree species in the north-east of the Southern MLR region. The regression outputs of *E. goniocalyx* agreed well with these findings. **Appendix 12** indicated that the presence area of *E. goniocalyx* was only avaiable in the north of the study area. The area representing high preference of *E. goniocalyx* was the smallest of all target species modelled. This tree was known to have a limited distribution

in the Mount Lofty Ranges; this may be due to its particular climate requirements. The modelling found that the standard deviation of the annual average temperature was a significant determinant that can directly influence the species (see Table 8 and Figure 6). However, the temperature changes of the whole study area were not obvious, and only little areas can satisfy *E. goniocalyx*‘s climate limits.

The area representing high preference of *E. fasciculosa* was the greatest of all target species modelled. This was because most of the study area can meet this tree’s environmental preference, which is middle level winter rainfall (100mm to 350mm) (see Table 9 and Figure 7). This variable played a decisive role in the distribution of trees. The other relatively important predictors, closer to water source and low or moderate elevations, contributed less than the average winter rainfall amount. **Appendix 13** displayed the widespread presence distribution areas of this species, which highly agreed with the observation of Dashorst and Jessop (1990); small disagreement areas were found in Summertown and Ashton.

E. obliqua was a rainfall sensitive tree species (Bonney 1997; Specht & Perry 1948; Stead 2008; Armstrong et al. 2003). Specht and Perry (1948) indicated that *E. obliqua*‘s majority distribution was mainly limited to high rainfall areas along the spine of the Mount Lofty Ranges. This was confirmed by the modelling analysis. The regression predicted the distribution of *E. obliqua* to be in the higher elevation areas (greater than 300 meters), provided its rainfall requirements were met (average winter rainfall > 300mm) (see Table 10 and Figure 8). **Appendix 14** indicated the likelihood of *E. obliqua* being present in the middle-upper areas, with another two relatively small patches to the south. This tree’s predictive surface had the highest degree of coincidence with the map observed by Croft (unpublished) (approximately 65.51%). This may be because *E. obliqua* was a dominant species that commonly occurred within the Mount Lofty Ranges. Provided its climate requirements are met, *E. obliqua* can survive in a variety of environments. Its distributions were widely spread from south to north of the MLR, but confined to climatic conditions. Thus, it was easy to analyse *E. obliqua*‘s distribution.

4.4 ROC test

The prediction probabilities of trees' distributions are on a continuous scale from 0 to 1, and the sensitivity (true positive rate) and specificity (true negative rate) of the model prediction will both change as the threshold alters (Davis & Goadrich 2006). Therefore, if the sensitivity of a regression is high while the specificity of the other method is high, it is difficult to compare these two with a specified threshold. The ROC plot analysis provides an independent model evaluation, without any reliance on thresholds (Anderson et al. 2003; Guisan & Zimmermann 2000). In this study, the ability of each regression algorithm to correctly predict its PA data was evaluated using the AUC values of the ROC plot analysis. The nature of a ROC plot is a dynamic analysis. The plot marks the positive and negative correction rate of different thresholds (Anderson et al. 2003; Guisan & Zimmermann 2000). The AUC values can then be used to evaluate the generality of models, which was the overall predictive ability of regressions (Fawcett 2006). This threshold-independent technique avoids any assumptions about the distribution characteristics of the model (Stead 2008; Fawcett 2006). However, the drawback of this method is that it does not consider the spatial distribution of classification errors (Barry & Elith 2006). For instance, AUC values recording regressions of *E. obliqua* all corresponded to a fair model performance, which was the worst of all target species modelled. This was because *E. obliqua* had an imbalanced PA data (P: 623; A: 1578, see Table 4). Increasing the number of samples is likely to improve the predictive power of the distribution models (Lethbridge et al. 2006). *E. obliqua* had a lack of presence samples, thus, the true positive accuracy may be dropped; meanwhile, there was sufficient absence data, and the true negative correction was relatively high, and this made up for the loss of true positive rate. The overall predictability of the model was acceptable, while its sensitivity was too low. In summary, the ROC plot can illustrate the general ability of one or more algorithms in the case where there are the same or a similar amount of negative and positive examples (Davis & Goadrich 2006). In the case of an uneven distribution of response, e.g., there are many more negatives than positives, AUC values cannot appropriately reflect the relatively poor performance of positive data (Davis & Goadrich 2006).

4.5 Threshold

As previously mentioned, a threshold can determine the information indicating presence and absence, moreover, it can alter the positive and negative correction rate of a continuous model prediction (Davis & Goadrich 2006). The optimal threshold of predictions is the one that has an acceptable true positive and negative correction (> 50%). 0.5 is a representative objective threshold and has been widely used in ecology researches. However, directly taking 0.5 as the fixed threshold to transform occurrence probability may give incomprehensible results for a characteristic of class imbalance dataset (Liu et al. 2005). In this study, only *E. fasciculosa* had a relatively even number of presence (201) and absence (121) samples (see Table 4). Its average and middle values of predicted probabilities were close to 0.5. Thus, their corresponding data corrections were similar (see Table 17). The presence data of the other three trees were much lower than the absence points. When a larger number of occurring sites were not available, the balance point of presence/absence was then lower than 0.5. If still selecting 0.5 as a threshold, it may cause a high data loss rate in predicting presence distribution. That is, a large threshold can significantly reduce the sensitivity (true positive rate). For example, *E. obliqua*'s PA data was seriously unbalanced (P: 623; A: 1578, see Table 4). Using 0.5 as the threshold of prediction probabilities, its true negative rate was relatively high with 86.82% while its true positive rate was only 11.72%. However, choosing mean or median values can make up for the data loss (see Table 18). In view of that situation, an appropriate threshold should be determined according the ratio of positive and negative samples.

4.6 Spatial autocorrelation

SpaceStat$_{TM}$ (BioMedware 2014) is an inclusive software package, which comprises various methods for spatial statistics, geo-statistics and spatial econometrics. It can create spatial weights set based on adjacency evaluations. The creation of spatial weights sets is then influenced by all the values of

datasets related to the geography, including missing values. Hence, if missing values occurred in a dataset, fewer neighbours than those were initially established may be employed to compute the weight set. Therefore, this study only considered the spatial effect of all samples (30m PA dataset) due to the missing points of the spacing sample dataset.

In the aspect of landscape ecology, spatial autocorrelation analysis has become the main method to study the spatial pattern of landscape (Zhang & Zhang 2003). Geographic datasets may lose independent samples owing to spatial autocorrelation (Anselin 1988). In order to resist spatial influence in modelling, a method was used to largely space the distance between sample points. An alternative way to ensure that a model is not biased by spatial autocorrelation is to introduce spatial terms into model analysis. In modelling analysis, the spatial relationship of species may contribute to the interaction between response and predictors. As a result of that, spatial weights set may alter the model's structure (Anselin 1988). The importance of the spatial weights term using bagged MARS models was zero, whereas for BRT models it increased (see Table 11 to Table 14). In other words, the spatial variable did not contribute to bagged MARS models, although it slightly altered BRT models.

In this study, bagging was integrated into MARS models to improve the prediction accuracy and robustness of individual models (Breiman 1996; Huang & Wang 2014; Xue 2016). Thus, the ranks and importance of predictor variables of a bagged MARS model were the average results of all sub-models. There are two possible reasons for a variable not contributing to the final bagged model. Firstly, bagging applies a bootstrap sampling technique to randomly selected samples with replacement (Breiman 1996; Efron & Tibshirani 1986; James et al. 2013). Each sample set has the same size as the original dataset; however, several samples may be repeated many times, while others may not be selected by any bagging process (Kuhn & Johnson 2013). Thus, the un-selected variable(s) is not modelled. Secondly, as previously mentioned, each MARS model is a sum of basis functions; if a predictor is never used in any sub-models, it means that this predictor did not contribute to the response, and its overall importance is zero in the final bagged MARS (Moisen & Frescino 2002;

Friedman 1991). Sufficient numbers of bootstrap samples (B = 500) were provided in this study to ensure that each variable associated with the response can be included in at least one bootstrap sample set. However, the random selection process is uncontrollable. Therefore, the zero importance of spatial variables using bagged MARS indicated that spatial weights tern may either not be selected by any of the bootstrap sample sets, or it was unimportant to the response. No matter what the above situation, all the changes in bagged MARS models in this study were independent of spatial autocorrelation, but related to the random bagging process and/or the large spacing sample distance.

Similarity, if a predictor is used in any splits in any individual trees of BRT models, this predictor has an important value corresponding to its contribution in the final boosted models. The ensemble learning technique used in regression trees in this study was boosting, which sequentially selected variables for training basic learners (Breiman 1996; Freund & Schapire 1996). Boosting evaluates the previous individual model and will select poorly performed predictors for the following sub-model (James et al. 2013; Huang & Wang 2014). Thus, boosting allowed those poor performance variables such as spatial weights term to have the opportunity to be modelled in order to further test their interactions with the dependent variable (Elith et al. 2008). Consequently, the strong related predictors were not the only ones modelled, but the poor performance predictors can also contribute to the final boosted models. This can potentially reduce the loss of model fit (Elith et al. 2008). Although, spatial variable contributed in the final BRT model, spatial autocorrelation was not a dominant factor (spatial variable's ranking was below 15 in 35 variables), and spatial weights term had a relatively small importance (less than 2) speaking for little contributions to the final fit (see Table 11 to Table 14).

As the spatial weights term did not contribute to bagged MARS models, the following discussion about spatial autocorrelation is then based on the impact of spatial weights term on the BRT models. As previously mentioned, two methods were used to combat spatial autocorrelation. Comparatively speaking, the spatial weights method had less of an impact on the BRT models

than the sample distance spacing method. For example, **Section 3.3** explained that the candidate predictors of BRTs with the greatest contribution (top three) were consistent after introducing the spatial weights term; moreover, **Section 3.6** found that BRT models' predictions were basically unchanged after including spatial weights term. However, using the SD spacing method to minimise or eliminate spatial autocorrelation may come at the cost of reducing the model fit. That is, a larger sample distance can potentially reduce spatial influence, but a spread out dataset had less sample data, possibly leading to an overfitted of the data. This finding is confirmed by the model prediction results in **Section 3.6**. After increasing the sample distance against spatial influence, the distribution of the prediction probabilities of *E. fasciculosa* shifted to the right (high probabilities), indicating over fitted (see Figure 12).

Data independence test results (see **Section 3.1**) indicated that there was a limited or no spatial relationship for *A. verticillata* and *E. obliqua*, while *E. goniocalyx* had spatial autocorrelation, but a strong relationship was not evident. Indeed, *E. fasciculosa* had a relatively stronger spatial relationship than the other three. Interestingly, the contribution of a spatial weights term on the model is not always related to the strength of the spatial autocorrelation. The importance of a variable represents the contribution of that variable on the model. The important values of spatial variables show that the spatial weights term of *E. fasciculosa* contributed the most to BRTs, while the second place is the spatial influence of *E. obliqua* (see Table 13 to Table 14). It should be noted that, for modelling *E. goniocalyx*, the top candidate predictor (Temperature standard deviation) had 65.777 importance while the spatial weights term's importance was only 0.107, and spatial influence can be ignored (see Table 8 and Table 12). This means that, even if the spatial autocorrelation is weak (such as *E. obliqua*), the spatial weights term can also have a certain influence on the model. On the contrary, the spatial variable does not necessarily affect the model fit even if there is a spatial autocorrelation (such as *E. goniocalyx*). The strength of spatial autocorrelation does not determine its impact on the modelling analysis.

4.7 Limitations

Species' distribution models are based on mathematical algorithms that generate a habitat suitability or distribution probability map in GIS by employing survey data and available resource (Lethbridge et al. 2006). However, resource availability is complex because variables may not only change over time but also interact with some other species (Ford et al 2001; Lethbridge et al. 2006). Such process uncertainty can reduce the predictive power of the static models. This study only explored the interactions between response and needed environmental predictors, other non-environmental issues, such as birds, human activities, etc., were not considered. For example, *A. verticillata* (dropping she oak) can assist the bird's survival, its seed is the primary food source for several cockatoo species (Pepper et al. 2000). Thus, bird species may also influence its distribution. Moreover, urban areas such as roads, housing, etc., were not removed from the final prediction. Although such problems may also affect the tree's occurrence, this study mainly focused on the realised environmental requirements of target trees.

Regressions approaches such as MARS and BRT, are not able to incorporate dynamic processes. Within the Mount Lofty Ranges, human activities such as large areas of vegetation clearance or deforestation have been stopped for environment protection, thus the remained trees were considered to be stable. This study assumed that the modelled trees and their surrounding environments maintained relevant for a long-term period. Thus, dynamic process such as succession was ignored in this study.

The target species in this study were either dominant or co-dominant. Hence, they are often considered representative of the overall biodiversity of the region (Lambeck 1997; Brooker 2002; Watson et al. 2001). Nevertheless, the environment requirements of trees are dynamic, except for seasonal preference, trees can be influenced by the environment during all differ phases of growth (Haferkamp 1988). This study did not take this problem into account, but rather on the general relation between the trees' occurrence with selected environmental variables. If the purpose of regressions is to explore the

influence of one or more specified environment data for a period of growth, the response curve may help to indicate their interactions.

Due to limited surveying data availability, resampling techniques was used to test different sample data combination to select the one with a better model performance. For instance, integrated bagging can improve the prediction accuracy and robustness, and boosted models can potentially reduce the loss of performance (Breiman 1996; Huang & Wang 2014; Xue 2016; Elith et al. 2008). Bagging and boosting are re-selected process, thus, the sample dataset may repeated use several samples many times, while not use others for modelling (Kuhn & Johnson 2013). This process may increase the space distance between samples through dropping several samples. That is to say, resample process can potentially improve the final fit of data, but it may alter the dataset and then affect spatial autocorrelation.

Another limitation is the spacing distance that can influence spatial autocorrelation. A random sampled quadrat survey method was used in field surveying to obtain PA data. The sample quarters' size can be controlled, but the distance between sample quarters cannot be determined. According to the nearest neighbour distances in Table 4, the original data spacing was large. Therefore, the spatial autocorrelation of target trees of all samples was weak and even can be ignored. This resulted in that this study cannot understand the spatial influence on the model more comprehensively.

4.8 Summary

The potential habitat preferences closer to the realised environmental niche of four Australia native trees in the Mount Lofty Ranges of SA have been modelled. The performance of models was evaluated and the overall model fit of data was good. Additionally, regressions using BRT algorithm had a better prediction ability than using bagged MARS.

Data independent tests can explain the strength of spatial autocorrelation of the target species. In order to combat spatial influence, the first method is to re-sample the data. This is compared with the spatial weights method. Because the spatial weights term has little impact on the model, it is hard to comprehensively compare these two methods. Interestingly, there is a potential problem of using the first method. That is the sample distance spacing method may result in over fitted.

The regression model accessed the realised environment requirements of the target trees. The preference of *A. verticillata* and *E. obliqua* obtained from the models were in good agreement with literatures, while partly of findings of *E. fasciculosa* were consistent with previous studies; yet the predictive dominant variables of *E. goniocalyx* were in great deviation from those in the literature.

The predictive PA surface created by the regression are highly consistent with priori studies. *A. verticillata* indicates a large presence area to the middle-right with other scattered areas to the north and south. The habitat niche areas of *E. goniocalyx* are limited and concentrated in the north. *E. fasciculosa* has a widespread distribution though the whole study area. The presence areas of *E. obliqua* are along the spline of the Mount Lofty Ranges. These predictive presence areas were generally agreed with the existing maps of Tim Croft (unpublished). The extrinsic comparison between them indicates that, the approximate presence locations of the target trees are consistent, but the predictive areas are broad.

4.9 Recommendations for further research

The modelling results of this study suggest that the spatial autocorrelation of the sample data does not necessarily affect the model analysis. However, largely increasing the sample distance to reduce or remove spatial relationship will lead to the loss of model fit. For further studies, spatial autocorrelation can be ignored in some cases, there is no need to always combat spatial influence.

The problem caused by largely increasing the spacing may be more serious than the impact of spatial autocorrelation.

In this study, resample techniques including bagging and boosting are combined to individual regression. This is to enhance the mode performance using limited data size. Integrated bagging is an effective approach to improve the accuracy and robustness of individual models (Breiman 1996; Huang & Wang 2014; Xue 2016). Combining boosting to regression trees can overcome the poor predictive performance issues associated with single tree models (Abeare 2009; Freund & Schapire 1996; Li et al. 2014; Jiao et al. 2015). The fit of the final models depends on how bagging and boosting selects variables for modelling (Elith et al. 2008). However, the resampling process is uncontrollable. This is particularly problematic when the data size is very small. Resampling process, especially random selection, may become unstable. Thus, if there are sufficient data available, resampling is not needed. An appropriate algorithm can also accurately describe the data.

The relationship between species and environment is a central important topic in ecology research. Regression models can use the presence/absence data and given variables to explore the interactions between species and its environment. SDMs using regressions can be used to predict the likely realised habitat niche or environmental suitability of a target species (Guisan et al. 2013). More and more studies try to solve practical problems through SDMs (Elith & Leathwick 2009). And increasingly ecologists need accurate and appropriate SDMs to access and evaluate the progress of conservation and biodiversity programs (Guisan & Zimmermann 2000). Each model has its own unique advantages and disadvantages. Even if using the consistent environmental predictor variables, the calculation results of the model may vary greatly due the different modelling algorithms (Lehmann et al. 2002b). A decision maker should select a suitable model method based on the purpose of the study and the characteristics of available data (Moisen & Frescino 2002).

Species' distribution models are limited in framing conservation planning because they only consider a single species rather than multi-species (Lethbridge et al. 2006). There are evidences to suggest that single species'

model is limit due it ignores the interactions between species such as mutualism and competition (Burgman et al. 1993). For further studies, the multi-species' models that can incorporate data of more than one species are considered more useful in conservation planning and biodiversity programs. A multi-species' model can provide comprehensive information of the whole community rather than focusing on the specified species (Young et al. 2005; Stead 2008).

BIBLIOGRAPHY

Abeare S. M., 2009. *Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico Lonline Fishery,* Doctoral dissertation, University of Pretoria.

Abraham A. & Steinberg D., 2001. MARS: Still an Alien Planet in Soft Computing? Publication of the Society for Computer Simulation International, Prague.

Anderson, R. P., D. Lew, and P. A.T. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162:211-232.

Adamson R. S. & Osborn T. G. B., 1924. *The ecology of the eucalyptus forests of the Mount Lofty Ranges (Adelaide district)*, South Australia. Gillingham & Company, printers.

Ahmed S. E., McInerny G., O'Hara K., Harper R., Salido L., ..., Joppa L. N., 2015. Scientists and software - surveying the species distribution modelling community. *Diversity and distributions*, 21(3), 258-267.

Anselin L., 1988. Spatial Econometrics, Methods and Models. Kluwer Academic., Dordrecht.

Anselin L., 1995. Local indicators of spatial association—LISA. *Geographical analysis*, 27(2), 93-115.

Archer K.J. & Kimes R.V., 2008. "Empirical characterization of random forest variable importance measures", Computational Statistics and Data Analysis, vol. 52, pp. 2249-2260.

Armstrong D. M., Croft S. J. & Foulkes J. N., 2003. A Biological Survey of the

Southern Mount Lofty Ranges South Australia. *Department for Environment and Heritage, Adelaide*.

Austin M., 2002. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological modelling*, 200(1), 1-19.

Baker R. T. & Smith H. G., 1902. *A research on the eucalyptus: especially in regard to their essential oils* (No. 24). Authority of the government of the state of New South Wales.

Ball N. M. & Brunner R. J., 2010. Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19 (07), 1049-1106.

Bar-Ness Y. D., Kirkpatrick J. B. & McQuillan P. B., 2012. Crown structure differences and dynamics in 100-year-old and old-growth Eucalyptus obliqua trees. *Australian forestry*, 75(2), 120-129.

Barry K. M., Janos D. P., Nichols S. & Bowman D. M., 2015) Eucalyptus obliqua seedling growth in organic vs. mineral soil horizons. *Frontiers in plant science*, 6, 97.

Bassett O. D. & White G., 2001. Review of the impact of retained overwood trees on stand productivity. *Australian Forestry*, 64(1), 57-63.

Bellman R. E., 1961. *Adaptive Control Processes*, Princeton University Press.

Berkinshaw T.,2010. *Native vegetation of the Eyre Peninsula, South Australia: a field guide to native communities, plant species and environmental weeds of the Eyre Peninsula, South Australia*. Greening Australia.

Bishop C., 1995. *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Bolker B. M., Brooks M. E., Clark C. J., Geange S. W., Poulsen J. R., Stevens M. H. H. & White J. S. S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3), 127-135.

Bonney N. B., 1997. *Economic native trees and shrubs for South Australia*. Greening Australia.

Bonney N. B., 2010. *What south east native plant is that?: identifying and growing native vegetation in the south east of South Australia*. Mount Gambier: South East Natural Resources Management Board.

Boomsma C. D. & Lewis N. B., 1980. The native forest and woodland vegetation of South Australia. *The native forest and woodland vegetation of South Australia*, (25).

Borra S. & Di Ciaccio A., 2002. Improving non-parametric regression methods by bagging and boosting. *Computational Statistics & Data Analysis*, 38(4), 407-420.

Boulesteix A. L., Janitza S., Kruppa J. & König I. R., 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6), 493-507.

Bradshaw C. J., 2012. Little left to lose: deforestation and forest degradation in Australia since European colonization. *Journal of Plant Ecology*, 5(1), 109-120.

Breiman L., 1996. Bagging Predictors, *Machine Learning*, 26, No. 2, 123-140

Breiman L., Friedman J. H., Olshen J. & Stone C., 1984. Classification and Regression Trees. Chapman and Hall, New York.

Brito J. C., Crespo E. G. & Paulo O. S., 1999. Modelling wildlife distributions: Logistic Multiple Regression vs Overlap Analysis. Ecography 22:251-260.

Broadhurst L., 2015. Pollen Dispersal in Fragmented Populations of the Dioecious Wind-Pollinated Tree, Allocasuarina verticillata (Drooping she-oak, Drooping She-Oak; Allocasuarinaceae). PloS one, 10(3), e0119498.

Brooker L., 2002, 'The application of focal species knowledge to landscape design in agricultural lands using the ecological neighbourhood as a template.' *Landscape and Urban Planning* vol. 60, pp.185-210.

Brooker M.I.H. & Slee A.V., 1996. Eucalyptus. In: Walsh, N.G.; Entwisle, T.J. (eds), Flora of Victoria Vol. 3, Dicotyledons Winteraceae to Myrtaceae. Inkata Press, Melbourne.

Burgman M. A., Ferson S. & Akacakaya H. R., 1993. Risk Assessment in Conservation
Biology. In Population and Community Biology. (series ed. M. B. Usher). Chapman and Hall,
London.

Burrough P. A.& McDonnell R. A., 1998. *Principles of Geographical Information Systems* 2nd edition (Oxford University Press, Oxford) ISBN 0 19 823366 3, 0 19 823365 5.
Caswell H., 1987. Theory and models in ecology: a different perspective. Ecological Modelling 43:33 - 44.

Chen Y., 2011. The explore of species estinction and species diversity protection. *Journal of Ningxia Agriculture & Foresty Science & Technology*, 52(8), 70-73.

Chu W. W., 2014. Data mining and knowledge discovery for Big Data. *Studies in Big Data*, 1.

CITES E. C. O. S. O. (2011). Convention on International trade in endangered species of wild fauna and flora. URL: *http://www. cites. org/esp/app/appendices.*

shtml.

Cliff A.D. and Ord J.K., 1973. Spatial autocorrelation. London : Pion. Google Scholar

Cramer J. S., 2003. *Logit models from economics and other fields*. Cambridge University Press.

Craven P. & Wahba G., 1979. Smoothing noisy data with spline functions. Numerische Mathematik 31, 377–403.

Cutler D. R., Edwards T. C., Beard K. H., Cutler A., Hess K. T., Gibson J., & Lawler J. J., 2007. Random forests for classification in ecology.Ecology, 88(11), 2783-2792.

Dashorst G. R. M. & Jessop J. P., 1990. Plants of the Adelaide plains and hills. *Kenthurst: NSW, Kangaroo Press 224p.-col. illus., maps.. ISBN, 864173237.*

De Oliveira, M. E. D., Vaughan, B. E., & Rykiel, E. J., 2005. Ethanol as fuel: energy, carbon dioxide balances, and ecological footprint.BioScience, 55(7), 593-602.

Dean N. & Ian, R. (illustrator.) & Dean N., 2013. Native Eucalyptus of South Australia. Melrose Park, South Australia.

DEH - Department for Environment and Heritage (2001), *Provisional List of Threatened Ecosystem of South Australia*.

Deng T., Huang Y., Gu J., Yu C. & Xiao J., et al. (2013). Spatial autocorrelation in spatial analysis. *Chinese Journal of health statistics*, 30 (3), 343-346.

Dobbertin M & Biging G. S., 1998. Using the non-parametric classifier CART to model forest tree mortality. For Sci 44(4):507–16.

Efron B. & Tibshirani R., 1986. “Bootstrap Methods for Standard Errors, Confidence Intervals and Other Measures of Statistical Accuracy”. *Statistical Science*, pp.54-75.

Elith J. & Leathwick J. R., 2009. Species distribution models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution and Systematics*, 40(1), 677-697.

Elith J. & Leathwick J., 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and distributions*, 13(3), 265-275.

Elith J., Graham H. C., Anderson P. R. Dudik M., Ferrier S., Guisan A., ..., Zimmermann N. E., 2006. Novel methods improve prediction of species’ distribution from occurrence data. Ecography, 2nd January, 129-151. doi:10.1111/j.2006.0906-7590.04596.x.

Elith J., Leathwick J. R. & Hastie T., 2008. A working guide to boosted regression trees. J*ournal of Animal Ecology*, 77(4), 802-813.

Elith J., Phillips S. J., Hastie T., Dudík M., Chee Y. E. & Yates C. J., 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, 17(1), 43-57.

Elliot W. R. & Jones D. L., 1990. *Encyclopaedia of Australian plants suitable for cultivation. Volume 5*. Lothian Publishing Company Pty Ltd.

ESRI 2017, *ArcGIS*, Environmental Science Research Institute, Redlands, California.

Facelli J. M., Williams R., Fricker S. & Ladd B., 1999. Establishment and growth of seedlings of Eucalyptus obliqua: interactive effects of litter, water, and pathogens. *Australian Journal of Ecology*, 24(5), 484-494.

Fawcett T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.

Fielding A. H. & Bell J. F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, 24 (01), 38-49.

Folke C., Holling C. S. & Perrings C., 1996. Biological diversity, ecosystems, and the human scale. *Ecological applications*, 6(4), 1018-1024.

Franco J. L. D. A., 2013. The concept of biodiversity and the history of conservation biology: from wilderness preservation to biodiversity conservation. *História* (São Paulo), 32(2), 21-48.

Franklin J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in physical geography*, 19(4), 474-499.

Franklin J., 2009. *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge, UK.

Freund Y. & Schapire R. E., 1996. Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference, pages 148–156.

Friedman J. H., 1991. Multivariate adaptive regression splines. Annals of Statistics 19:1-141.

Getis A. & Ord J. K., 1996. Local spatial statistics: an overview. *Spatial analysis: modelling in a GIS environment*, 374.

Green P. S., 1994. *Vegetation Ecology of the Central Mount Lofty Ranges, Department of Botany*. The University of Adelaide, Adelaide.

Guerin G. R. & Lowe A. J., 2013. Systematic monitoring of heathy woodlands

in a Mediterranean climate—a practical assessment of methods. *Environmental monitoring and assessment*, 185(5), 3959-3975.

Guerin G. R., Biffin E., Baruch Z., & Lowe A. J., 2016. Identifying Centres of Plant Biodiversity in South Australia. *PloSone*, 11(1), e0144779.

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

Guisan A. & Thuiller W., 2005. Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.*, 8, 993-1009.

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

Guisan A., Lehmann A., Ferrier S., Austin M., OVERTON J., Aspinall R. & Hastie T., 2006. Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, 43(3), 386-392.

Guisan A., Tingley R., Baumgartner J. B., Naujokaitis-Lewis I., Sutcliffe P. R., Tulloch A. I., ... & Martin T. G., 2013. Predicting species distributions for conservation decisions. *Ecology letters*, 16(12), 1424-1435.

Gutiérrez Á. G., Schnabel S., & Contador J. F. L., 2009. Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecological modelling*, 220(24), 3630-3637.

Haferkamp M. R., 1988. Environmental factors affecting plant productivity. *Achieving efficient use of rangeland resources. Bozeman: Montana State University Agricultural Experiment Station*, 27-36.

Harrington P., 2012. *Machine learning in action* (Vol. 5). Greenwich, CT: Manning.

Hastie T., Friedman J. & Tibshirani R., 2001. *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Hastie T. & R. J. Tibshirani 1990. *Generalized Additive Models*. Chapman and Hall.

He X.H, Wen Z.M. & Wang J.X, 2008. Predict spatial distribution of main grassland species and its relationship with environmen along river basin with GAMs. *Journal of ecology*, 27(10), 1718-1724.

Hijmans R. J. & Elith J., 2013. Species distribution modeling with R. *R package version 0.8-11*.

Hilborn R. & Mangel, M., 1997. *The ecological detective: confronting models with data* (Vol. 28). Princeton University Press.

Huang W. & Wang Z., 2014. *Data Mining: R Practice*. Electronic Industry Press.

James G., Witten D. & Hastie T., 2014. An Introduction to Statistical Learning: With Applications in R.

James G., Witten D., Hastie T. & Tibshirani R., 2013. *An introduction to statistical learning* (Vol. 6). New York: springer.

Jiao L.L., Chang Y., Shen D., Hu Y.M, Li C.L. & Ma J., 2015. Using boosted regression trees to analyze the factors affecting the spatial distribution patern of wildfire in China. *Chinease Journal of Ecology*, 34(8), 2288-2296.

Joseph L., 1982. The glossy black-cockatoo on Kangaroo Island. *Emu*, 82(1), 46-49.

Kantvilas G. & Jarman, S. J., 2004. Lichens and bryophytes on Eucalyptus obliqua in Tasmania: management implications in production forests. *Biological*

Conservation, 117(4), 359-373.

Kinnear A. J., Overton I. C., & Hyde, M. J., 2001. Brownhill Creek recreation park vegetation management plan.

Kong X.Q, 2015. *Stability Assessment of Species Distribution Models and Application Software*, Master thesis, Anqing Teacher College.

Kuhn M. & Johnson K. ,2013. *Applied predictive modeling* (pp. 389-400). New York: Springer.

Kuhn M., 2013. Predictive Modeling with R and the caret Package.

Kuhnert P. M., Pagendam D., Cox J., Fleming N., He Y., Thomas S., ... & van der Linden, L., 2015. An improved water quality model for the Onkaparinga Catchment. *Goyder Institute for Water Research Technical Report Series*, (15/6).

Lambeck R., 1997, 'Focal species: a multi-species umbrella for nature conservation'. *Conservation Biology* vol 11, pp 849-856.

Laut P, Heyligers P C, Keig G, Löffler E, Margules C, Scott R M & Sullivan M E, (1977), Environments of South Australia Province 3 Mt Lofty Block. Division of Land Use Research Commonwealth Scientific and Industrial Research Organisation, Canberra.

Leathwick J. R., Elith J. & Hastie T., 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological modelling*, 199(2), 188-196.

Lehmann A., Krauss W., Hinrichsen H., 2002a. Effects of remote and local atmospheric forcing on circulation and upwelling in the Baltic Sea. Tellus Series A 54:299-316.

Lehmann A., Overton J. M. & Austin M. P., 2002b. Regression models for spatial prediction:
Their role for biodiversity and conservation. Biodiversity and Conservation 11, 2085-2092.

Lemon S. C., Roy J., Clark M. A., Friedmann P. D. & Rakowski W., 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, 26(3), 172-181.

Lethbridge M., 2004. Flinders University of South Australia.

Lethbridge M., 2005. *Sort PA*, Flinders University of South Australia.

Lethbridge M. & Green P., unpublished. 'Modelling tree distributions in the Mount Lofty Ranges using Multivariate Adaptive Regression Splines', Flinders University of South Australia.

Lethbridge M., Wijk E. V., Harper M. & Best J., 2006. *Modelling of Bird and Plant Distributional Data in the Mount Lofty Ranges (MLR) for Conservation Planning*, Report to the Department for Environment & Heritage.

Levine N., 2004. CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0). Houston (TX): *Ned Levine & Associates/Washington, DC: National Institute of Justice*.

Lewis P.A.W & Stevens J.G., 1991. "Nonlinear modeling of time series using multivariate adaptive regression splines (mars)," Journal of the American Statistical Association, vol. 86, no. 416, pp. 864-877.

Li C.L., Liu M., Hu Y.M., Xu Y.Y. & Sun F., 2014. Driving forces analysis of urban expansion based on boosted regression trees and Logistic regression. *Acta Ecol*. Sin, 34, 727-737.

Lim T., Loh W. & Shih Y., 2000. 'A comparison of prediction accuracy, complexity, and

Liu C., Berry P. M., Dawson T. P. & Pearson R. G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28(3), 385-393.

Loh W. Y., 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14-23.

Loh W. Y., 2014. Fifty years of classification and regression trees. *International Statistical Review*, 82(3), 329-348.

Lutze M. T., Campbell R. G. & Fagg P. C., 1999. Development of silviculture in the native State forests of Victoria. *Australian Forestry*, 62(3), 236-244.

McCallum B. T., 2000. *Theoretical analysis regarding a zero lower bound on nominal interest rates* (No. w7677). National bureau of economic research.

Neagle N., 1995. An Update of the Conservation Status of the Major Plant Associations of South Australia. Native Vegetation Conservation Section, Department of Environment and Natural Resources, South Australia.

Malanson G. P., Westman W. E. & Yan Y. L., 1992. Realized versus fundamental niche functions in a model of chaparral response to climatic change. *Ecological Modelling*, 64(4), 261-277.

Menut C., Molangui T., Lamaty G. E., Bessiere J. M. & Habimana J. B., 1995. Aromatic Plants of Tropical Central Africa. 23. Chemical Composition of Leaf Essential Oils of Eucalyptus goniocalyx F. Muell. and Eucalyptus patens Benth. Growth in Rwanda. *Journal of Agricultural and Food Chemistry*, 43(5), 1267-1271.

Milborrow S., 2013. Derived from mda: mars by Trevor Hastie and Rob

Tibshirani. earth: Multivariate Adaptive Regression Spline Models, 2011. *R package http://CRAN. R-project. org/package= earth*. Cited on, 4.

Miller G. F. & Franklin J., 2002. Modelling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling 157:227-247.

Moisen G. G. & Frescino T. S., 2002. Comparing five modelling techniques for predicting forest characteristics. *Ecological modelling*, 157(2), 209-225.

Nakagawa S. & Cuthill I. C., 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4), 591-605.

Nyström Sandman A., 2011. *Modelling spatial and temporal species distribution in the Baltic Sea phytobenthic zone* (Doctoral dissertation, Department of Systems Ecology, Stockholm University).

Obertello M., Santi C., Sy M. O., Laplaze L., Auguy F., Bogusz D., & Franche C., 2005. Comparison of four constitutive promoters for the expression of transgenes in the tropical nitrogen-fixing tree Allocasuarina verticillata. *Plant cell reports*, 24(9), 540-548.

Paton D. C., Prescott A. M., Davies R. J. P. & Heard L. M., 2000. The distribution, status and threats to temperate woodlands in South Australia. *Temperate eucalypt woodlands in Australia: Biology, conservation, management and restoration*, 57-85.

Paton D. & O'Connor J., 2010. The state of Australia's birds 2009: restoring woodland habitats for birds.

Pepper J. W., Male T. D. & Roberts G. E., 2000. Foraging ecology of the South Australian glossy black-cockatoo (Calyptorhynchus lathami halmaturinus). Austral Ecology, 25(1), 16-24.

Pierzynski G. M., Vance G. F. & Sims J. T., 2005. *Soils and environmental quality*. CRC press.

Ridgeway G., 2006. gbm: Generalized boosted regression models. R package version, 1(3), 55.

Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., Sanchez J. C. & Müller M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12(1), 77.

Rokach L. & Maimon O., 2014. *Data mining with decision trees: theory and applications*. World scientific.

Schapire R., 2003. The boosting approach to machine learning – an overview. *MSRI Workshop on Nonlinear Estimation and Classification, 2002* (eds D.D. Denison, M. H. Hansen, C. Holmes, B. Mallick & B. Yu). Springer, New York.

Simard Y., P. Legendre, G. Lavoie & D. Marcotte, 1992. Mapping, estimating biomass, and optimizing sampling programs for spatially autocorrelated data: case study of the northern shrimp (Panda/us borealis). Canadian Journal of Fisheries and Aquatic Sciences 49:32-45.

Sinclair R., 1980. Water potential and stomatal conductance of three Eucalyptus species in the Mount Lofty Ranges, South Australia: responses to summer drought. *Australian journal of botany*, 28(6), 499-510.

Skidmore A. K., Gauld A. & Walker P. W., 1996. A comparison of GIS predictive models for mapping kangaroo habitat. *International Journal of Geographical Information Systems*, 10, 441-454.

Slippers B., Fourie G., Crous P. W., Coutinho T. A., Wingfield B. D., Carnegie A. J. & Wingfield M. J., 2004. Speciation and distribution of Botryosphaeria spp. on native and introduced Eucalyptus trees in Australia and South Africa. Studies in Mycology, 50(343), e358.

Sharma B. K., 2000. *Industrial chemistry (including chemical engineering)*. Chapter 11: Hydrocarbons from petroleum, p.p. 440. GOEL Publishing House.

Smyth G. K., 1989. Generalized Linear models with varying dispersion. Journal of the Royal Statistical Society 51:47-60.

Specht R. L. & Perry R. A., 1948. *Plant ecology of part of the Mount Lofty Ranges (1)*. Department of Botany, University of Adelaide.

Stead M. G., 2008. *Niche area sensitivity of tree species in the Mount Lofty Ranges to climate change*. Flinders University, School of Geography, Population and Environmental Management.

Steinberg D, Colla P. L. & Martin K., 1999. MARS user guide. San Diego (CA): Salford Systems.

Sumathi S. & Sivanandam S. N., 2006. *Introduction to data mining and its applications* (Vol. 29). Springer.

Swets J., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

Szabo J. K., Vesk P. A., Baxter P. W. & Possingham H. P., 2011. Paying the extinction debt: woodland birds in the Mount Lofty Ranges, South Australia. *Emu*, 111(1), 59-70.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Tufféry S., 2011. Data mining and statistics for decision making.

Valavanis V. D., Pierce G. J., Zuur A. F., Palialexis A., Saveliev A., Katara I. &

Wang J., 2008. Modelling of essential fish habitat based on remote sensing, spatial analysis and GIS. *Hydrobiologia*, 612(1), 5-20.

Vesk P. & Mac Nally R., 2006, 'Changes in vegetation structure and distribution in rural
landscapes: implications for biodiversity and ecosystem processes'. *Agriculture, Ecosystems and Environment*, vol. 112, pp 356-366.

Vittinghoff E., Glidden D. V., Shiboski S. C. & McCulloch C. E., 2011. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media.

Vold T. & Buffett, D. A., 2000. *Ecological concepts, principles and applications to conservation*. Biodiversity BC.

Wang Z., 2015. Studies on the ecological effects of planting Eucalyptus species in large areas. *Agriculture and Technology,* 35(7), 86-87.

Ward M. J., 2007. Ecology of box mistletoe Amyema miquelii dispersal in pink gum Eucalyptus fasciculosa woodlands.

Watson J., Freudenberger D. & Paull D., 2001, 'An assessment of the focal-species approach for conserving birds in the variegated landscapes in southeastern Australia'. *Conservation Biology* vol 15, pp1364-1373.

White G. C., 2000. Population Viability Analysis: Data Requirements and Essential Analyses in L. Boitani, and T. K. Fuller, editors. Research Techniques in Animal ecology: Controversies and Consequences. Columbia University Press, New York.

White T. C. R., 2015. Lerp insect (Cardiaspina densitexta) outbreaks on pink gum (Eucalyptus fasciculosa) in the southeast of South Australia. *Austral Ecology*.

Whittington J., & Sinclair R., 1988. Water relations of the mistletoe, Amyema miquelii, and its host Eucalyptus fasciculosa. *Australian journal of botany*, 36(3), 239-255.

Williams K. J. & Cary J., 2002. Landscape preferences, ecological quality, and biodiversity protection. *Environment and Behavior*, 34(2), 257-274.

Williams M., 1977. *The changing rural landscape of South Australia*. Heinemann Educational Australia.

Wilson E. O., 1999. *The diversity of life*. WW Norton & Company.

Wood G., 1986. Mt Lofty Ranges Watershed: impact of land use on water quality and implications for reservoir water quality management.

Xue W., 2016. *R Language and Data Mining: Methods and Applications*. Publishing House of Electronics Industry.

Yee T.W., Mitchell N.D., 1991. Generalized Additive Models in Plant Ecology. Journal of Vegetation Science 2:587-602.

Young T.P., Peterson D.A. & Clary J.J., 2005. 'The ecology of restoration: historical links, emerging issues and unexplored realms', *Ecology Letters*, vol. 8, pp. 662-673.

Yu-Wei C. D. C., 2015. *Machine learning with R cookbook*. Packt Publishing Ltd.

Zhang C., Luo L., Xu W. & Ledwith V., 2008. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of the total environment*, 398(1), 212-221.

Zhang C.S., Zhang S. & He J.B., 1998. Study on Spatial Distribution Characteristics of Heavy Metals in Sediments of the Yangtze River --- Spatial

autocorrelation and fractal method. *Journal of Geography*, (1), 87-96.

Zhang F. & Zhang X., 2003. Landscape spatial autocorrelation analysis of TM remote sensing data: A case study of Changping District, Beijing, China. *Acta Ecologica Sinica*, 24(12), 2853-2858.

APPENDIX

Appendix 1: Summary of predictor variables

Code	Description	Units	Sources
Topography			
elev	Elevation	m	Contours
envas	Environmental association	classes	DEWNR
ew	EW aspect	index	Elevation
ns	NS aspect	index	Elevation
slope	Surface slope	%	Elevation
flowac	Flow accumulation	Cell #	Elevation
wet	Soil wetness	index	Elevation
smrsrad	Summer solar radiation		Elevation
wntsrad	Winter solar radiation		Elevation
anlsrad	Annual solar radiation		Elevation
seasrad	Seasonal solar radiation	index	Elevation
major	Distance from major streams	m	DEWNR
minor	Distance from minor streams	m	DEWNR
geol	Geology	classes	PIRSA

Code	Description	Units	Sources
Soil			
comsoil	Common soil	classes	PIRSA
drain	Deep drainage	classes	
fertil	Inherent fertility	classes	
waterlog	Waterlogging	classes	
wtdepth	Water table depth	classes	
alkali	Alkalinity	classes	
acid	Acidity	classes	
awhc	Available water holding capacity	classes	
rkdepth	Depth to hard rock	classes	
srock	Surface rockiness	classes	

Code	Description	Units	Sources
Climate			
arain1k	Annual rainfall	mm	Australian Bureau of Meteorology
srain1k	Summer rainfall	mm	
wrain1k	Winter rainfall	mm	
snrain1k	Seasonal rainfall	index	
sdrain1k	Rainfall standard deviation	mm	
atmp1k	Annual temperature	^{o}C	
stmp1k	Summer temperature	^{o}C	
wtmp1k	Winter temperature	^{o}C	
sntmp1k	Seasonal temperature	index	
sdtmp1k	Temperature standard deviation	^{o}C	

Appendix 2: Relatively important variables

Table 19: Important predictors selected by bagged MARS for *A. verticillata*.

	Exclude spatial weights	Include spatial weights
Rank	**Predictors**	**Predictors**
1	Annual rainfall	Summer temperature
2	Temperature standard deviation	Common soil
3	Alkalinity	Summer rainfall

Table 20: Important predictors selected by BRT for *A. verticillata*.

	Exclude spatial weights	Include spatial weights
Rank	**Predictors**	**Predictors**
1	Winter rainfall	Winter rainfall
2	Summer temperature	Summer temperature
3	Temperature standard deviation	Temperature standard deviation

Table 21: Important predictors selected by bagged MARS for *E. obliqua*.

	Exclude spatial weights	Include spatial weights
Rank	**Predictors**	**Predictors**
1	Elevation	Annual rainfall
2	Winter temperature	Winter temperature
3	Environmental association	Environmental association

Table 22: Important predictors selected by BRT for *E. obliqua*.

	Exclude spatial weights	Include spatial weights
Rank	**Predictors**	**Predictors**
1	Winter rainfall	Winter rainfall
2	Elevation	Elevation
3	Summer rainfall	Summer rainfall

Table 23: Important predictors selected by bagged MARS for *E. goniocalyx*.

	Exclude spatial weights	Include spatial weights	500m PA dataset
Rank	**Predictors**	**Predictors**	**Predictors**
1	Temperature standard deviation	Temperature standard deviation	Temperature standard deviation
2	Summer temperature	Summer temperature	Summer temperature
3	Winter temperature	Annual temperature	Annual temperature

Table 24: Important predictors selected by BRT for *E. goniocalyx*.

	Exclude spatial weights	Include spatial weights	500m PA dataset
Rank	**Predictors**	**Predictors**	**Predictors**
1	Temperature standard deviation	Temperature standard deviation	Temperature standard deviation
2	Winter temperature	Winter temperature	Seasonal temperature
3	Seasonal rainfall	Seasonal rainfall	Seasonal rainfall

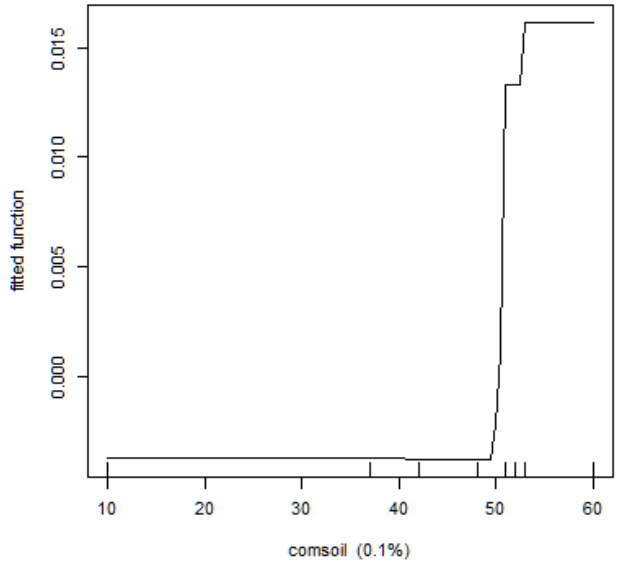
Table 25: Important predictors selected by bagged MARS for *E. fasciculosa.*

	Exclude spatial weights	Include spatial weights	1500m PA dataset
Rank	**Predictors**	**Predictors**	**Predictors**
1	Winter rainfall	Annual rainfall	Winter rainfall
2	Acidity	Acidity	Acidity
3	Distance from major streams	Available water holding capacity	Available water holding capacity

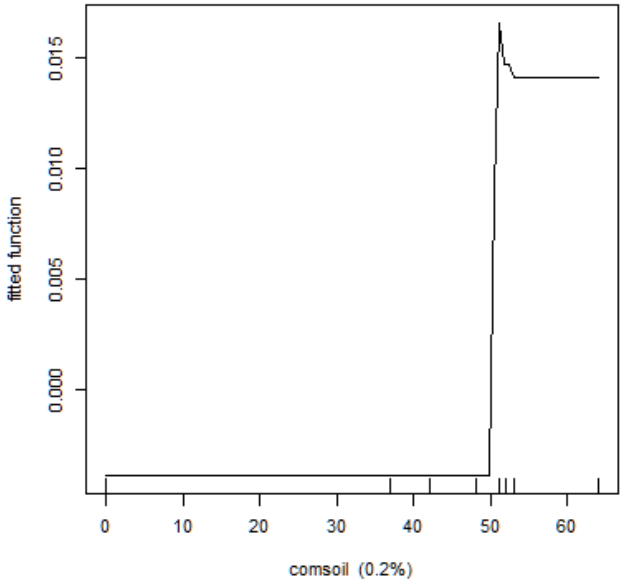
Table 26: Important predictors selected by BRT for *E. fasciculosa*.

	Exclude spatial weights	Include spatial weights	1500m PA dataset
Rank	**Predictors**	**Predictors**	**Predictors**
1	Winter rainfall	Winter rainfall	Winter rainfall
2	Temperature standard deviation	Temperature standard deviation	Distance from major streams
3	Distance from major streams	Distance from major streams	Elevation

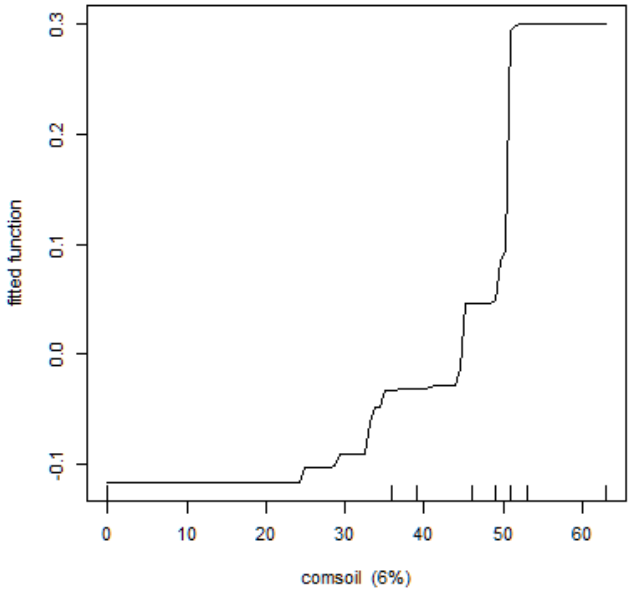
Appendix 3: Soil & Geology Requirements



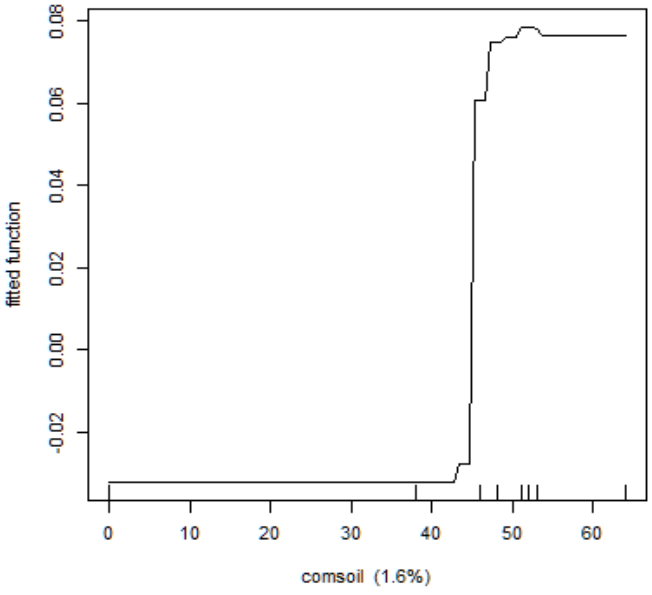
a) A. verticillata



b) E. goniocalyx

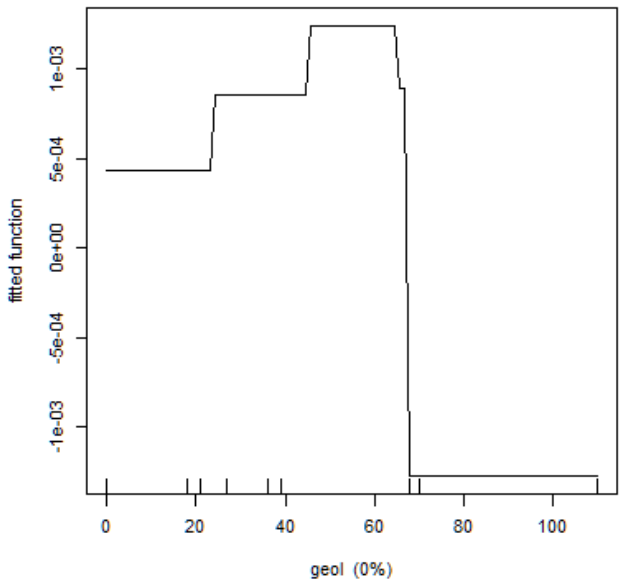


c) E. fasciculosa

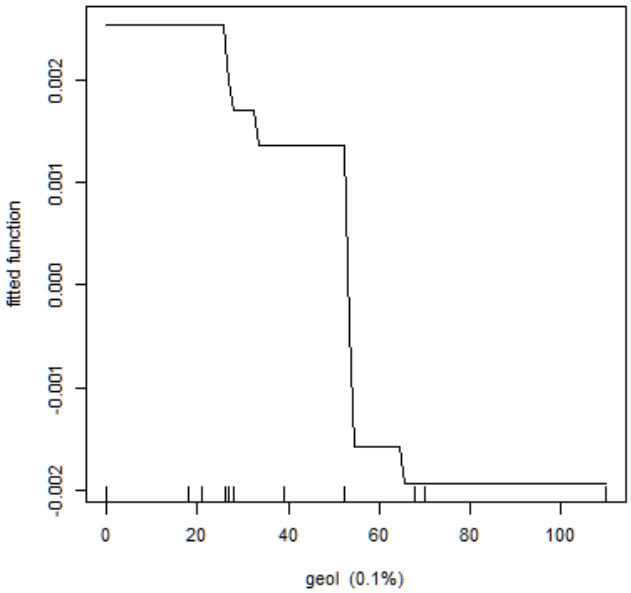


d) E. obliqua

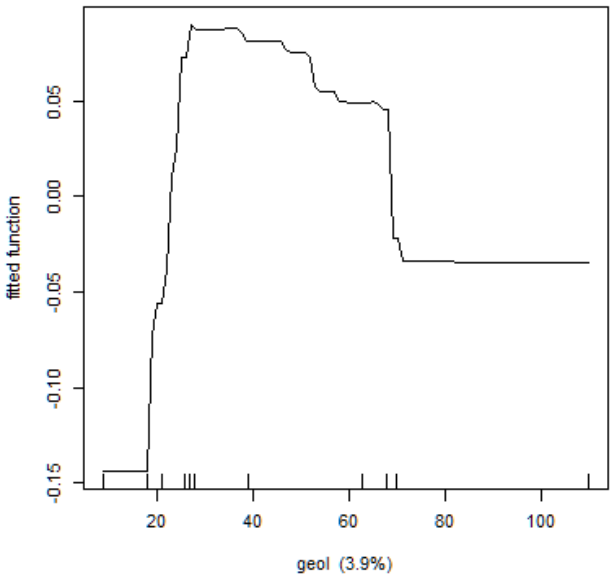
Figure 13: The interactions of the common soil with response.



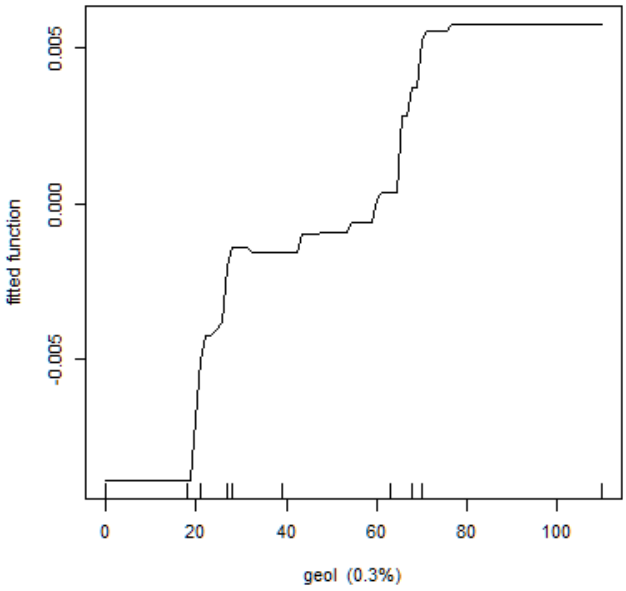
a) A. verticillata



b) E. goniocalyx



c) E. fasciculosa



d) E. obliqua

Figure 14: The interactions of the geology with response.

Appendix 4: Soil Types

Code	Description
0	Not applicable
1	Highly calcareous sandy loam
2	Calcareous loam on rock
3	Deep moderately calcareous loam
4	Deep (rubbly) calcareous loam
5	Rubbly calcareous loam on clay
6	Gradational calcareous clay loam
7	Calcareous clay loam on marl
8	Gypseous calcareous loam
9	Shallow highly calcareous sandy loam on calcrete
10	Shallow calcareous loam on calcrete
11	Shallow sandy loam on calcrete
12	Shallow red loam on limestone
13	Shallow dark clay loam on limestone
14	Shallow loam over red-brown clay on calcrete
15	Shallow sand over clay on calcrete
16	Shallow sand on calcrete
17	Shallow clay loam over brown or dark clay on calcrete
18	Gradational red-brown sandy loam
19	Gradational red-brown loam on rock
20	Friable gradational red-brown clay loam
21	Hard gradational red-brown clay loam
22	Gradational dark clay loam
23	Loam over clay on rock
24	Loam over red clay
25	Loam over poorly structured red clay
26	Loam over pedaric red clay
27	Hard loamy sand over red clay
28	Ironstone gravelly sandy loam over red clay
29	Loam over poorly structured red clay on rock
30	Black cracking clay
31	Red cracking clay

32	Grey or brown cracking clay
33	Loam over brown or dark clay
34	Sandy loam over poorly structured brown or dark clay
35	Sand over sandy clay loam
36	Bleached sand over sandy clay loam
37	Thick sand over clay
38	Sand over poorly structured clay
39	Sand over acidic clay
40	Carbonate sand
41	Siliceous sand
42	Bleached siliceous sand
43	Highly leached sand
44	Wet highly leached sand
45	Ironstone soil with calcareous lower subsoil
46	Ironstone soil
47	Shallow soil on ferricrete
48	Acidic gradational loam on rock
49	Acidic loam over clay on rock
50	Acidic sandy loam over red clay on rock
51	Acidic sandy loam over brown or grey clay on rock
52	Acidic gradational sandy loam on rock
53	Shallow soil on rock
54	Deep sandy loam
55	Deep friable gradational clay loam
56	Deep gravelly soil
57	Deep hard gradational sandy loam
58	Peat
59	Saline soil
60	Wet soil (non to moderately saline)
61	Volcanic ash soil
63	Rock
64	Water

Appendix 5: Geology

Code	Description
0	water
1	Amphibolite, undifferentiated.
2	Barite, undifferentiated.
3	Breccia, undifferentiated.
4	Dolerite, undifferentiated.
5	Haematite bodies/veins/orebodies, undifferentiated
6	Pegmatite, undifferentiated.
7	Pyrite, undifferentiated.
8	Quartz veins/bodies, undifferentiated
9	Quartz-haematite veins: ADELAIDE/BARKER digital database
10	Undifferentiated acid intrusive
11	Undifferentiated basic igneous rocks
12	Lower dolomite member: ADELAIDE/BARKER digital database
13	Basal quartzite unit: ADELAIDE/BARKER digital database
14	Topmost quartzite unit: ADELAIDE/BARKER digital database
15	Sandstone; dolomitic siltstone interbeds. Based on lower part of Skillogalee as on ANDAMOOKA
16	Lower member, typified by pale dolomite. BURRA: interim unit for compilation.
17	Calcsilicate gneiss (Houghton "Diorite"): ADELAIDE/BARKER digital database
18	Undifferentiated Quaternary rocks
19	Undifferentiated aeolian sediments
20	Undifferentiated alluvial/fluvial sediments
21	Undifferentiated Tertiary rocks
22	Hindmarsh Clay, Carisbrooke Sand, Ochre Cove Fmn, Seaford Fmn: ADELAIDE/BARKER digital database
23	Quartzite, slightly feldspathic, fine to medium grained, pale pinkish grey, clay intraclasts, flaggy to medium bedded, heavy mineral lamination, minor siltstone.
24	Marble, coarse-grained, amphibolitic, white, crystalline.
25	Siltstone, reddish, thin bedded; interbeds of dolomite and minor grey-green shale; pisolitic and algal limestone.
26	Sandstone, laminated, thick bedded, slumped, crossbedded, with minor siltstone interbeds. Widespread siltstone unit at base.
27	Metamorphic rocks with retrograde metamorphism; metasediments, strongly banded parallel to gneissic foliation; minor intrusive granitic, pegmatitic and amphibolitic dykes.
28	Siltstone, dark grey, laminated with minor sandstone, dolomite interbeds; quartzite, fine to coarse, feldspathic, cross bedded, minor siltstone interbeds; slate
29	Shale, black, carbonaceous, lenticular
30	Mudstone, glauconitic, calcareous; spicular chert; calcareous mudstone and spongolite

31 Clay, greenish grey, sandy; limestone, thin; and quartzsand; clay, green-grey, mottled, sandy.

32 Siltstone, shale, red-brown and olive green, laminated, flaggy to medium bedded; alternating with sandstone, fine grained, occasionally coarse grained. All lithologies calcitic in part.

33 Bioclastic barrier shoreline deposits, silica rich, with heavy minerals, shallow sub-tidal. Coastal, cross-bedded aeolian calcarenite with palaeosol horizons and capped by calcrete.

34 Sandstone, fine to coarse grained, feldspathic, quartzitic, to arkosic, ripple marks, cross bedding, lenticular, minor pale grey to greenish siltstone, minor pale grey dolomite

35 Siltstone, shale, grey-red to grey-green, partly calcitic, minor fine grained sandstone; dolomite, grey; limestone, grey, lenses, thin beds

36 Glacio-marine and fluvioglacial sediments and residual erratics.

37 Sandstone, grey, thick bedded, with thinly bedded, muddy, siltstone interbeds. Minor cross-bedding, ripples, rare trace fossils.

38 Sandstone, arkosic, medium grained, red-brown, slumped, ripple cross laminated; siltstone, sandy, red, dropstones and minor beds of diamictite with cobble to boulder size clasts of dolomite, basalt, dolerite, tuff.

39 Quartzite, sandstone, dolomite, conglomerate.

40 Granite, megacrystic and even-grained, blue quartz, metasediment xenoliths, metasomatic albitisation. Hybrid phases as inclusions. I-type to marginally S-type. Possibly syn-DD1, pre DD2. Age 504+/-8Ma (IR = 0.717)

41 Limestone, blue-grey, clean, massive, archaeocyathid-rich, biohermal, lower member. Limestone, sparsely fossiliferous, massive, mottled, upper member.

42 Sand, yellow-red, ferruginous

43 Clay, mottled, shelly; calcarenite, skeletal, coquina. Geochron age 132 000+/-6 000 years Bp on TL

44 Quartzite, arenaceous, with conglomerate lenses.

45 Sandstone, arkosic.

46 Sandstone, calcareous; sandy limestone. Transgressive, shallow marginal marine.

47 SHALE, blue-black, grey, pyritic, calcareous; LIMESTONE, blue-black, pyritic, nodular and phosphatic. Rare trilobites, hyolithids, sponge spicules, gastropods, worm tracks. Tuff horizon: 526+/-4Ma (U-Pb).

48 Clay, smectite-rich, grey-green, with red or yellow mottling and rare sand lenses

49 Metasandstone, fine grained, grey, quartzose, large-scale tabular crossbeds with heavy mineral laminations, slumped tops to foresets, rare angular shale clasts. Relatively shallow marine

50 Limestone, white, grey, recrystallised, garnetiferous in part, calcsilicate in part.

51 Siltstone, with very rare pebbles of sandstone, quartzite and limestone.

52 Sandstone, massive, gritty, highly feldspathic; quartzite with pebbles.

53 Dolomite, cherty, magnesitic.

54 Arkose, cross-bedded, coarse-grained to conglomeratic, Basal part, fluviatile? pyritic and glauconitic sandstone, minor shale siltstone and dolomite.

55 Limestone; dolomite; sandstone.

56 Limestone; sandstone; shale; volcanics

57 Dolomite, thin, laminated, micritic, with interbedded shale near the top.

58 Siltstone and sandy siltstone, sparse granule to boulder erratics, pale grey or

greyish green, massive or bedded, often calcareous. Minor lenses and interbeds of massive and laminated calcareous sandy siltstone and calcareous sandstone.

59 Grey-black Bouma-like sandstone-mudstone couplets. Climbing ripples, ball and pillow structures, channelled.

60 Clay, sand and carbonate earth, silty, with gravel lenses.

61 Siltstone, red, gritty, glacigenic?

62 Sandstone, coarse-grained, feldspathic, conglomeratic.

63 Mudstone; siltstone; shale, partly carbonaceous.

64 Quartzite; sandstone; siltstone.

65 Sandstone, arkosic, burrowed; silty ribbon limestone with intraformational breccias; bioherms and phosphatised hardgrounds

66 Dolomite; marble, with magnesite mud-pellet conglomerates.

67 Sand, marginal marine, glauconitic, carbonaceous and pyritic at depth.

68 Quartzite, feldspathic, with shale interbeds; silty sandstone in part schistose and calcareous.

69 Tillite; diamictite; shale; siltstone.

70 Sandstone to greywacke, fine to coarse-grained, dark grey, thick-bedded to laminated; interbedded with laminated siltstone and thin, sulphidic siltstone and lenticular grit to conglomerate beds. scour-and fill channels, rare cross-bedding.

71 Siltstone, grey to black, dolomitic and pyritic grading upwards to calcareous, thinly laminated, locally cross-bedded; dolomite, grey, flaggy to massive; limestone conglomerate, intraformational; greywacke.

72 Siltstone, sandy, flaser bedded.

73 Sand, bryozoal, ferruginous; limestone,fossiliferous, glauconitic. Shallow water, transgressive.

74 Andesite, dark grey, amygdaloidal; trachybasalt and andesite, greenish, calcite-filled amygdales; hawaiite pillow lava; interbedded volcanic breccia, conglomerate, volcaniclastic sediment, tuff, phyllite, sandstone, limestone. Age 526+/-4Ma (U-Pb)

75 Marble, white medium-grained crystals; Calc-silicate, grey,nodular

76 SILTSTONE, blue-black, laminated, sulphidic, partly limonitic, upper and lower horizons; SANDSTONE, medium to coarse-grained, dark grey, siltstone and phyllite interbeds. PEBBLE CONGLOMERATE at base. Worm casts, bioturbation, very rare trilobites.

77 Siltstone; shale, green-grey and purple.

78 Quartzite, white to cream, medium-grained, well bedded, feldspathic; interbeds of sandy, carbonaceous and pyritic shale.

79 LIMESTONE: dark to pale grey, mottled; oolitic; intraclastic and fenestral structured; SANDSTONE, calcareous, grauconitic and bioturbated; SILTSTONE, calcareous.

80 Sandstone, red, grey,with grit bands and gritty limestone; siltstone, red, gritty.

81 Shale, black; dolomitic siltstone; dolomite; grey laminated siltstone.

82 Siltstone; sandstone; diamictite.

83 Undifferentiated calcrete

84 Limestone, pale grey, micritic,poorly bedded, with clasts of Hallett Cove Sandstone at base.

85 Tillite; quartzite; siltstone. Massive, grey.

86 Sandstone and siltstone, laminated, graded bedding, flame structures and ripple drift crossbedding. Channelling.

87 Arkosic siltstone, blue, flaggy and thinly bedded, lenticular

88 Marble, white, blue, pink, amphibolitic.

89 Quartz sand; quartz gravel. Braided river system, fluviolacustrine in part.

90 Siltstone, laminated calcareous, light and dark grey (lighter bands being more calcareous), massive or laminated, local lenticular sandstone interbeds; sulphidic siltstone bands.

91 Undifferentiated Holocene coastal marine sediment.

92 Limestone, massive, oolitic, stromatolitic, ripple marks, overlain by dolomite with teepee structures. Colour from blue-grey at base to reddish-grey at top.

93 Limestone, sandy, grey, oolitic, stromatolitic, trough cross bedding; interbedded with siltstone, grey-green. Local diapir derived conglomerate.

94 Quartzite; arkose

95 Sandstone; siltstone, occasionally sulphidic; metamorphosed.

96 Limestone, echinoidal, bryzoal, crinoidal; sandstone,calcareous, minor carbonaceous clay and silt

97 Siltstone, green. Lower third is fine grained, includes glacial dropstones; middle unit is medium to coarse sandstone; upper unit is siltstone with minor sandstone. Minor diamictite, sandy and pebbly dolomite.

98 Middle quartzite member: ADELAIDE/BARKER digital database

99 Undifferentiated lacustrine/playa sediments

100 Brown coal, carbonaceous clay, silt and sand.

101 Calcarenite, bryozoal,calcrudite, glauconitic, silt and sand, Spicular mudstone, bryozoal marl.

102 Quartzite or sandstone interbeds.

103 Granitic, strongly foliated with lineation parallel to the country rock. Quartz, plagioclase(Oligiclase), and biotite, minor microcline. Pre to syn tectonic

104 Gneiss, coarse-grained, porphyritic; strongly foliated with well-defined biotite lineation. Pre- to syn-tectonic. Related to Rathjen Gneiss?

105 Unconsolidated white bioclastic quartz-carbonate sand of modern beaches and transgressive dune fields.

106 Undifferentiated Upalinna and Yerelina Subgroups; includes the superseded Willochra Subgroup

107 GRANITE, fine-grained; metasomatically altered to albitite. Syn-tectonic.

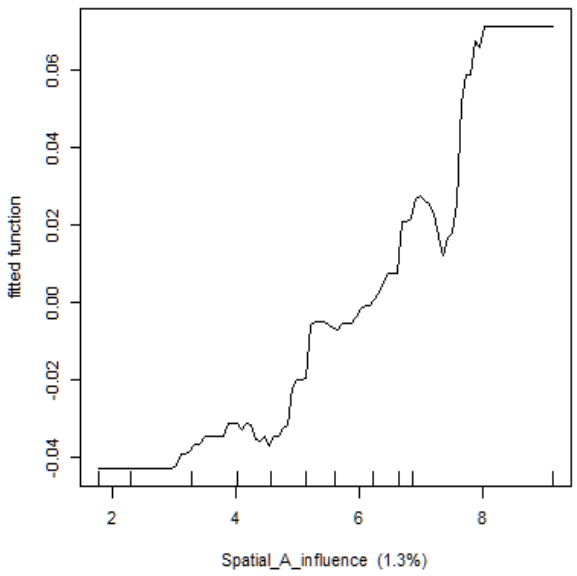
108 Dolomite

109 Dolomicrite, dark to medium grey, flaggy, laminated to medium bedded, occasional chert blebs, some thin chert layers

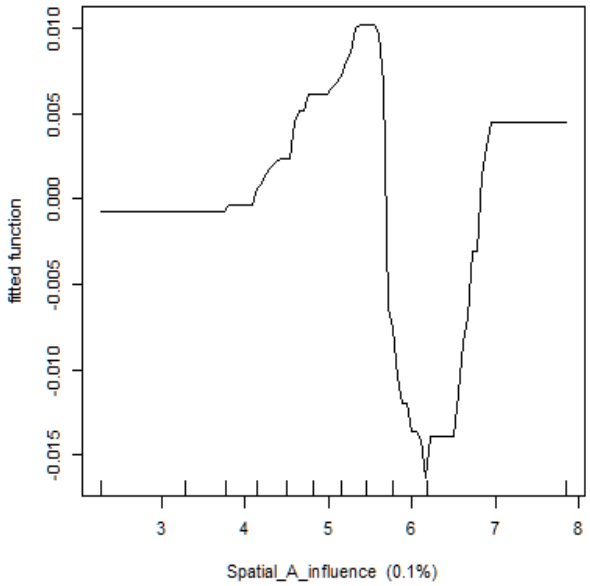
110 Siltstone, fine, sandy, cross bedding, minor thin dolomite lenses, local slumped siltstone beds.

111 Dolomite; sandstone; siltstone; quartzite.

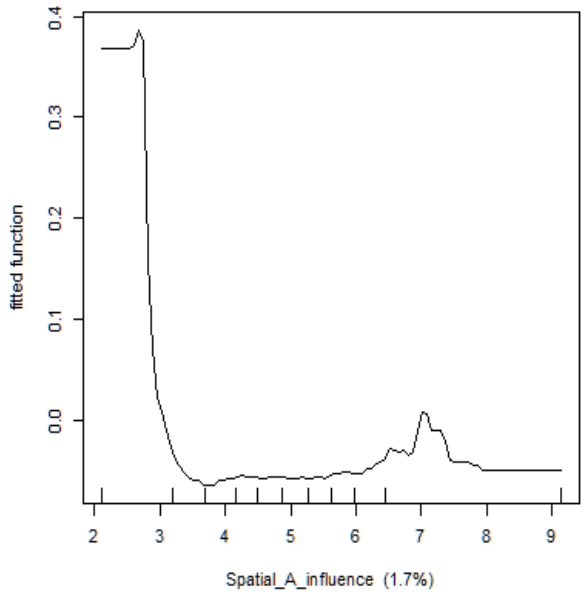
Appendix 6: Spatial interactions



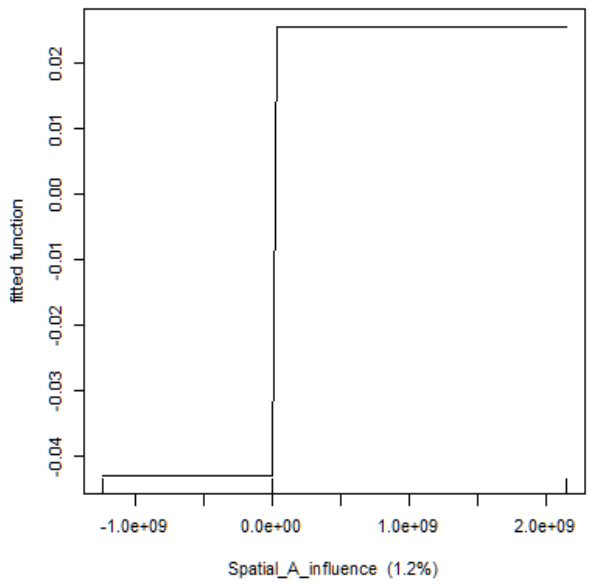
a: *A. verticillata*



b: *E. goniocalyx*



c: *E. fasciculosa*

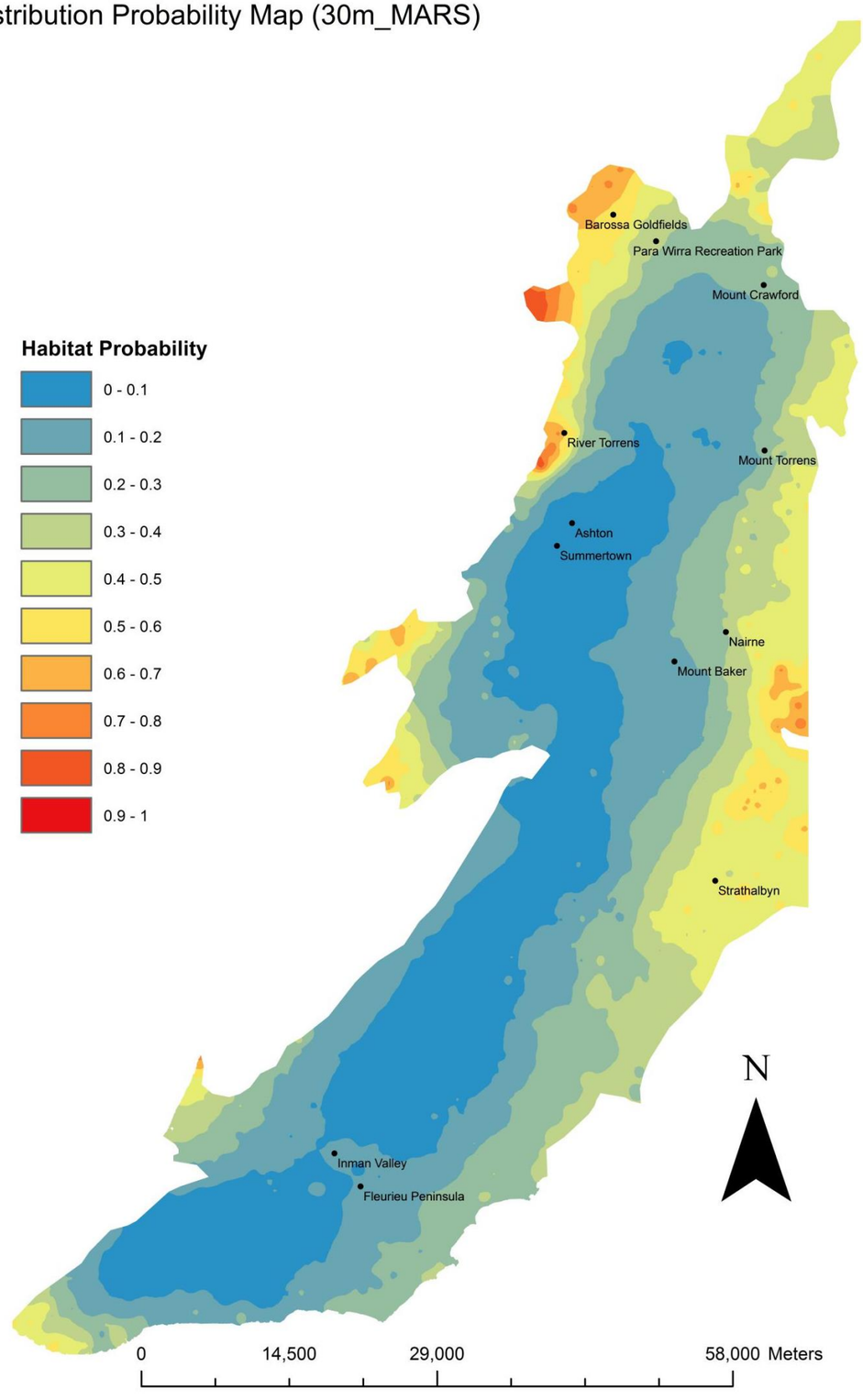


d: *E. obliqua*

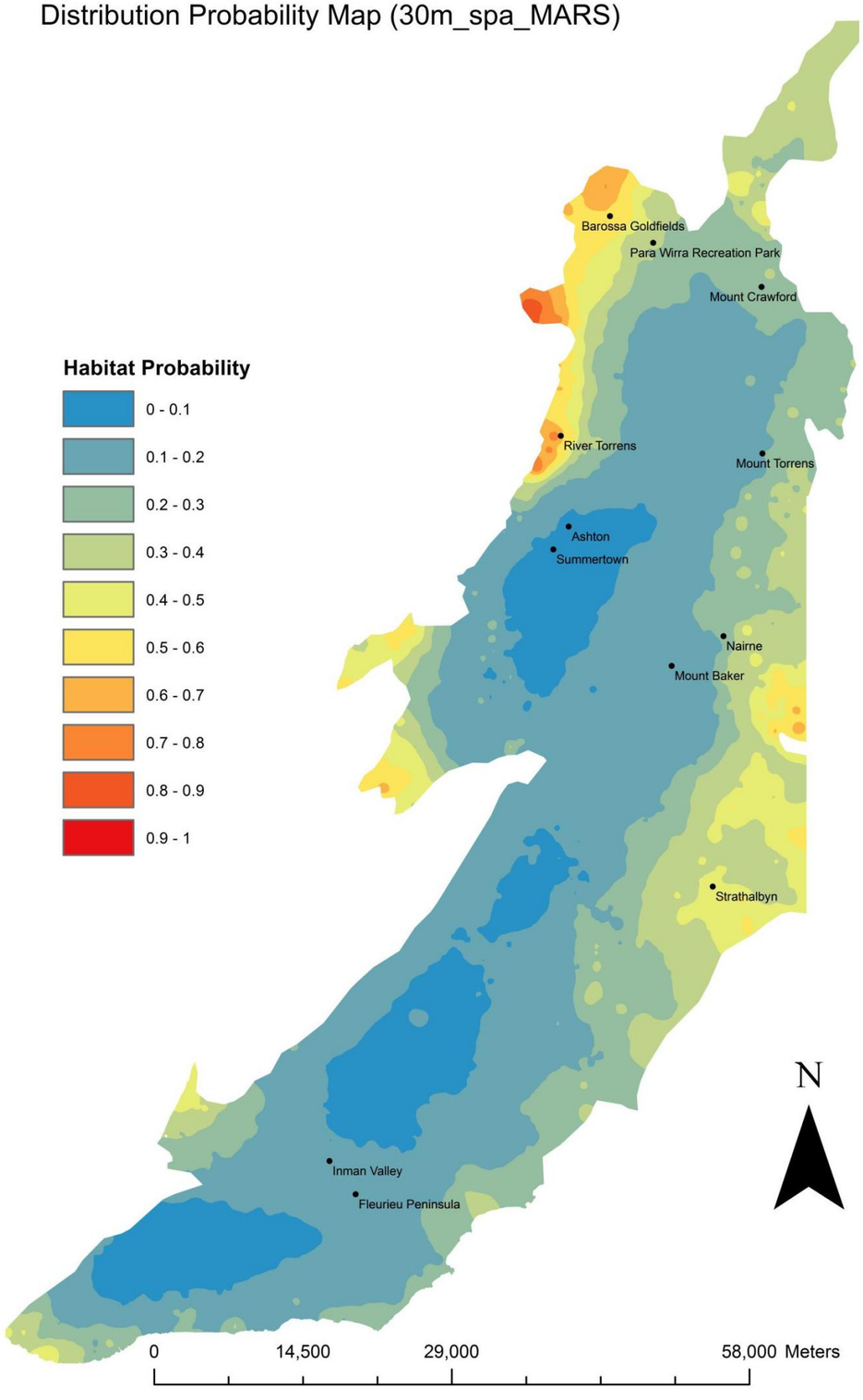
Figure 15: Response-predictors curve of spatial weights.

Appendix 7: Distribution probability of *A. verticillata*

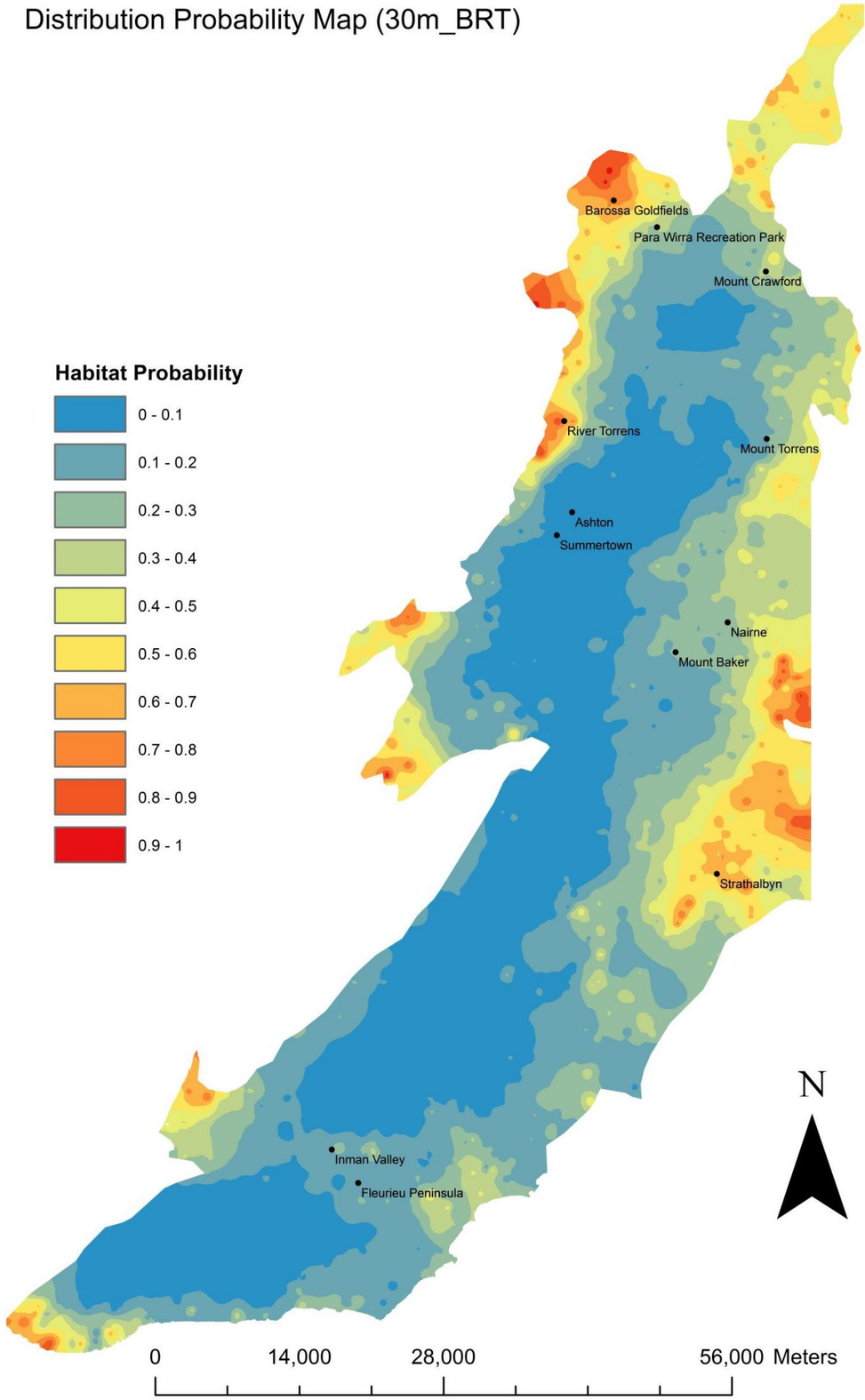
Distribution Probability Map (30m_MARS)



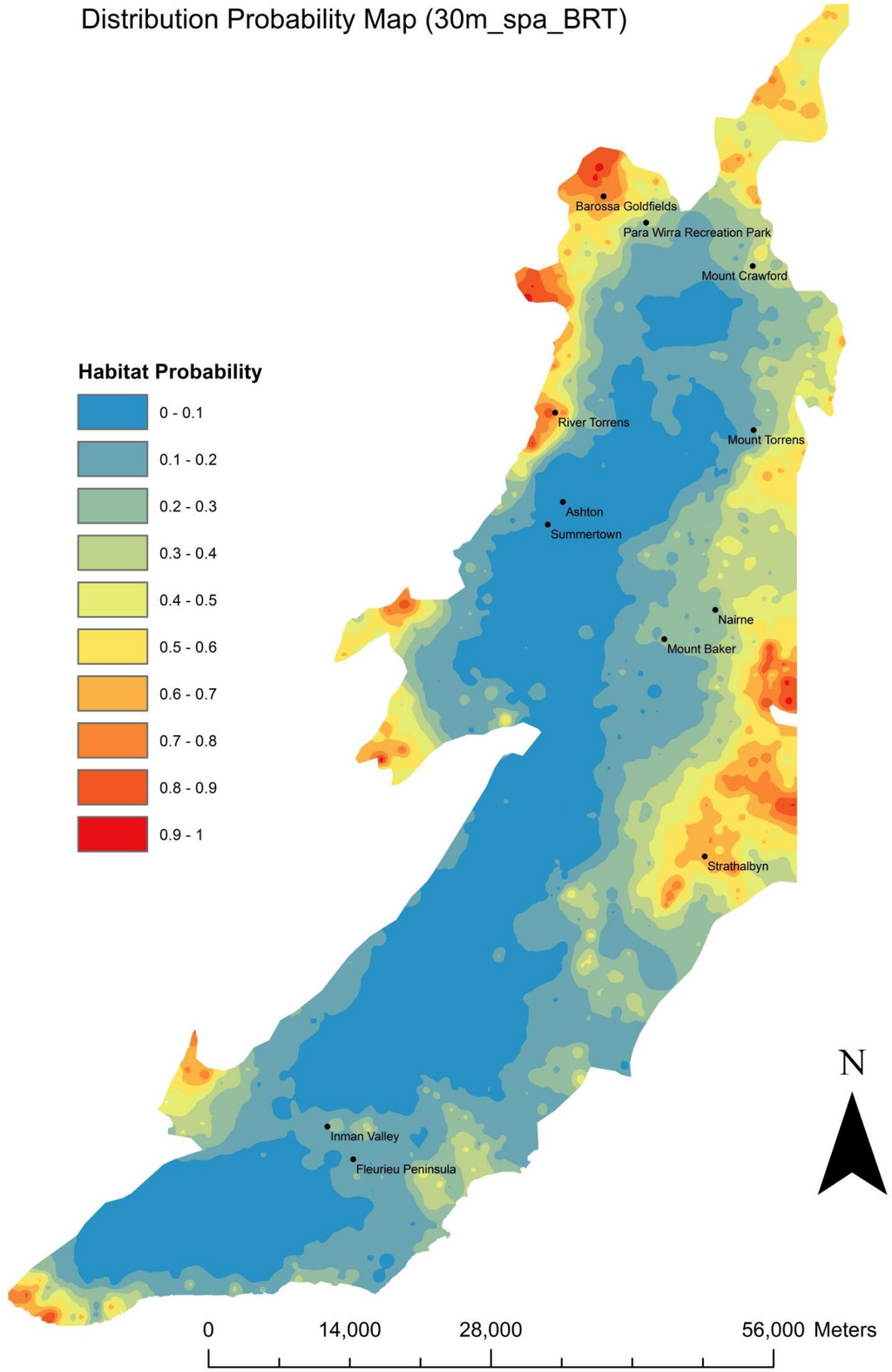
Distribution Probability Map (30m_spa_MARS)



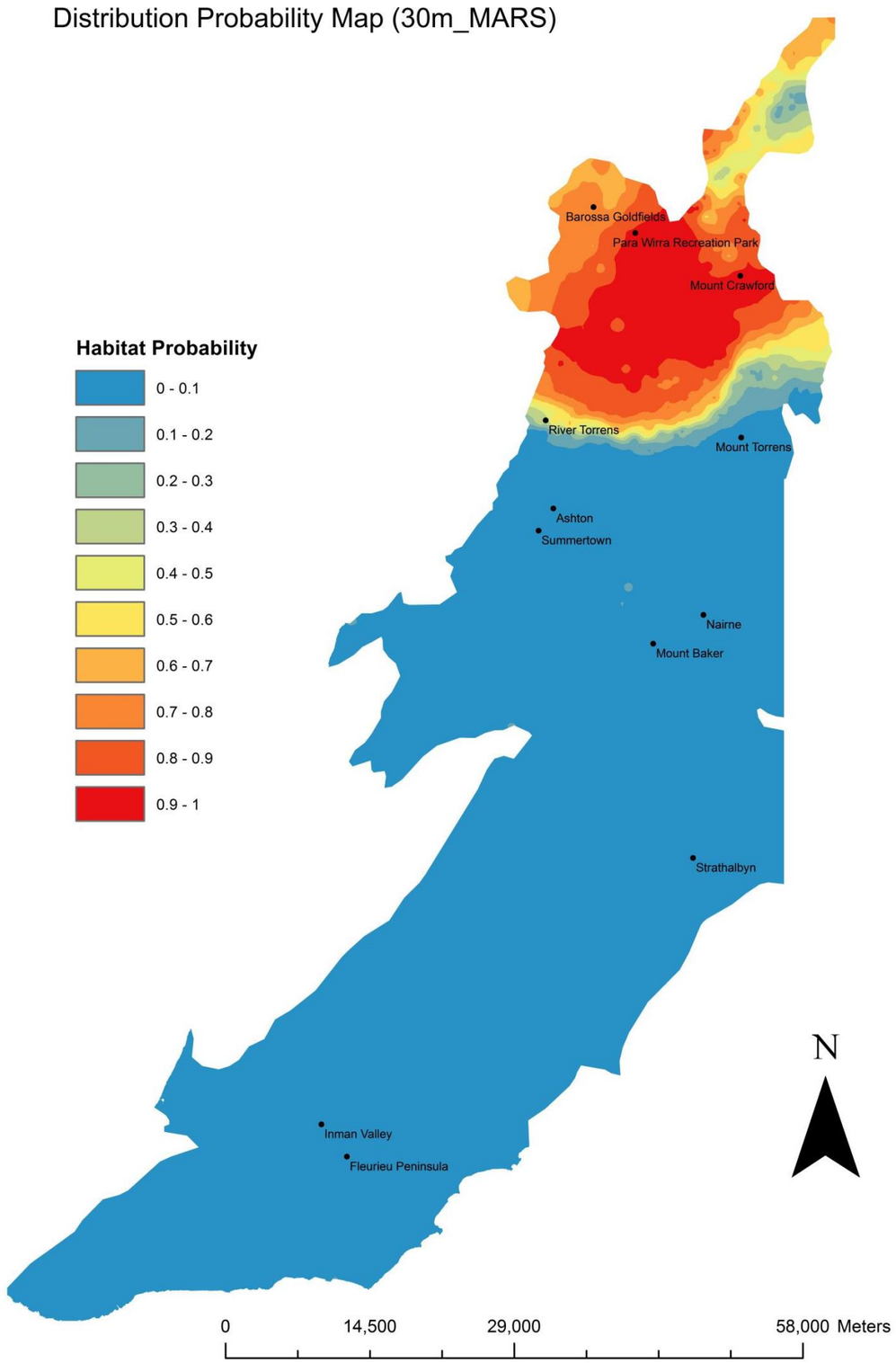
Distribution Probability Map (30m_BRT)

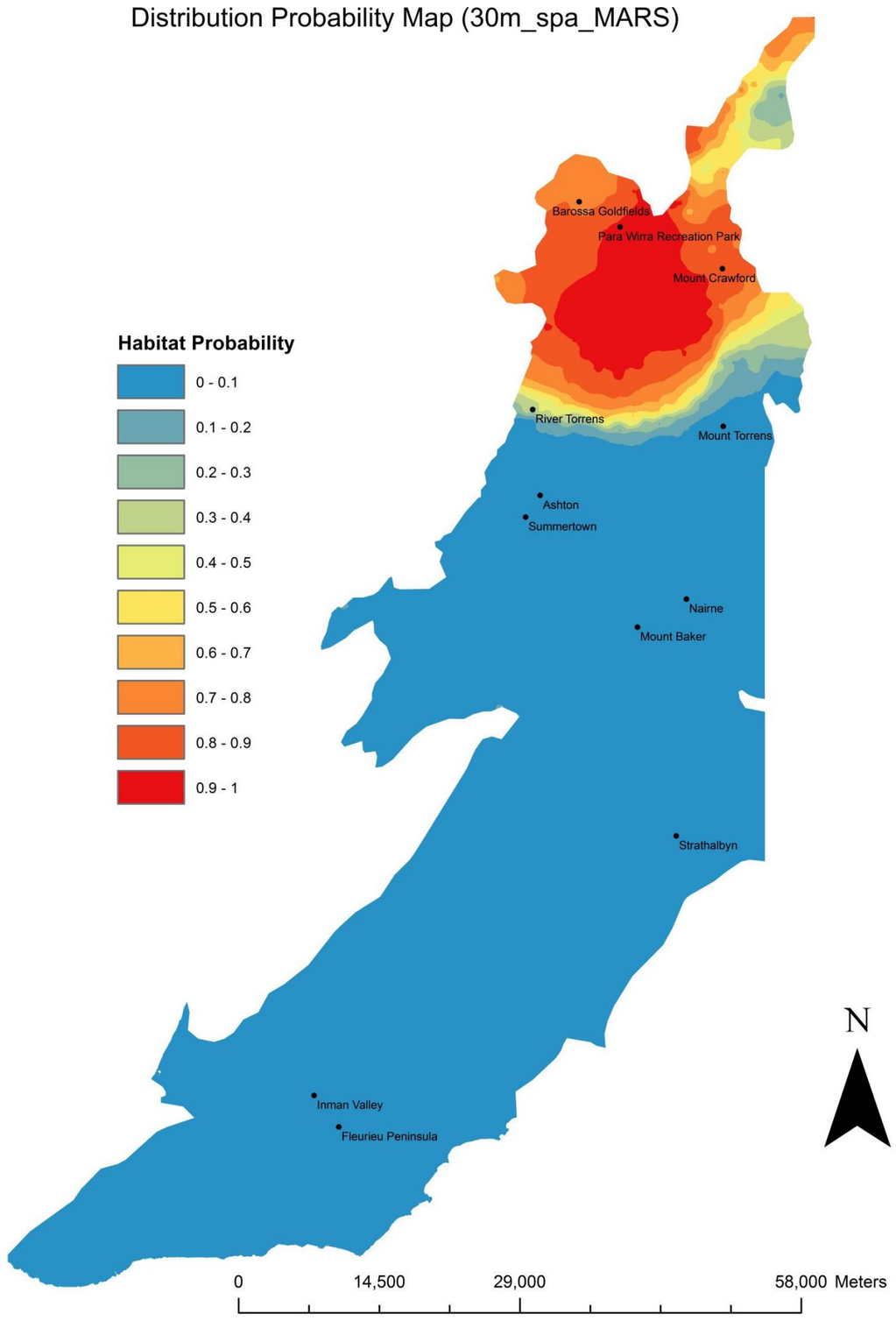


Distribution Probability Map (30m_spa_BRT)

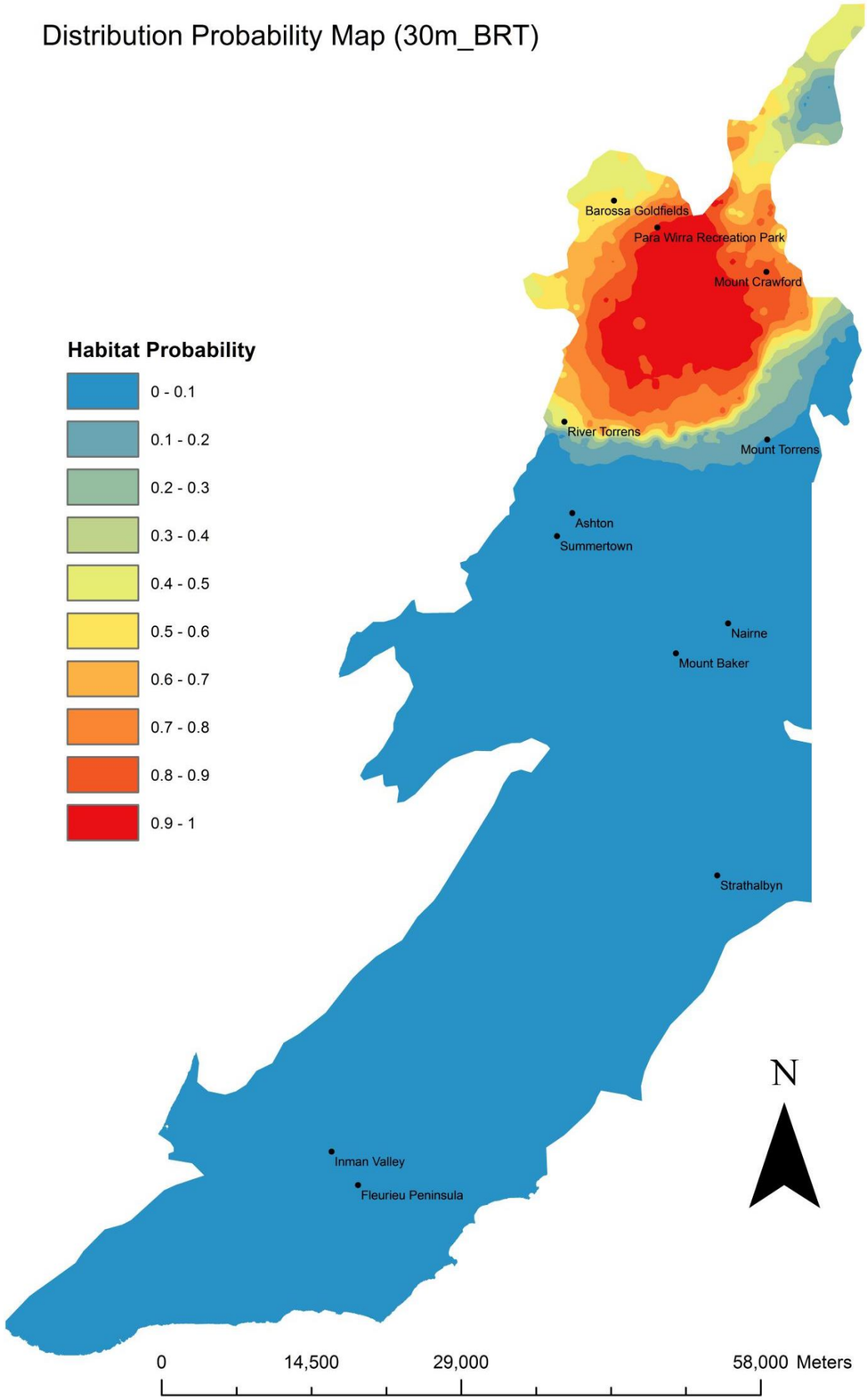


Appendix 8: Distribution probability of *E. goniocalyx*

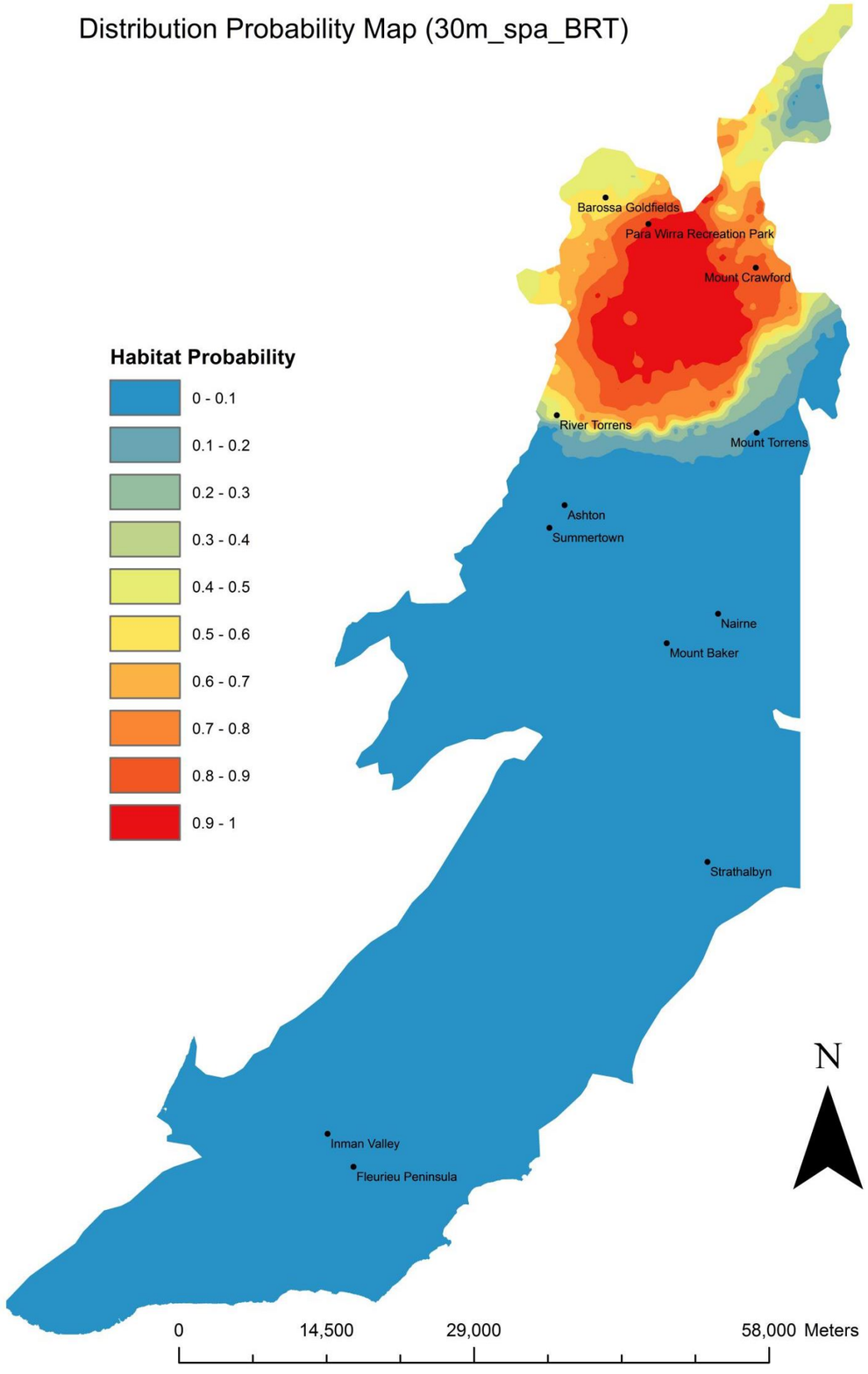




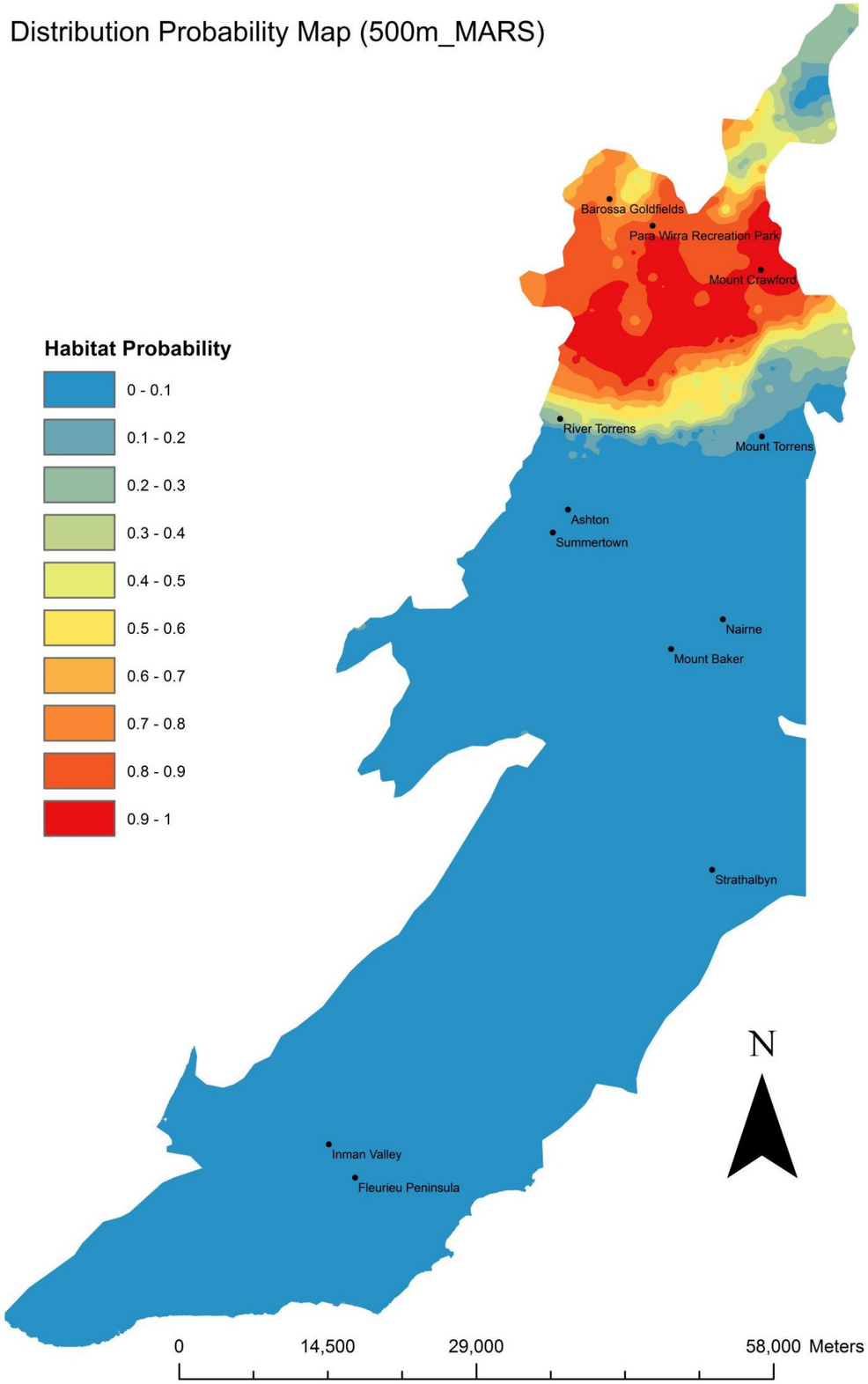
Distribution Probability Map (30m_BRT)



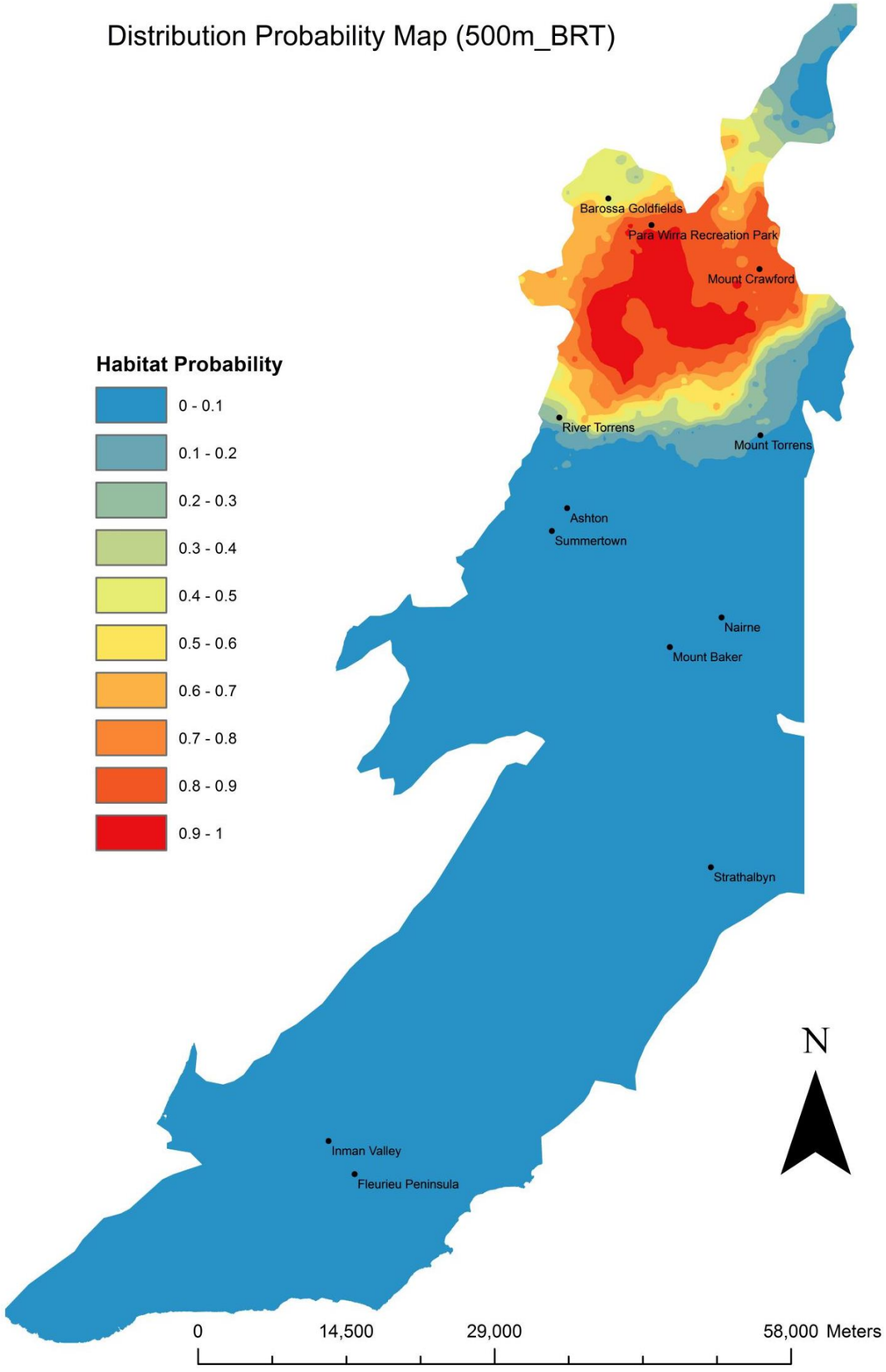
Distribution Probability Map (30m_spa_BRT)



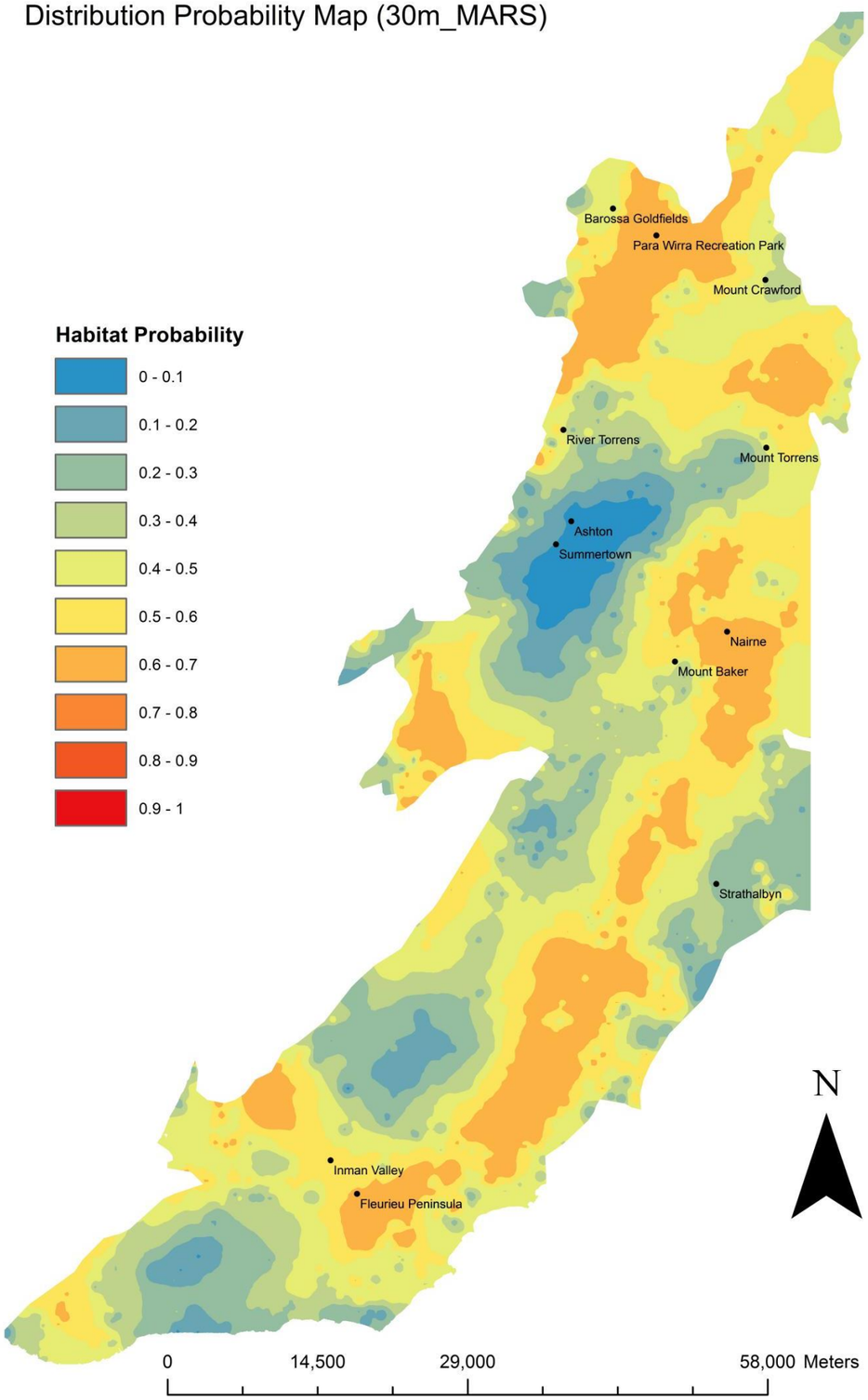
Distribution Probability Map (500m_MARS)

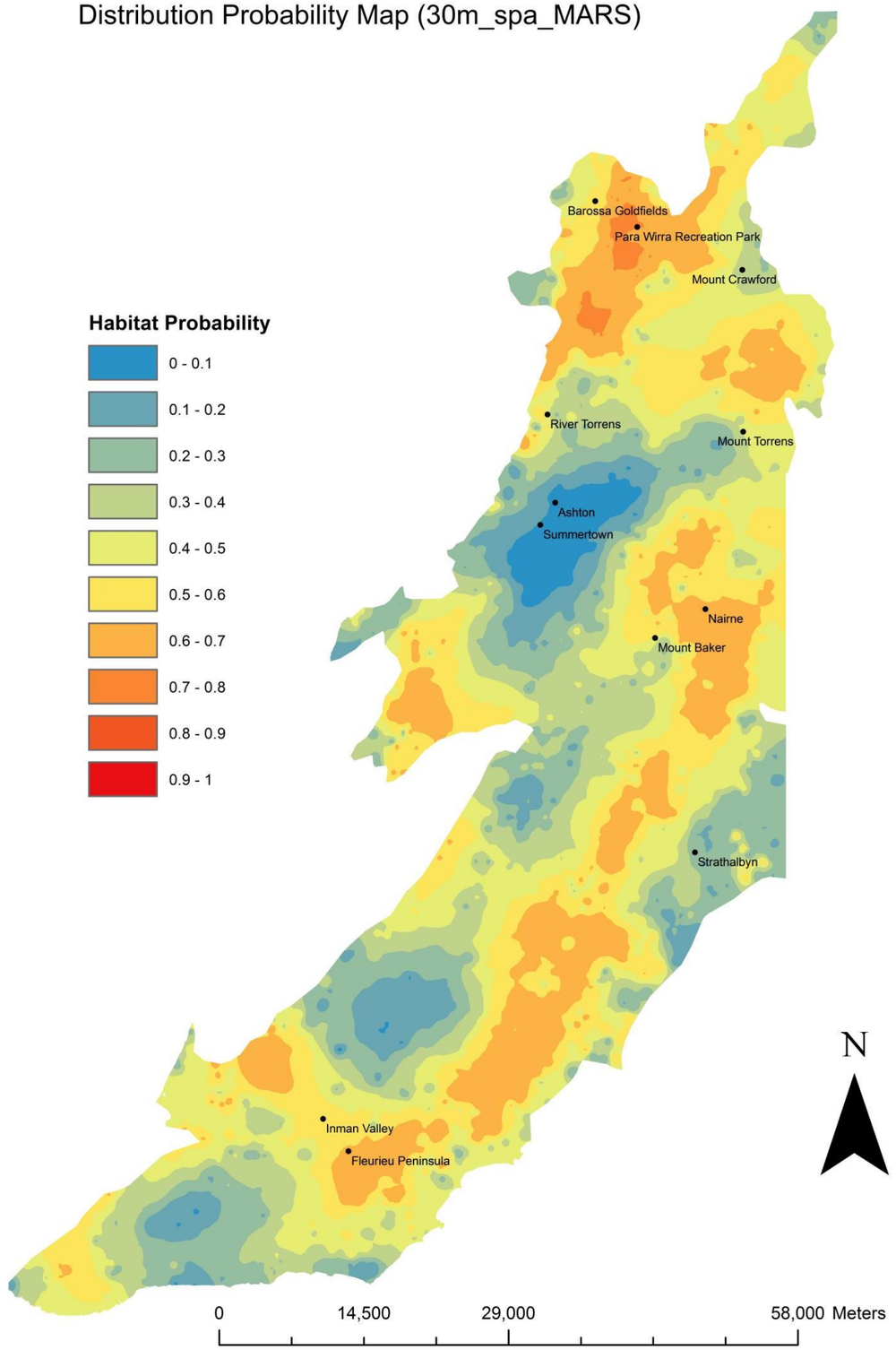


Distribution Probability Map (500m_BRT)

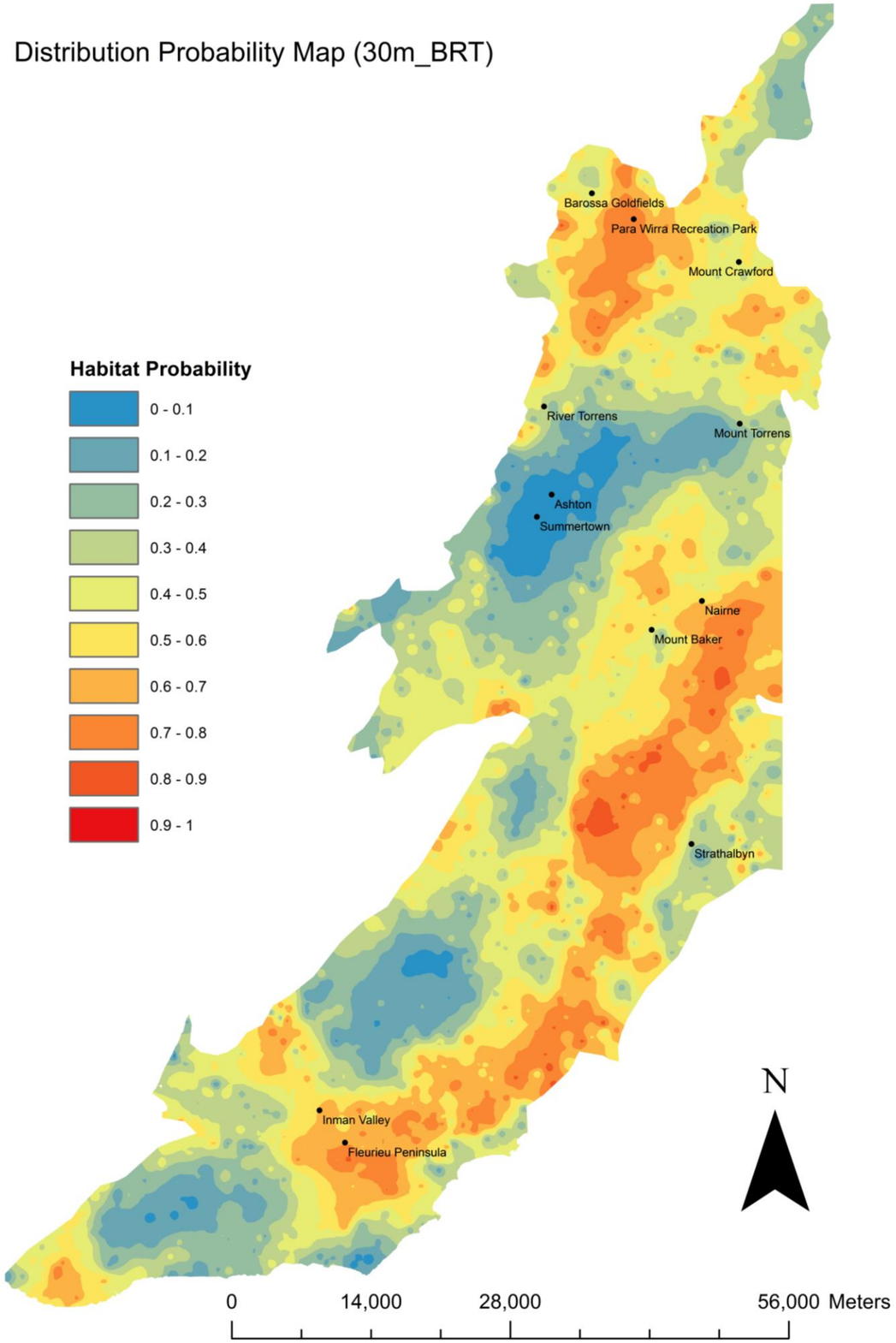


Appendix 9: Distribution probability of *E. fasciculosa*

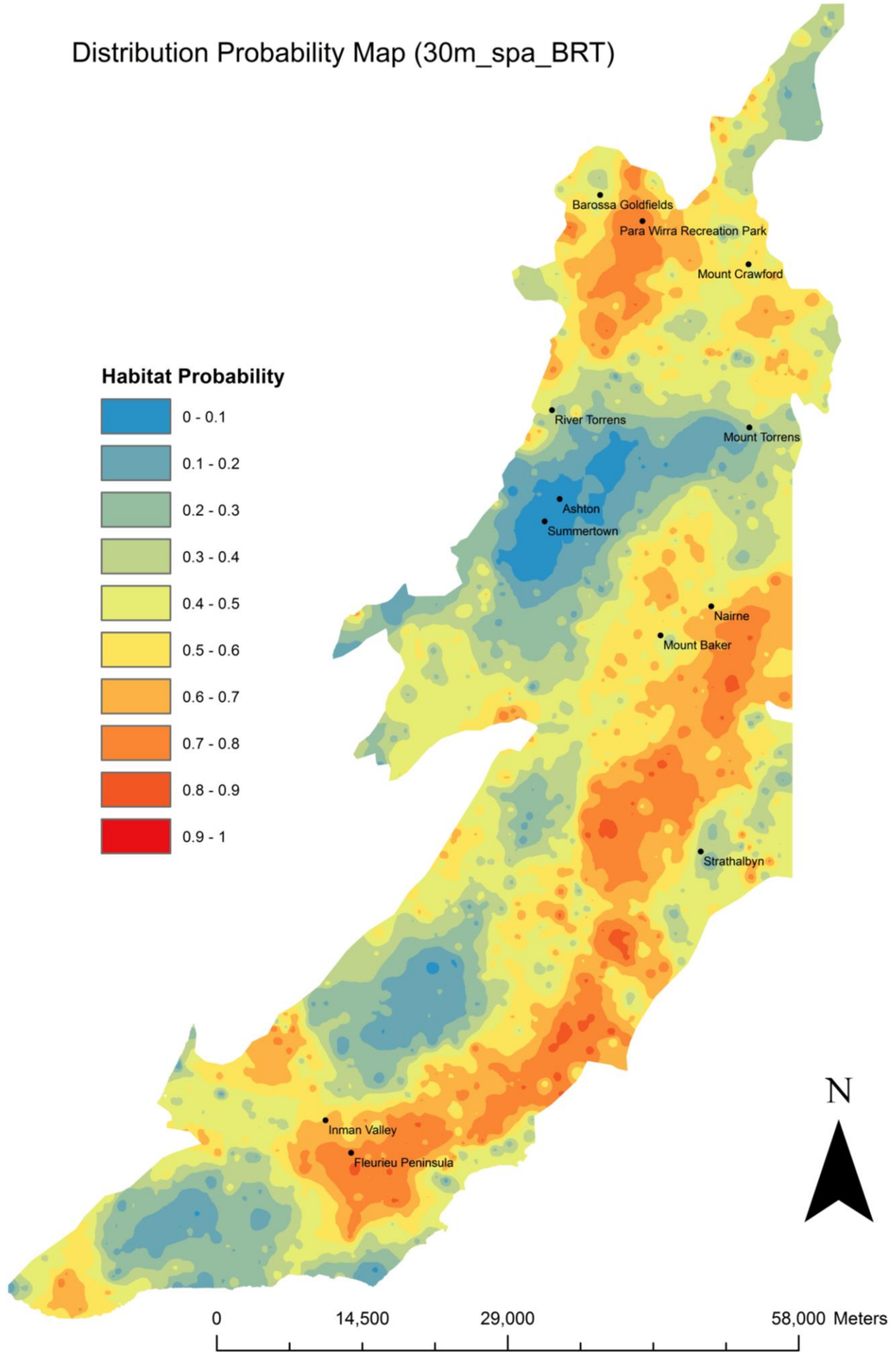




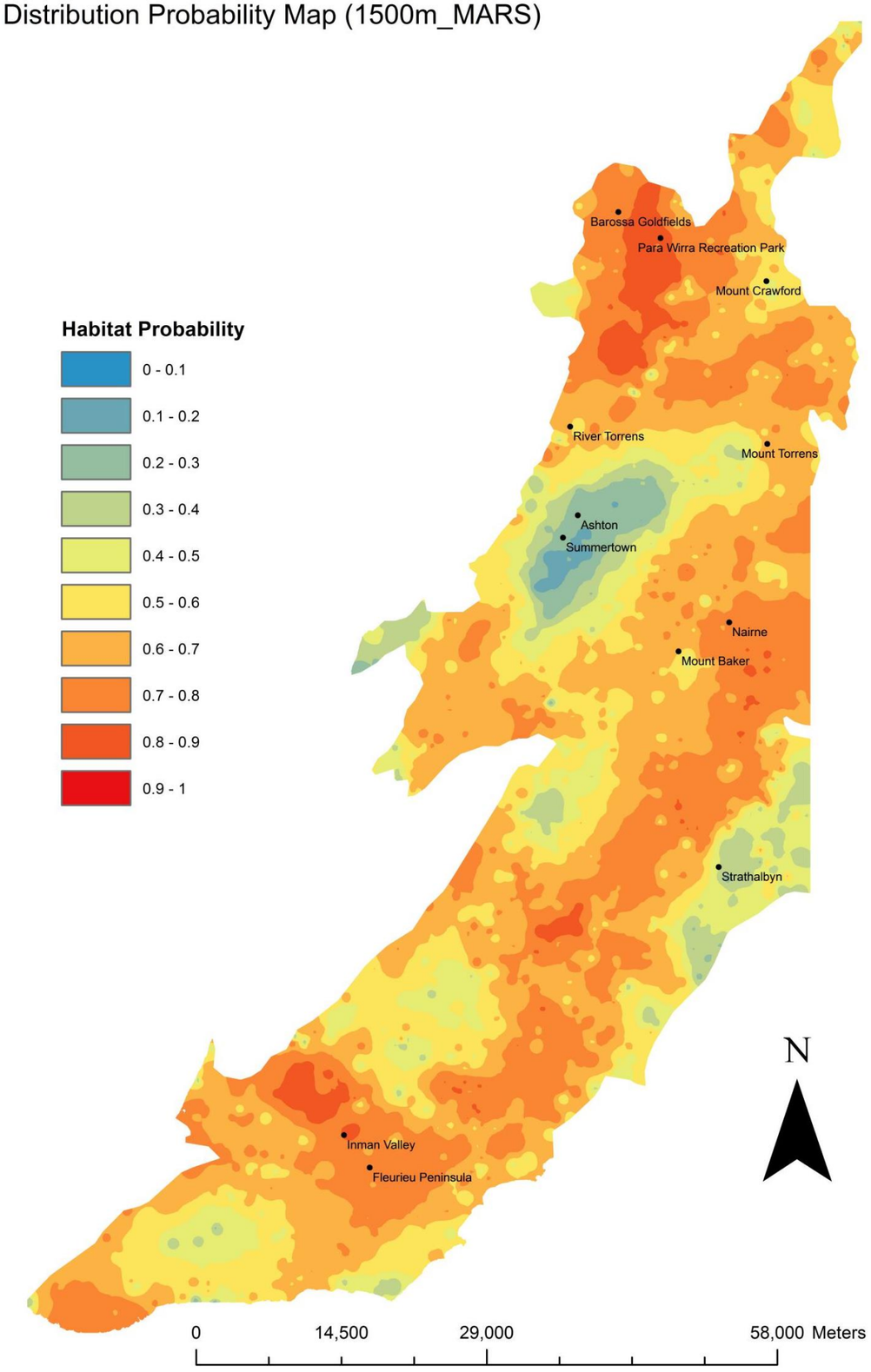
Distribution Probability Map (30m_BRT)



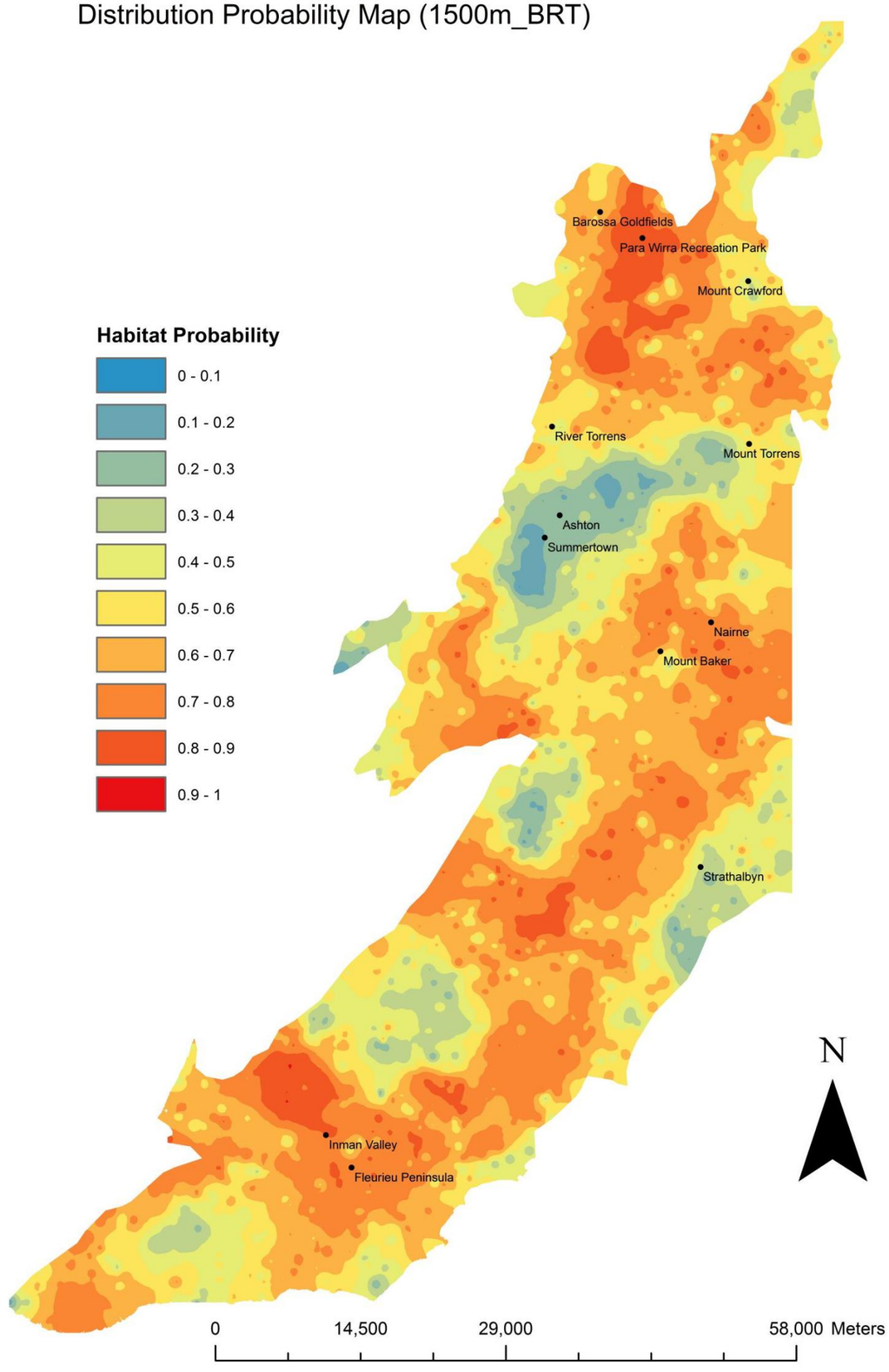
Distribution Probability Map (30m_spa_BRT)



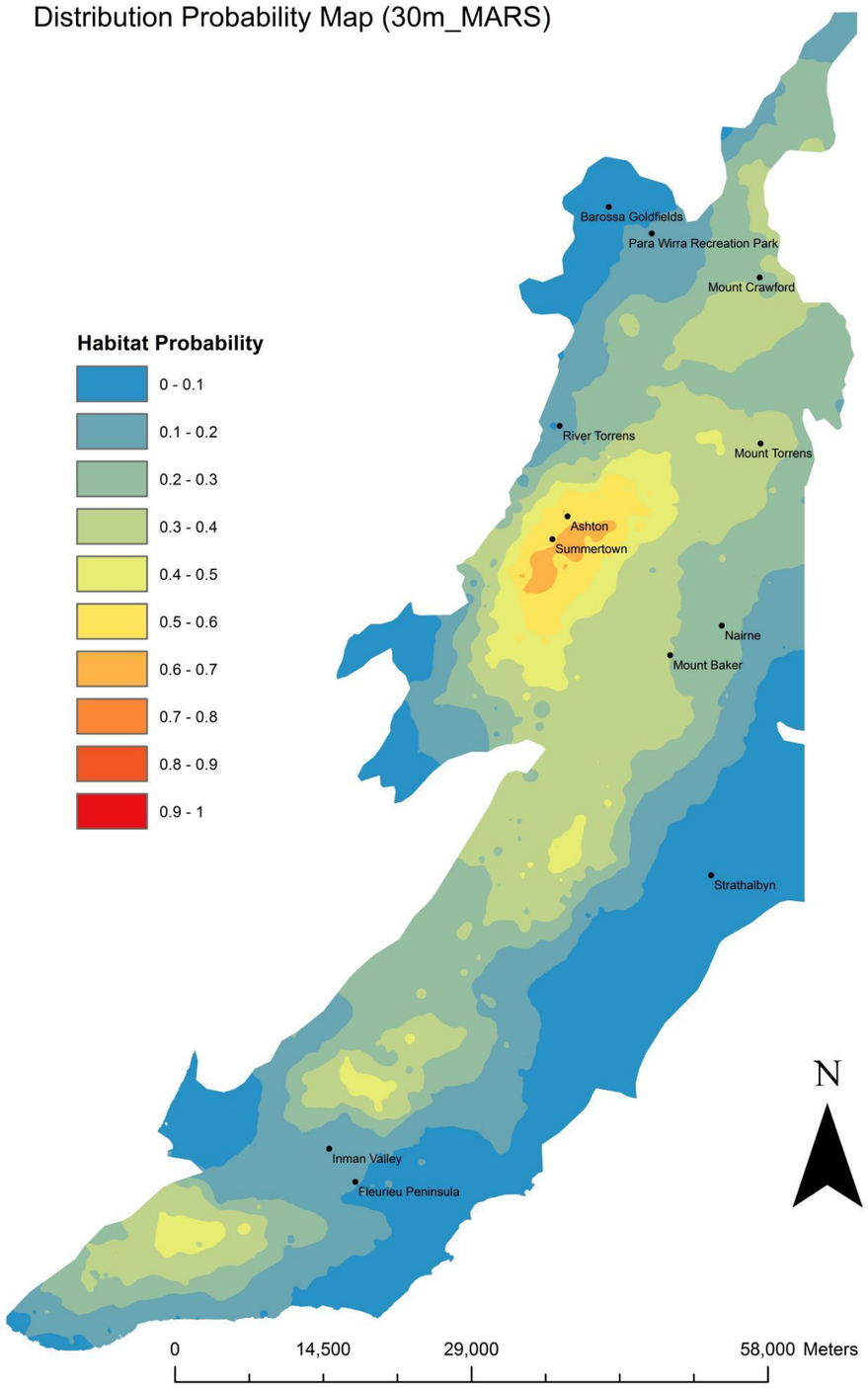
Distribution Probability Map (1500m_MARS)

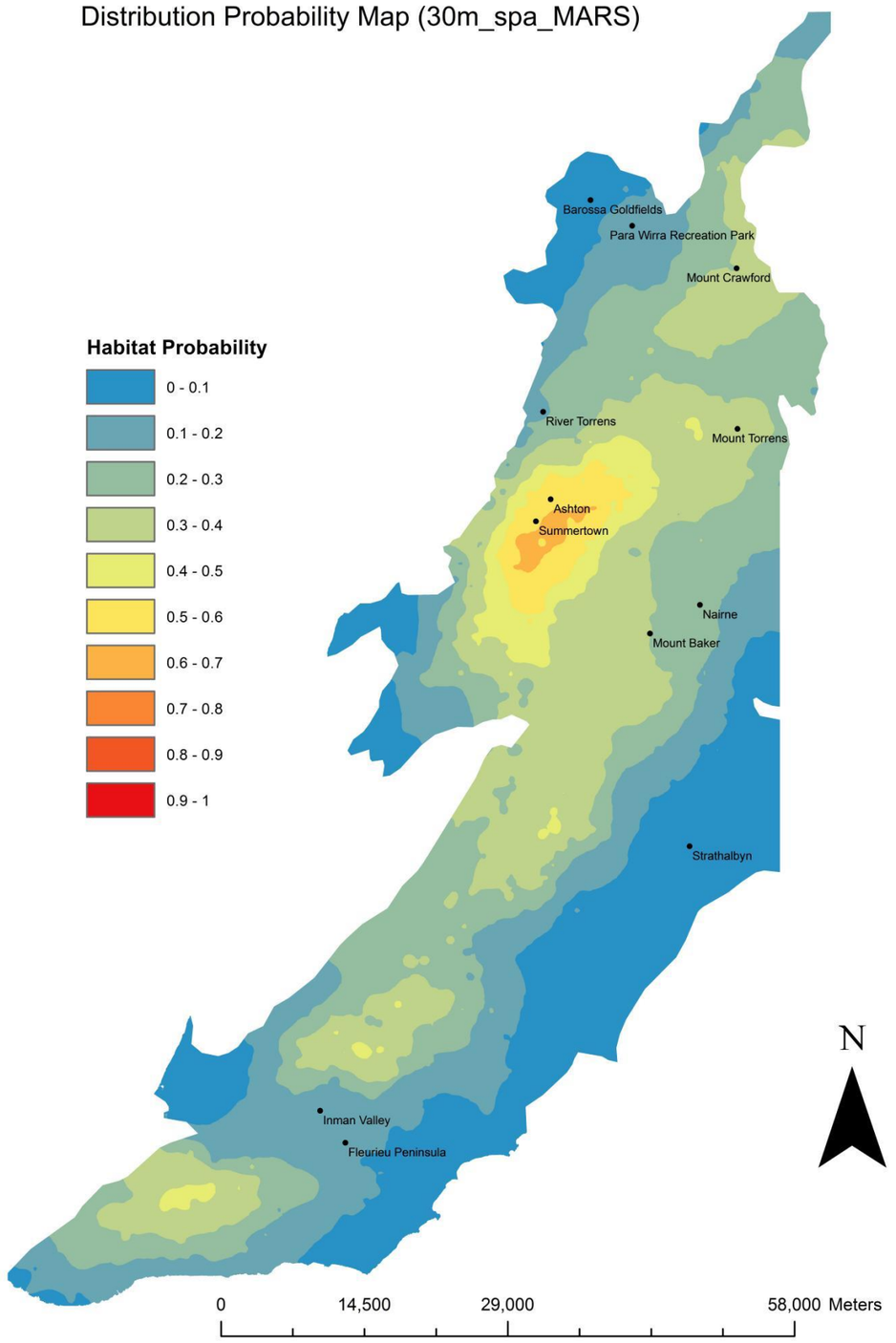


Distribution Probability Map (1500m_BRT)

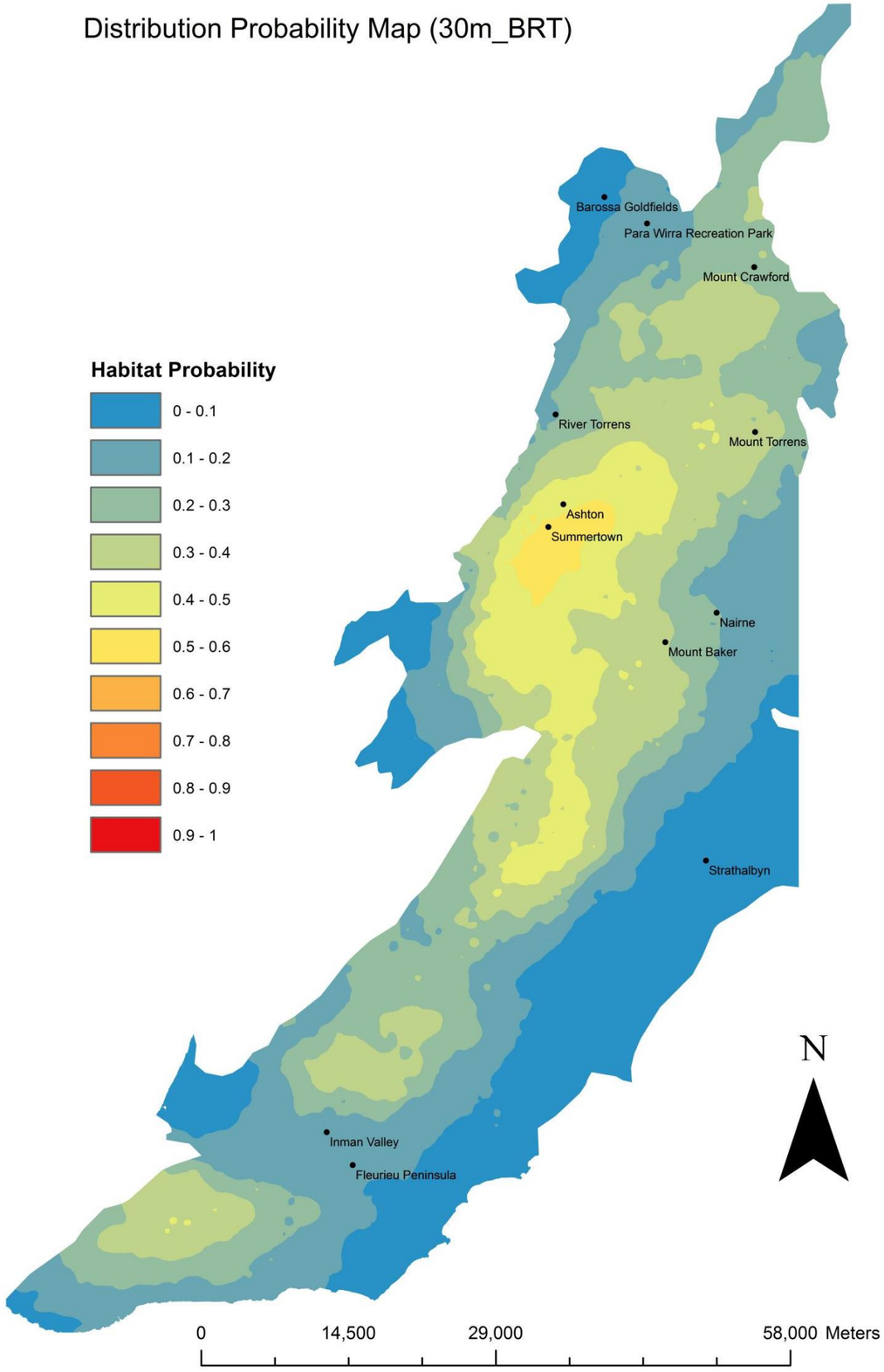


Appendix 10: Distribution probability of *E. obliqua*

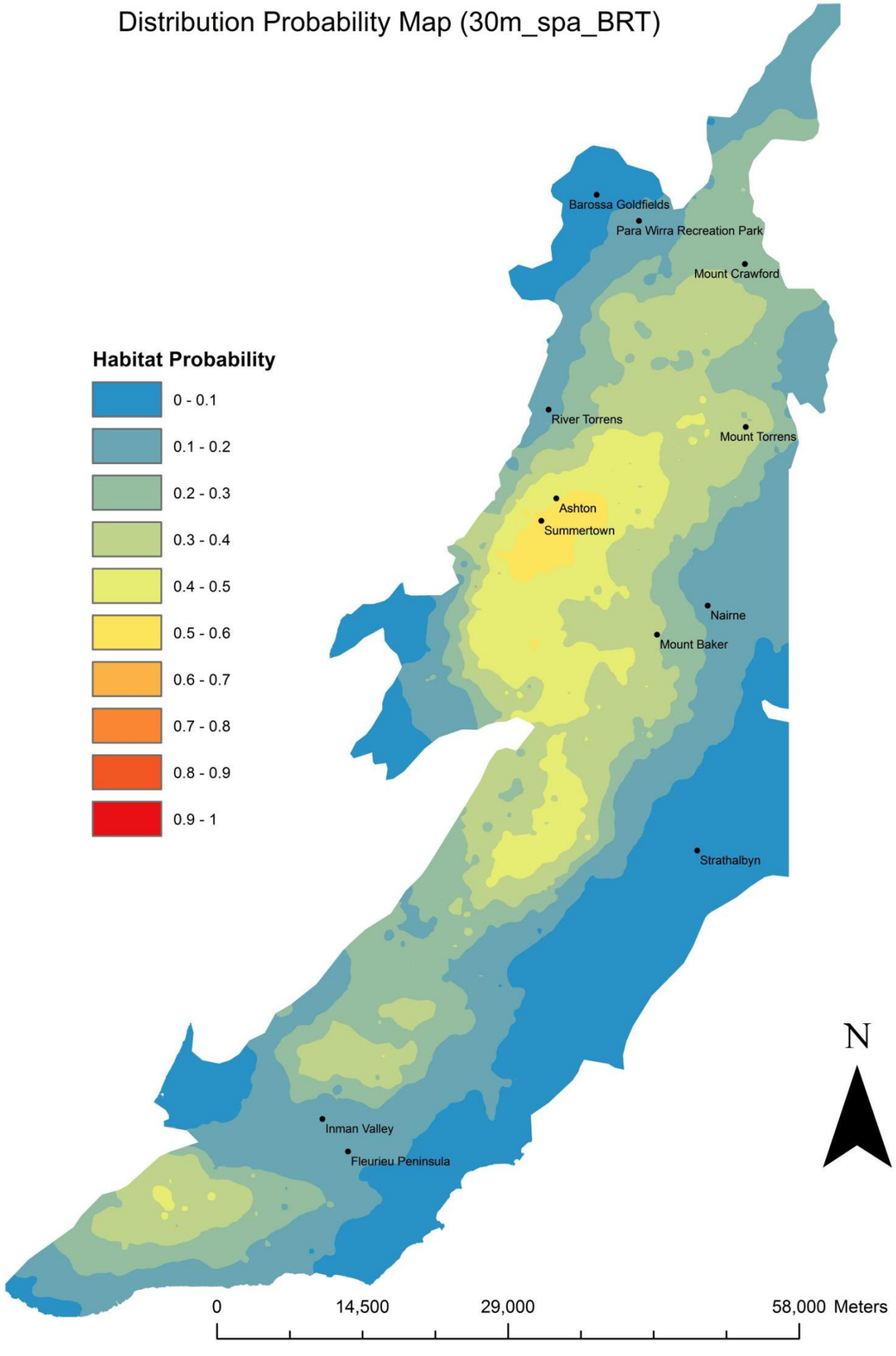




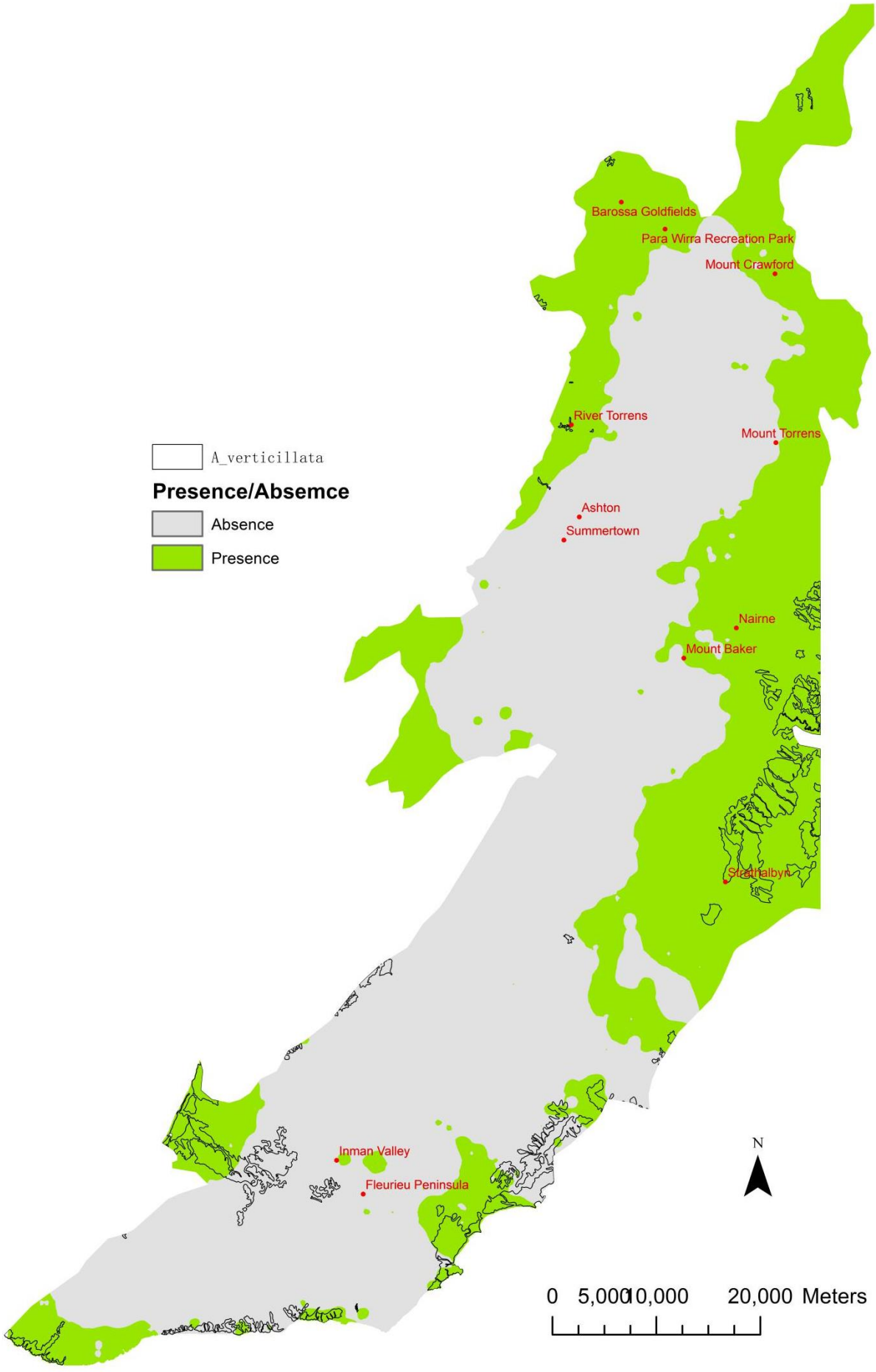
Distribution Probability Map (30m_BRT)



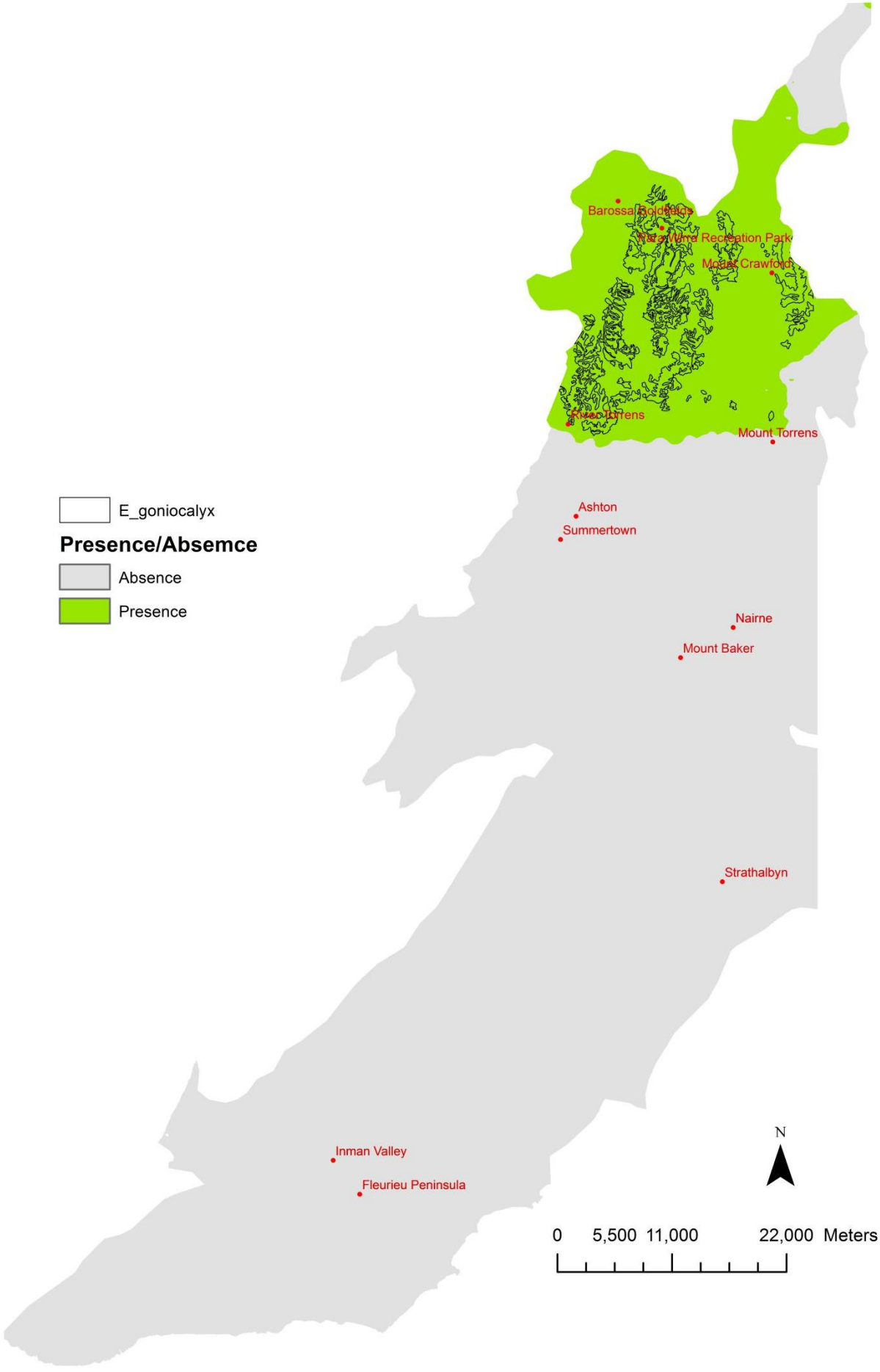
Distribution Probability Map (30m_spa_BRT)



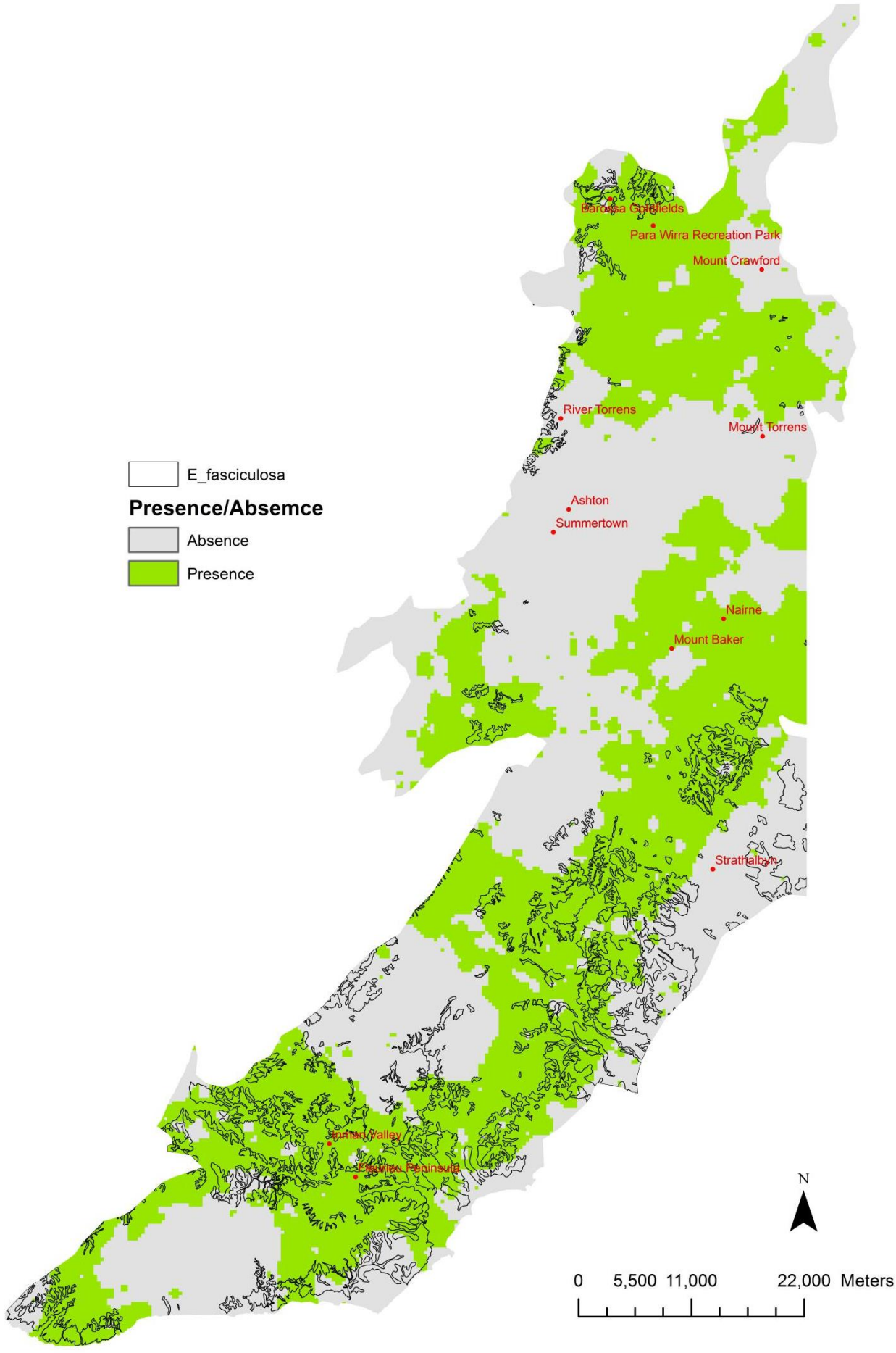
Appendix 11: Presence/Absence Map of A. Verticillata



Appendix 12: Presence/Absence Map of *E. goniocalyx*



Appendix 13: Presence/Absence Map of *E. fasciculosa*



Appendix 14: Presence/Absence Map of *E. obliqua*