

國立臺灣大學生物資源暨農學院森林環境暨資源學系
碩士論文

School of Forestry and Resource Conservation

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

優勢度於樣點資料對物種分布模型之影響

The Effects of Dominance in Sampling Data on Species

Distribution Modelling

The seal of National Taiwan University is a circular emblem. It features a central bell (the 'University Bell') flanked by two traditional Chinese lanterns. The seal is surrounded by the university's name in Chinese characters: '國立臺灣大學' at the top and '林政道' at the bottom. The outer ring contains the motto '愛國 勵學 愛人' (Love Country, Encourage Learning, Love People).

林政道

Lin, Cheng-Tao

指導教授：邱祈榮博士

Advisor: Chiou, Chyi-Rong, Ph.D.

中華民國 98 年 6 月

June 2009

In memory to my father, who has left me



Acknowledgement

I would like to thank my advisor and mentor Dr. Chiou, Chyi-Rong (邱祈榮). He has supported my professional researches and academic activities. He also encouraged me to attend the international symposium and gave me the confidence to finish my thesis after my father passed away. I would also express my appreciation to my colleague and mentor Dr. Song, Guo-Zhang (宋國彰) who has taught me scientific writings and gave me constructive suggestions in my experiment and thesis composition. I would like to thank Li, Ching-Feng's (五木學長) comments and field work experiences to my experiment. I would like to appreciate the assistance obtained from Dr. David Zelený who has made my experiment more complete. I would also like to express my enormous appreciation to Professor Ladislav Mucina, Oppenheimer foundation and School of Forestry and Resource Conservation of NTU. Without their support, I would not have a chance to attend the International Association for Vegetation Science annual symposium and publish my thesis topic. I would like to thank my committee Professor Lin, Yu-Pin (林裕彬) and Professor Chen, Tze-Ying (陳子英) who gave me various comments and suggestions, especially in spatial analysis.

This research is my first attempt to ecological research and I would like to thank my colleagues and friends who help me to finish the work: Chi-Rong (志融), Da-Pang (大胖), Dun-Chung (惇淳學姊), Kiki, Yu-Chi(玉琦), Dun-Wei (惇為), Chen-Wei (阿邱), Hsiao-Lung (孝隆學長), Kai-Ru (愷如學姊), Velen, Kie-Tsung (台哥), Wen-Wei (文韉), Jing-Su (錦淑), Yeng-Hsiang (燕翔), Yu-Ching (郁青) and our vegetation group members: uncle Song (宋叔叔), Yao Chiang (姚強), Chien-Rong (建融), ssdot (世鐸) and Rita (睿涵). We have had many beautiful memories in reading papers, discussing weekly topics and writing symposium papers. I would like to thank citagar who helped my father go through the illness during his last six months and finally I want to thank my mother who supported my study and decisions and my grandma who taught me land ethics and gave me sufficient knowledge about agriculture and nature.

The end is another beginning. "*Progressus ad futurum*", and I would dedicate my thesis to the nature, the mother earth.



Abstract

It has been reported that the performance of species distribution models are related with properties of data and species traits. The dominance of a species in a habitat represents the successfulness of regeneration of a population there and thereby may be associated with the probability of species occurrence. Habitats with low dominance of a species may be a noise for modelling, which might reduce the accuracy of SDMs. Here we would like propose two questions: Does removal of low dominance data increase the accuracy of SDMs? Is species dominance an influential factor for SDMs?

Tsuga chinensis var. *formosensis*, a native conifer species which is widely distributed in habitats ranging from 2000 m to 3100 m above sea level in Taiwan, was selected for modelling. Two scenarios were evaluated for testing the dominance effects in sampling data. The first scenario used IVI to select presence data according to the dominance and the sampling plots were divided into ascendant and descendant accumulative datasets. The second scenario used logarithm basal area to select presence data and the sampling plots were also divided into ascendant and descendant accumulative datasets. GAM and MAXENT were both used for building the models.

In the first scenario, AUC values of the two models decrease while gradually removing higher dominance datasets in the descendant accumulative datasets. In contrary, removal of low dominance data in ascendant accumulative datasets does not increase the accuracy of the two models. Similarity, in the second scenario, there are no significant differences amongst ascendant and descendant datasets of the two models. Regardless of various dominance levels of data, the accuracy of prediction of MAXENT is slightly higher than that of GAMs. Our result shows dominance in sampling data would affect the performance of species distribution modelling.

Keywords: Dominance, *Tsuga chinensis* var. *formosensis*, GAM, MAXENT, species distribution models (SDM).



摘要

近年來許多研究報告指出物種分布模型會受到資料以及物種特性的影響。而一個物種的優勢度往往表示這個物種的族群能夠成功建立更新的指標，此外低優勢度的棲地也許對於分布模型來說具有雜訊，並影響物種分布模型的表現。因此在本篇研究中，我們將探討以下兩個問題：將較低優勢度資料移除是不是會增加物種分布模型的表現？在取樣資料中的優勢度對於物種分布模型來說是不是一個具有影響力的因素？

臺灣鐵杉在臺灣主要是分布在海拔兩千至三千公尺廣泛分布的物種，而本研究中將臺灣鐵杉選定為目標物種。在本篇研究中，總共有兩個測試模式，第一個測試模式使用重要值指數(IVI)來選擇出現點位，並根據優勢度將樣點切分為遞增及遞減兩個累積相對優勢度資料集。第二個測試模式則是根據對數胸高斷面積來選擇出現點位，並根據優勢度將樣點切分為遞增及遞減兩個累積相對優勢度資料集。切分完測試模式後，使用 GAM 和 MAXENT 來進行物種分布模式的測試。

在第一個測試模式中，在遞減累積相對優勢度資料集內若逐漸去除較高的優勢度資料，兩個物種分布模式的 AUC 值會逐漸下降。相對於遞減累積相對優勢度資料集，遞增累積相對優勢度資料集則無此趨勢。在第二個測試模式中，不管是遞增或遞減的資料集則無明顯的差異。儘管不同的優勢度之下，整體來說 MAXENT 的表現比 GAM 還要來的好一些。我們的研究結果並指出在樣點資料的優勢度會對物種分布模型的表現造成影響。

關鍵字：優勢度、臺灣鐵杉、GAM、MAXENT、物種分布模型(SDM)



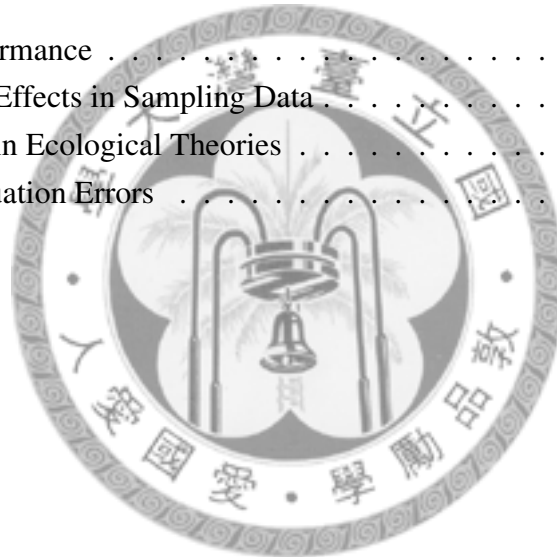
Table of Contents

Abstract	i
Chinese Abstract	iii
Table of Contents	v
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Background	1
1.2 Purpose of Research	2
2 Literature Review	4
2.1 Dominance in Analytic Concepts	4
2.1.1 Analytic Method	5
2.1.2 Combination of Analytic Characters	8
2.2 Species Distribution Models	9
2.2.1 Generalized Additive Models	10
2.2.2 Maximum Entropy Principles	13
2.3 Comparison of Different Models	16
2.4 Influential Factors to SDM	18
2.4.1 Species Traits	18
2.4.2 Data Characteristics	18
2.5 Model Performance and Evaluation	19
3 Material and Methods	23
3.1 Data Preprocess and Preparation	23



TABLE OF CONTENTS

3.1.1	Target Species	23
3.1.2	Occurrence Data	25
3.1.3	Environmental Variables	26
3.1.4	Datasets Preparation	28
3.2	Model Building	31
3.3	Model Evaluation	32
3.4	Implementation of Experiment	33
3.5	Analyses of Dominance Effects	38
4	Results	40
4.1	Experimental Test	41
5	Discussion	47
5.1	Model Performance	48
5.2	Dominance Effects in Sampling Data	49
5.3	Dominance in Ecological Theories	51
5.4	Model Evaluation Errors	53
6	Conclusion	58
	References	61
A	Demo program	66
B	NPMC results	79



List of Figures

2.1	Concept of maximum entropy applied in prediction of species distribution.	15
2.2	Sample ROC curves from \mathbf{R} SIM3DATA	21
3.1	Overall experiment flowchart	39
4.1	Jackknife analysis of training gains of <i>Tsuga chinensis</i> var. <i>formosensis</i> .	41
4.2	Scenario 1 - Ascendant accumulative relative dominance datasets. Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.	43
4.3	Scenario 1 - Descendant accumulative relative dominance datasets. Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.	44
4.4	Scenario 1 - Ascendant accumulative relative dominance datasets. Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.	45
4.5	Scenario 2 - Descendant accumulative relative dominance datasets Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.	46
5.1	Diagram of removal of higher dominance datasets in range of tolerance .	50
5.2	Diagram of removal of lower dominance datasets in range of tolerance . .	50
5.3	Histogram of basal area (scenario 2)	52

List of Tables

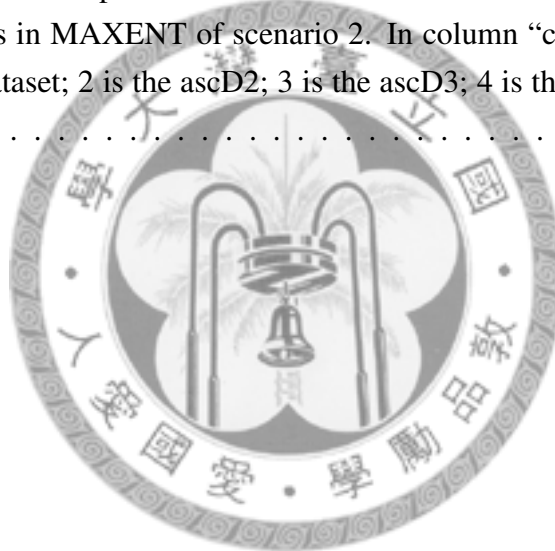
2.1	Braun-Blanquet coverage classes and values	7
2.2	Some exponential family distributions.	11
2.3	A 2×2 confusion matrix	20
2.4	Measurements derived from a 2×2 confusion matrix	20
3.1	Environmental variables used for modelling of species distribution	27
3.2	Scenario 1: octave scale IVI datasets.	29
3.3	Scenario 1: ascendant and descendant accumulative datasets	30
3.4	Scenario 2: dataset cut points and sampling plots quantity	30
4.1	Multivariate correlation matrix	40
B.1	Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in GAM of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo8); 2 indicates the dataset2 (RDo7-8); 3 is RDo6-8 4 is RDo5-8; 5 is RDo4-8; 6 is RDo3-8.	79
B.2	Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in MAXENT of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo8); 2 indicates the dataset2 (RDo7-8); 3 is RDo6-8 4 is RDo5-8; 5 is RDo4-8; 6 is RDo3-8.	80
B.3	Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets in GAM of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo3-8); 2 indicates the dataset2 (RDo3-7); 3 is RDo3-6 4 is RDo3-5; 5 is RDo3-4.	80
B.4	Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets in MAXENT of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo3-8); 2 indicates the dataset2 (RDo3-7); 3 is RDo3-6 4 is RDo3-5; 5 is RDo3-4.	81

B.5 Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in GAM of scenario 2. In column “cmp”, 1 indicates the ascD1 dataset; 2 is the ascD2; 3 is the ascD3; 4 is the ascD4 and 5 is the ascD5 . 81

B.6 Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in MAXENT of scenario 2. In column “cmp”, 1 indicates the ascD1 dataset; 2 is the ascD2; 3 is the ascD3; 4 is the ascD4 and 5 is the ascD5 82

B.7 Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets of scenario 2. In column “cmp”, 1 indicates the descD1 dataset; 2 is the descD2; 3 is the descD3; 4 is the descD4 and 5 is the descD5 82

B.8 Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets in MAXENT of scenario 2. In column “cmp”, 1 indicates the ascD1 dataset; 2 is the ascD2; 3 is the ascD3; 4 is the ascD4 and 5 is the ascD5 83

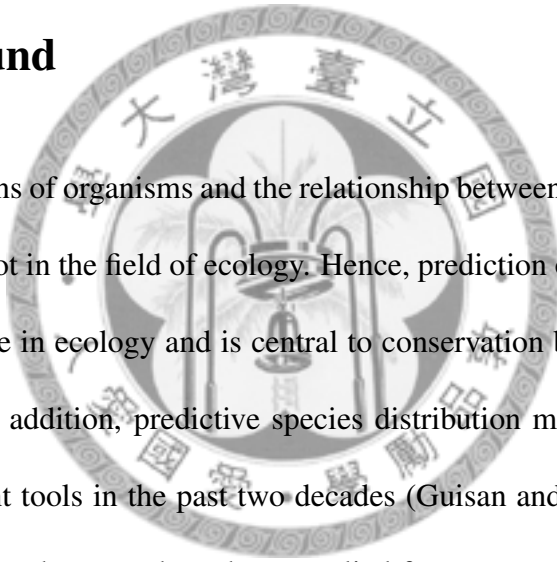




Chapter 1

Introduction

1.1 Background



The distribution patterns of organisms and the relationship between environmental factors and species is a hot spot in the field of ecology. Hence, prediction of species' distribution plays an important role in ecology and is central to conservation biology (Austin, 2007; Elith et al., 2006). In addition, predictive species distribution modellings (SDM) have becoming an important tools in the past two decades (Guisan and Zimmermann, 2000). A wide variety of research papers have been applied for conservation biology, biogeography and climate change research (Galparsoro et al., 2009; Pearson et al., 2007; Tanaka et al., 2006; Wilson et al., 2005; Matsui et al., 2004). Yet the various impact and threats to natural have also upraised the study of SDM for prediction of potential distribution in conservation management. Through the prediction of threaten species, the ecologist and conservationist could use this tool to build protection project against environmental changes and preserve the biodiversity.

Species distribution modelling is based on the species-environment relationship which

proposes the environmental variables associated with the distribution of a given species. The SDM based on the species-environment relationship is used to predict the potential distribution of a given species without obtaining the complete field data. Therefore, the SDM has the ability to predict the potential distribution of a given species with partial field data and can be used for conservation of rare species or the inaccessibility area.

A wide variety statistical-based and procedure-based techniques on species distribution modelling including generalized linear models, generalized additive models, boosting regression trees, neural networks, maximum entropy principles, genetic algorithms for rule-sets production, etc (Kriegler, 2007; Heglund, 2003; Hastie and Tibshirani, 1990).

1.2 Purpose of Research

Song et al. (2007) have compared three different models applied in *Tsuga chinensis* var. *formosensis* with two datasets. The vegetation of first dataset is dominant by *Tsuga chinensis* var. *formosensis* and the other is the habitat where *Tsuga chinensis* var. *formosensis* is present. The overall AUC values are higher in first dataset than second dataset. Their study implies that the SDM performance of higher dominance datasets would be better than lower ones. Therefore, this study tries to build different possible scenarios to find if the dominance would be an possible influential factors to species distribution models and the higher dominance data would increase the SDM performance.

Before trying to discuss the dominance effects, the concept of dominance in vegetation ecology should be declared. Vegetation ecologists usually use phytosociological parame-

ters to indicate the quantitative object to find the correlation amongst the phytosociology and environment. These phytosociological parameters which include density, frequency and dominance play an role to represent the characters of phytosociology. General speaking, “dominance” describes an abstract concept of habitat adaptability in phytosociology. In other words, high dominance plants can control the most resources in the habitat and have large coverage or quantity (Liu and Su, 1983). In order to quantify the abstract concept, dominance could be calculated by biomass, volumes and other methods. In this study, logarithm basal area and importance value index are both used in data partition for evaluating the dominance effects in species distribution modelling. The main subject in this study is trying to find if the dominance would be an influential factors of SDM and if the dominance effect exists, does higher dominance in sites increase the SDM performance.

This thesis is structured as follows: chapter 2 consists the overview of species distribution models and the concept of dominance. Modelling techniques include maximum entropy principles and generalized additive models. Partitioning datasets according to different dominance measures and detailed evaluating procedures are in chapter 3 and chapter 4 consists the boxplot results and significance tests of our testing procedures. In chapter 5, we will discuss the dominance effects in sampling data on species distribution models. We conclude the final remarks in chapter 6, and appendix chapter consists the demonstration of computational code with supplemental explanation.

Chapter 2

Literature Review

This chapter has five parts: the first section will discuss dominance in ecological meanings and its evaluation methods; the second section reviews two species distribution models; the third and fourth parts review the recent comparison of different species distribution models and the possible influential factors to SDM; and the last section discuss the evaluation methods of model performance.

2.1 Dominance in Analytic Concepts

The object in phytosociology analysis can be divided into two methods according to survey: (1) analytic method, (2) synthetic method (Daubenmire, 1968; Cain and de Oliveira Castro, 1959). Analytic method is to choose a representative stand to set up relevé for inventory. In other words, analytic method regards a plant community as a representative stand (Liu and Su, 1983). In contrary, synthetic method is based on the analytic method and extends a stand to different stands. Based on our study subject, only analytic method will be introduced for explanation.

2.1.1 Analytic Method

The phytosociological analytic characters describing the plant community are major divided into quantitative and qualitative characters (Braun-Blanquet, 1932). Qualitative characters include stratification, vitality and periodicity and quantitative characters include abundance, density, dominance, gregariousness and frequency (Braun-Blanquet, 1932). The purpose of quantitative characters is to find the importance, number of individuals and the extent of dominant species for indicating the correlation amongst plant communities and environmental factors (Liu and Su, 1983). Quantitative characters are introduced for elucidating the importance value index (IVI):

Abundance and density

Abundance usually describes the quantity (number of individuals) of species in phytosociology. Braun-Blanquet (1932) defines the abundance as five classes: (1) very rare, (2) rare, (3) infrequent, (4) abundant, (5) very abundant. Abundance is a very subjective and has arbitrary limitations due to its connotation to estimated number (Cain and de Oliveira Castro, 1959). Hence, density is objectively to express the actual abundance or the number of individuals of each species. Density is also intended to imply the dynamic trends of each species, for example, higher density of saplings usually indicates that the species will gradually become the dominant in the future (Liu and Su, 1983).

However, it exists two limitations of density (Daubenmire, 1968):

1. Density can not be applied in plant communities contain vegetative reproduced species, prostrate plants (grasses, shrub branches). For example, rhizomatous species such as *Yushania niitakayamensis* (Poaceae). It can not count exactly number of in-

2. Literature Review

dividuals and is not satisfactory as a basis for comparing different species.

2. Maturity of one species may have different growing conditions. Therefore, it provides very little biologic information of two individuals of a given species per meter.

Another measure of density is relative density (RDe) which indicates the proportion of species in a plant community and is intended to compare the proportion in different plant community or stands (Liu and Su, 1983). RDe of a species (SP) is calculated as:

$$RDe = \frac{\text{Density of SP}}{\text{Total density}} \quad (2.1)$$

Frequency

Frequency provides the uniformity and regularity of the distribution throughout a plant community (Cain and de Oliveira Castro, 1959) without indicating how many or how much (Daubenmire, 1968). Frequency is the problem of pattern which defined as the percentage of occurrence of a species amongst different relevé in a stand. It is intended to express the evenness of plants species and indicate the homogeneity in a plant community. Raunkiaer (1934) developed a method to calculate the frequencies of species in five classes (Raunkiaer's law of frequency) as follows:

Class A 0 to 20%

Class B 21 to 40%

Class C 41 to 60%

Class D 61 to 80%

Class E 81 to 100%

Raunkiaer's law of frequency is affected by relevé size and debated by many phytosociologist (Liu and Su, 1983) but it is a simple method to express the homogeneity of a stand. However, relative frequency (RF) of a species (SP) in a stand is as:

$$RF = \frac{\text{Frequency of SP} \times 100}{\text{Summation of total species frequency}} \quad (2.2)$$

Dominance

Dominance is in terms of the extent of a plant community control or occupancies in an area (Liu and Su, 1983; Cain and de Oliveira Castro, 1959). The concept of dominance is trying to indicate the prevalence or adaptability of an organism and the dominance can be represented as area (coverage), volume or biomass. Because it is difficult to measure the biomass or volume, coverage is used to represent the dominance. The coverage means the projection of the canopy or leaves and branches which usually uses basal area for indication. The coverage has different measurement systems such as Braun-Blanquet, Hult-Sernander and Lagerberg-Raunkiaer system (Cain and de Oliveira Castro, 1959). Braun-Blanquet (1932) system is commonly used in field survey of phytosociology and with six classes as in table 2.1 (Cain and de Oliveira Castro, 1959; Braun-Blanquet, 1932)

In forestry, the basal area calculation is based on cross section area at breast height (dba). The dominance has the following calculations (Liu and Su, 1983):
 quadrat dominance (QDo):

Table 2.1: Braun-Blanquet coverage classes and values

class	coverage percentage
x	< 1
1	1 – 5
2	6 – 25
3	26 – 50
4	51 – 75
5	76 – 100

$$QDo = \frac{\sum \text{dba in all quadrat}}{\text{number of quadrat}} \quad (2.3)$$

relative dominance:

$$RDo = \frac{\text{dominance(coverage) of a species}}{\text{dominance(coverage) of all species}} \quad (2.4)$$

2.1.2 Combination of Analytic Characters

In order to represent the quantity in plant communities, single analytic characters are combined to express importance. For example, density-frequency-dominance index (DFD) and importance value index (IVI) are both used for indicating the importance of species in a plant community.

Importance value index

DFD is developed by Curtis (1947) and IVI (Curtis and McIntosh, 1951) is based on DFD but uses relative value for indication. IVI is defined the summation of relative density, relative frequency and relative dominance. It is applied for evaluating the dominant species

in a stand.

In this study, dominance is regarded as ecological dominance which indicates the adaptability of species. Single measure (basal area) and weighted measure (IVI) are both used to represent the dominance for evaluating the performance of species modelling distribution.

2.2 Species Distribution Models

There are many modelling techniques having applied in predicting potential species distribution. The modelling approaches can be generally classified as method-driven and statistical-driven. The method-driven or mechanistic techniques try to make the potential species distribution via the method itself, such as machine learning methods. Particularly, machine learning techniques emphasize on the features of method, for example, the genetic algorithms simulate the mutation of genes and employed in several modelling techniques such as genetic algorithms for ruleset production (GARP). The recent mechanistic methods applied in prediction of species distribution are GARP, MAXENT and neural networks (Phillips et al., 2006). In contrary, statistical methods focus on statistical techniques to predict the potential distribution and most of the statistical approaches are regression-based. Generalized linear models, generalized additive models, multivariate adaptive regression splines (MARS) and boosted regression trees are most recent tech-

niques applied in species distribution modelling (Austin, 2007; Guisan et al., 2007b; Leathwick et al., 2006). In this study, both statistical and mechanistic approaches were used for evaluation of species distribution modelling.

2.2.1 Generalized Additive Models

The concept of generalized additive models (GAMs) are derived from generalized linear models (GLMs) and GLMs are derived from linear models.

Essentially, generalized additive models (GAMs) (Hastie and Tibshirani, 1990) are extensions of regression-based models and follows from additive models, as generalized linear models (GLMs) follow from linear models (Wood, 2006). GAMs allow variables which do not normally distribute and their linear predictor contains a sum of smoothing functions of covariates which uses penalized regression methods to estimate penalized regression splines (Tsao, 2007; Wood, 2006).

Generalized linear models

A GLM is combined systematic and random components with the response variables other than normal distributions (Nelder and Wedderburn, 1972). The characters of GLMs include:

1. Exponential family

A general function of exponential family in distribution of Y is as equation 2.5 (Faraway, 2006):

$$f(y|\theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\theta)} + c(y, \phi) \right] \quad (2.5)$$

where the θ is the canonical parameter which indicates the location and ϕ is the dispersion parameter. The response variable in GLMs is a member of the exponential family distribution. In equation 2.5, a , b and c functions could be specified normal, Poisson, Binomial, Gamma and inverse Gaussian distribution (Faraway, 2006).

2. Link function

A Link function describes the relationship between the expected value Y and the linear predictors. The general form of a link function η is shown as equation 2.6.

$$\eta = g(\mu_i) = g(\mathbb{E}(Y_i)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \mathbf{X}_i \boldsymbol{\beta} \quad (2.6)$$

where μ_i is expected value of Y_i , and

The functions and canonical links of some exponential family distributions in GLMs are as table 2.2.

Generalized additive models

A GAM is a generalized linear model whilst the linear predictors contain non-parametric function and comprise a sum of smooth functions of covariates (Wood, 2006). GAMs use a series of smooth functions (e.g. cubic splines, P-splines, etc.) to avoid detailed parametric relationships, hence the GAMs are more flexible and convenient than GLMs

2. Literature Review

Table 2.2: Some exponential family distributions.

Family	Function(f(y))	Link(θ)	Variance(V(μ))
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\eta = \mu$	1
Poisson	$\frac{\mu^y \exp(-\mu)}{y!}$	$\eta = \log \mu$	μ
Binomial	$\binom{n}{y} \left(\frac{\mu}{n}\right)^y \left(1 - \frac{\mu}{n}\right)^{n-y}$	$\eta = \log(\mu/(1 - \mu))$	$\mu(1 - \mu/n)$
Gamma	$\frac{1}{\Gamma_\nu} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} \exp\left(-\frac{\nu y}{\mu}\right)$	$\eta = \mu^{-1}$	μ^2
Inverse Gaussian	$\sqrt{\frac{\gamma}{2\pi y^3}} \exp\left[\frac{-\gamma(y\mu)^2}{2\mu^2 y}\right]$	$\eta = \mu^{-2}$	μ^3

Note: The table is revised from Wood, 2006.

(Wood, 2006). A general GAM model is as following equation

2.7 (Wood, 2006) :

$$g(\mathbb{E}(Y_i)) = \mathbf{X}_i^* \boldsymbol{\theta} + f_i(X_i) = \mathbf{X}_i^* \boldsymbol{\theta} + f_i(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \quad (2.7)$$

where $f_i(X_i)$ contains a series of smooth functions.

Implementation of generalized additive models

There are some implementation of generalized additive models, such as `mgcv`, `gam`, `gss`, `gamlss` and `vgam` packages in R. Following example shows the `gam` in `mgcv` package in R:

```
> ptsuga <- gam(tsuga ~ s(slope) + s(wetness) + s(wi),
  family=binomial(link=log), data=tsuga_data)
> ptsuga

Family: binomial
Link function: log
```

Formula:

```
tsuga ~ s(slope) + s(wetness) + s(wi)
```

Estimated degrees of freedom:

```
1 1 2.292953 total = 5.292953
```

```
UBRE score: -0.6768139
```

where the original data is `tsuga_data` with column `tsuga` (presence/absence), environmental variables: `slope`, `wetness` and `warmth index (wi)`. The distribution family is binomial and link function is `log`. The total degree of freedom is 5.292953 and both `slope` and `wetness` contribute 1 *df*, `wi` contributes 2.292953 *df*. UBRE score is -0.6768139.

Generalized regression analysis and spatial prediction

Lehmann et al. (2002) implemented generalized regression analysis and spatial prediction (GRASP) using statistical models such as GLMs and GAMs. The GRASP function can predict the species abundance/presence (response variables) with spatial coverages of environmental variables (predictors).

2.2.2 Maximum Entropy Principles

Maximum entropy (maxent) originally introduced by Jaynes (1957) in statistical mechanics. Maxent is a general-purpose method for inferences from incomplete information in the field of information theory and applied for astronomy, statistical physics, image reconstruction and signal processing (Phillips et al., 2006). In other words, maxent provides

a constructive criterion for setting up probability distributions on the partial knowledge (Jaynes, 1957). Phillips et al. (2006) introduced maximum entropy as a method for presence only modelling of species distribution. The concept of maxent is trying to estimate an unknown target probability distribution through finding the best probability distribution of maximum entropy under a set of constraints (Phillips et al., 2006). For example, if we want to know about the distribution of velocity in the gas at a given temperature, we can find the maximum entropy distribution under the temperature constraint. The maximum entropy distribution formula is as equation 2.8

Maximum entropy distribution

Let $f^*(x) = f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x)}$, $x \in S$, where $\lambda_0, \lambda_1, \dots, \lambda_m$ are chosen so that all probability densities f^* satisfying the following

$$\int_S f(x) r_i(x) = \alpha_i \quad \text{for } 1 \leq i \leq m \quad (2.8)$$

Then f^* uniquely maximizes entropy $h(f)$ over all f satisfying these constraints (Cover and Thomas, 2006). When maxent applied in species distribution modelling, study area is regarded as a set of pixels to make up the space on which the maximum entropy probability distribution is defined. The pixels of species occurrence records constitute the sample points and the features are environmental variables, such as climatic or topographical variables (Phillips et al., 2006). Figure 2.1 illustrates the concept of maximum entropy applied in prediction in species distribution. Green area is the study area which would constitute the maxent probability distribution. Each cell (grid) indicates the pixel which contains a set of environmental variables and the red dots are the species occurrence records.

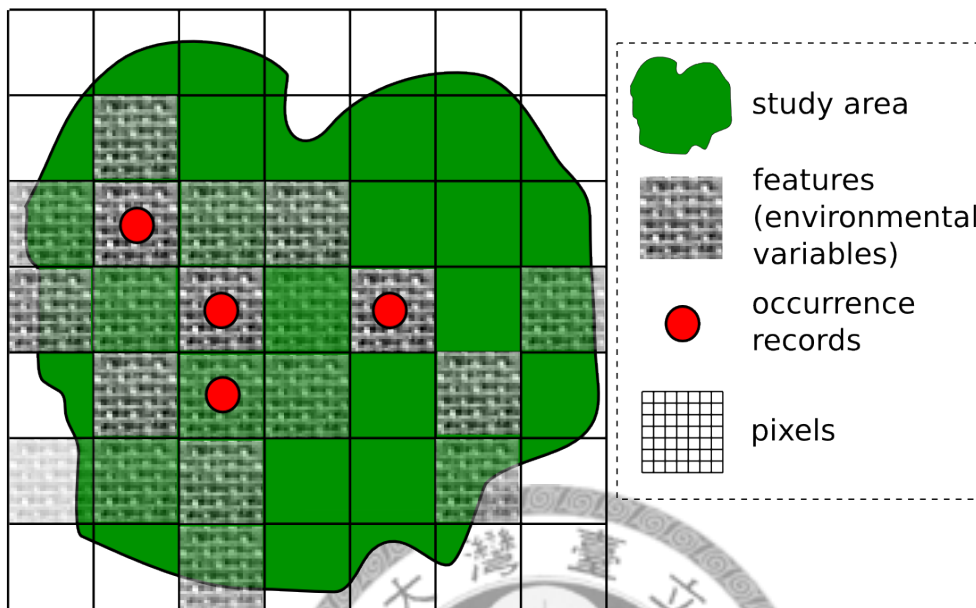


Figure 2.1: Concept of maximum entropy applied in prediction of species distribution.

The implementation of principle of maximum entropy applied in species habitat modelling is MAXENT software (Phillips et al., 2004). The MAXENT software also implements six features: linear, quadratic, product, threshold, hinge and category indicator (Phillips and Dudík, 2008; Phillips et al., 2006). Environmental variables (as known as features) in MAXENT are continuous and categorical. In contrary to GAMs, MAXENT can use presence only occurrence records rather than presence and absence records.

2.3 Comparison of Different Models

Comparison of different model is a critical issue when new approaches introduced and applied in prediction of species distribution modelling. No single method is applied for universal prediction, but multiple comparisons amongst modelling techniques, species characteristics and evaluation methods are necessary to find a best method in different conditions.

Segurado and Araújo (Segurado and Araújo, 2004) assess nine modelling techniques and sixteen environmental variables using 9939 occurrence records for 44 species of reptiles and amphibians in Portugal. Their study represented species with low marginality and high ecological tolerance had lower overall performances (Segurado and Araújo, 2004). For the most part performances were better with non-parametric techniques and neural networks using NNET library (NNETW) had the highest model performance whilst DOMAIN and BIOMAP showed the lowest performance. Their study also mentioned data quality was strongly influential to model performances and suggested two choices to model the species distribution. The first one is to use expert systems such as GARP or the second choice is to use a single robust technique like NNETW.

Elith et al. (2006) made a comprehensive comparisons amongst 16 modelling techniques and 226 species. Boosted regression trees (BRT), generalized dissimilarity modelling (GDM) and MAXENT have the best performance, followed by multivariate adaptive regression splines (MARS), GLM, GAM, openModeller GARP (OM-GARP) and the poorest ones are desktopGARP, BIOCLIM and multivariate distance model (DOMAIN).

The evaluation is based on three methods: the area under the Receiver Operating Characteristic curve (AUC), correlation and Kappa statistic (Elith et al., 2006).

Meynard and Quinn (2007) use artificial species to make comparison of the most common statistical models. Eighteen artificial species are generated from three environmental gradients for evaluating four models (GAM, GLM, classification trees and GARP). Their results recommend to use GAM or GLM over classification trees or GARP because the later two methods perform poorly and tend to over-predict the area of occupancy especially in low prevalence. The model varies in model performance for low prevalence species and their results suggest the performance can be improved through targeted sampling (Meynard and Quinn, 2007).

Guisan et al. (2007b) have also evaluated a comprehensive study discussing influential factors of species distribution followed up by Elith et al. (2006). They focus on 31 native tree species in Switzerland and compare 10 modelling techniques in terms of map resolution, predictive power and sensitivity to location error and sample sizes, and try to elucidate variation in model performance (Guisan et al., 2007b). The ranking of overall model performance based on AUC is: MAXENT, GAM, MARS, GDM single species (GDMSS), BRUTO, GLM, OM-GARP, DOMAIN and BIOCLIM. The generalized linear mixed models (GLMM) analyses results of modelling techniques can be divided into three groups, BRT and MAXENT ranked first, regression-based techniques (GAM, MARS, BRUTO and GLM) ranked second and profile-based techniques (BIOCLIM, DOMAIN) ranked last.

2.4 Influential Factors to SDM

There have been numerous studies in the literature dealing with the influential factors of species distribution modelling. The influential factors can be the data characteristics or the species traits.

2.4.1 Species Traits

Araújo and Williams (2000) extrapolate widespread species had higher sensitivity and lower specificity. In contrary, restricted-range species had lower sensitivity and higher specificity. Rare species usually ease to predict because they are sensitive to certain environmental variables and widespread species are not easy to predict due to unclear species-environment relationships.

2.4.2 Data Characteristics

Data characteristic would be also an important factor to prediction of species distribution models.

Hernandez et al. (2006) proposed the effect of sample size and species characteristics on species distribution modelling performances. Their study has compared six different sample sizes and evaluated four model techniques whilst the model accuracy increased with larger sample sizes for all modelling methods (Hernandez et al., 2006). However, they confirm ecological low tolerance species are easier to model than widespread species. The result also indicates multiple evaluation is necessary to examine the accuracy of mod-

els with presence-only data.

Sample selection can be also an influential factor to SDM. Reddy and Davalos (2003) assess the patterns of species richness and find the intensity of collecting have been heavily influenced by human accessibility. If the sample plots is collected around cities, road sides, rivers, etc., it may have significant sampling bias. In addition, most sampling data in herbarium are collected from accessible area, especially aggregation on certain regions or along the route and it may cause the intensity of sampling bias on species distribution models.

2.5 Model Performance and Evaluation

The model accuracy in presence/absence prediction can be considered as four possible conditions: true positive/negative and false positive/negative. True positive implies that the species occurs and our prediction is true. False positive implies that the species does not exist, but the models prediction is incorrect. In contrary, true negative means model prediction is correct and the species does not exist. False negative means that model prediction is wrong and the species does not exist. However, many model measurements are derived from true positive/negative and false positive/negative such as sensitivity and specificity (Fielding, 1999). An error matrix or confusion matrix as table 2.3 summarizes the model accuracy and provides an effective way to represent the commission errors (the errors of inclusion, i.e. false positive) and omission errors (the errors of exclusion, i.e. false negative) in an overview (Congalton and Green, 1999). In addition, more measure-

2. Literature Review

ments based on a 2×2 confusion matrix are listed in table 2.4 (Fielding, 1999).

Table 2.3: A 2×2 confusion matrix

		Observed	
		Presence	Absence
Predicted	Presence	a	b
	Absence	c	d

Note: a means true positive, b means false positive c means false negative, d means true negative. a and d are correct results; c and d are incorrect prediction (revised from Fielding, 1999).

Table 2.4: Measurements derived from a 2×2 confusion matrix

Measurement Index	Calculation or Description
Sensitivity	$a/(a + c)$
Specificity	$d/(b + d)$
Positive predictive power	$a/(a + b)$
Negative predictive power	$d/(c + d)$
False positive rate	$b/(b + d)$
Odds ratio	ad/cb
Kappa statistic	$\frac{(a+d) - (((a+c)(a+b) + (b+d)(c+d)) / (a+b+c+d))}{(a+b+c+d) - (((a+c)(a+b) + (b+d)(c+d)) / (a+b+c+d))}$
Receiver operating characteristics (ROC)	In a ROC curve, the vertical axis indicates the sensitivity and indicates the 1-specificity.
Area under ROC curve	The area under the ROC curve is AUC.

Note: revised from Fielding, 1999.

Sensitivity is true positive over the actual presence which indicates the true positive rates. Specificity indicates the true negative rates. Odds ratio means the ratio amongst correct prediction and incorrect prediction. Receiver operating characteristic (ROC) curve and area under ROC (AUC) is widely used for recent species distribution model performance assessment (Elith et al., 2006; Hernandez et al., 2006; Guisan et al., 2007b,a). The value of AUC is between 0.5 to 1, In Swet's (1988) research, if the AUC values is be-

tween 0.5 to 0.7, the discrimination of models is regarded as low. If the values of AUC fall between 0.7 to 0.9, it implies the prediction performance is good. When the AUC values larger than 0.9, the performance result is excellent whilst a score of AUC is 0.5 implies random predictive discrimination. Figure 2.2 shows typical ROC curves, Predicted1 indicates the model is exactly as the reality (AUC=1); Predicted2 has very good model performance and the performance of Predicted3 is good.

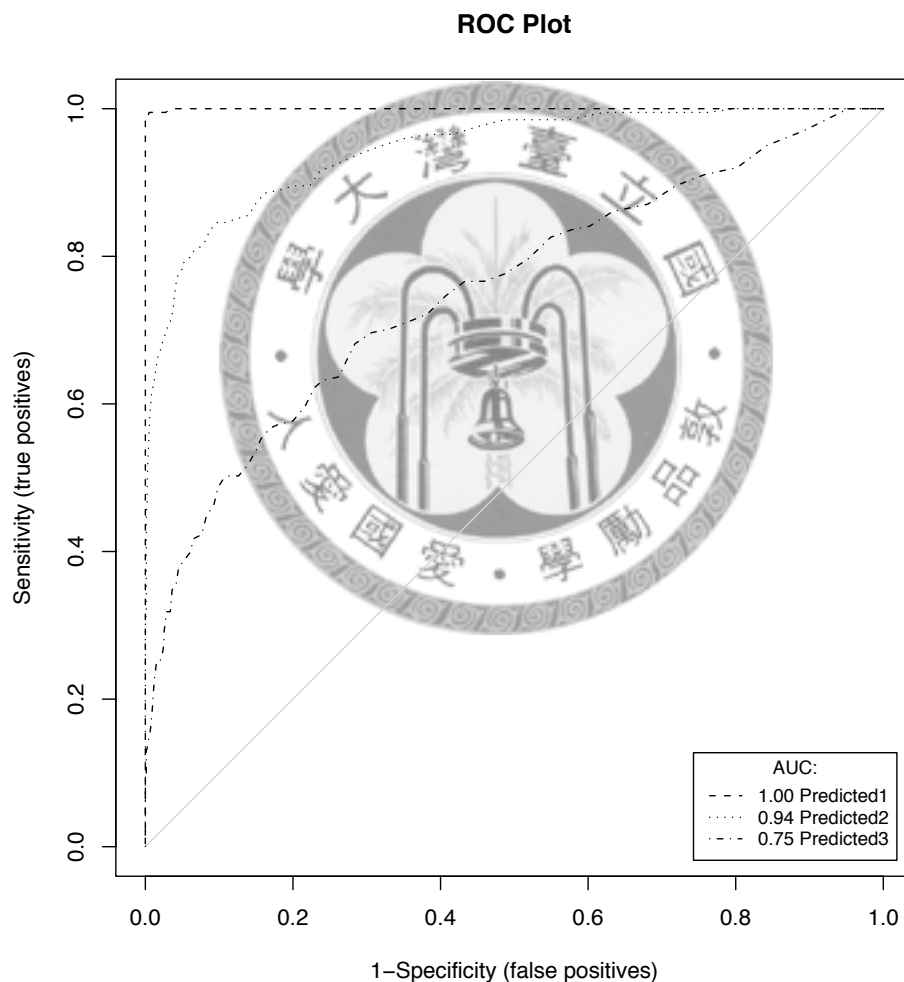


Figure 2.2: Sample ROC curves from R SIM3DATA

A Kappa statistic is often used to assess improvement over chance (Fielding, 1999).

2. Literature Review

The Kappa statistic value smaller than 0.4 indicates the poor agreement; a Kappa value between 0.4 and 0.75 is good (Landis and Koch, 1977).

Although sensitivity and specificity are good error indicators to model performance, they are not robust and reliable to evaluate the species distribution models. Prevalence (occurrence records over total sample plots in all relev ) is an influential factor to both sensitivity and specificity in which Manel et al. (2001) reported higher prevalence will increase the value of sensitivity and decrease the specificity. In other words, if we try to predict a rare species in a wide area (low prevalence), the sensitivity would be low and specificity would be relatively high. However, both sensitivity and specificity could be used for evaluating if the model is over-prediction or under-prediction. Over-prediction means that the model prediction shows the species exist but the species does not exist in observed data. In contrary, under-prediction indicates the species does not exist in model prediction but exists in real world.

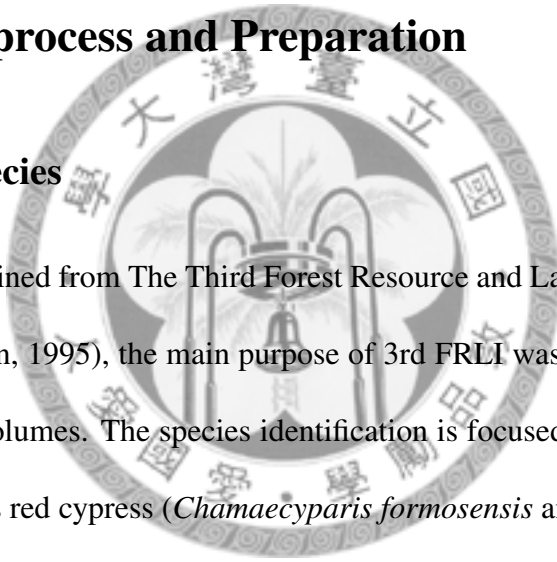
The predictive power of Kappa is better than sensitivity and specificity due to the low affected by prevalence (Manel et al., 2001). Moreover, ROC analysis and AUC value are also more better than sensitivity and specificity.

Chapter 3

Material and Methods

3.1 Data Preprocess and Preparation

3.1.1 Target Species



The raw data was obtained from The Third Forest Resource and Land-Use Inventory (3rd FRLI) (Guan and Chen, 1995), the main purpose of 3rd FRLI was to investigate the land-use type and stand volumes. The species identification is focused on the artificial trees, industrial trees such as red cypress (*Chamaecyparis formosensis* and *Chamaecyparis obtusa* var. *formosana* (Cupressaceae)).

There are two major backwards in 3rd FRLI data:

1. Species identification:

As mentioned above, the main purpose of 3rd FRLI is focused on use of woods. The trees without industrial use would be neglected or combined as a set, for example:

- The species other than *Machilus thunbergii* (Lauraceae), *Machilus zuihoensis* (Lauraceae) and *Machilus japonica* var. *kusanoi* (Lauraceae) would be

3. Material and Methods

regarded as *Machilus*.

- Most species in Fagaceae would be regarded as *Castanopsis* or *Pasania*.
- Some species like *Euonymus laxiflorus* (Celastraceae), *Bridelia balansae* (Euphorbiaceae) and *Prunus phaeostica* (Rosaceae) would be regarded as “other woody species” because they do not have any economical benefits.

2. Insufficient of relevé:

The usable occurrence data does not sufficient. For example, there is only 45 records of *Abies kawakamii* (Pinaceae), 62 records of *Picea morrisonicola* (Pinaceae).

However, the target species selection should be considered to meet the following prerequisites:

1. Sufficient occurrence data for data partition:

The datasets are partitioned by different relative dominance criteria so the occurrence records should as more as possible. We assume occurrence records of each dataset are about 40 plots and the total number of records is more than 200 plots.

2. Less disturbance from human beings:

It is reported occurrence records are often biased toward human population centres and roads (Reddy and Davalos, 2003). Therefore, sample selection should avoid the human impaction areas especially the side of roads.

Although there are 235 occurrence records of *Chamaecyparis formosensis*, it has been severe exploitation in cutting *Chamaecyparis formosensis* and *Chamaecyparis obtusa* var. *formosana* during the past 50 years.

Thus, removal of the reasons of insufficient occurrence records and human disturbance, *Tsuga chinensis* var. *formosensis* is selected for target species for the modelling framework.

Tsuga chinensis var. *formosensis* is the dominant species in the *Tsuga* belt (Su, 1984) or *Tsuga-Abies* belt and *Tsuga-Chamaecyparis* belt (Lin, 2009) which distributed from 1400m to 3400m in altitude (Chiou et al., 2006; Chen, 2004), such as Central Mountain Range, HsuehHsan Range, Yushan Range and DaWu Mountain, etc. The optimal range is 2800m to 3000m in altitude. *Tsuga chinensis* var. *formosensis* is often mixed with *Chamaecyparis obtusa* var. *formosana* (Cupressaceae), *Chamaecyparis formosensis* (Cupressaceae), *Trochodendron aralioides* (Trochodendraceae), *Pinus armandii* var. *masteriana* (Pinaceae) and *Pinus taiwanensis* (Pinaceae) in lower altitude. In higher altitude, *Tsuga chinensis* var. *formosensis* often becomes pure stands or mixed with *Picea morrisonicola* (Pinaceae) and *Abies kawakamii* (Pinaceae). It is reported *Tsuga chinensis* var. *formosensis* prefers the south-facing slopes (sun-facing side in north hemisphere), dry lands, cliffs or mountain ridges (Chen, 2004; ?).

3.1.2 Occurrence Data

The original data used 2-degree transverse Mercator projection grid and the local datum is the TWD67 projection. Raw data was obtained from 3rd FRLI administered by the Forestry Bureau, Council of Agriculture. Investigation of 3rd FRLI was systematic sampling by 3 km and the starting point coordinate is 302000, 277000. The total num-

3. Material and Methods

ber of relevé is 3996. There were three types of datasets in the database. The first dataset (DATASET1) contained the species data and related elevation, coverage, density, etc; the dataset 2 (DATASET2) was relevé data which included map number, relevé plot id, abscissa, ordinate and environmental variables. And the last dataset (DATASET3) was orthophotos that recorded the species occurrence data based on the aerial photographs. For the prerequisite of the modelling techniques, in addition to species presence data, absence data were required. However, due to the aims of forestry management, all of the datasets only comprised the species presence data. Instead of real absence data, modellings can still be conducted with pseudo-absence data (Phillips et al., 2009; Araújo and Guisan, 2006; Elith et al., 2006).

3.1.3 Environmental Variables

The implicit theory assumes potential species distribution is determined by physical environments (such as temperature, precipitation, altitude, etc.) (Austin, 2007) or biotic factors (competition and other biotic interactions). We compiled topographic and climatic layers of the Taiwan island. Climatic environmental predictors were included warmth index (WI) (Chiou et al., 2004), wetness index (WET), and solar radiation in each month (RD01 12). Topographic environmental predictors were aspect (ASP), altitude (ALT), sediment transport capacity index (STCI), openness of the forest stands (OPEN), plan curvature (PLAN), profile curvature (PROF), relative stream power (RSP), slope (SLP), tangential curvature (TANG) and sky view factor (SVF). The original environmental variables were 40 meters in resolution and obtained from the Laboratory of Resource Inves-

tigation and Analysis, School of Forestry and Resource Conservation, National Taiwan University. Jackknife analysis was used to select the most influential variables from environmental variables mentioned above. Detailed information and abbreviations of these environmental variables were shown in Table 3.1.

Table 3.1: Environmental variables used for modelling of species distribution

Code	Description	Variable	Unit	Processing Software type
ASP	Aspect	D	Degree	ArcGIS
ALT	Altitude	C	Meter	ArcGIS
WI	Warmth index	C	Celcius	ArcGIS
STCI	Sediment transport capacity index	C	Unitless	TAS
OPEN	Openness	C	Degree	SkyRatio
PLAN	Plan curvature	C	deg/m	TAS
PROF	Profile curvature	C	deg/m	TAS
RD01	Solar radiation of January	C	MJ/m ² /day	CLIRAD-SW
RD02	Solar radiation of February	C	MJ/m ² /day	CLIRAD-SW
RD03	Solar radiation of March	C	MJ/m ² /day	CLIRAD-SW
RD04	Solar radiation of April	C	MJ/m ² /day	CLIRAD-SW
RD05	Solar radiation of May	C	MJ/m ² /day	CLIRAD-SW
RD06	Solar radiation of June	C	MJ/m ² /day	CLIRAD-SW
RD07	Solar radiation of July	C	MJ/m ² /day	CLIRAD-SW
RD08	Solar radiation of August	C	MJ/m ² /day	CLIRAD-SW
RD11	Solar radiation of September	C	MJ/m ² /day	CLIRAD-SW
RD10	Solar radiation of October	C	MJ/m ² /day	CLIRAD-SW
RD11	Solar radiation of November	C	MJ/m ² /day	CLIRAD-SW
RD12	Solar radiation of December	C	MJ/m ² /day	CLIRAD-SW
RSP	Relative stream power	C	Unitless	TAS
SLP	Slope	D	Degree	ArcGIS
SVF	Sky view factor. From 0 to 1	C	unitless	
TANG	Tangential curvature	C	deg/m	TAS
WET	Wetness index	C	Unitless	TAS

Note: Variable type C means continuous and D means discrete.

3.1.4 Datasets Preparation

Data cleaning

In the field of knowledge discovery, or data mining, the process consists an iterative sequence to extract the knowledge from raw data (Han and Kamber, 2006). Preprocess of data is important because the raw data may contain incomplete, noisy and inconsistent data. The noise, incompleteness, and inconsistency in raw data may lead to misunderstandings to the real pattern and character during data analysis. Meanwhile, occurrence data of the 3rd FRLI contains a wide variety of landuse types, and some of them were farmland, bamboo forests, artificial forests and disturbed by human activities. Since the stands of these human disturbed landuse type may plant the trees which are not natural to the habitat, the results of modelling would be over or under prediction. Furthermore, if the prevalence of a species in a study area is too high or too low, the performance of sensitivity or specificity would approximate the ideal value. Therefore the data containing such landuse types have to be removed to avoid such data noise.

Data selection and partitioning

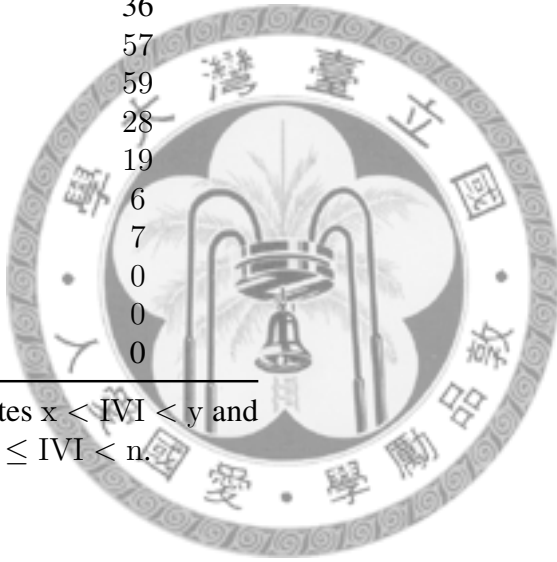
In order to comprehend how dominance may affect the performance of the species distribution models, raw occurrence data were divided into different datasets according the relative dominance of the focal species in the local communities to meet our aims. Therefore, the concept of dominance could be evaluated via two scenarios: scenario 1: important value index (IVI) (Curtis and McIntosh, 1951) and scenario 2: basal area (BA). The IVI has been originally calculated by relative density, relative frequency and relative

dominance (coverage), but due to the limitation of raw data we only used relative density and relative dominance. The datasets of IVI have been divided into nine sub datasets according to Gauch (1982)'s octave scale method. Table 3.2 shows the datasets and detailed sample plots of octave scale IVI datasets.

Table 3.2: Scenario 1: octave scale IVI datasets.

Datasets	IVI value	Number of plots
RDo9	[64, 100)	36
RDo8	[32, 64)	57
RDo7	[16, 32)	59
RDo6	[8, 16)	28
RDo5	[4, 8)	19
RDo4	[2, 4)	6
RDo3	[1, 2)	7
RDo2	(0.5, 1)	0
RDo1	(0, 0.5)	0
RDo0	0	0

Note: (x, y) indicates $x < IVI < y$ and $[m, n)$ indicates $m \leq IVI < n$.



To make sure the number of sample plots is enough for model evaluation, datasets were combined with following criteria: ascendant accumulative relative dominance datasets (ascRDo) and descendant accumulative dominance datasets (descRDo). In the ascRDo datasets, sample plots with low dominance were gradually removed. For example, dataset RDo3-8 contained RDo3, RDo4, RDo5, RDo6, RDo7 and RDo8; dataset RDo5-8 contained RDo5, RDo6, RDo7 and RDo8. In the descRDo datasets, sample plots with high dominance were gradually removed. For instance, dataset RDo7-3 contained RDo3, RDo4, RDo5, RDo6 and RDo7; dataset RDo6-3 contained RDo3, RDo4, RDo5 and

3. Material and Methods

RDo6. Table 3.3 shows the datasets and number of sampling plots.

Scenario 2 calculated the basal area from diameter at breast height (DBH) and took logarithm to the base ten. Processed data were also divided into ascendant and descendant accumulative relative dominance datasets. Considering the number of sampling plots in each dataset, datasets were divided into different criteria based on logarithm of basal area.

Table 3.4 shows the overall datasets and cut points.

Table 3.3: Scenario 1: ascendant and descendant accumulative datasets

Datasets ascRDo		Datasets descRDo	
Dataset	Number of sampling plots	Dataset	Number of sampling plots
RDo3-8	176	RDo8-3	176
RDo4-8	169	RDo7-3	119
RDo5-8	163	RDo6-3	60
RDo6-8	144	RDo5-3	32
RDo7-8	116	RDo4-3	13
RDo8	57		

Table 3.4: Scenario 2: dataset cut points and sampling plots quantity

Datasets ascRDo			Datasets descRDo		
Dataset	Cut points ($\log(BA)$)	Number of sampling plots	Dataset	Cut points ($\log(BA)$)	Number of sampling plots
ascD1	> 3.50	65	descD1	< 3.86	80
ascD2	> 3.59	53	descD2	< 3.77	73
ascD3	> 3.68	47	descD3	< 3.68	63
ascD4	> 3.77	37	descD4	< 3.59	57
ascD5	> 3.86	30	descD5	< 3.50	45

3.2 Model Building

Generalized linear models (GAMs) were performed with a package of R software, `mgcv` (Wood, 2006), and GRASP (generalized regression analysis and spatial prediction) combined spatial prediction and GAM analysis (Lehmann et al., 2002). GRASPER (GRASP in R) were used for model building and stepwise model selection family was quasibinomial; degree of freedom was four and one hundred modelling steps; spatial resolution was 1 km.

Before execution of the model building, relevé data were divided into training and evaluating datasets. In scenario 1, RDo9 dataset was used for evaluation dataset. Three types of datasets were needed in GRASPER for model building: response variables (Y), predictor variable (X) and data for prediction (Xpred). Response variables contained species names and its presence (zero for absence, one for presence); predictor variables included coordinates (abscissa and ordinate), environmental variables. In GAM, occurrence data should contain presence and absence data. Although we did not have “real absence” data, data not used for presence regard as absence. Xpred included those in predictor variable dataset but provided the data for prediction. Presence data was randomly selected from each datasets as in ascRDo and descRDo datasets and picked the same amount with presence data from absence database which was natural forest layers (NFL) in Taiwan island. In scenario 2, NFL was random separated into two parts: one for absence in training and the other for evaluation. In contrary to scenario 1, selection of presence data in scenario 2 would not delete the plots but regarded the plots as absence. Total plots in training datasets were fixed at 718. The training and evaluation datasets used

3. Material and Methods

for MAXENT model building were the same with those in GRASPER. The input datasets also contained species, coordinates, correspond environmental variables (in MAXENT program, it is named sample with data (SWD) format). All of the environmental predictors were converted into ascii format for model building in MAXENT program.

Prepared data would convert to the GRASP file input format and predicted the potential distribution maps. We tried to repeat one hundred times for each model building in both scenarios.



3.3 Model Evaluation

After finished the building process, results would be imported into GRASS GIS for pre-process of the model evaluation. Nearest neighbour method were used sampling the raster point value, and evaluation points were also imported using for extracting the raster values from generated results in model building part. Model performance could be evaluated with many criteria, such as alkaike information criteria (AIC), Bayesian information criteria (BIC), area under ROC curve (AUC), Kappa, etc. We used AUC for evaluating the model performances. AUC values were calculated by `PresenceAbsence` package in R. Figure 3.1 shows the overall experiment flowchart.

3.4 Implementation of Experiment

In this section, we would elucidate the detail information about the experiment implementation, such as how to set up the scenarios and step-by-step experiment flow. The design of experimental program was trying to modulate all of the components and making it easy to find possible experiment errors and any blind spot in our experiment. There were three principles in the program:

1. **Reliability:** Store intermediate files during execution of the program.

No matter what the program robustness, they would be minor or trivial errors during the execution. Hence, we tried to store everything after program started, each randomly generated datasets were stored separately in each directory by attribute like modelling methods and experiment repeat times. For example,

`Scenario1/ascD1/ascD183`

means that this is the Scenario 1, ascendant accumulative relative dominance dataset 1, the 83th repeat experiment. The relevant data will store following this method.

2. **Flexibility:** Portable and easy to modify.

Considering the convenience and portability, we used bash script program language which is easy to write and can be executed under Unix-like operating systems such as GNU/Linux, FreeBSD, MacOSX and this program can also work under windows in Cygwin environment. Although the efficiency of scripting language is slower than compiled C or C++ language, the code is easy for maintenance and editing. This is important for such try-and-error experiment in which frequent modification

is needed.

3. Modularity

There are numerous statistical and geographical information system (GIS) programs in the research field. Each program has its features and peculiar functions. It is not necessary to reinvent the wheels but peculiar functions in other programs are applied in the program on demands. R(R Development Core Team, 2008) program is employed for running the statistical regression models and evaluating the accuracy of modelling results; GRASS GIS (GRASS Development Team, 2006) as used for spatial processing; PostgreSQL (PostgreSQL Global Development Group, 2008) database management system was used for the preprocessing of relevé and datasets. Thus the program's component could be separately executed and combined by modulized functions and options. The strength of modularity is that if you had a problem during a certain step, it was not necessary to re-run the whole program but only re-run the component in specific step.

Following steps are the execution in detail:

1. Preliminary: Set up scenes and local variables

Before running the program, some dependency programs and configuration should be set up properly. GRASS GIS, java runtime environment, PostgreSQL database, R and required packages for R such as `PresenceAbsence`, `GRASPER` should be installed first. Raw occurrence data should be imported to PostgreSQL database

and environmental layers should be converted into ascii format. In our program (see Appendix I), line 32 to 39 indicates the local variables. DB means the database name; BASE means the base directory we would run the program and experiment; PGSQL indicates the binary name of PostgreSQL database; JAVA indicates the binary name of java program; MAXENT_JARFILE indicates the path of MAXENT jar file; MAXIMUM_MEMORY shows the maximum memory usage in MAXENT program; and ENVLAYERS_DIR indicates the directory where put the environmental layers.

2. Declare the functions

There are five functions in our program, PGSQL_QUERY(), SWD_PREPROCESS(), MAXENT_MODELLING() and GRASS_SAMPLE().

- PGSQL_QUERY()

The function PGSQL_QUERY() executes the database query to prepare the required datasets for model testing. The first part of PGSQL_QUERY() was to create datasets for model training and cross-evaluation. Datasets from table natural forest relevé (nf_releve) are selected randomly and separated in half. Half is for training and the other half is for cross-validation. Occurrence data in sample plots with human disturbed is removed to reduce the noise.

Since presence/absence data are necessary for GAM modelling, presence and absence data had been selected according to the dominance dataset cut point, in this case, we used SQL condition WHERE to choose the occurrence data.

3. Material and Methods

The statement is as below:

```
SELECT * FROM $TABLE WHERE ba ${CR} ${CUTPOINTS};
```

where $\{CR\}$ means greater or smaller; $\{CUTPOINTS\}$ indicates the cut points; and ba means basal area which logarithm is taken. For example, the SQL statement describing $ba > 3.58$ is written as:

```
SELECT * FROM $TRAINING_TABLE WHERE ba > 3.58;
```

After selecting presence data, the occurrence points which does not match the selecting criteria is regarded as absence. They will be also merged with absence data which is selected in `nf_releve` table as final absence data. For example, choosing $ba > 3.58$ would select only logarithm basal area greater than 3.58, and others would not be selected. SQL statement UNION is used to merge absence data:

```
(SELECT * FROM $TRAINING_TABLE WHERE tsuga=0) UNION  
(SELECT * FROM $DATA_DIDNOT_MATCH_CRITERIA
```

where `$DATA_DIDNOT_MATCH_CRITERIA` indicates the data which did not match the criteria.

- SWD_PREPROCESS ()

SWD_PREPROCESS () uses regular expression to modify the sample file to match the sample with data (SWD) format in MAXENT. (refer to Appendix I. line 100-107)

- MAXENT_MODELLING ()

This function MAXENT_MODELLING () iteratively executes the MAXENT program. (refer to Appendix I. line 109-112).

- GRASS_SAMPLE ()

GRASS_SAMPLE () function will extract raster values from given evaluation points in GRASS GIS. There are four steps in this function:

- (a) Using `r.in.ascii` to import ascii file into grass raster format.
- (b) Using `v.in.ascii` to import evaluation xy coordinate calculated from the preprocess step in `PGSQL_QUERY ()` function.
- (c) Extracting raster value from given evaluation points using `v.sample`, and the extraction method is nearest neighbour.
- (d) Exporting the sample output to comma separate values (CSV) file.

- EVALUATE ()

EVALUATE () function calculates and output results with `PresenceAbsence` package in R.

3. Main function

3. Material and Methods

-p — Preprocess option

This option enable the preprocess execution: data cleaning, data integration and partition relative dominance datasets

-gam — GAM model building option

Execute the GAM model building.

-maxent — MAXENT option

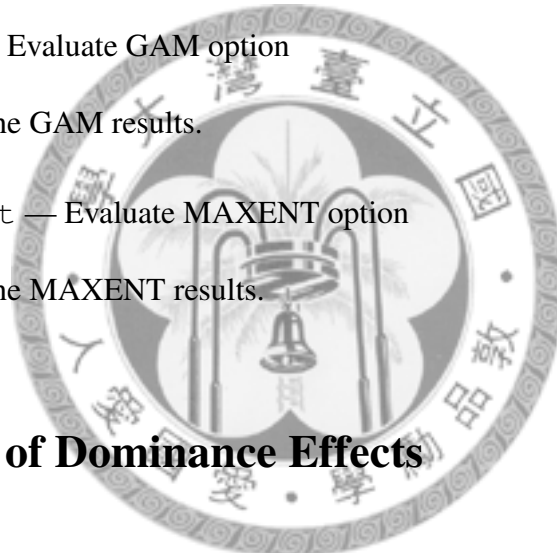
Execute the MAXENT model building.

-vgam — Evaluate GAM option

Evaluate the GAM results.

-vmaxent — Evaluate MAXENT option

Evaluate the MAXENT results.



3.5 Analyses of Dominance Effects

To assess the effects of dominance, we tried to visualize the experimental results by box-plots comparisons and used non-parametric multiple comparison with `npmc` package in

R.

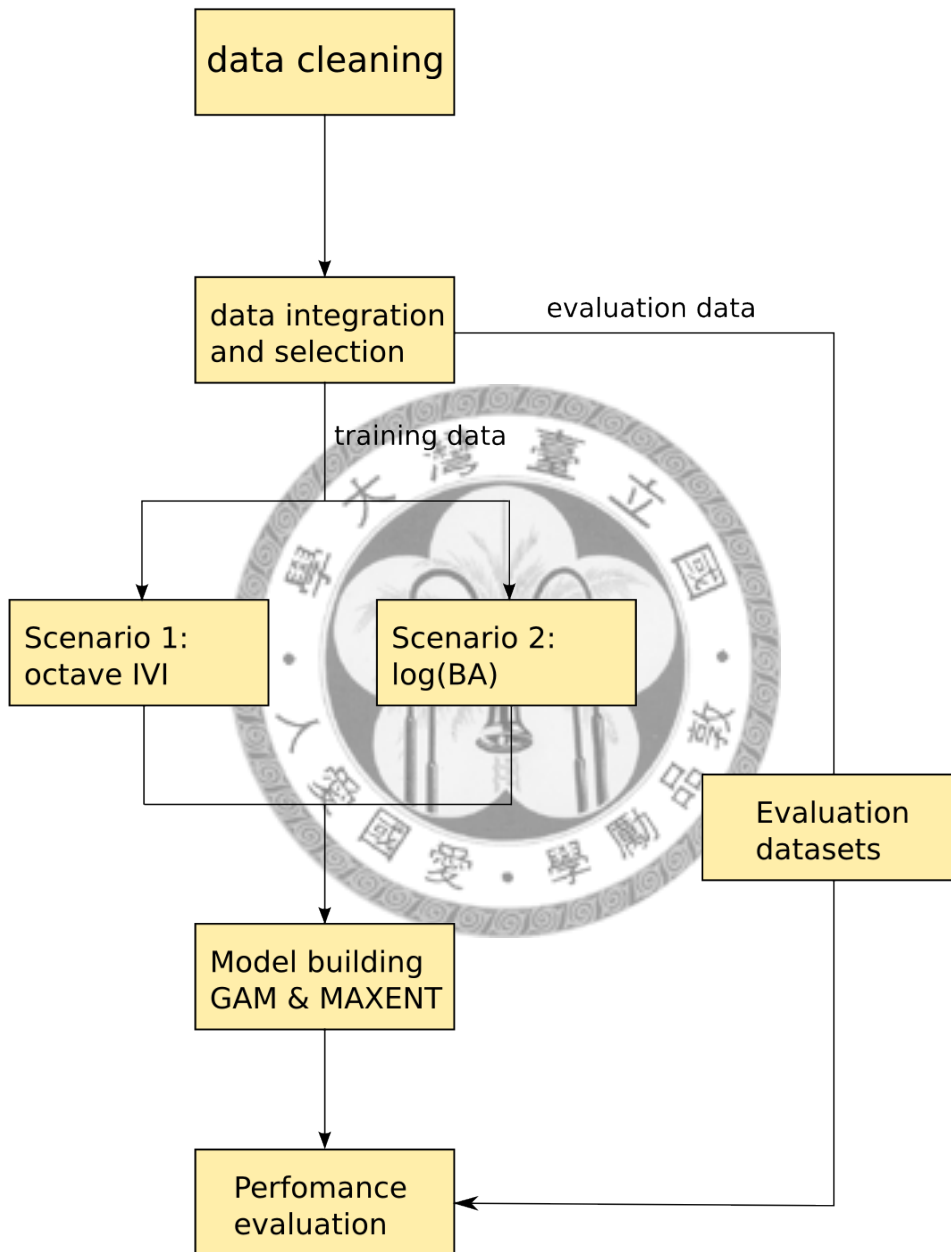


Figure 3.1: Overall experiment flowchart

Chapter 4

Results

The result of Jackknife analysis is as Figure 4.1. The most influential top five environmental variables are warmth index (WI), altitude, slope, wetness index (WET) and sediment transport capacity index (STCI). Considering the multi-collinearity presence, the top five environmental variables have been made multi-collinearity test. The result is as table 4.1. In table 4.1, warmth index has high correlation with altitude and the covariance is -0.9939; sediment transport capacity index has also high correlations with slope and wetness index. Therefore, considering the multivariate correlations, warmth index, slope and wetness index were used for the predictor variables.

Table 4.1: Multivariate correlation matrix

	WET	ALT	SLP	STCI	WI
WET	1.0000	0.0700	-0.6570	-0.5621	-0.0588
ALT	0.0700	1.0000	0.0792	-0.2071	-0.9939
SLP	-0.6570	0.0792	1.0000	0.7821	-0.1206
STCI	-0.5621	-0.2071	0.7821	1.0000	0.1570
WI	-0.0588	-0.9939	-0.1206	0.1570	1.0000

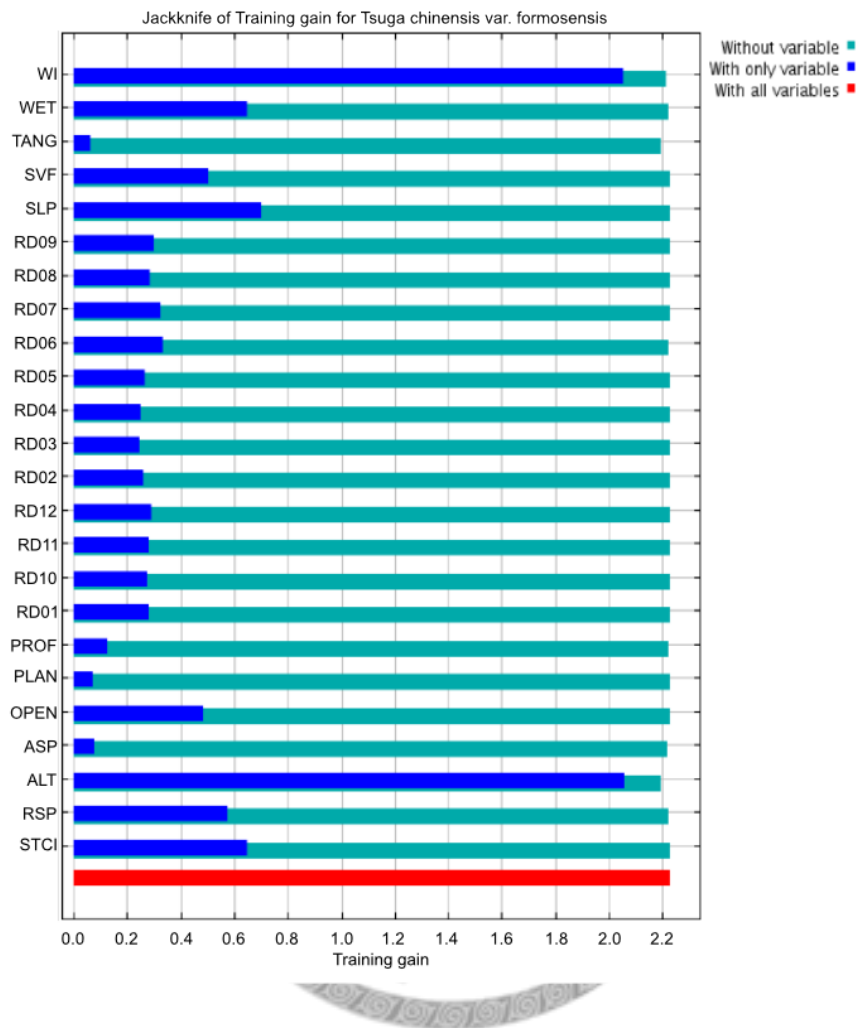


Figure 4.1: Jackknife analysis of training gains of *Tsuga chinensis* var. *formosensis*

4.1 Experimental Test

Figure 4.2 and 4.3 showed the scenario 1 results; figure 4.4 and 4.5 showed the results of scenario 2. From the boxplot of ascendant accumulative RDo datasets comparisons in scenario 1 and the results of multiple Behren-Fisher-Test, we could find that there was no significant difference between the datasets. In contrary, result of the descendant accu-

4. Results

ulative RDo datasets appeals to a slightly trend amongst the datasets. If we gradually removed the higher dominance dataset, the AUC value decreased. Clearly, in the first three datasets (RDo3-8, RDo3-7, RDo3-6) appealed very similar. Their maximum values, minimum values, lower quartile, upper quartile and median were differ not too much. But the AUC values amongst RDo3-6, RDo3-5 and RDo3-4 datasets showed significant difference from the figure 4.3, especially the lower and upper quartile, median and minimum observation. The multiple Behren-Fisher-Test also carried out the similar results.

In scenario 2, both ascendant and descendant datasets show that there is no significant trend amongst different relative dominance datasets. The average AUC locates around 0.90 and the interval between lower and upper quartile is smaller than scenario 1. Therefore, gradually removing lower relative dominance datasets (ascRDo) does not affect the AUC values of GAM and MAXENT. But gradually removing higher relative dominance datasets in scenario 1 does affect the AUC values, especially in GAM modelling. In contrary to scenario 1, gradually removing higher relative dominance datasets in scenario 2 does not affect the AUC values of GAM and MAXENT.

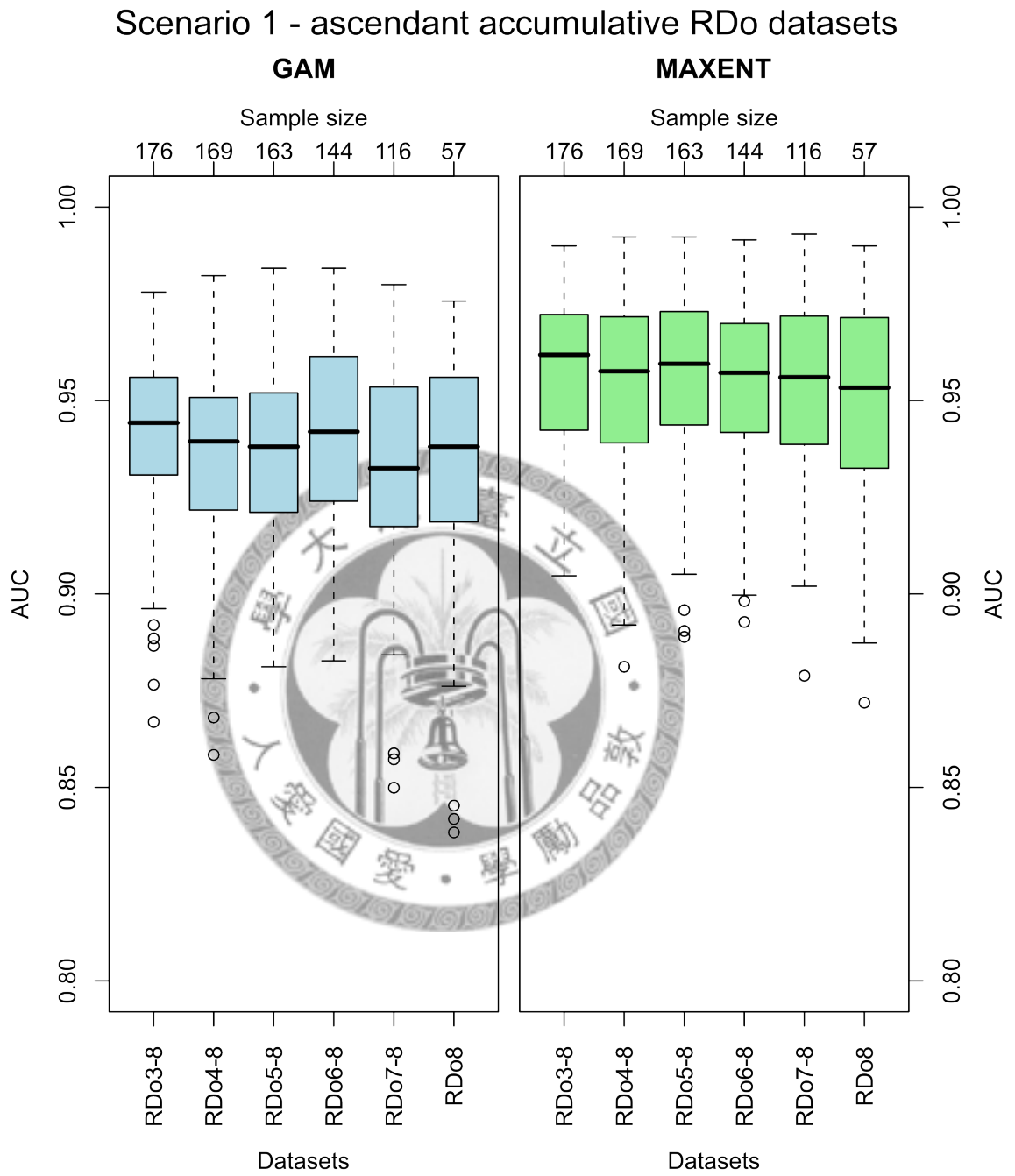


Figure 4.2: Scenario 1 - Ascendant accumulative relative dominance datasets. Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.

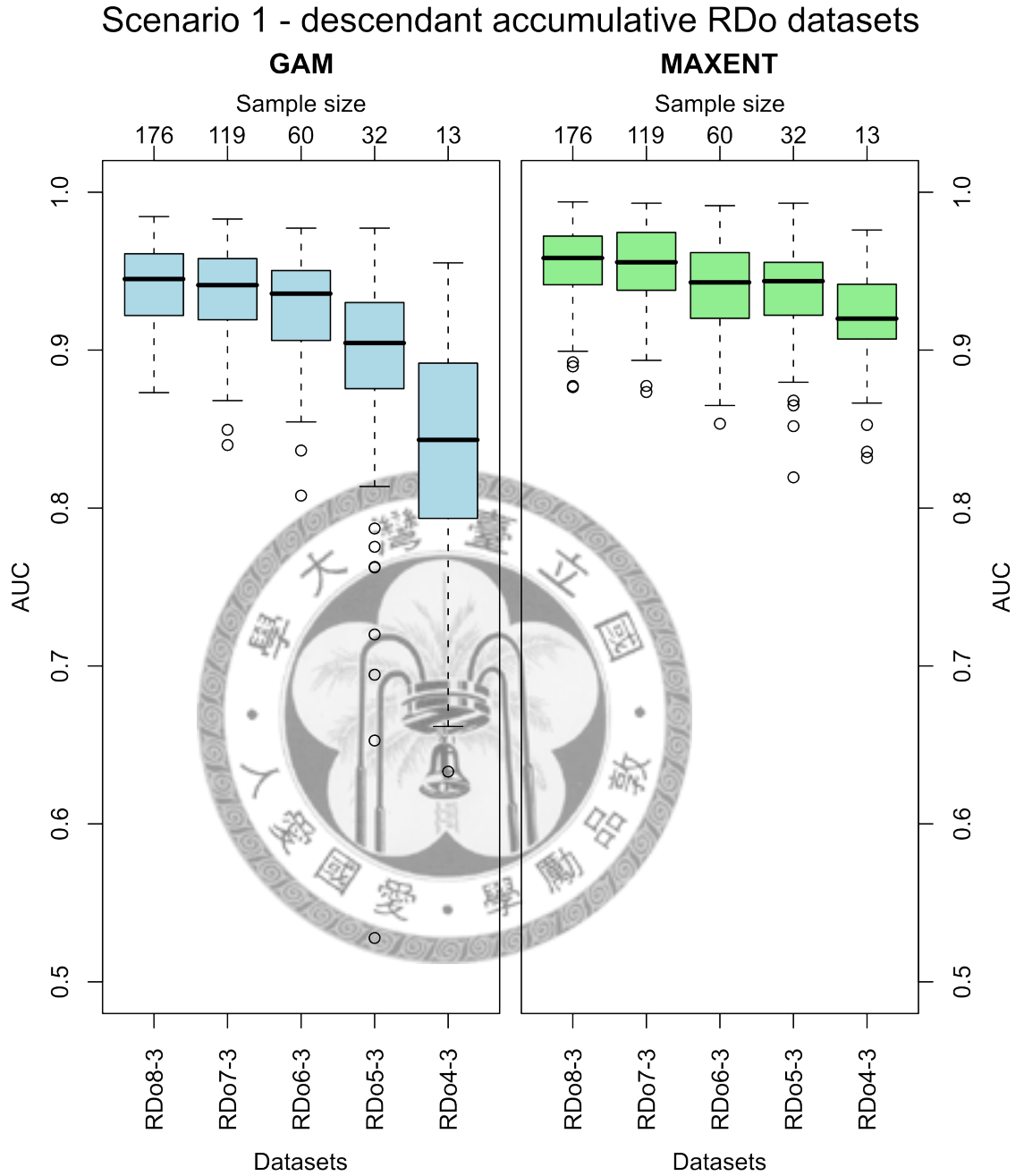


Figure 4.3: Scenario 1 - Descendant accumulative relative dominance datasets. Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.

Scenario 2 - Ascendant accumulative RDo datasets

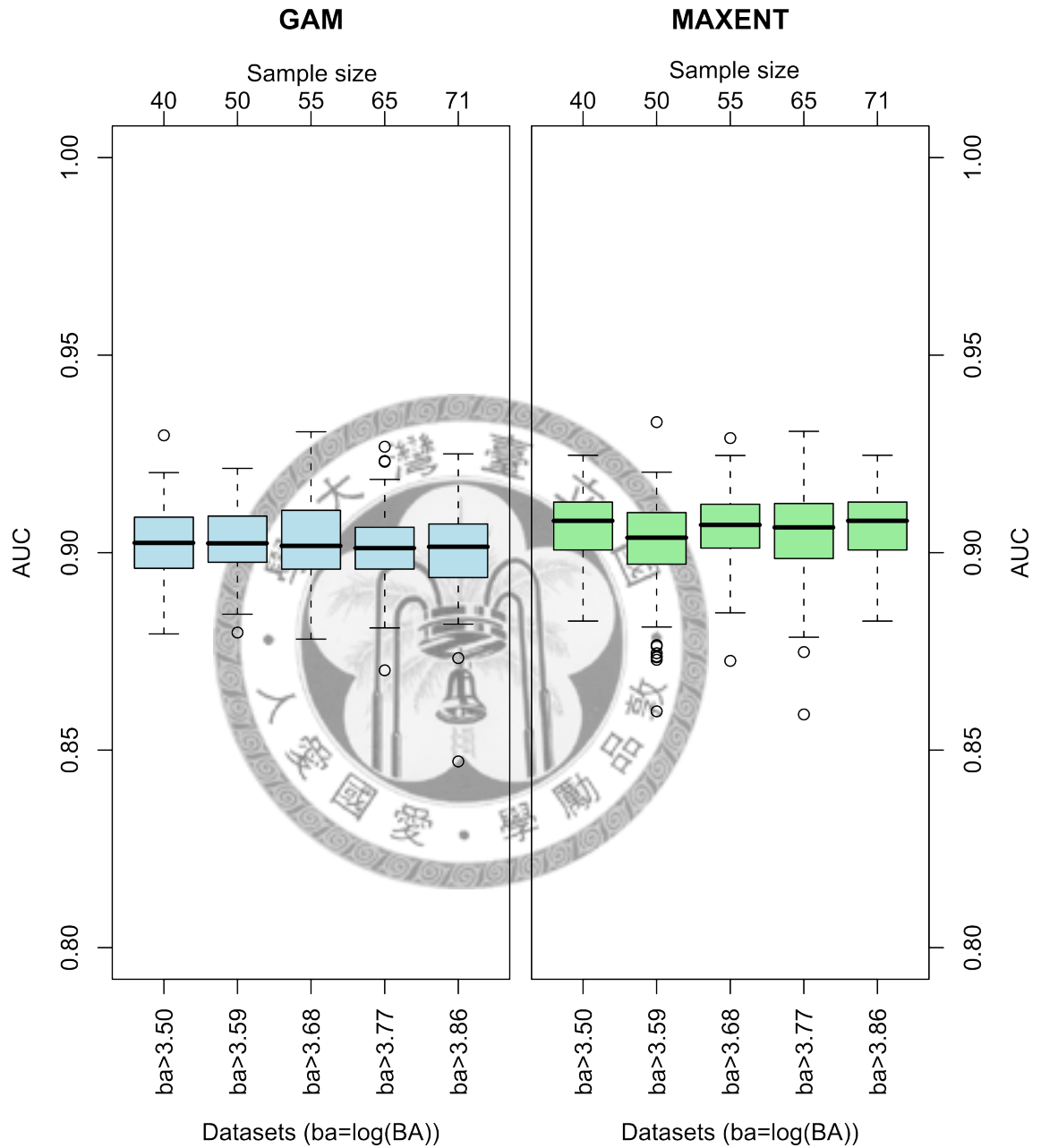


Figure 4.4: Scenario 1 - Ascendant accumulative relative dominance datasets. Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.

Scenario 2 - Descendant accumulative RDo datasets

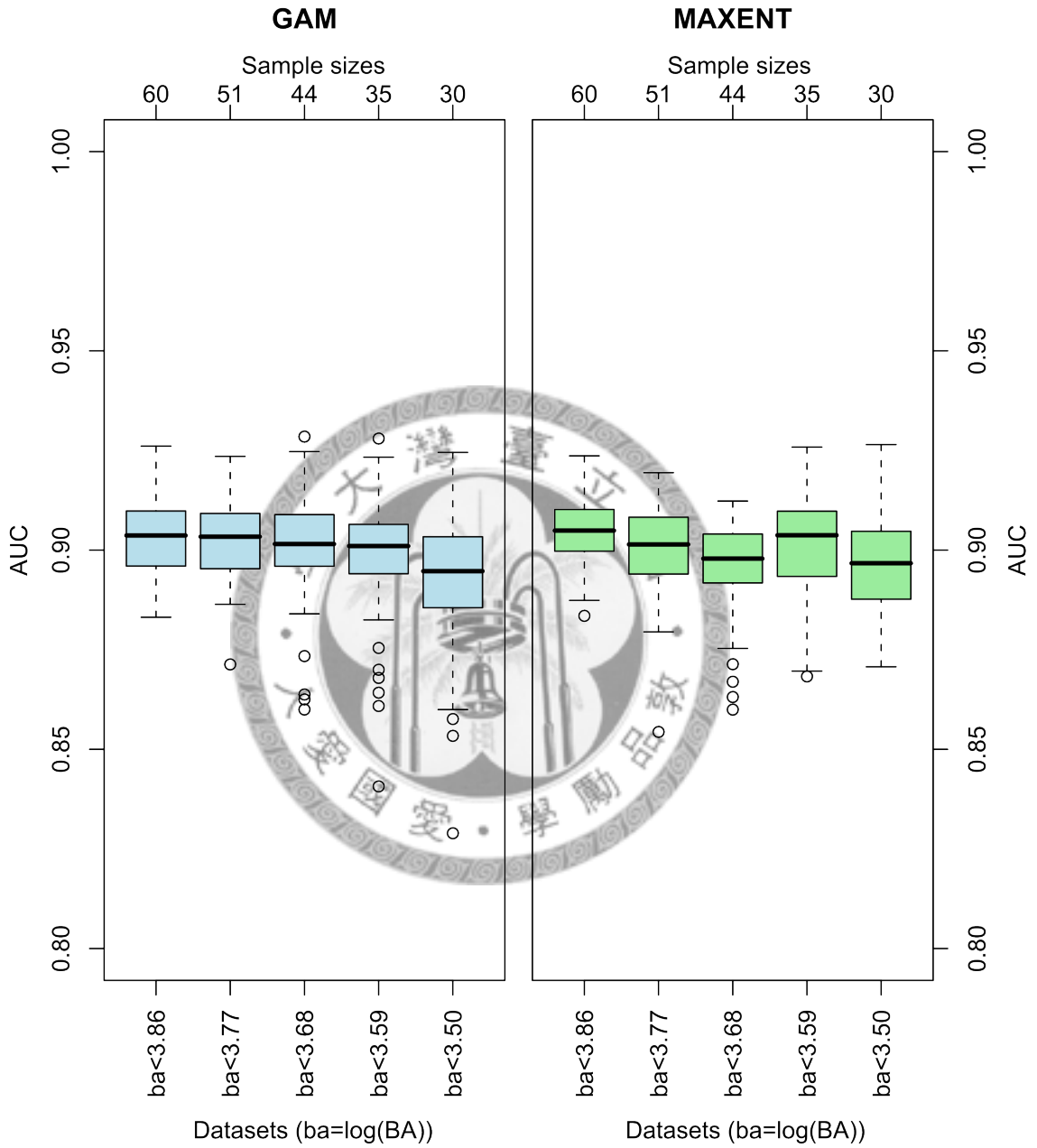


Figure 4.5: Scenario 2 - Descendant accumulative relative dominance datasets Vertical axis shows the area under ROC curve (AUC) values and horizontal axis shows the experimental datasets.

Chapter 5

Discussion

This chapter is constituted by five parts: the first section discuss the environmental variables; the second section discuss the model performance; the third part discuss the dominance effect in sampling data; and the last two parts in this chapter discuss the “dominance” in ecological theories and possible errors in experiment samplings.

There are five most influential environmental factors in jackknife analysis: (1) warmth index (2) wetness index (3) sediment transport capacity index (4) slope (5) altitude. These environmental variables also reflect the ecological meanings of *Tsuga chinensis* var. *formosensis* distributions. As mentioned before, *Tsuga chinensis* var. *formosensis* preferably distributes the mountain ridges, cliffs and drier south-facing slopes in high altitude of Taiwan. ALT reflects that higher altitude (2800-3000m of optimum) is more suitable for *Tsuga chinensis* var. *formosensis*'s distribution. WI is highly correlated with ALT because WI formula is generated from altitude. SLP and STCI reflect the preference of steep or precipitous environment such as cliffs and mountain ridges. The natural habitat of *Tsuga chinensis* var. *formosensis* is also more drier than other places, for example,

it often has less humidity in the mountain ridges and cliffs which *Tsuga chinensis* var. *formosensis* usually becomes pure stands and is dominant species. But due to the high multi-collinearity revealed in table 4.1, SLP, WI and WET are considered for modelling.

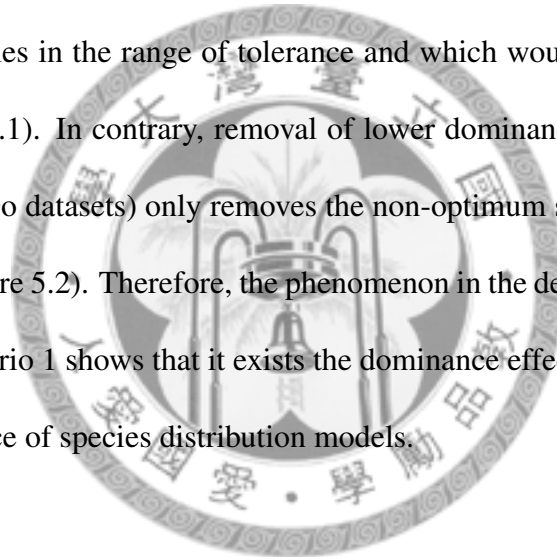
5.1 Model Performance

Elith et al. (2006) compared 16 modelling techniques and tried to make comprehensive research about the species distribution modelling. And Guisan et al. (2007b) extended the topic and also compared the robustness of 10 species modelling techniques to various data issues. From Elith et al. (2006)'s study, they found that there existed some trends for more variation across modelling techniques for species which were harder to model (Elith et al., 2006). The situation also existed in recent species distribution research articles (Guisan et al., 2007b). However, Guisan et al. (2007b) suggested that variation in SDM performance is greater amongst species than amongst modelling techniques and different modelling techniques produced consistently distinctive model performance (Guisan et al., 2007b). This study also confirmed that the species traits effects are greater than modelling techniques because there is no distinctive differences amongst model techniques.

The recent comparative studies (Guisan et al., 2007b; Elith et al., 2006; Segurado and Araújo, 2004; Thuiller et al., 2003) conclude that both MAXENT and GAM are good models for predicting the species distribution. In addition, Meynard and Quinn (2007) have approved the GAM and GLM are good trade-off models between model performance and complexity. The results of this study also reveal that the average model performance amongst MAXENT and GAM are very good.

5.2 Dominance Effects in Sampling Data

In the descendant accumulative RDo datasets of scenario 1, the results reveal that removal of higher dominance datasets would decrease the model accuracies. But in the accumulative RDo datasets do not have the same trend. The possible reason which made the trend would be the dominance effects. Removal of higher dominance datasets mean removal of the optimum samples in the range of tolerance and which would decrease the overall performance (figure 5.1). In contrary, removal of lower dominance datasets (i.e. ascendant accumulative RDo datasets) only removes the non-optimum samples in the range of tolerance (refer to figure 5.2). Therefore, the phenomenon in the descendant accumulative RDo datasets of scenario 1 shows that it exists the dominance effect in sampling data and affects the performance of species distribution models.



5. Discussion

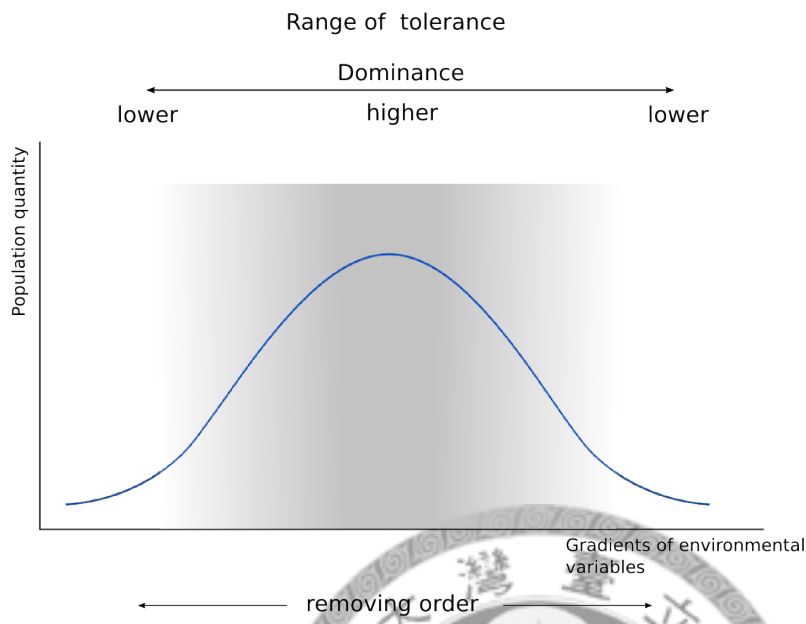


Figure 5.1: Diagram of removal of higher dominance datasets in range of tolerance

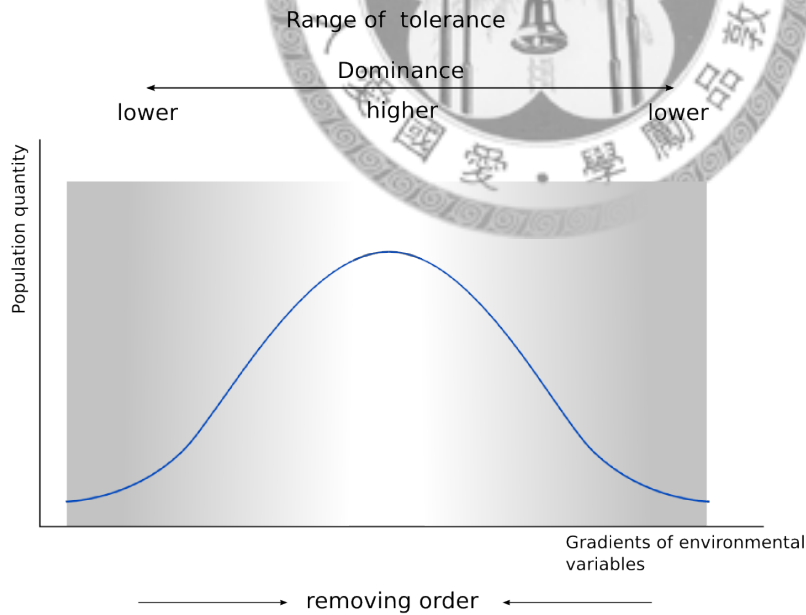


Figure 5.2: Diagram of removal of lower dominance datasets in range of tolerance

However, the results of scenario 2 do not reveal the dominance effects in sampling data on species distribution models. The reasons may be the sample selection bias, sample sizes or other unknown factors. We would discuss the possible reasons in the following sections and there are two subjects proposed to elucidate and discuss the dominance effects in ecological theories and the distribution modelling evaluation errors.

5.3 Dominance in Ecological Theories

Datasets separation by dominance

Dominance is analytic artificial concept to describe the adaptability and control of natural habitat. Dominance itself is also an abstractive concept to express the “real dominance”. Few methods which is mentioned in section 2.1 have interpreted the concept of dominance, such as coverage or basal area in Zürich-Montipeller school or IVI in Wisconsin school. Both basal area and IVI are applied in this study for evaluating the dominance effects of species distribution modelling. The scenario 2 (i.e. logarithm basal area) results in that the overall AUC values have no significant difference. Considering the data distribution of basal area in figure 5.3, the shape is inverse-J which indicates the data aggregate on left hand side. The distribution is log-normal distribution and it is divided into five parts for model training. Although I try to separate the numbers of datasets equally to decrease the sample size effect, the separation may have two problems:

1. The separation method is based on the numbers of logarithm basal area but not exactly match ecological meanings.

5. Discussion

2. In the sense of ecology, the dominance level of *Tsuga chinensis* var. *formosensis* may be the same.

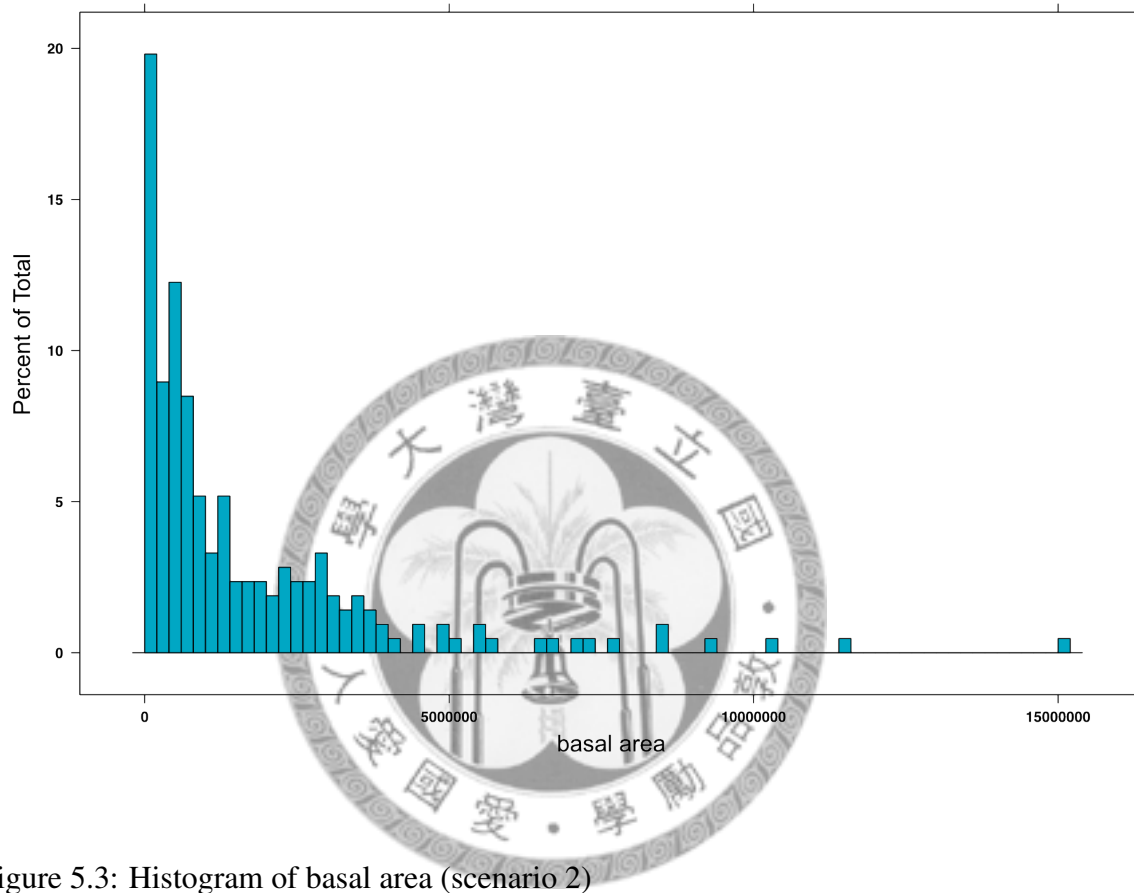


Figure 5.3: Histogram of basal area (scenario 2)

However, the separation method based on IVI (scenario 1) provides a better way to divide the different dominance datasets because the IVI uses relative dominance, relative density and relative frequency to evaluate dominance comprehensively. But some of the IVI datasets have small sample plots (it is only 13 sample plots in RDo4–3 and 32 in RDo5–3) and the sample sizes are influential to species distribution models (Hernandez et al., 2006).

Beyond dominance

Another point beyond dominance is that the *Tsuga chinensis* var. *formosensis* does not have specific habitat, and the distribution of *Tsuga chinensis* var. *formosensis* could be randomly distributed in the range of certain altitude. The phenomenon implies that the dominance of the *Tsuga chinensis* var. *formosensis* is not clear and the selection of different datasets may be random selections.



5.4 Model Evaluation Errors

Sample selection bias

Sample selection bias and implication of pseudo-absence data are another influential issue to species distribution modelling (Phillips et al., 2009; VanDerWal et al., 2009). Low prevalence of raw data also makes the prediction in high specificity. In scenario 1, I equally selected presence and absence data in each training and evaluation datasets and the sample plots did not match relative dominance selection criteria were deleted. The

selection method applied in scenario 1 assured if the model running by chance would not dramatically affect the accuracies, In other words, over-prediction and under-prediction were avoided. For example, if I selected a large number of pseudo-absence plots in training datasets, no matter what the robustness of models, the model always resulted in high specificity and led to over-prediction. Selecting a large number of presence plots also resulted in high sensitivity. In contrary to scenario 1, scenario 2 try to model fixed training datasets. Those sample plots did not match relative dominance selection criteria were regarded as absence.

The overall performance in scenario 1 - ascRDo datasets are much similar to scenario 2 but the variation of scenario 1 is slightly higher than scenario 2. Consequently, it shows that the variation in flexible sample sizes (scenario 1) is higher than fixed sample sizes (scenario 2). The overall sensitivity of GAM in scenario 2 is approximately zero and specificity is approximately 1 may due to the data selection problem, too many absence plots and few presence plots lead to under-prediction. The MAXENT sensitivity in scenario 2 is relative higher than GAM but it is low (about 0.5) and may also lead to under-prediction.

Sample sizes

Hernandez et al. (2006)'s study delivered that the smaller sample size in training datasets would decrease the model accuracies. Both the results of descendant datasets in scenario 1 and 2 show that smaller sample sizes performs poorly (refer to figure 4.3 and 4.5). RDo4-3 dataset in scenario 1 of GAM has the poorest performance, the max-

imum and minimum AUC value exceed over 0.2. In general, the performance gradually decreases whilst the sample size decreases especially the sample size smaller than 60. The performances amongst different relative dominance datasets of MAXENT seem to be insignificant different but appeal to a very slight trend (figure 4.3). If we select the datasets in which sample sizes are larger than 40, the performances amongst different relative dominance datasets seem to be no significant difference.

Comparing to descendant datasets, ascendant relative dominance datasets do not have significant difference amongst different relative dominance datasets. The results may due to the sample size effect because the smallest sample size in ascendant datasets is 40 and the sample sizes of last two datasets in descendant datasets are both smaller than 40. However, the results of descendant datasets need further examinations to elucidate the performance is impacted by sample sizes or by dominance effects.

Number of species

The mostly recent studies for evaluating species distribution modelling use more than one species, for example, VanDerWal et al. (2009) used 12 vertebrate species in Australian Wet Tropics; Guisan et al. (2007b) used 30 native tree species in Switzerland; Randin et al. (2006) used 54 plant species in Austria and Swiss Alps; and Phillips et al. (2009) even used 226 species from diverse regions of the world. Only one species was applied in this study may be problematic and would be the Achilles heel. Elith et al. (2006) mentioned that variation in species characteristics is greater than modelling techniques. Therefore, the results of our study may only conclude relative dominance would not be a

major influential factor to species distribution models. On the other hand, if I applied the method to other species, the results may differ.

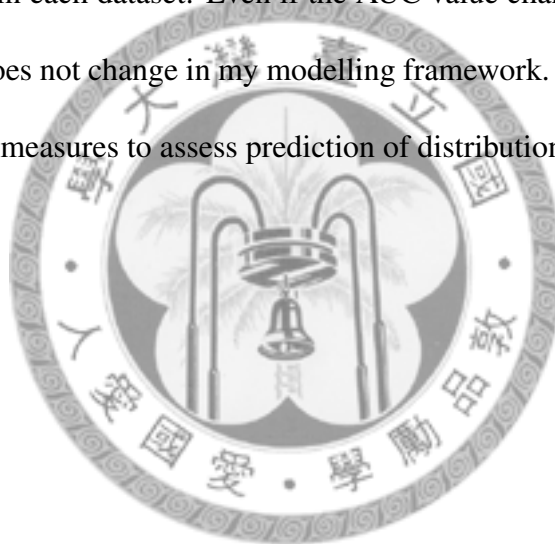
Besides the species characteristics mentioned above, data quality and consistency would be influential factors to the model accuracy. There are many field work data in Taiwan but the resolution and data quality are not good for prediction in species distribution. Furthermore, the consistence of data are different amongst different projects, for example, the 3rd FRLI is systematic sampling, but the National Vegetation Diversity Inventory and Mapping Project is not systematic sampling.

However, Austin et al. (2006) used artificial species generated from real environmental gradient in a real landscape to evaluate species distribution models. Meynard and Quinn (2007) also suggested that artificial species could be used for prediction of rare species. The strength of artificial data is assured the niche-based theory in species distribution models and avoids uncertainty existed in real species. Since Austin (2007) has argued that different statistical models are used in prediction of species distribution without explicit ecological theory. This suggests the species occurrence data can be generated from environmental gradients to make artificial relative dominance datasets for further evaluating dominance effects.

ROC analysis and AUC issues

AUC has been widely used in many SDM research articles (Guisan et al., 2007b; Meynard and Quinn, 2007; Elith et al., 2006; Hernandez et al., 2006; Randin et al., 2006) due to its good performance and ease of computation (Lobo et al., 2008), and it is recom-

mended to use of the AUC from a ROC plot because it is threshold independent (Fielding, 2002). Lobo et al. (2008) had different opinions against AUC in assessment of distribution models. They suggest uncertainty and spatial dimensions of distributions are specific characteristics of species occurrence data to prevent the use of AUC in distribution modelling (Lobo et al., 2008) and multiple accuracy measures should be used for evaluating the model performance, such as combination of AUC, sensitivity and specificity. As the AUC issues argued in above literature, only one accuracy measure (AUC) is used in this study may result bias in each dataset. Even if the AUC value changes, the overall trends of different datasets does not change in my modelling framework. However, it is worth to use multiple accuracy measures to assess prediction of distribution models.



Chapter 6

Conclusion

From the Jackknife analysis, it could be concluded that warmth index is the most influential environmental factors for the distribution of the *Tsuga chinensis* var. *formosensis*. Follow up is the altitude, slope, wetness index and sediment transport capacity index. Due to the multicollinearity, the results show that warmth index, slope, wetness index are considered to be the predictive environmental variables for *Tsuga chinensis* var. *formosensis*.

In terms of AUC values, the overall performance of either GAM or MAXENT are reasonably good. Ascendant accumulative relative dominance in scenario 1 appeared not to be an influential factor to accuracy of SDM. In contrary, the descendant accumulative relative dominance revealed a trend and the result showed that dominance effects. Removal of lower dominance would affect the model performances but removal of higher dominance datasets would affect the overall performances of the two models. However, both ascendant and descendant accumulative datasets in scenario 2 revealed that there were no significant differences amongst high and low relative dominance.

Sample selection bias, species traits and model criteria are *a posteriori* the possible influential reasons to the overall performance of species distribution models. However,

the concept of “dominance” remains the critical issue in this study. The “dominance” separation in this case cannot exactly interpret the species supremacy in ecological habitat. Although it seems a trend existence in descendant datasets of scenario 1, it may mix with sample size effect and dominance effect.

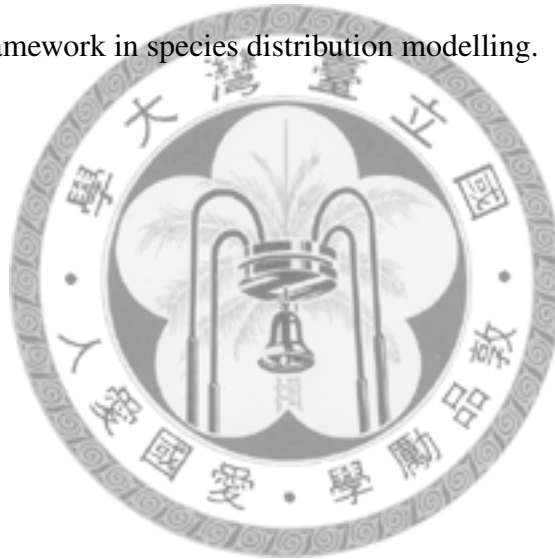
Rethinking the sample selection criteria in this study, half presence/absence and randomly selection may avoid bias. But it also needs more attention to set up the selection criteria, I would suggest:

1. In presence-absence modelling techniques (regression-based statistical models), absence plots should be selected carefully. I suggest to use half presence and half absence in both GAM and MAXENT modelling.
2. Selection should be avoided human disturbance areas.
3. Using multiple evaluation criteria detect the possible errors. Both sensitivity and specificity can be used for error rate prediction and the AUC and Kappa statistic can examine the model performances.
4. Using the data with better quality.
5. Using herbaceous (annual or perennial) plants for target species. The life cycle of herbaceous plant is shorter than woody species and this could be avoided time effects.
6. Multiple species should be used for evaluation. Since the sample sizes of raw data are not sufficient, there are two cases are considered:

6. Conclusion

- (a) Artificial species is suggested for evaluation.
- (b) Species of similar ecological habitat can be combined as pseudo-species for evaluation and this method can increase sample sizes. For example, *Machilus zuihoensis* (Lauraceae) and *Machilus japonica* (Lauraceae) and *Machilus thunbergii* (Lauraceae) can be grouped as one pseudo-species.

Despite the results rejected assumption, the methods in this study provide a theme for such exploratory ecological experiment. The concept inherited from data mining also dispense a possible framework in species distribution modelling.



References

- Araújo, M. and Williams, P. (2000). Selecting areas for species persistence using occurrence data. *Biological Conservation*, 96(3):331–345.
- Araújo, M. B. and Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10):1677–1688.
- Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1-2):1–19.
- Austin, M. P., Belbin, L., Meyers, J. A., Doherty, M. D., and Luoto, M. (2006). Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecological Modelling*, 199(2):197–216.
- Braun-Blanquet, J. (1932). *Plant Sociology: the Study of Plant Communities*. Hafner Publishing Company, Inc. Translated by Fuller, George D. and Conard, Henry S. pp. 439.
- Cain, S. A. and de Oliveira Castro, G. M. (1959). *Manual of Vegetation Analysis*. Harper and Brothers Publishers, New York. pp. 325.
- Chen, Y.-F. (2004). *Taiwan Tsuga Belt (I)*, volume 1, page 95. Avanguard Publisher, Taipei, Taiwan.
- Chiou, C.-R., Lai, Y.-J., Li, C.-F., and Liang, Y.-C. (2004). The application of GIS on the simulation of climate change impact on forest - a case study of Taiwan cypress forest. *Greater China GIS Conference and Exhibition 2004*.
- Chiou, C.-R., Lin, C.-J., and Li, C.-F. (2006). Analysis of distribution characteristics of Taiwan hemlock communities. In *Proceedings of Fourth Symposium of Vegetation Diversity in Taiwan Vegetation Mapping Series*, volume 8, pages 280–305.
- Congalton, R. G. and Green, K. (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, page 137. Lewis Publishers, Boca, Raton.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons, Inc., Publication, New Jersey. pp. 748.

References

- Curtis, J. and McIntosh, R. (1951). An upland forest continuum in the prairie-forest border region of Wisconsin. *Ecology*, 32(3):476–496.
- Curtis, J. T. (1947). The palo verde forest type near Gonivaves, Haiti. and its relation to the surrounding vegetation. *Caribbean Forestry*, 8:1–26.
- Daubenmire, R. F. (1968). *Plant Communities: a Textbook of Plant Synecology*. Harper & Row Publishers, New York. pp. 300.
- Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberon, J., Williams, S., Wisz, M. S., and Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151.
- Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, Boca Raton. pp. 301.
- Fielding, A. H. (1999). How should accuracy be measured. In Fielding, A. H., editor, *Machine Learning Methods for Ecological Applications*, chapter 8, pages 209–223. Kluwer Academic Publishers, Norwell, Massachusetts.
- Fielding, A. H. (2002). What are the appropriate characteristics of an accuracy measure? In Scott, J. M., Heglund, P. J., and Morrison, M. L., editors, *Predicting Species Occurrences: Issues of Accuracy and Scale*, chapter 21, page 271. Island Press, Washington, DC.
- Galparsoro, I., Borja, A., Bald, J., Liria, P., and Chust, G. (2009). Predicting suitable habitat for the European lobster (*Homarus gammarus*), on the Basque continental shelf (Bay of Biscay), using ecological-niche factor analysis. *Ecological Modelling*, 220(4):556–567.
- Gauch, H. G. (1982). *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge, UK. pp. 298.
- GRASS Development Team (2006). *Geographic Resources Analysis Support System (GRASS GIS) Software*. ITC-irst, Trento, Italy. <http://grass.itc.it>.
- Guan, L.-H. and Chen, J.-H., editors (1995). *The Third Survey of Forest Resources and Land Use*. Bureau of Forestry, Council of Agriculture, Executive Yuan, Taipei, Taiwan. (in Chinese).
- Guisan, A., Graham, C. H., Elith, J., and Huettmann, F. (2007a). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, 13(3):332–340.

- Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147–186.
- Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., and Peterson, A. T. (2007b). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, 77(4):615–630.
- Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufman Publishers, San Francisco, second edition. pp. 770.
- Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall/CRC, Boca, Raton. pp. 335.
- Heglund, P. J. (2003). Foundations of species-environment relations. In Scott, J. M., Heglund, P. J., and Morrison, M. L., editors, *Predicting Species Occurrences: Issues of Accuracy and Scale*, chapter 1, pages 35–41. Island Press, Washington, DC.
- Hernandez, P. A., Graham, C. H., Master, L. L., and Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5):773–785.
- Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.
- Kriegler, B. (2007). *Cost-sensitive Stochastic Gradient Boosting Within a Quantitative Regression Framework*. PhD thesis, University of California, Los Angeles.
- Landis, J. R. and Koch, G. C. (1977). Measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., and Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding new zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, 321:267–281.
- Lehmann, A., Overton, J. M., and Leathwick, J. R. (2002). GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling*, 157(2-3):189–207.
- Lin, C.-J. (2009). Conferring the Composition and Distribution of Vegetation Diversity in Taiwan. Master thesis, School of Forestry and Resource Conservation, National Taiwan University, Taipei, Taiwan.
- Liu, T.-S. and Su, H.-J. (1983). *Forest Plant Ecology*. Commercial Press Taiwan, Taipei, Taiwan. pp. 462.
- Lobo, J. M., Jimenez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151.

References

- Manel, S., Williams, H., and Ormerod, S. (2001). Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931.
- Matsui, T., Nakaya, T., Yagihashi, T., Taoda, H., and Tanaka, N. (2004). Comparing the accuracy of predictive distribution models for *Fagus crenata* forests in Japan. *Japanese Journal of Forest Environment*, 46(2):93–102.
- Meynard, C. N. and Quinn, J. F. (2007). Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, 34(8):1455–1469.
- Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.
- Pearson, R. G., Raxworthy, C. J., Nakamura, M., and Peterson, A. T. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34(1):102–117.
- Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259.
- Phillips, S. J. and Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175.
- Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197.
- Phillips, S. J., Dudík, M., and Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 472–486, New York. ACM Press.
- PostgreSQL Global Development Group (2008). *PostgreSQL Database Management System*. PostgreSQL Foundation, California, USA. <http://www.postgresql.org>.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Randin, C. F., Dirnbock, T., Dullinger, S., Zimmermann, N. E., Zappa, M., and Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography*, 33(10):1689–1703.
- Raunkiaer, C. (1934). *The Life Forms of Plants and Statistical Plant Geography*. Clarendon Press, Oxford.

- Reddy, S. and Davalos, L. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 30(11):1719–1727.
- Segurado, P. and Araújo, M. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography*, 31(10):1555–1568.
- Song, G.-Z., Lin, C.-T., Chiou, C.-R., and Lu, Y.-C. (2007). Comparing three species distribution models - applied in *Tsuga chinensis* distribution in Taiwan. In *Proceedings of the Fifth Symposium of Vegetation diversity in Taiwan*, volume 9, Taipei, Taiwan.
- Su, H.-J. (1984). Studies on the climate and vegetation types of the natural forests in Taiwan. (II). Altitudinal vegetation zones in relation to temperature gradient. *Quarterly Journal of Chinese Forestry*, 17(4):57–73.
- Swets, J. (1988). Measuring the accuracy of diagnostic system. *Science*, 240:1285–1293.
- Tanaka, N., Matsui, T., Yagihashi, T., and Taoda, H. (2006). Climatic controls on natural forest distribution and predicting the impact of climate warming: Especially referring to Buna (*Fagus crenata*) forests. *Global Environmental Research*, 10(2):151–160.
- Thuiller, W., Araujo, M., and Lavorel, S. (2003). Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, 14(5):669–680.
- Tsao, L.-S. (2007). Using Generalized Additive Models to Establish the Relationships Between Distribution Ranges and Climatic Factors for Six Conifer Species in Taiwan. Master's thesis, School of Forestry and Resource Conservation, National Taiwan University, Taipei, Taiwan. pp. 76.
- VanDerWal, J., Shoo, L. P., Graham, C., and Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, 220(4):589 – 594.
- Wilson, K., Westphal, M., Possingham, H., and Elith, J. (2005). Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, 122(1):99–112.
- Wood, S. N. (2006). *Generalized Additive Models - an Introduction with R*. Chapman & Hall/CRC, Boca Raton. pp. 392.

Appendix A

Demo program

There are two scenarios in our study, following demo program is Scenario2-BA which executes the experiment.

Scenario2-BA



```
1 #!/usr/bin/env bash
2 # Copyright (c) 2008-2009 Lin, Cheng-Tao <r96625028@ntu.edu.tw>
3 # All rights reserved.
4 #
5 # Redistribution and use in source and binary forms, with or
6 # without modification, are permitted provided that the
7 # following conditions are met:
8 #     1 Redistributions of source code must retain the above
9 #       copyright notice, this list of conditions and the
10 #      following disclaimer.
11 #     2 Redistributions in binary form must reproduce the
12 #       above copyright notice, this list of conditions
13 #       and the following disclaimer in the documentation
14 #       and/or other materials provided with the distribution.
15 #
16 # THIS SOFTWARE IS PROVIDED BY LIN, CHENG-TAO ''AS IS'' AND ANY
17 # EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO,
18 # THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A
19 # PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL LIN,
20 # CHENG-TAO BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
```

```

21 # SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT
22 # NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES;
23 # LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION)
24 # HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN
25 # CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE
26 # OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS
27 # SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
28
29
30
31 ## Local variables
32 DB=forestsurvey3rd
33 BASE=/home/db/pgsql/Tsuga/Scenario2
34 XXXpred=${BASE}/Xpredlk
35 PGSQL=pgsql
36 JAVA=java
37 MAXIMUM_MEMORY=1024
38 MAXENT_JARFILE=/home/db/pgsql/Tsuga/maxent.jar
39 ENVLAYERS_DIR=/home/db/pgsql/Tsuga/EnvLayers
40
41 #####
42 ##### FUNCTIONS #####
43 #####
44
45 PGSQL_QUERY(){
46 ##### pgsql SQL query #####
47
48 # 1. random selection -> 2. create training table -> 3. create
   cross-validation table
49
50 ${PGSQL} -q -d ${DB} -c "
51     -- 1.1 create temp table rnd${rnd}, and order by random
52     CREATE TEMP TABLE rnd${rnd} AS (SELECT * FROM nf_releve ORDER
   BY random());
53     -- 1.2 add rsn (type: serial) column to table rnd${rnd}
54     ALTER TABLE rnd${rnd} ADD COLUMN rsn serial;
55     -- 1.3 create training table
56     CREATE TEMP TABLE r${rnd}t AS (SELECT * from rnd${rnd} where
   rsn <= 718);
57     -- 1.4 create cross-validation table
58     CREATE TEMP TABLE r${rnd}v AS (SELECT * from rnd${rnd} where
   rsn > 718);
59
60     -- 2.1 select presence data, according to the ba cut point
61     CREATE TEMP TABLE r${rnd}ta${d}p
62     AS (SELECT * FROM r${rnd}t WHERE ba ${cr} ${cut_point});
63     -- 2.2 other presence data regard as absence
64     CREATE TEMP TABLE r${rnd}ta${d}a AS
65     (SELECT * FROM r${rnd}t WHERE ba ${icr} ${cut_point});
66     -- 2.3 union with other absence data
67     CREATE TEMP TABLE r${rnd}ta${d}au AS ((SELECT * FROM r${rnd}t
   WHERE tsuga=0)

```

A. Demo program

```
68     UNION (SELECT * FROM r${rnd}ta${d}a));
69     -- 2.4 set all of the value in tsuga column to 0
70     UPDATE r${rnd}ta${d}au SET tsuga=0;
71     -- 2.5 combine presence and absence table and make a table
72     CREATE TEMP TABLE r${rnd}ta${d}f AS
73     ((SELECT * FROM r${rnd}ta${d}p) UNION (SELECT * FROM r${
74         rnd}ta${d}au));
75     -- 2.6 add index column (for grasper)
76     ALTER TABLE r${rnd}ta${d}f ADD COLUMN index serial;
77     -- 2.7 export all
78     COPY (SELECT index,mapno,x,y,landuse,slope,wetness,wi,tsuga,ba
79         FROM r${rnd}ta${d}f) TO
80         '${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}/$
81         {SAMPLING_CODE}${d}${rnd}-raw.csv' DELIMITER AS ',' CSV
82         HEADER;
83     -- 2.8 export YYY
84     COPY (SELECT index,tsuga FROM r${rnd}ta${d}f)
85     TO '${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}
86     }/YYY'
87     DELIMITER AS ',' CSV HEADER;
88     -- 2.9 export XXX
89     COPY (SELECT index,x,y,slope,wetness,wi FROM r${rnd}ta${d}f)
90     TO '${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}
91     }/XXX'
92     DELIMITER AS ',' CSV HEADER;
93     --2.10 export to sample.csv
94     COPY (SELECT tsuga,x,y,slope,wetness,wi FROM r${rnd}ta${d}f)
95     TO '${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}
96     }/sample.csv'
97     DELIMITER AS ',' CSV HEADER;
98     -- 3. create validation format
99     COPY (SELECT tsuga as presence,x,y FROM r${rnd}v)
100     TO '${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}
101     }/r${rnd}v.csv'
102     DELIMITER AS ',' CSV HEADER;
103     "
104 }
105
106 SWD_PREPROCESS(){
107     sed -e 's/^0/background/g'  ${BASE}/${SAMPLING_CODE}${d}/${S
108     AMPLING_CODE}${d}${rnd}/sample.csv |
109     sed -e 's/^1/Tsuga/g' > sample.csv
110     if [ -d output ]; then
111         echo "output directory exists!"
112     else mkdir output
113     fi
114 }
115
116 MAXENT_MODELLING(){
117     ${JAVA} -mx${MAXIMUM_MEMORY}m -jar ${MAXENT_JARFILE} -o
118     output -a -e ${ENVLAYERS_DIR} \
```

```

111         -s sample.csv redoifexists
112     }
113
114
115
116 GRASS_SAMPLE() {
117     ##### 5. Extract value from points via GRASS GIS #####
118
119     if [ `pwd` = ${BASE}/${SAMPLING_CODE}${d}/${
120         SAMPLING_CODE}${d}${rnd}/gam ]; then
121         MODEL=gam
122         RASTER=pred_tsuga.asc
123     elif [ `pwd` = ${BASE}/${SAMPLING_CODE}${d}/${
124         SAMPLING_CODE}${d}${rnd}/maxent ]; then
125         MODEL=maxent
126         RASTER=output/Tsuga.asc
127     else
128         echo "Exception caught! (GRASS_SAMPLE)"
129     fi
130
131     cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${
132         rnd}/${MODEL}
133     # 1 import asc into grass (RASTER)
134     r.in.arc input=${BASE}/${SAMPLING_CODE}${d}/${
135         SAMPLING_CODE}${d}${rnd}/${MODEL}/${RASTER} \
136         output=${SAMPLING_CODE}${d}r${rnd} type=FCELL mult
137         =1.0 --o &&
138
139     # 2 import validation xy coordinate (VECTOR)
140     v.in.ascii input=${BASE}/${SAMPLING_CODE}${d}/${
141         SAMPLING_CODE}${d}${rnd}/r${rnd}v.csv \
142         format=point fs="," output=${SAMPLING_CODE}${d}v${
143         rnd} \
144         skip=1 'columns=presence int,x int,y int' x=2 y=3
145         z=0 cat=0 --o &&
146
147     # 3 sample (nearest neighbor)
148     v.sample input=${SAMPLING_CODE}${d}v${rnd} column=
149         presence \
150         output=${SAMPLING_CODE}${d}s${rnd} rast=${
151             SAMPLING_CODE}${d}r${rnd} z=1.0 --o &&
152
153     # 4 export to csv
154     v.out.ogr input=${SAMPLING_CODE}${d}s${rnd} type=point
155         \
156         dsn=${SAMPLING_CODE}${d}s${rnd} olayer=${
157             SAMPLING_CODE}${d}s${rnd} \
158         layer=1 format=CSV --o
159 }
160
161 EVALUATE() {
162     # 5 modify csv file to meet the input format of

```

A. Demo program

```
PresenceAbsence
151 cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${
    rnd}/${MODEL}/${SAMPLING_CODE}${d}s${rnd}/ &&
152 cat ${SAMPLING_CODE}${d}s${rnd}.csv | sed -e 's/cat/
    PlotID/g' |
153 sed -e 's/pnt_val/Observed/g' | sed -e 's/rast_val/
    Predicted/g' |
154 awk -F',' '{ print $1,$2,$3 }' > ${SAMPLING_CODE}${d}
    a${rnd}
155
156 cat > s${rnd}.R << _EOF
157
158 library(PresenceAbsence)
159 ${SAMPLING_CODE}${d}a${rnd} <- read.table("${SAMPLING_CODE}${d}a${
    rnd}", header=T)
160 p.a.accuracy <- presence.absence.accuracy(${SAMPLING_CODE}${d}a${
    rnd})
161 write.table(cbind(auc(${SAMPLING_CODE}${d}a${rnd})\$AUC, p.a.
    accuracy\$Kappa),
162 file="${BASE}/results/${MODEL}-${SAMPLING_CODE}${d}r-${rnd}",
    row.names=F, col.names=F)
163
164 _EOF
165 # execute calculation
166 cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${
    rnd}/${MODEL}/${SAMPLING_CODE}${d}s${rnd}
167 R CMD BATCH s${rnd}.R
168 }
169
170
171
172 #####
173 ##### Main #####
174 #####
175
176 case $1 in
177 -p|--preprocess)
178
179 # cut points
180 # [ >3.50(65) >3.59(53) >3.68(47) >3.77(37) >3.86(30)
181 # [ <3.50(45) <3.59(57) <3.68(63) <3.77(73) <3.86(80) ]
182
183 # check for results directory
184 if [ -d ${BASE}/results ] ; then
185 echo "${BASE}/results exists!"
186 else
187 mkdir -p ${BASE}/results
188 fi
189
190 # dataset type: 1, ascendant; 2, descendant
191 for (( type=1 ; type<3 ; type++))
192 do
```

```

193
194     if [ ${type} = 1 ] ; then
195         SAMPLING_CODE=ascD
196         tp=a
197         cr=">"
198         icr="<" # inverse criteria
199     elif [ ${type} = 2 ] ; then
200         SAMPLING_CODE=descD
201         tp=d
202         cr="<"
203         icr=">" # inverse criteria
204     else
205         echo "Exception caught! (type)"
206     fi
207
208     # 5 datasets
209     for (( d = 1 ; d < 6 ; d++ ))
210     do
211
212         if [ ${d} = 1 ] ; then
213             cut_point=3.50
214         elif [ ${d} = 2 ] ; then
215             cut_point=3.59
216         elif [ ${d} = 3 ] ; then
217             cut_point=3.68
218             p
219         elif [ ${d} = 4 ] ; then
220             cut_point=3.77
221         elif [ ${d} = 5 ] ; then
222             cut_point=3.86
223         else
224             echo "Exception caught! (5 datasets)"
225         fi
226
227         for (( rnd = 1 ; rnd < 101 ; rnd++ ))
228         do
229
230             # check the working directories
231             if [ -d ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd} ] ; then
232                 echo "${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd} exists!"
233             else
234                 mkdir -p ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}
235                 chmod 777 ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}
236             fi
237
238             # change to the working directory
239             cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}

```

A. Demo program

```
240
241             # execute the database query
242             PGSQL_QUERY
243
244         done
245
246     done
247
248 done
249 ;;
250
251 -gam)
252
253 for (( type=1 ; type<3 ; type++))
254 do
255
256     if [ ${type} = 1 ] ; then
257         SAMPLING_CODE=ascD
258         tp=a
259         cr=">"
260         icr="<" # inverse criteria
261     elif [ ${type} = 2 ] ; then
262         SAMPLING_CODE=descD
263         tp=d
264         cr="<"
265         icr=">" # inverse criteria
266     else
267         echo "Exception caught! (type)"
268     fi
269
270     # 5 datasets
271     for (( d = 1 ; d < 6 ; d++ ))
272     do
273         for (( rnd = 1 ; rnd < 101 ; rnd++ ))
274         do
275             echo "
276             #####
277             "
278             echo "##### ${type}RDo training
279             #####"
280             echo "#### The ${d}-${rnd} repeat processing
281             ....####"
282             echo "
283             #####
284             "
285             # check the working directories
286             # if working directory
287             if [ -d ${BASE}/${SAMPLING_CODE}${d}/${
288             SAMPLING_CODE}${d}${rnd} ]; then
289                 echo "${BASE}/${SAMPLING_CODE}${d}/${
290                 SAMPLING_CODE}${d}${rnd} exists!"
291             else
292                 echo "I will use it for modelling"
```

```

284         else
285             echo "You have to run preprocess first!"
286         fi
287
288         # create gam working directory
289         if [ -d ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}/gam ]; then
290             echo "${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}/gam exists!"
291             echo "I will use it for storing the gam results."
292         else
293             mkdir -p ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}/gam
294         fi
295
296         # go to the working directory
297         cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}/gam
298         #### 4. R execution
299         #####
300         cat > tr${SAMPLING_CODE}${d}${rnd}.R <<_EOF
301 # tr${SAMPLING_CODE}${d}${rnd}
302 # load grasper library
303 library(grasper)
304 # Data input
305 XXX <- read.table("../XXX", header=T, sep=",")
306 YYY <- read.table("../YYY", header=T, sep=",")
307 XXXpred <- read.table("$XXXpred", header=T)
308
309 # (1) set grasp options
310 OPT <- list()
311 OPT$TITLE <- as.character("R-GRASP: ")
312 OPT$LAYOUT <- eval(parse(text = as.character("c(3,3)")))
313 OPT$NBBARS <- as.integer(10)
314 #OPT$WEIGHTS <- as.character(WEIGHTS)
315 OPT$RESOLUTION <- as.numeric(1000)
316 OPT$SEP <- as.character(",")
317 print(OPT)
318 apply.ok <- TRUE
319
320 # (2) grasp import
321 grasp.in(YYY,XXX,XXXpred)
322
323 # select response
324 gr.Yi <- 2
325 # select predictors
326
327         selX <- c(4,5,6)
328 # grasp GAM model family=quasibinomial, F

```

A. Demo program

```
328 grasp.model(gr.Yi, trace=TRUE, df=4, calcdf=FALSE, stepfam = "
    quasibinomial()")
329 grasp.scope(gr.selX, df = 4, calcdf = FALSE)
330 grasp.step.gam(direction = "both", steps = 1000, trace = TRUE,
    limit = 0.05, test = "F")
331 grasp.pred()
332 grasp.pred.plot(gr.predmat, resolution = 1000)
333
334 grasp.ascii(gr.Yi, resolution=1000)
335 _EOF
336
337         R CMD BATCH tr${SAMPLING_CODE}${d}${rnd}.R
338
339
340
341         done
342     done
343 done
344
345
346
347     ;;
348
349 -maxent)
350
351     for (( type=1 ; type<3 ; type++))
352     do
353
354         if [ ${type} = 1 ] ; then
355             SAMPLING_CODE=ascD
356             tp=a
357             cr=">"
358             icr="<" # inverse criteria
359         elif [ ${type} = 2 ] ; then
360             SAMPLING_CODE=descD
361             tp=d
362             cr="<"
363             icr=">" # inverse criteria
364         else
365             echo "Exception caught! (type)"
366         fi
367
368         # 5 datasets
369         for (( d = 1 ; d < 6 ; d++ ))
370         do
371             for (( rnd = 1 ; rnd < 101 ; rnd++ ))
372             do
373
374                 # check the working directories
375                 if [ -d ${BASE}/${SAMPLING_CODE}${d}/${
                    SAMPLING_CODE}${d}${rnd} ]; then
376                     echo
```

```

377         else
378             echo "You have to run preprocess first! (
                option -p or --preprocess)"
379         fi
380
381
382         # create gam working directory
383         if [ -d ${BASE}/${SAMPLING_CODE}${d}/${
                SAMPLING_CODE}${d}${rnd}/maxent ]; then
384             echo ""
385         else
386             mkdir -p ${BASE}/${SAMPLING_CODE}${d}/${
                SAMPLING_CODE}${d}${rnd}/maxent
387         fi
388         # change to the working directory
389         cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}$
                {d}${rnd}/maxent
390
391         # preprocess the input file sample.csv
392         SWD_PREPROCESS
393         # maxent modelling
394         MAXENT_MODELLING
395
396         done
397     done
398 done
399 ;;
400
401 -vgam)
402     for (( type=1 ; type<3 ; type++))
403     do
404
405         if [ ${type} = 1 ] ; then
406             SAMPLING_CODE=ascD
407             tp=a
408             cr=">"
409             icr="<" # inverse criteria
410         elif [ ${type} = 2 ] ; then
411             SAMPLING_CODE=descD
412             tp=d
413             cr="<"
414             icr=">" # inverse criteria
415         else
416             echo "Exception caught! (type)"
417         fi
418
419         # 5 datasets
420         for (( d = 1 ; d < 6 ; d++ ))
421         do
422             for (( rnd = 1 ; rnd < 101 ; rnd++ ))
423             do
424

```

A. Demo program

```
425         # check the working directories
426         if [ -d ${BASE}/${SAMPLING_CODE}${d}/${
           SAMPLING_CODE}${d}${rnd} ]; then
427             echo ""
428         else
429             echo "You have to run preprocess/modelling
           first! (option -p or --preprocess)"
430             exit 1
431         fi
432
433
434         # create gam working directory
435         if [ -d ${BASE}/${SAMPLING_CODE}${d}/${
           SAMPLING_CODE}${d}${rnd}/gam ]; then
436             echo ""
437         else
438             echo "You cannot evaluate before modelling"
439             exit 1
440         fi
441         # change to the working directory
442         cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}$
           {d}${rnd}/gam
443
444         # sample with grass
445         GRASS_SAMPLE
446         EVALUATE
447
448     done
449 done
450 done
451 ;;
452
453 -vmaxent)
454     for (( type=1 ; type<3 ; type++))
455     do
456         if [ ${type} = 1 ] ; then
457             SAMPLING_CODE=ascD
458             tp=a
459             cr=">"
460             icr="<" # inverse criteria
461         elif [ ${type} = 2 ] ; then
462             SAMPLING_CODE=descD
463             tp=d
464             cr="<"
465             icr=">" # inverse criteria
466         else
467             echo "Exception caught! (type)"
468         fi
469
470         # 5 datasets
471         for (( d = 1 ; d < 6 ; d++ ))
472         do
```

```

473         for (( rnd = 1 ; rnd < 101 ; rnd++ ))
474         do
475
476             # check the working directories
477             if [ -d ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd} ]; then
478                 echo ""
479             else
480                 echo "You have to run preprocess/modelling
481                     first! (option -p or --preprocess)"
482                 exit 1
483             fi
484
485             # create gam working directory
486             if [ -d ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}/maxent ]; then
487                 echo ""
488             else
489                 echo "You cannot evaluate before modelling"
490                 exit 1
491             fi
492             # change to the working directory
493             cd ${BASE}/${SAMPLING_CODE}${d}/${SAMPLING_CODE}${d}${rnd}/maxent
494
495             # sample with grass
496             GRASS_SAMPLE
497             EVALUATE
498         done
499     done
500 done
501 ;;
502
503 -h|--help)
504     cat << EOF
505 Preprocess
506 =====
507 -p           preprocess the datasets
508             this step will create sampling datasets locate at ${BASE}, for example:
509             ${BASE}/ascD1, and prepare the prerequisite files for modelling and
510             validation.
511             Following file will be created:
512             ${SAMPLING_CODE}${dataset}raw.csv [raw data]
513             XXX [ predictor variables ---> for gam ]
514             YYY [ response variables ---> for gam ]
515             sample.csv [ sample with data ---> for maxent]
516
517
518 Modelling

```

A. Demo program

```
519 =====
520 -gam          generalized additive models
521 -maxent       maximum entropy principles
522
523 Evaluation
524 =====
525 -vgam         evaluate the gam results
526 -vmaxent     evaluate the maxent results
527
528 -h | --help  : display this help
529
530 EOF
531     exit 0
532     ;;
533
534 *)
535     echo "-p -maxent -gam -vgam -vmaxent or -h for more
           information"
536     exit 1
537     ;;
538
539 esac
```

Listing A.1: Scenario2-BA



Appendix B

NPMC results

Table B.1: Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in GAM of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo8); 2 indicates the dataset2 (RDo7-8); 3 is RDo6-8 4 is RDo5-8; 5 is RDo4-8; 6 is RDo3-8.

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.47020	0.3516111	0.5887889	1.0000000	0.9819211	N/A
2	1-3	0.55835	0.4410956	0.6756044	0.5175224	0.7119003	N/A
3	1-4	0.50285	0.3842914	0.6214086	0.9969894	1.0000000	N/A
4	1-5	0.49325	0.3744033	0.6120967	0.9997751	0.9999990	N/A
5	1-6	0.56655	0.4491364	0.6839636	0.3988447	0.5851660	N/A
6	2-3	0.58540	0.4688037	0.7019963	0.1756152	0.2882897	N/A
7	2-4	0.53845	0.4201327	0.6567673	0.8050105	0.9425640	N/A
8	2-5	0.52805	0.4085328	0.6475672	0.9086242	0.9869894	N/A
9	2-6	0.59635	0.4789512	0.7137488	0.1019124	0.1763965	N/A
10	3-4	0.44295	0.3252928	0.5606072	1.0000000	0.7362566	N/A
11	3-5	0.42745	0.3101454	0.5447546	1.0000000	0.4843283	N/A
12	3-6	0.50780	0.3888456	0.6267544	0.9920453	0.9999966	N/A
13	4-5	0.49405	0.3752672	0.6128328	0.9997093	0.9999997	N/A
14	4-6	0.57240	0.4548383	0.6899617	0.3218797	0.4891300	N/A
15	5-6	0.57880	0.4622213	0.6953787	0.2391153	0.3783997	N/A

B. NPMC results

Table B.2: Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in MAXENT of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo8); 2 indicates the dataset2 (RDo7-8); 3 is RDo6-8 4 is RDo5-8; 5 is RDo4-8; 6 is RDo3-8.

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.53100	0.4129206	0.6490794	0.8784263	0.9764527	N/A
2	1-3	0.55700	0.4394237	0.6745763	0.5386762	0.7342307	N/A
3	1-4	0.57235	0.4553479	0.6893521	0.3167277	0.4830739	N/A
4	1-5	0.54050	0.4226605	0.6583395	0.7744258	0.9244208	N/A
5	1-6	0.57790	0.4612967	0.6945033	0.2497722	0.3937401	N/A
6	2-3	0.52305	0.4048770	0.6412230	0.9383052	0.9942870	N/A
7	2-4	0.54145	0.4238174	0.6590826	0.7607830	0.9163710	N/A
8	2-5	0.50940	0.3912733	0.6275267	0.9887428	0.9999647	N/A
9	2-6	0.55145	0.4341314	0.6687686	0.6183859	0.8091645	N/A
10	3-4	0.51860	0.4001965	0.6370035	0.9613392	0.9981084	N/A
11	3-5	0.48805	0.3698263	0.6062737	0.9999660	0.9998379	N/A
12	3-6	0.52665	0.4084266	0.6448734	0.9142579	0.9884255	N/A
13	4-5	0.47020	0.3522769	0.5881231	1.0000000	0.9803350	N/A
14	4-6	0.50835	0.3901002	0.6265998	0.9905586	0.9999846	N/A
15	5-6	0.53890	0.4211841	0.6566159	0.7929353	0.9356845	N/A

Table B.3: Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets in GAM of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo3-8); 2 indicates the dataset2 (RDo3-7); 3 is RDo3-6 4 is RDo3-5; 5 is RDo3-4.

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.48135	0.36782433	0.5948757	0.9997811	9.976e-01	N/A
2	1-3	0.39405	0.28370192	0.5043981	1.0000000	6.793e-02	N/A
3	1-4	0.21420	0.12603469	0.3023653	1.0000000	4.773e-15	***
4	1-5	0.06225	0.01941706	0.1050829	1.0000000	0.000e+00	***
5	2-3	0.41370	0.30233027	0.5250697	1.0000000	2.166e-01	N/A
6	2-4	0.22975	0.13834774	0.3211523	1.0000000	1.459e-13	***
7	2-5	0.06955	0.02328466	0.1158153	1.0000000	0.000e+00	***
8	3-4	0.29965	0.19779913	0.4015009	1.0000000	1.180e-06	***
9	3-5	0.10255	0.04384222	0.1612578	1.0000000	0.000e+00	***
10	4-5	0.24260	0.14778631	0.3374137	1.0000000	8.252e-12	***

Table B.4: Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets in MAXENT of scenario 1. In column “cmp”, 1 indicates the dataset1 (RDo3-8); 2 indicates the dataset2 (RDo3-7); 3 is RDo3-6 4 is RDo3-5; 5 is RDo3-4.

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.49535	0.38101271	0.6096873	0.9979746	9.999e-01	N/A
2	1-3	0.36320	0.25426643	0.4721336	1.0000000	6.092e-03	**
3	1-4	0.31975	0.21483574	0.4246643	1.0000000	4.381e-05	***
4	1-5	0.17700	0.09489415	0.2591058	1.0000000	0.000e+00	***
5	2-3	0.37430	0.26537849	0.4832215	1.0000000	1.472e-02	*
6	2-4	0.33810	0.23237529	0.4438247	1.0000000	3.319e-04	**
7	2-5	0.19735	0.11124963	0.2834504	1.0000000	0.000e+00	N/A
8	3-4	0.47045	0.35663549	0.5842645	0.9999997	9.697e-01	N/A
9	3-5	0.31580	0.21126374	0.4203363	1.0000000	1.287e-05	***
10	4-5	0.33855	0.23055308	0.4465469	1.0000000	5.766e-04	**

Table B.5: Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in GAM of scenario 2. In column “cmp”, 1 indicates the ascD1 dataset; 2 is the ascD2; 3 is the ascD3; 4 is the ascD4 and 5 is the ascD5

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.51625	0.4030617	0.6294383	0.9285366	0.9958472	N/A
2	1-3	0.50050	0.3871368	0.6138632	0.9900539	1.0000000	N/A
3	1-4	0.45730	0.3442176	0.5703824	1.0000000	0.8395583	N/A
4	1-5	0.45610	0.3432678	0.5689322	1.0000000	0.8244191	N/A
5	2-3	0.47940	0.3661877	0.5926123	0.9999485	0.9887558	N/A
6	2-4	0.43565	0.3235118	0.5477882	1.0000000	0.5142199	N/A
7	2-5	0.44150	0.3291592	0.5538408	1.0000000	0.6079554	N/A
8	3-4	0.46100	0.3480062	0.5739938	1.0000000	0.8793439	N/A
9	3-5	0.45990	0.3469964	0.5728036	1.0000000	0.8676399	N/A
10	4-5	0.49510	0.3813400	0.6088600	0.9964294	0.9999946	N/A

B. NPMC results

Table B.6: Results of the multiple Behren-Fisher-Test in ascendant accumulative RDo datasets in MAXENT of scenario 2. In column “cmp”, 1 indicates the ascD1 dataset; 2 is the ascD2; 3 is the ascD3; 4 is the ascD4 and 5 is the ascD5

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.45525	0.3427027	0.5677973	1.0000000	0.81280593	N/A
2	1-3	0.49520	0.3819564	0.6084436	0.9967566	0.99999390	N/A
3	1-4	0.38970	0.2803226	0.4990774	1.0000000	0.04787775	N/A
4	1-5	0.45030	0.3376270	0.5629730	1.0000000	0.74614759	N/A
5	2-3	0.53745	0.4247303	0.6501697	0.7154295	0.89498942	N/A
6	2-4	0.43910	0.3270473	0.5511527	1.0000000	0.56935505	N/A
7	2-5	0.49420	0.3811446	0.6072554	0.9974066	0.99997815	N/A
8	3-4	0.39660	0.2869319	0.5062681	1.0000000	0.07621253	N/A
9	3-5	0.45440	0.3417848	0.5670152	1.0000000	0.80238462	N/A
10	4-5	0.55700	0.4447511	0.6692489	0.4360458	0.63248874	N/A

Table B.7: Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets of scenario 2. In column “cmp”, 1 indicates the descD1 dataset; 2 is the descD2; 3 is the descD3; 4 is the descD4 and 5 is the descD5

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.63250	0.5233923	0.7416077	4.577828e-03	8.714e-03	**
2	1-3	0.67190	0.5662663	0.7775337	5.016986e-05	7.371e-05	***
3	1-4	0.69250	0.5887779	0.7962221	1.069885e-05	3.173e-06	***
4	1-5	0.70060	0.5982262	0.8029738	8.102625e-07	8.468e-07	***
5	2-3	0.55030	0.4378501	0.6627499	5.260283e-01	7.415e-01	N/A
6	2-4	0.57710	0.4652972	0.6889028	1.940149e-01	3.248e-01	N/A
7	2-5	0.58340	0.4722328	0.6945672	1.407128e-01	2.421e-01	N/A
8	3-4	0.51790	0.4043139	0.6314861	9.156844e-01	9.952e-01	N/A
9	3-5	0.53680	0.4237044	0.6498956	7.191173e-01	9.060e-01	N/A
10	4-5	0.51045	0.3971580	0.6237420	9.587799e-01	9.997e-01	N/A

Table B.8: Results of the multiple Behren-Fisher-Test in descendant accumulative RDo datasets in MAXENT of scenario 2. In column “cmp”, 1 indicates the ascD1 dataset; 2 is the ascD2; 3 is the ascD3; 4 is the ascD4 and 5 is the ascD5

	cmp	effect	lower.cl	upper.cl	p.value.1s	p.value.2s	significance
1	1-2	0.45525	0.3427797	0.5677203	1.0000000	0.81344993	N/A
2	1-3	0.49520	0.3820339	0.6083661	0.9967753	0.99999389	N/A
3	1-4	0.38970	0.2803975	0.4990025	1.0000000	0.04654412	N/A
4	1-5	0.45030	0.3377041	0.5628959	1.0000000	0.74733016	N/A
5	2-3	0.53745	0.4248074	0.6500926	0.7141939	0.89351607	N/A
6	2-4	0.43910	0.3271240	0.5510760	1.0000000	0.56919599	N/A
7	2-5	0.49420	0.3812220	0.6071780	0.9974271	0.99997816	N/A
8	3-4	0.39660	0.2870069	0.5061931	1.0000000	0.07578694	N/A
9	3-5	0.45440	0.3418619	0.5669381	1.0000000	0.80190561	N/A
10	4-5	0.55700	0.4448279	0.6691721	0.4357837	0.63194231	N/A