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

資料探勘技術應用於 Maxent 物種分布模式之變數篩選

—以台灣鐵杉為例

Applying data mining approach to variable selection for Maxent: Taiwan Hemlock case study

研究生：姚強

By：Yao, Chiang

指導教授：邱祈榮 博士

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

中華民國 97 年 6 月

June, 2008

Acknowledgement

終於到了寫謝誌這一天，除了摘要之外不用再寫英文了，想當初在剛開始著手論文寫作時，一心想往著寫到這一頁，現在的心情還算平靜，只是離論文繳交日期也不遠矣，在我進入研究所的期間，承蒙邱老師祈榮的熱情與細心的灌溉、學術領域市場分析、小組討論與論文指導，惠我良多，使我從大學時代的提問題世代轉變成研究所教育後的具有發現和解決問題的本事，邱老師除了在學術上的指導外，在日常生活中的生活態度更是值得我學習的待人接物處事良方，這本論文我想信只是我這一年半年所學到東西的一小部分，在此特別感謝邱老師。另外要感謝在我論文寫作其間，邱老師旗下的六人植群小組也是讓我有機會和大家分享、報告、討論激盪的夥伴，學習到以不同的方式來分系資料，包括帶頭的宋叔叔、建融、政道、睿涵和世鐸，尤其是政道在我跑 30 次重複取樣的機率抽取上幫了大忙，雖然很不幸的那些結果最後沒有放到論文上，但是那些學習過程仍是值得的。另外還要感謝我們邊遠地帶的 204 同學們珍汝、小花、莉坪、昱光、麻吉、依霖，讓我在電腦前面並不感覺孤單，因為不式只有我一個人在趕報告。感謝我父母的養育之恩才使得我有今天的論文產出，感謝台大拳修社讓我在寫作不順時有地方休閒。最後，好酒藏甕底，因為怕寫在前面會讓人無法堅持看到最後啦，我要感謝我最心愛的女友美瑤，她在我論文寫作期間容忍我不修邊幅的外貌，情緒低落時的脾氣和種種細數不盡的抱怨或白目行為，並且為被論文占去陪妳的時間而感到抱歉，雖然我這段期間好像心中只有論文似的，少了些甜言蜜語和浪漫氣氛，成天只有與 R 和統計為伍，每個禮拜固定要 meeting，準備報告資料，讓妳好不是滋味，但是我始終把你放在心裡的第一位，在我心中你永遠是我最美好的瑤兒，沒有你的陪伴我也無法也無因走到這一段，在這論文即將完成之際，感謝妳也委屈妳這段日子，我終於要解脫回到你身邊啦！

此論文由衷的獻給關心我愛護我祝福我的人們

姚強

2008/7/28

摘要

由於人為活動增加過量的溫室氣體，導致氣候變遷下環境也可能發生改變，在未知的變動下植群社會要如何面對氣候變遷所造成的衝擊，了解植物社會是如何適應自然環境，將是首要的任務。近年來物種分布模式(Species Distribution Models, SDM)被廣泛的使用在了解物種與環境之間的關係，並且應用在生物多樣性保育與經營上。本研究的目標物種為台灣鐵杉(*Tsuga chinensis* var. *formosana* Li and Keng)出現樣點，以 16 個環境因子(包括大尺度的氣候因子與中尺度的地形因子)為 Maxent 物種分布預測模式的輸入，並測試三種不同的輸入各是如何影響預測模式的表現：(1)以種成分分析法(principal component analysis, PCA)與分類樹(classification and regression tree, CART)和條件推論樹(conditional inference tree, CIT)分析種環境因子與台灣鐵杉的關係當作預測模式環境因子選擇的依據，(2)比較所有台灣鐵杉出現的樣點數與以矩陣群團分析法分類之台灣鐵杉次植群型單位的樣點數，(3)不同的環境因子解析度。並分析植群與優勢物種分布和環境因子對模式的貢獻程度，進一步以 Maxent 物種分布模式預測出機率分布圖，預測之結果以受試者工作特徵曲線面積(AUC)值來評估台灣鐵杉植群型分布模式的準確性。應用 2 種合併模式的方法結合機率模式的結果與門檻值的篩選產生台灣鐵杉的潛在植群圖(potential vegetation map)並以誤差矩陣(confusion matrix)來評估潛在植群圖的準確性。植群分析結果產生四群台灣鐵杉次植群型，環境分析和模式預測結果顯示影響台灣鐵杉的空間分布為主要的環境因子為海拔，次之為雨量，都屬於氣候因子；地形因子及對預測模式沒有主要的貢獻，但是仍然使預測模式更加精確。樣點數較小較且均質的植群型單位模式有著比樣點數較多的物種單位模式還高的模式預測能力。本研究中環境圖層的解析度對模式的預測能力沒有特別顯著的影響，預測的區域因為受到樣本數跟著改變的影響來無法突顯預測範圍的大小是否影響模式的表現，潛在植群圖的合成有助於應用的決策和考量，使得物種分布模式的應用更具有彈性。最後預測植群圖的可適用性能需要進一步的實驗預測的環境條件是否真的是和目標物種的生存來加以支持預測物種的空間分布。

關鍵字：台灣鐵杉、分類樹、條件推論樹、Maxent、AUC、誤差矩陣、潛在植群。

Abstract

To know the adaptation of plant society under climate change impacts is based on knowledge of the potential distribution of vegetation distributions. Vegetation is a society of plant species. Applying combination of species distribution models (SDMs) results to establish potential vegetation maps (PVMs) need determination strategies. This article firstly analyzes the relationship between Taiwan Hemlock (*Tsuga chinensis* var. *formosana* Li and Keng) and 16 topographical and climatic variables and then to generate a probability map by Maxent to test how 3 different situations of model input affects the model performance: (i) selection and analysis of suitable environmental variables by principal component analysis (PCA), classification and regression tree (CART) and conditional inference tree (CIT) method, (ii) sample size and homogeneity of species and vegetation sub-unit occurrence data (iii) resolution for environmental layers. Model evaluated by area under receiver-operating characteristic (ROC) curve (AUC) and Kappa statistic. 2 model combination approaches is also applied in this study to aid to generate the potential vegetation map (PVM) of Taiwan Hemlock. PVM is evaluated by error matrix and its derived indices. The result of vegetation analysis by cluster analysis classified Taiwan Hemlock into 4 sub-unit vegetation type. The result of environmental analysis and modeling revealed that the environmental variable that is affecting spatial distribution of Taiwan Hemlock most is majorly elevation gradient and the secondary is precipitation and both are climatic variables. Topographical showed minor contribution to the model. Sample size test showed more accurately when input the smaller size and more homogeneous samples. Resolution of environmental layers showed no sigibificant effect on model performance in this case. Overlaying Taiwan Hemlock vegetation sub-unit probability maps with 2 deterministic combination approaches synthesizes a potential vegetation map of Taiwan Hemlock. Modification of strategy for predicting PVMs is according to local ecological theory and further study on testing the potential ability from the environmental variable is really suitable for the target species.

Keywords: *Tsuga chinensis*, CART, CIT, Maxent, AUC, confusion matrix, PVM.

Table of Content

Table of Figures

Table of Tables

Chapter 1: Introduction

1.1 Background

The question about plants and animals' current distribution is discussed for a long history and makes many ecologists find the explanation. Many modelers root in species-environment relationships to establish many modeling approaches for solving this question (Guisan and Thuiller, 2005). Analysis of the species' geographic distribution has always been an important issue in vegetation science, and is currently focused by other sub-disciplines such as biogeography and landscape ecology. The relationship between environmental gradients and vegetation distribution is one of the most important issues examined in vegetation science (Miller *et al*., 2007). The ability to quantify the relationship leads to predict potential distribution of vegetation and is applicable for predicting spatial distribution under changing environmental conditions, such as climate change occurring (Miller *et al*., 2007).

The purpose of potential vegetation field survey is to understand plant ecology and apply to ecological conservation, landscape restoration, and landscape planting (Yang, 1997). Survey data for estimating or predicting potential vegetation according to plant ecology, or plant geography, explain forest composition, structure, and function, further more, the relationship with environmental variables and its succession stage. Those data information allow us to predict the next succession, current or future distribution of species and vegetation and are useful for forest ecosystem management, biological conservation, and landscape restoration. On the contrary to traditional survey analysis, Chiou *et al*. (2006) introduced GIS technique and several models are rapidly developed in recent years. Combining vegetation mapping and analysis of satellite images with GIS generate predictive vegetation model, a new approach for analysis species/vegetation and environmental variables (such as Gu *et al.,* 2006; Tsao, 2007; Yen, 2007) (Franklin, 1995; Guisan and Zimmermann, 2000; Scott *et al*., 2002) and this study is also based on the new approach.

Climate change has become an important focused issue in recent years, as a basis for assessing whether anthropogenic greenhouse effect has enhanced climate change and how the continuingly growing greenhouse gas concentration will lead to an unknown future climate. According to the IPCC's Third Assessment Reports (TAR, IPCC, 2001), the average temperature of global surface has increased by about 0.6 °C in the past century, and to the IPCC's Forth Assessment Report (AR4), warming in the last 100 years has increased by 0.74 °C in global average temperature. This is above the 0.6 °C increase in the 20th century prior to the Third Assessment Report (IPCC, 2007). Taiwan has been moving toward a warmer and drier climate. Enhanced precipitation is observed in the limited areas in the limited times, however, a systematic trend (or change) is not observed (Hsu, 2002). In general, under constant climate change and global warming conditions possibly forces the distribution area of current vegetation diminishing, and increases the risk of species extinction. To predict the change of distribution of species under different climate scenarios is essential to assess the risk of species extinction under climate change (Thomas *et al*., 2004). Thus, the first mission is to establish predictable models for species potential distribution range (Tsao, 2007).

Previous ecologist in Taiwan mainly focused on classifying the plant communities and identifying the relationships between plant societies and environmental variables (i.e. Su, 1984a; 1984b). Island of Taiwan has a great diversity of fauna and flora due to a high degree of topographical complexity and an about 4000 m variation in altitude (elevation). A large proportion of Taiwan is not easy to access for field survey due to its hilly topography. Available field data might not be completely enough to support decision making of conservational or environmental policies (Song *et al.*, 2007). Species distribution models provide a possible way to fill up the gap of incomplete vegetation data (Franklin, 1998).

Projections of species distribution under climate and environmental change are of great scientific and social relevance, and basing on species distribution models (SDMs) make some assumptions such as species not adopt to global dispersal in evolution and consistency of limiting factors (Dormann, 2007). Although some of the assumptions are ecologically untenable, the predictions of the SDMs are still a useful reference to policy maker for climate change impact assessment and conservation management. This study examines the relationships between distributions of dominant species, Taiwan Hemlock (*Tsuga chinensis* (Franch.) Pritz. ex Diels *var. formosana* (Hayata) Li and Keng) of alpine forest in Taiwan, and climatic and topographical environmental variables. Not only traditional statistical approaches, but importance of new direction of data mining approaches in analyzing relationships between species and environmental factors will lead more precise insight and performance on SDMs prediction.

1.2 Objective

Although SDM are widely spread in many fields, the analysis of relationship in species-environment is still methodologically not well organized. This study will focus on comparing statistical and data mining methods to find the suitable environmental layers for building the spatial distribution of Taiwan Hemlock vegetation. Additionally, some studies of SDM application can be found in Taiwan (Gu *et al.*, 2006; Tsao, 2007; Yen, 2007) but fewer studies in Taiwan considers the comparison of difference model setting (like Song, 2007). Smaller grid size of the predicted background can reflect the more detail of topographical variables than climatic variables but is time consuming due to a very large data size for model estimation. Contrarily, larger grid size of background is much faster when calculating but lacks of or reduces detail information of the meso-environmental variables. Thus, this study also takes grid size, predicted area, locality units, and environmental selection of the model input into account.

Pervious studies for SDMs (mentioned later in Ch. 2 literature review) were using multiple model combination to improve the predictive accuracy for each species' spatial distribution, however, combining different hierarchy units of species and vegetation to increase the predictive accuracy are seldom seen in resent researches. Therefore, comparing model techniques combination and species-vegetation unit combination is to compare and combine different SDMs to synthesize potential vegetation map of Taiwan Hemlock another goal of this study. The combination criteria are utilized to determine the binomial potential distribution area that Taiwan Hemlock may occur. There are 4 main objectives listed as follow:

1. Using data mining approach (classification and regression tree, CART and conditional inference tree, CIT) compared to statistical approaches (detrened correspondence analysis, DCA, principal component analysis, PCA and correlation analysis, CA) for analysis of relationship between species distributions and environmental variables.
2. Assess how localities inputs of different vegetation and species based sub-units on distribution modeling relate to environmental variables.
3. Evaluate how difference grid resolution affects the model performance and relationship between target species and environmental variables.
4. Model combination and synthesis of potential nature vegetation maps.

Chapter 2: Literature review

2.1 Climatic Factor and Vegetation Distribution in Taiwan

Alpine ecosystem's unique characteristic environment, such as strongly wind blowing, shallow soil, low temperature, snow cover, is unsuitable for growing and is sensitive to climate change (Luckman, 1990; Walther, 2004). Therefore, monitoring alpine ecosystem for climate change impact assessment on alpine forest is a very important approach worldwide (Luckman, 1990).

Vegetation Zone differentiation in Taiwan changes along with altitude gradient and variation of vegetation in Taiwan is specified by vegetation zones or vegetation biomes divided by different elevation (Liu, 1962; Su, 1984b; Su, 1992). Su (1984a; 1984b) described the relationships between vegetation zones and climate factors and tried to divide the range of vegetation distribution by temperature factor. Su (1984b) investigated vegetation of Chou-Shui river basin in mid-Taiwan and established the relationships between elevation and temperature by regression analysis and determined the up and low limits of elevation for every vegetation zone. Su's vegetation zone (1984b) is a high hierarchy classification unit and each zone may contain different forest types due to differentiation of topography, soil, or succession stage.

Many ecologists considered latitude and elevation are the main factor to species distribution (Kellman, 1980; Su, 1985; Su, 1987; Guissan *et al.*, 1998, Liang, 2004; Yen *et al.*, 2007). Previous studies on vegetation science in Taiwan restricted to financial support and thus most vegetation survey sites limited within local area like boundary of watershed, naturally conserved area, or administrative area, such as Su’s (1988) study site in Single-seed Juniper conserved area, Su (1984b) in Chou-shui river basin, Liu *et al.* (1999) in Sha Li Shian watershed, and Fu (2002) in Dan-Da area. Those local studies with varied purposes, methods, and location, which boundary doesn’t represent the real boundary of species’ distribution, hardly integrate the vegetation distribution in Taiwan (Chiou *et al.*, 2006).

Chiou *et al.* (2006) analyzed the distribution characteristic of Taiwan Hemlock community in Taiwan using cluster analysis and compared with environmental conditions, altitude, latitude, and warmth index. Two methods of the comparison are order and inter-specific association. Yen *et al.* (2007) and Tsao (2007) firstly introduced the model techniques for modeling species distribution over the whole Taiwan Island. Yen *et al.* (2007) used second-order logistic regression in generalized linear model (GLM) to estimate the probabilistic distribution of Taiwan Hemlock in Taiwan by two of the most important environmental variables, latitude and altitude. Tsao (2007) used generalized additive model (GAM) to establish the relationships between distribution ranges and environmental variables for six conifer species, *Chamaecyparis obtusa* var. *formosana*, *Chamaecyparis formosensis*, *Abies kawakamii*, *Tsuga chinensis*, *Picea morrisonicola*, and *Pinus taiwanensis*, of Taiwan. Tsao’s result shows all of the six GAM models select the variable of mean annual temperature for building model, in other words, distribution of six plant species is affected by mean annual temperature. Annual precipitation, however, is not selected by any of the six models. Therefore, the probable explanation may be Taiwan Island receives abundant precipitation all year round, so precipitation is not a limiting factor for vegetation distribution in Taiwan (Kuo, 1978; Yen, 2007).

2.2 Data Mining Approach in Environmental Factor Analysis

2.2.1 CART

Classification and regression tree (CART) is a kind of decision tree. Breiman *et al.*'s (1984) CART is a common basis for some ensemble procedures such as bagging (Breiman, 1996), random forest (Breiman, 2001), and stochastic gradient boosting (Friedman, 2001a). Kriegler (2007) mentioned four key aspects of CART are (i) Splitting criteria for regression tree, (ii) Pruning and knowing when to stop making splits, (iii) Costs and the relation to priors, and (iv) Obtaining fitted values. CART can handle both numeric (regression tree) and categorical (classification tree) predictor and response variable.

De'Ath and Fabricius (2000) described CART is ideally suitable for analyzing the complex ecological data, which is usually strongly non-linear, involving higher order iteration, unbalanced, and containing missing values. Because CART is flexible and robust analytical methods and can handle for such complex data. Furthermore, CART results are simple to understand and easily interpretable by its graphical representation with root (undivided data) at the top and branches and leaves (final groups) underneath. The explanation for variation of a single response variable for trees is using one or more explanatory variables (continuous/categorical) to repeatedly split the data into more homogeneous groups, which is defined by a single rule based on single explanatory variable, but to remain the tree reasonably small. Splitting is continued until an overlarge tree is grown and then pruning reshapes it back to the desired size. Each group/leaf is characterized by a typical value of the response variable (mean value for numeric response and distribution for categorical response), the number of observations in the group, and the values of the explanatory variables that define it. The tree is represented graphically, and this aids exploration and understanding.

Trees are interactive exploration and both descriptive and predictive for patterns and processes. De’Ath and Fabricius (2000) stated advantages of CART including: (i) the flexibility to handle different types of response variables, numeric, categorical, ratings, and survival data; (ii) invariance to monotonic transformations of the explanatory variables; (iii) the capacity for interactive exploration, description, and prediction; (iv) ease and robustness of construction; (v) ease of interpretation by graphical representation; and (vi) the ability to handle missing values in both response and explanatory variables. Therefore, CART is an alternative to many traditional statistical techniques, such as multiple regression, analysis of variance, logistic regression, log-linear models, linear discriminant analysis, and survival models for complement or representation (De’Ath and Fabricius, 2000).

2.2.2 CIT

CIT is an abbreviation of conditional inference tree, which can deal with recursive partitioning for continuous, censored, ordered, nominal and multivariate response variables and the implementation utilizes a unified framework for conditional inference developed by Strasser and Weber (1999). Hothorn *et al.* (2006) described conditional inference framework for recursive binary partitioning can be solve two fundamental problems of exhaustive search procedures, (i) over fitting and (ii) a selection bias towards covariates with many possible splits or missing values, by (i) pruning procedures and (ii) embedding tree-structured regression models into a well defined theory of conditional inference procedures, based on invariant p-value.

Roughly, the algorithm of CIT works as follow steps: (i) Test the global null hypothesis of independence between any of the explanatory and the response variables and stop if this hypothesis cannot be rejected (i.e. the explanatory and response variables in a specific splitting node are not independent to each other). Otherwise select the explanatory variable with strongest association to the response variable with measuring a p-value corresponding to a test for the partial null hypothesis of a single explanatory variable and the response variable. (ii) Split binaurally in the selected explanatory variable. (iii) Recursively repeat steps (i) and (ii).

The stop criterion in step (i) is either based on multiplicity adjusted or univariate p-values and it is shown that the predictive performance of the resulting trees is as good as the performance of established exhaustive search procedures (Hothorn *et al.,* 2006). This statistical test ensures that the right side of the tree is grown and no form of pruning or cross-validation. The selection of the explanatory variable to split in is based on the univariate p-values preventing a variable selection bias from explanatory variables with too many possible splitting points. Moreover, the prediction accuracy of trees with early stopping is equivalent to the prediction accuracy of pruned trees with unbiased variable selection.

2.3 Ecological Niche and Species Distribution Model

2.3.1 Predicted Vegetation Modeling

Maps of vegetation composition have traditionally been produced by field survey and photo interpretation, but these methods are costly and inefficient. Predictive vegetation modeling (PVM) can be defined as predicting the distribution of vegetation across a landscape based on the relationship between the spatial distribution of vegetation and relevant environmental variables (Franklin, 1995). Fraklin (1995) provided the relationship between environmental variables and their process affecting distribution of potential natural and actual vegetation (Figure 1).

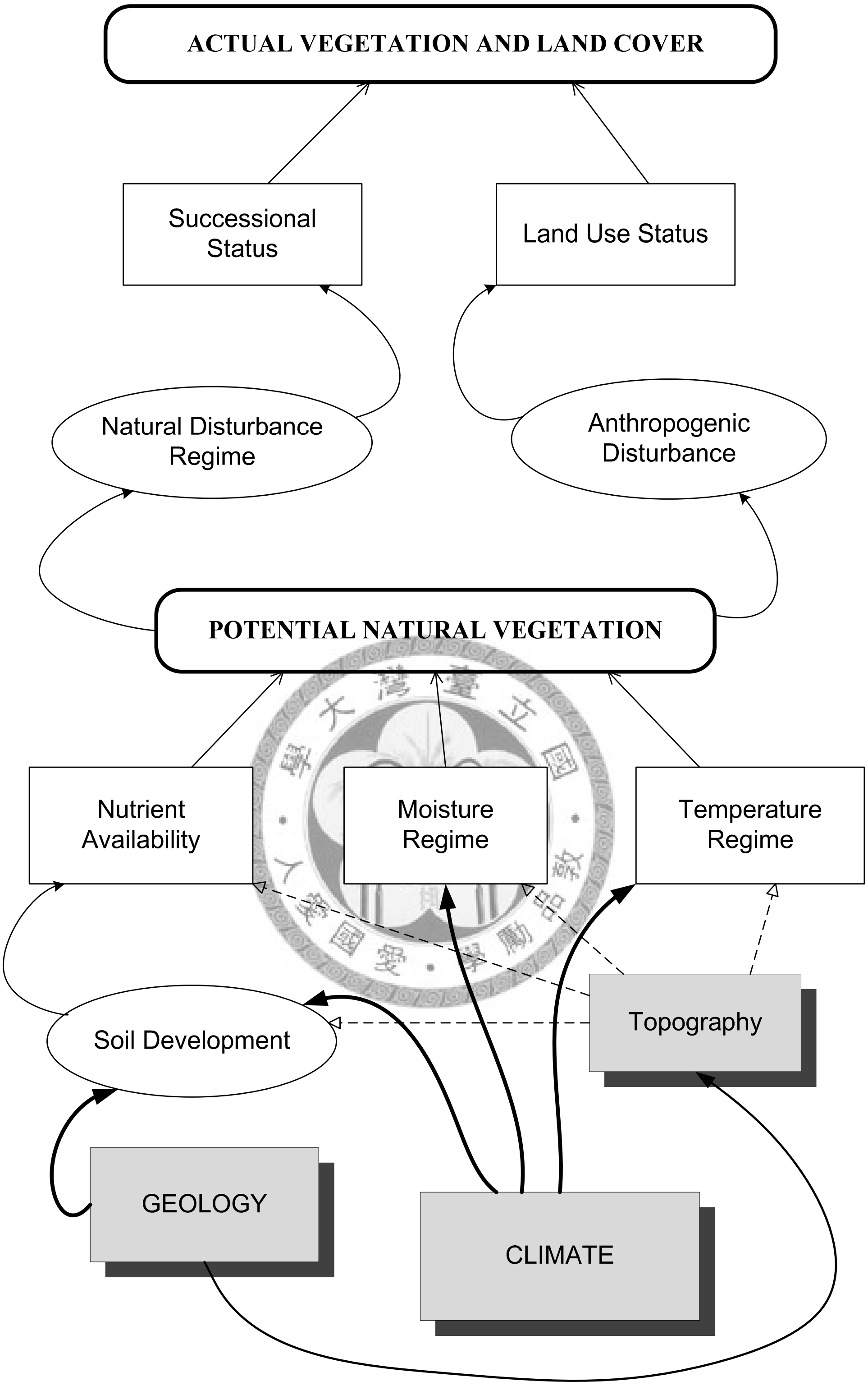


Figure 1. Conceptual model showing relationships and processes between climatic determinants, direct gradients, potential natural vegetation and actual vegetation (revised from Franklin, 1995)

PNV is determined by environmental determinants (climatic, geographic, and topographic factors) which changes local soil nutrients, moisture, and temperature. After natural and/or anthropogenic disturbance, like competition, succession, land-use, actual vegetation and land-use type form as mosaics on landscape.

To evaluate potential vegetation, predicting vegetation mapping (PVM) is firstly considered. Three steps described by Franklin (1995) and Chen (1997) for PVM are (i) Traditional approach: explanation by aerial photos or numeric vegetation map from geographic mapping combines with geographic information system (GIS) for decision making. (ii) Numeric approach: establishment of mathematical relationship between environmental variables (such as temperature, precipitation, soil type) and traditional field survey records helps to understand the distribution of vegetation. (iii) Predicting vegetation mapping. Step 1 actually not real analyze the data as step 2 does, in the other words, the spatial pattern of species is not only simply digitizing the survey data, but also establishes the statistically or mechanistically mathematical relationship with environmental variables.

Development methodology of PVM traced to Kessell's (1976) series studies on connecting real spatial object with abstract spatial model by gradient modeling with GIS in Glacier National Park, USA; Box (1981) used empirical model to generate distribution of global vegetation with global plant communities and macro-scaled climatic variables. After this point plenty of relevant studies followed. Table 1 integrates the methods of recent 20 years studies for PVM.

Climate-Vegetation classification model, one of PVM, assumes that major vegetation of any site is the result of environmental factors and considers climatic variables playing an important role on it (Tuhkanen, 1980). Köppen (1931) and Holdridge (1967) established the classification of global vegetation and its life zone, Walter (2002) divided 9 zonal biomes in macro-scale mapping of vegetation, Chiu *et al.* (2005) mapped Holdridge's life zones at Taiwan. Ecologists also consider the dominant vegetation type is response by reaction of climatic variables and every climatic zone should have its vegetation type (Whittaker, 1975).

Table 1. Predicted vegetation modeling (PVM) techniques with continuous and categorical data (revised from Franklin, 1995; Chen, 1997)

Dependent variable	Independent variable		
	Continuous	Mixed	Categorical
Continuous	RM	ANCOVA	ANCOVA
	RT	MANCOVA	MANCOVA
	GLM	RT	RT
		GLM)	GLM
Categorical	MLC	MLC with priors	Contingency Table
	Logit (GLM)	Logit (GLM)	Logit (GLM)
	DA	GAM	GAM
	GAM	CART	CART
	CART	NN	NN
		GA	GA
		ES	ES

Notes: ANCOVA: Analysis of Co-Variance; CART: Classification and Regression Tree; DA: Discriminant analysis; ES: Expert System; GA: Genetic Algorithms; GAM: Generalized Additive Models; GLM: Generalized Linear Models; MANCOVA: Multivariate Analysis of Co-Variance; MLC: Maximum Likehood Classification; NN: Neural Network; RM: Regression Tree; RT: Regression Tree.

Chen (1997) described two scales for studying PVM. (i) Regional scale and (ii) Local scale. Study range of regional scale is considering global and continental area and the environmental variable is mainly climatic variable. In this scale the climate-vegetation model belongs to static model (Lowell, 1991) and is based on assumption of the equilibrium between distribution of vegetation and environmental variables. (Leniha and Neilson, 1993)，and this assumption is acceptable for larger area and loner time (Cramer and Leemans, 1993). Study range of local scale is smaller than regional scale such as a watershed and the environmental variables are selected for this scale such like topographic variable, slope, aspect, and soil type. In this scale, predicting of PVM is not only the static model, but also considers the topographic and micro-climate variables to project the active procedures of species' birth, growth, and death (Urban *et al.*, 1991). However, the complex background knowledge about climate-vegetation interaction is needed for the dynamic models (Brovkin, 2002). Unfortunately, few dynamic mechanisms of interaction of vegetation and ecosystem are well developed (Foley *et al.*, 1998).

Chen (1997) introduced three stage of model building in GIS: (i) establish spatial database, digitalizing the survey records and environmental variables for spatial analysis and further application in GIS. (ii) Set up the mathematical relationship between species and environmental variables. (iii) Combine the mathematical model and GIS tool and database to output and display the results.

Potential natural vegetation maps are applied to communicate the natural baseline conditions for assessing ecosystem health, predictions of vegetation distribution which is caused by responding to environmental factors to management, and determination of potential resource value (Jansen *et al.,* 2002). Danijela (2003) applied predictive vegetation model to manage and conserve the developed area based on the PNV conception. PNV concept is not only for vegetation mapping, but also for land-use development and as a potential and basic reference for describing and integrating ecosystems (Hardtle, 1995; Seibert and Conrad-Brauner, 1995). For example, PNV represents the climax stage of vegetation under the stable environmental condition and can be used as a guide for ecosystem restoration (Danijela, 2003). In Japan, many cities have their own actual and potential vegetation maps for land-use planning and integrate the PNV studies concluding which native species is better for planting on the area. (Miyawaki *et al*., 1987; Miyawaki and Fujiwara, 1988; Miyawaki, 1988). PNV is also applied in simulating global and local climate change impacts, such as Cha (1998) estimated potential change of forest area under 2 times CO_2 concentration.

2.3.2 SDM/ENM

Species distribution models (SDMs) or ecological niche-based models (ENMs) are two kinds of the PVM techniques and are empirical models relating field survey observations to environmental predictor variables, which is based on statistically or theoretically derived response layers (Guisan and Zimmermann, 2000). SDMs describes the spatial distribution of a species or species groups, as a function of environmental predictor variable such as latitude, longitude, altitude, climate, topography, land-use type, vegetation type, soil conditions, and so on. Species presence-only and presence-absence data are two major formats of the SDMs’ explanatory variables and the former data type is usually easier to obtain by historical records such as museum specimens, private collections, or field surveys. In other words, the potential species distribution model (PSDM) is developed from a set of environmental variables for a set of rasters, together with a set of data localities where the species are observed and predicts the suitability for the target species as functions of environmental variables. (Phillips *et al.*, 2006; Prates-Clark *et al.*, 2008).

Guisan and Thuiller (2005) describe three phases of SDMs by author's personal communication to S. Ferrier:

(i) Non-spatial statistical quantification of species–environment relationship based on empirical data,

(ii) expert-based (non-statistical, non-empirical) spatial modeling of species distribution.

(iii) Spatially explicit statistical and empirical modeling of species distribution.

Species distribution models (SDMs) of plants and animals are interested widely in the last two decades and applied many issues in ecology, biogeography, evolution and, in conservation biology and climate change research(Guisan and Thuiller, 2005), such as predicting species distributions from museum and herbarium records (Elith and Leathwick, 2007), predicting future range of species distributions under climate change impacts (Thomas *et al.*, 2004), mapping species ranges and species richness (Graham and Hijmans, 2006), predicting the invasive spread of a cactus species in Australia (Johnson, 1989) (quoted in Pearson and Dawson, 2003), assessing the climatic determinants of the distribution of several European species (Hengeveld, 1990) (quoted in Pearson and Dawson, 2003), enhancing a regional vegetation map (Franklin, 2002), biodiversity conservation (Rodríguez *et al.*, 2007) …etc. Table 2 shows the application of SDMs in ecology fields.

Table 2. Some application of SDMs in ecology fields (revised from Guisan and Thuiller, 2005)

Type of use	References
Quantifying the ecological niche of species	Austin *et al*. (1990), Peterson *et al.* (2002), Vetaas (2002), Sattler *et al*. (2007), Rissler and Apodaca (2007), Raxworthy *et al*. (2008)
Testing biogeographical, ecological and evolutionary hypotheses	Leathwick (1998), Anderson *et al*. (2002), Graham *et al*. (2004b)
Assessing species invasion and proliferation	Beerling *et al*. (1995), Peterson (2003), Sanchez-Flores (2007), Wang *et al.* (2007)
Assessing the impact of climate, land use and other environmental impacts on species distributions	Thomas *et al*. (2004), Thuiller (2004), Early *et al.* (2007), Dormann (2007)
Suggesting unsurveyed sites of high potential of occurrence for rare, endemic, threatened species	Elith and Burgman (2002), Raxworthy *et al.* (2003), Engler *et al.* (2004), Zimmermann *et al.* (2007)
Supporting appropriate management plans for species recovery and mapping suitable sites for species reintroduction	Pearce and Lindenmayer (1998)
Supporting conservation planning and reserve selection	Ferrier (2002), Arau´jo *et al.* (2004), Pape and Gaubert (2007
Modeling species assemblages (biodiversity, composition)/vegetation from individual species predictions	Guisan and Theurillat (2000), Cairns (2001), Ferrier *et al.* (2002), Graham and Hijmans (2006), Rodrguez (2007), Saatchi *et al.* (2008)
Predicting distribution of high value trees	Prates-Clark *et al.* (2008)
Vegetation mapping support	Scott *et al.* (2001), Franklin (2002), Cawsey *et al.* (2002), Tatsuhara and Antatsu (2007)

ENM is slightly difference in the definition to SDM. Models of ecological niches are designed to estimate the potential niche's area of the target species, and thus ENM predicts broader range than actual distribution (Phillips *et al.*, 2006; Peterson *et al.*, 2008).

James and McCulloch (1990) stated all parametric statistical models face to the problem with highly non-Gaussian distribution data such as most environmental variables. Stocktwell (2006) described:

"The ideal ENM method will (1) be capable of modeling a wide range of responses, (2) allow critical examination of assumptions, (3) be a simple approach that will not fit inappropriate functions, but (4) will handle extremely non-linear data, and (5) will efficiently turn an increasing flood of data from satellites, geographic information systems and climate model outputs into simple, scalable ENMs."

2.3.3 SDMs and Ecological Theory

Niche based models like some of SDM or ENM (Maxent, GARP, GAM …etc) representing the approximation of species' ecological niche in the examined environmental layers (Phillips *et al.*, 2006). ENM is based on the idea of ecological niches defined as the set of conditions under which a species is able to maintain populations without immigration (Grinnell, 1917; 1924; Hutchinson, 1957; Hutchinson, 1978; and Austin *et al.*, 1990). The ecological niche includes the fundamental niche, which consists of a set of conditions for species' long-term survival, and realized niche, which is subset of fundamental niche for species' actual occupation (Hutchinson, 1957).

Therefore the realized niche of a species may be smaller than its fundamental niche due to disturbances from human influence, biotic interaction (such as competition), geographic barriers, and/or natural disasters, and such factors are influential to its survival range and prevent the species from fully spreading its ecologically potential niche (Pulliam, 2000; Anderson and Mart´ınez-Meyer, 2004; Phillips *et al.*, 2006). Thus niche based models estimate the approximation of species' realized niche in environmental layers considered, however, the departure between realized and fundamental niche remains unknown in practice (Phillips *et al.*, 2006). Realized niche can be estimated by removing areas that species is known or inferred not to inhabit from the predictive distribution such as areas suitable for the target species without colonized due to geographic barriers (Peterson *et al.*, 1999; Anderson, 2003), biotic interactions (Anderson *et al.*, 2002), and human influences (Anderson and Mart´ınez-Meyer, 2004).

Phillips *et al.* (2006) described the ecological assumption of environmental variables used for modeling are (i) temporal correspondence, (ii) scale, (iii) space and time. Temporal correspondence will be existed when using locality record that investigated very long time age for current land-cover classification (Anderson and Mart´ınez-Meyer, 2004). Mackey and Lindenmayer (2001) defined environmental variables for different scale: (i) global and meso-scales: climatic variables such as temperature and precipitation, (ii) meso- and topo-scales: topographic variables such as elevation and aspect, and (iii) micro-scales: land-cover variables such as forest canopy. Su (1983) also introduced the classification of factors affecting species habitat: (i) direct/indirect factors, (ii) scales, (iii) affection, and (iv) sources. For source classification environmental variables influence habitat divided into 4 categories:

(i) Climatic factors: such as radiation, air temperature, precipitation, wetness.

(ii) Edaphic factors: also called soil factors, such as soil type, soil temperature.

(iii) Physiographic factors: also called topographic factors, such as aspect, altitude, slope, curvature.

(iv) Biotic factors: such as anthropogenic or biotic interactions or disturbances.

SDM has applied in Taiwan vegetation science for just a few years. Song *et al.* (2007) compared the model performance of three SDM techniques, Maxent, GARP, and GAM, by evaluating sensitivity, specificity, and area under receiver operating characteristic (ROC) curve. Tsao (2007) used GAM to establish the relationships between distribution ranges and environmental variables for six conifer species, *Chamaecyparis obtusa* var. *formosana*, *Chamaecyparis formosensis*, *Abies kawakamii*, *Tsuga chinensis*, *Picea morrisonicola*, and *Pinus taiwanensis*, of Taiwan.

2.3.4 MAXENT

Maxent program for maximum entropy based machine-learning modeling technique predicts species geographical distributions and is firstly introduced by Phillips *et al.* (2005). Maxent model’s estimation is based on a decision theoretic perspective as robust Bayes estimation (Phillips and Dudı´k, 2008) and simulates predictions from data with incomplete information to estimate a probability distribution by finding the probability distribution of maximum entropy (Della Pietra *et al.*, 1997) (i.e. the Maxent approach assumes that the occurrence data of incomplete empirical probability distribution can be approximated with a probability distribution of maximum entropy subject to environmental layer’s constraints, and use this approximated distribution for predicting a species potential geographic distribution (Phillips *et al.*, 2005). Phillips and Dudı´k (2008) described the Maxent model uses the species' occurrence data to define the region of probability with maximum entropy. The probability distribution π over the set X of plots is non-negative value and the sum of π(x) is one, where the x is the sample of the population X. The π is displayed in terms of "gain"—the log (the number of rasters) - the log (loss) (i.e. the average of the negative log (probabilities of the sample locations) (Prates-Clark *et al.,* 2008) and coincides with the potential distribution stated by biologists (Phillips *et al.*, 2004). The simple function of environmental variables are a set of real-valued variables and called features, and the constraints are the mean of predictive features required to be near the empirical average over the occurrence sites (Phillips *et al.*, 2006).

Initially, each environmental variable is treated as potentially an important predictor variable to develop the model. Jackknife test re-sampling method (Peterson and Cohoon, 1999) of Maxent's internal procedures reduces the bias of correlated environmental variables and to diagnose which environmental variables were the most important variables for building models. The environmental variables with the highest gain means higher the relative importance of variables that potentially, contribute to generating the SDM (Phillips *et al.*, 2004).

Maxent displays the influence of each environmental variable in response curve diagrams. As the Maxent model is an exponential model (Della Pietra *et al.*, 1997), the probability of prediction is proportional to the exponential contribution of each environmental variable (Phillips *et al.*, 2006). The response curves in version 3.2.1 are in logistic (probability) space, rather than exponent (linear) space, so they're easier to interpret. Statistical approaches for evaluating model performance such as dependent omission rate and independent AUC of ROC analysis are also including the internal procedures of the Maxent. Some other features of Maxent 3.2.1 can visit Maxent website (http://www.cs.princeton.edu/~schapire/maxent/) for more information. Maxent with pros and cons were reviewed by some study. The advantages of Maxent include the usage of both categorical and continuous environmental data (Prates-Clark *et al.*, 2008).

2.4 Model Performance Evaluation

Model performance can be evaluated by the accuracy of model predictions, the interpretability and rationality of the explanatory variables, and the validity of predicted shape of response curves (Pearce and Ferrier, 2000). A good prediction includes both reliable and discriminatory prediction. Reliable prediction means the accurate estimation of probability for a species' occurrence site and discriminatory prediction means the ability to discriminate the species occupied or unoccupied site in the study area. The model predicts each site from the study area with a probability π for species occurrence and the observation from each site consists of presence or absence of the target species χ. Murphy and Winkler (1987; 1992) factorized the joint distribution of π and χ into a conditional distribution ($p(\chi \mid \pi)$ or $p(\pi \mid \chi)$)and a marginal distribution ($p(\pi)$ or $p(\chi)$) as shown in Figure 2, where $p(\chi \mid \pi)$ and $p(\pi)$ reflect model calibration and refinement respectively; $p(\pi \mid \chi)$ and $p(\chi)$ represent the ability to discriminate and base rate (prevalence) respectively. If the model is well calibrated then the points should lie along a 45° line of the scatter plot for predicted probabilities comparing with observed occurrence and if the is well discriminated then little overlap between presence/absence distributions on the plot of frequency distribution of the predicted values for occupied sites comparing with unoccupied sites. The prevalence needs to be moderately large for examining the predictive performance of a model (Pearce and Ferrier, 2000).

	$\pi_1 \cdots \pi_k$	
$x = 0$	$n_{01} \cdots n_{0k}$	n_0
$x = 1$	$n_{11} \cdots n_{1k}$	n_1
	$n_1 \cdots n_k$	

Figure 2. Frequency table of observation χ and predictive value π from model for each evaluated site.

Two of factorizations introduced by Murphy and Winkler (1987) are equivalent:

$$p(\pi, \chi) = p(\chi \mid \pi) \cdot p(\pi) = p(\pi \mid \chi) \cdot p(\chi)$$

Therefore since the base rate (prevalence) is a constant, a model which has good calibration and refinement must also have a good discrimination, on the contrary, however, a good discrimination is not necessarily with good calibration and refinement. These two aspects of model performance, calibration/refinement and discrimination/base rate reflect the reliable prediction of absolute value about how closely the predicted probabilities match the occurrence proportions and the ability of prediction to discriminate the observed presence to absence of predictions.

2.4.1 Confusion Matrix for Measuring Discrimination Performance

2 × 2 classification table (Table 3) often examines the model performance by comparing predicted value and actual observation (Pearce and Ferrier, 2000). Generally thinking, greater numbers of both observed/predicted presence and absence (A and D in table 3) imply a good performance of the prediction, on the other hand, greater numbers of predicted presence and absence but actually absence and presence (B and C in Table 3) tell a bad performance of the prediction. False positive (B) and false negative (C) are also called omission (including unsuitable sites in the prediction) and commission (leaving out from distributional area) respectively (Peterson *et al.*, 2008). Predicted presence or absence is determined by predicted probability value which is higher or lower than the specific threshold. The four condition of the classification table can calculate four more indices: sensitivity, specificity, false positive fraction, false negative fraction, and other measures of model performance listed in Table 4.

Table 3. 2 × 2 classification table (confusion matrix), each of the values A to D represents the number of species observed (revised from Pearce and Ferrier, 2000; Wang *et al.*, 2007)

		Observed		
		Presence	Absence	
Predicted	Presence	A	B	A+B
	Absence	C	D	C+D
		A+C	B+D	A+B+C+D

Note: A: true positive, B: false positive, C: false negative, D: true negative

Table 4. Indices derived from confusion matrix of Table 3 (revised from Fielding and Bell, 1997, Pearce and Ferrier, 2000; Wang *et al.*, 2007; Tsao, 2007)

Index	Description and Formula		
Sensitivity	$\frac{\text{Number of positive sites correctly predicted}}{\text{Total number of positive sites}}$	=	$\frac{A}{A+C}$
Specificity	$\frac{\text{Number of negative sites correctly predicted}}{\text{Total number of negative sites}}$	=	$\frac{D}{B+D}$
False Positive Fraction	$\frac{\text{Number of false positive predictions}}{\text{Total number of positive sites}}$	=	$\frac{C}{A+C}$
False Negative Fraction	$\frac{\text{Number of false negative predictions}}{\text{Total number of negative sites}}$	=	$\frac{B}{B+D}$
Accuracy (Correct classification rate)	$\frac{\text{Number of total sites correctly predicted}}{\text{Total number of sample sites}}$	=	$\frac{A+D}{A+B+C+D}$
Misclassification rate	$\frac{\text{Number of total misclassified sites}}{\text{Total number of sample sites}}$	=	$\frac{B+C}{A+B+C+D}$
Overall diagnostic power	$\frac{\text{Total number of negative sites}}{\text{Total number of sample sites}}$	=	$\frac{B+D}{A+B+C+D}$
Prevalence	$\frac{\text{Total number of positive sites}}{\text{Total number of sample sites}}$	=	$\frac{A+C}{A+B+C+D}$
Positive predict power (PPP)	$\frac{\text{Number of positive sites correctly predicted}}{\text{Total number of predicted positive sites}}$	=	$\frac{A}{A+B}$
Negative predict power (NPP)	$\frac{\text{Number of negative sites correctly predicted}}{\text{Total number of predicted negative sites}}$	=	$\frac{D}{C+D}$

Note: A: true positive, B: false positive, C: false negative, D: true negative

Table 4. Indices derived from confusion matrix of Table 3 (revised from Fielding and Bell, 1997, Pearce and Ferrier, 2000; Wang *et al*., 2007; Tsao, 2007) (cont.)

Index	Description and Formula
Odds-ratio	Ratio between total correctly predicted and total errors $= \frac{AD}{CB}$
Kappa	$\frac{(A+D)-\{[(A+C)(A+B)+(B+D)(C+D)]/(A+B+C+D)\}}{(A+B+C+D)-\{[(A+C)(A+B)+(B+D)(C+D)]/(A+B+C+D)\}}$
Normalized mutual information (NMI)	$\frac{-A\ln(A)-B\ln(B)-C\ln(C)-D\ln(D)+(A+B)\ln(A+B)+(C+D)\ln(C+D)}{(A+B+C+D)\ln(A+B+C+D)-((A+C)\ln(A+C)+(B+D)\ln(B+D))}$
True Skill Statistic (TTS)	Sensitivity + Specificity – 1
Area Under ROC Curve (AUC)	In ROC curve, 1- specificity values are plotted on X axis and sensitivity values are plotted on Y axis respectively.

Note: A: true positive, B: false positive, C: false negative, D: true negative

The sensitivity represents true positive rates. A greater true positive rate indicates model has higher ability to predict species presence when observed presence occurs. On the contrary, the value of specificity represents the true negative rate which indicates model ability to predict spices absence when observed absence occurs. Landis and Koch (1977) have suggested 3 ranges of agreement for Kappa statistic K: (i) poor; $K < 0.4$, (ii) good; $0.4 < K < 0.75$, (iii) excellent; $K > 0.75$.

2.4.2 Threshold Independence AUC

The abbreviation of AUC means area under ROC curve. ROC means receiver operating characteristic analysis which is firstly introduced in evaluation the ability to receive radar signals and applied to medical field (Wang, 2007) and the broad application in many ENM and SDM studies (take Elith *et al*., 2006, Guisan *et al*., 2007 for instance) happened in resent ten years. Figure 3 is an example demonstrated the ROC curve an AUC value. ROC analysis plots “sensitivity” (equal to 1 - omission error rate) against “1 minus specificity” (equal to commission error rate) (Cantor *et al.*, 1999) and calculates the area under ROC curve (AUC), and then compare the predicted AUC against null expectation (the area under the line from origin to the upright corner of the graph) probabilistically (Peterson *et al.*, 2008). Figure 3 is an example of ROC analysis. Y-axis is sensitivity of Table 3, which is calculated by A/(A+C), and X-axis is 1-specificity, which is calculated by B/(B+C). The procedure of ROC analysis is using threshold to generate points on ROC plots. For a continuous probability distribution, larger threshold means smaller distribution area than smaller threshold does, thus a specific threshold selection leads to a proportion of presence/absence’s distribution area. The specific threshold selection implies selecting different threshold for dividing the continuous probability distribution into binomial presence/absence parts and leads to changing the values of the evaluated indices in Table 3 such as sensitivity, specificity, and accuracy. The feature of ROC analysis is threshold independent and from prevalence and often used for evaluating accuracy of diagnostic tests (Swets, 1988; Tsao, 2007). To achieve this independency, ROC analysis estimates all thresholds of the probability distribution (from 0 to 1) to plot each value of sensitivity against 1 – specificity generated by specific threshold on the scatter plot of ROC and joints each points to become the ROC curve and the area under this curve is AUC. The ROC analysis represents the tradeoffs between the omission and commission error and AUC represents a specific metric for evaluating diagnostic procedures because it is a representation of the average sensitivity over all possible specificities (Prates-Clark *et al.*, 2008). If a larger threshold is selected then the area of predicted presence contains partial observed presence points and area of predicted absence contains almost observed absence points, and therefore, the ROC algorithm almost doesn’t falsely identifies absence, but fails to indentify most presence and generates a point with larger omission and smaller commission plotted near down-left corner (0, 0) of the plot. Continuously diminishing the threshold to a smaller one, the area of predicted presence contains almost observed presence points and area of predicted absence contains fewer observed absence, and thus, the algorithm indentifies most true presence correctly, but misclassifies most absence as positive and generates a point with smaller omission and larger commission plotted near the up-right comer (1, 1) of the plot. Ideally the top-left corner (0, 1) of ROC plot means the algorithm correctly indentifies every true presence and never misclassifies a true absence as a presence (Peterson *et al.*, 2008).

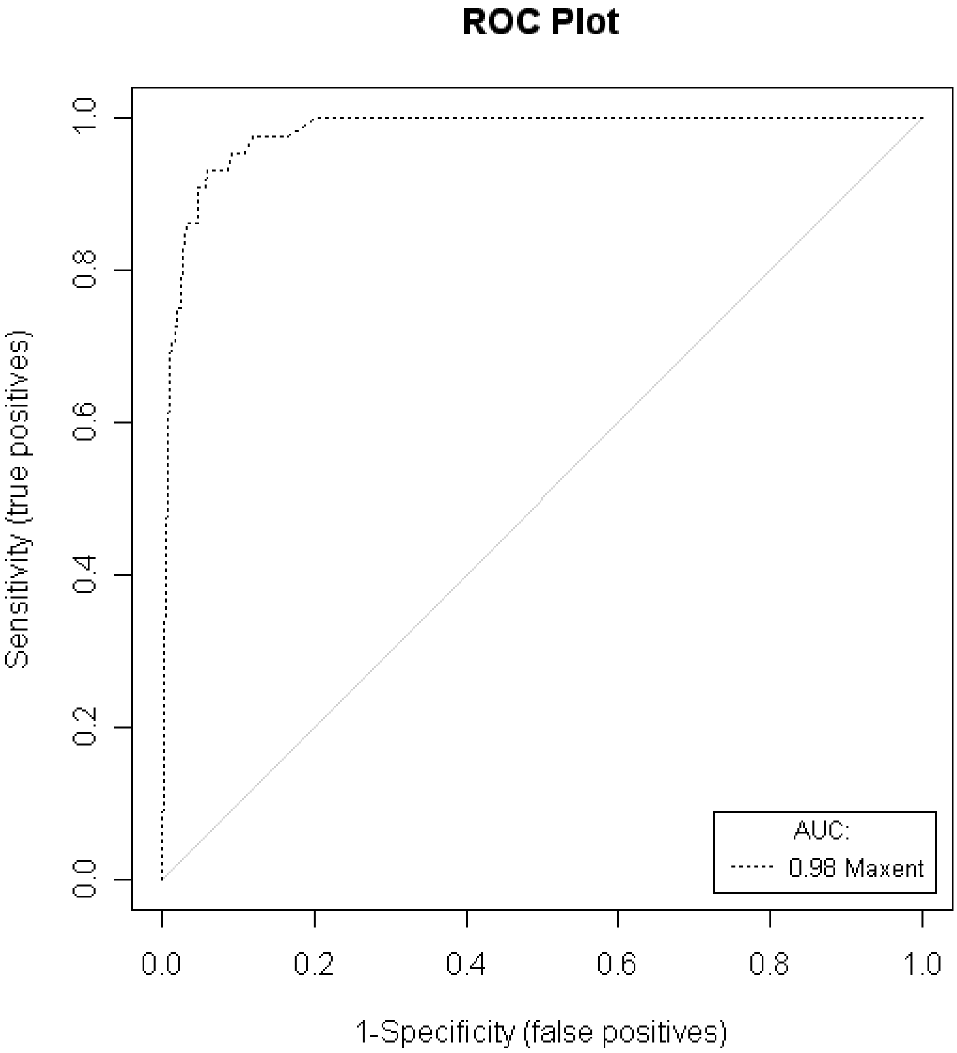


Figure 3. ROC analysis by `PresenceAbsence` package in R. where Y-axis is sensitivity of Table 3, which is calculated by A/(A+C) , and X-axis is 1-specificity, which is calculated by B/(B+C).

Prates-Clark *et al.* (2008) described 2 data sets for evaluating predicted models: (i) a training data set for model building, and (ii) a test data set for model validation. A low omission rate (high sensitivity) of species presence is essential for predicting predicted range of distribution (Anderson *et al.*, 2003). After selecting a threshold, model performance can be evaluated using both: (i) the extrinsic omission rate (using test dataset); (ii) the proportional predicted area (Prates-Clark *et al.*, 2008).

Unlike sensitivity and specificity, area under ROC curve (AUC) value is independent from prevalence and often used for evaluating accuracy of diagnostic tests (Swets, 1988; Tsao, 2007). AUC value combines sensitivity and specificity to estimate model performance and ranging from 0.5-1. According to Swets (1988), AUC value is 0.5, that means accuracy of model happen by chance; AUC value falling between 0.5-0.7 means the discrimination of model is low; AUC value falling between 0.7-0.9 means the prediction is responsible good and can be applied to other researches; AUC value is grater than 0.9 representing very good model accuracy.

2.5 Model Comparison and Combination

As motioned formerly, SDM has become an expanding tool in the areas of conservation biology, climate change research, land-use/land-cover change assessment, and biodiversity estimation (Guisan and Zimmermann, 2000). Although there are many available statistical methods, previous model comparison studies show that the prediction accuracy from different models was little in difference (Franklin, 1998; Vayssières *et al.*, 2000; Cairns, 2001; Thuiller *et al.*, 2003; Muñoz and Felicísimo, 2004). Moisen and Frescino (2002) compared predictive performance of five methods, linear models (LM), generalized additive models (GAM), classification and regression trees (CART), multivariate adaptive regression splines (MARS), and artificial neural networks (ANN), however, still found little difference among those methods (Moisen and Frescino, 2002). And besides, Elith and Burgman (2003) found greater disparities in accuracy among the plant species being modeled than among the four modeling methods that were compared. Guisan *et al.* (2007) compared 10 model techniques, BIOCLIM, BRUTO, BRT, DOMAIN, GDMSS GAM, GLM, MAXENT, MARS, and OM-GARP, 30 tree species in Switzerland, and found the greater difference in model accuracy among species than model techniques and also found that location error and sample size reduced predictive performance of many models, whereas resolution of environmental grids had little effect on most model techniques, and no model technique is able to rescue difficultly predictive target species. Therefore, to maximize accuracy of multiple model performances is needed since there is no study founding a best model (Gilmer, 2007).

Model combination (also known as consensus modeling, composite models, forecast aggregation, forecast synthesis and forecast combination) is one of the alternative ways to improve predictive accuracy of multiple models (Gilmer, 2007). Clemen (1989), Reid (1968), Bates and Granger (1969), and Batchelor and Dua (1995) suggested model combination is optimal and can yield greatest benefits for predictive accuracy. In niche model predictions, multiple models can be created for each species and the model outputs combined to determine locations present or absent of each species (Anderson *et al.*, 2002a; Lim *et al.*, 2002; Anderson *et al.*, 2003; Araújo *et al.* 2006). Olmeda and Fernández (1997) combined models by a simple voting scheme, called “majority-vote criterion”, to determine the presence/absence of locations and founded that less accurate models combination produced less predictive accuracy than the single models. Araújo *et al.* (2005) also suggested model averaging gave best predictive performance and accuracy. Clemen (1989) concluded model combination as:

“Combining forecasts has been shown to be practical, economical and useful. Underlying theory has been developed, and many empirical tests have demonstrated the value of composite forecasting. We no longer need to justify this methodology. We do need to find ways to make the implementation of the technique easy and efficient.”

Gilmer (2007) used three kinds of model combination approaches: (i) Composite (Anderson *et al.*, 2002a; and Lim *et al.*, 2002), (ii) Averaging (See and Abrahart, 2001), (iii) Summation (Anderson *et al.*, 2003). Composite (i.e. majority vote criteria) uses conditional statement to determine the final prediction. For instance, if there three binary outputs from individual models, any location are given value 2 representing presence, otherwise absence. Averaging means averaging standardized probabilistic outputs from different individual models and determined presence/absence by threshold. Summation gives useful visual explanation by summing the binary outputs from individual models (i.e. the higher number the location gets, the more model supports).

Chapter 3: Materials and Methods

3.1 Study Area

Taiwan Island expends 394 km from north to south (ca. 25°20' N to 21°55' N), stretching 140 km from east to west (ca. 22° E to 20° E) and measures about 35800 km^2 . Peaks above 3000 m in elevation are about 200 in number, locating in Central Range. (Huang *et al.*, 1994).

Although climatic zones of Taiwan Island are range widely, the area has distinct oceanic and subtropical monsoon climate. Constant wind from the sea and frequent rains and typhoons make climate in Taiwan mild and with a high humidity (Huang *et al.*, 1994). Frost is rare in the lowlands where most of the population is concentrated. Mean monthly temperatures range from 15°C to 20°C in the winter to around 28°C in the summer. The highest (40.2°C in May 2004) and lowest (-1.0°C in February 1901) urban temperatures were recorded in Taitung and Taichung, respectively. Taiwan's surface temperature has increased about 1.4°C in the past 100 years, about twice the global mean (0.6°C) (TGIO, 2008).

3.2 Target Species

Elith *et al.* (2006) described the distribution patterns of rare species are hard to predict and hilly complexity topography of Taiwan might increase the difficulty for modeling. Thus, a conifer species, which is widely spread in habitat ranging from 2400-3100 m in elevation (Su, 1984b), minimal anthropogenic impacts, endemic species of Taiwan, filling with ecological meanings, and sensitive to climate change impacts, was selected for the target of the model input. *Tsuga chinensis* commonly called Taiwan Hemlock is an evergreen large tree native to Taiwan Island, southern, central, and eastern China, and this variety is endemic to Taiwan, up to 50 m tall and 2 m in diameter, in altitudes of 2000 to 3500 m, in association with other trees or forming pure stand, (Huang *et al.*, 1994), especially mixing with the Taiwan Spruce (*Picea morrisonicola* Hayata), Taiwan Cypress (*Chamaecyparis formosensis* Matsu), Taiwan Red Pine (*Pinus taiwanensis* Hayata), Masters Pine (*Pinus armandii* Franchet var. *masteriana* Hayata), and *Quercus* zone (Su, 1984b; Su, 1991; Ou *et al.*, 1994, Liu and Tseng, 1999; Lu, 2003; Chiou *et al.*, 2006; Yen *et al.*, 2007; Song, 2007).

According to recent studies of Chiou *et al.* (2006), the distribution of Taiwan Hemlock along the elevation is from 1400 to 3400 m (similar to Huang *et al.*, 1994) and the optimal range is from 2800 to 3000 m (converted from warmth indices 30 to 140 ℃ and similar to Chen, 2004).

3.3 Data Preparation and Preprocessing

Methodological flow chart of this study is listed in Figure 4 as followed. Four parts of them are (i) data preparation, (ii) environmental variable analysis, (iii) model building, and (iv) model assessment. Data preparation phase prepares the localities for model inputs. Environmental variable analysis phase sieves and selects the environmental variables for model inputs. Model building phase attempts to compare how four types of model inputs, environmental selection, sample size, resolution, and predicted area, will affect model performance. Final phase exams the model performance of each approach: Principal component analysis (PCA), classification and regression tree (CART), and conditional inference tree (CIT).

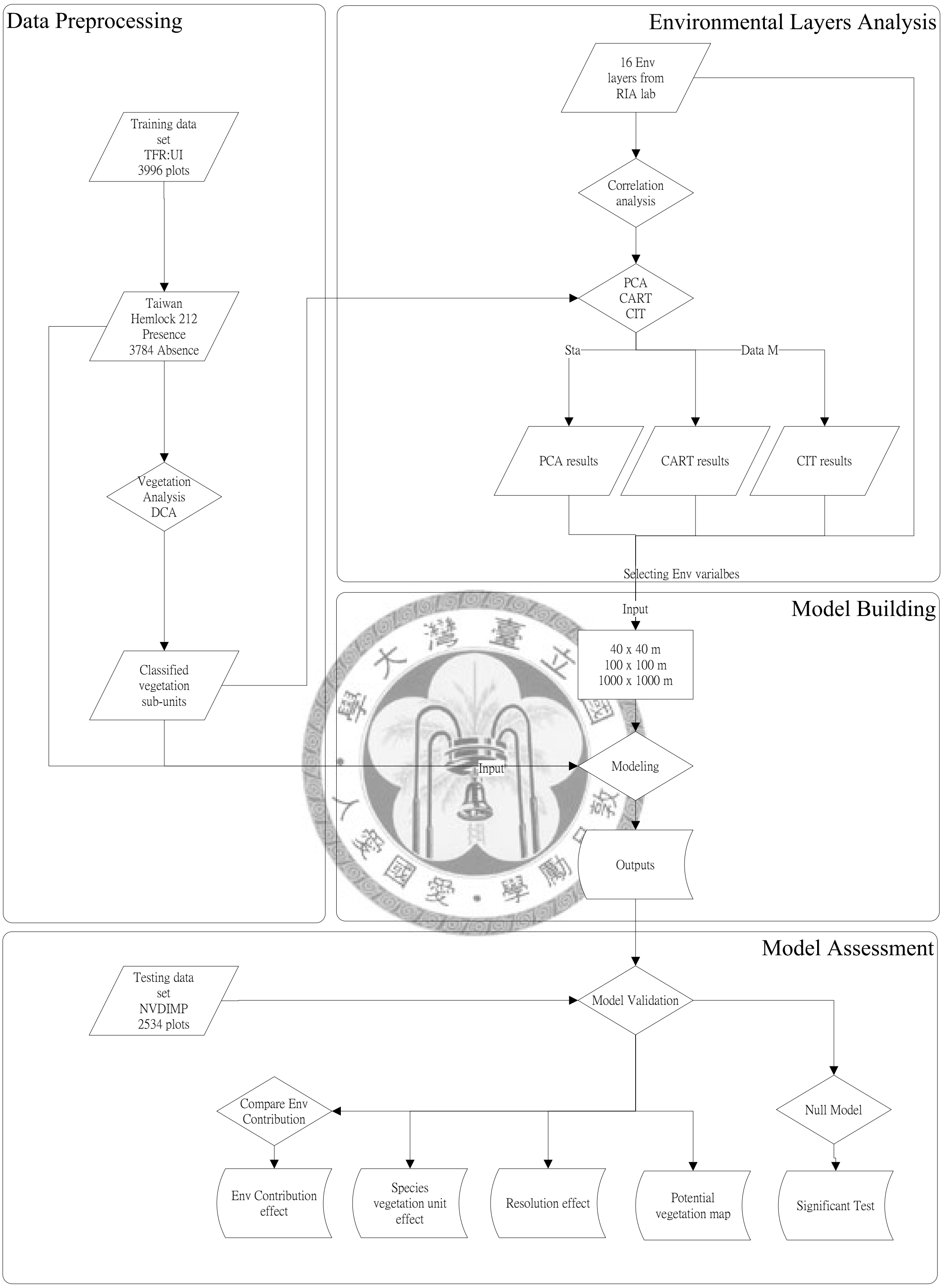


Figure 4. Flowchart of methodology in this study

3.3.1 Occurrence Data

Data of *T. chinensis* presence and absence are consisted of field survey samples from the National Vegetation Diversity Inventory and Mapping Project (NVDIMP) and Third Forest Resource and Land-Use Inventory (TFRLUI), conducted by Forestry Bureau, Council of Agriculture, Taiwan (R.O.C.). The TFRLUI records were compiled from aerial photographs by systematic sampling method, in which a plot was sampled every 500 by 250 m (Forest Bureau, 1995; Yen, 2007) and the dataset is established on Taiwan Vegetation Information System (Chiou *et al.*, 2005). 212 samples of Taiwan Hemlock presence from TFRLUI are used for model building and the rest 3784 absence samples from TFRLUI (total 3996 samples, Figure 5) are selected together with 408 samples of Taiwan Hemlock presence localities from NVDIMP (Figure 6) for model evaluation. Elith and Leathwick (2007) described the inventory pseudo-absence strongly outperforms the random pseudo-absence. Elevation range of Taiwan Hemlock from the two data sets is listed in Table 5.

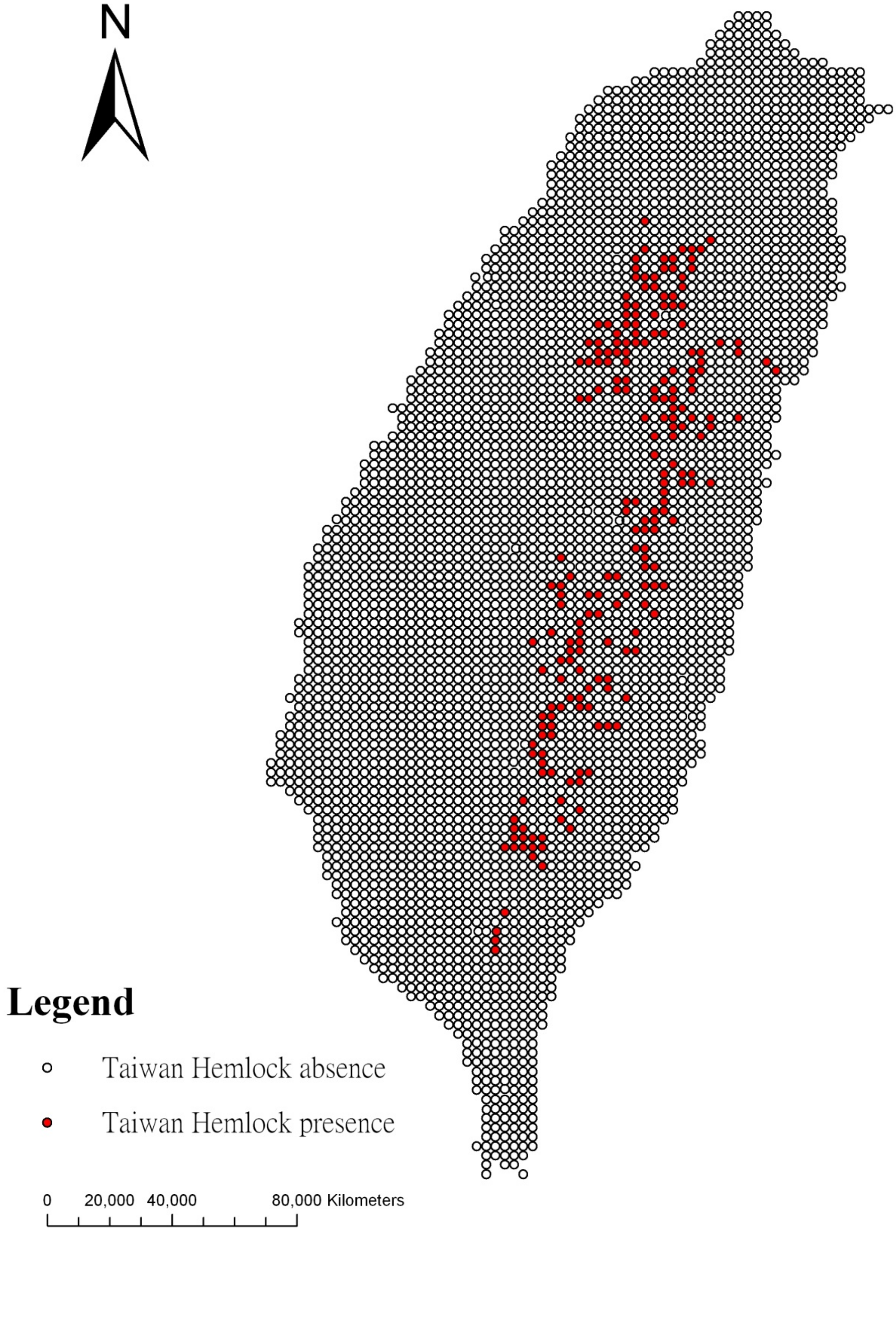


Figure 5. 212 occurrence and 3784 absence samples of Taiwan Hemlock from Third Forest Resource and Land-Use Inventory (TFRLUI)

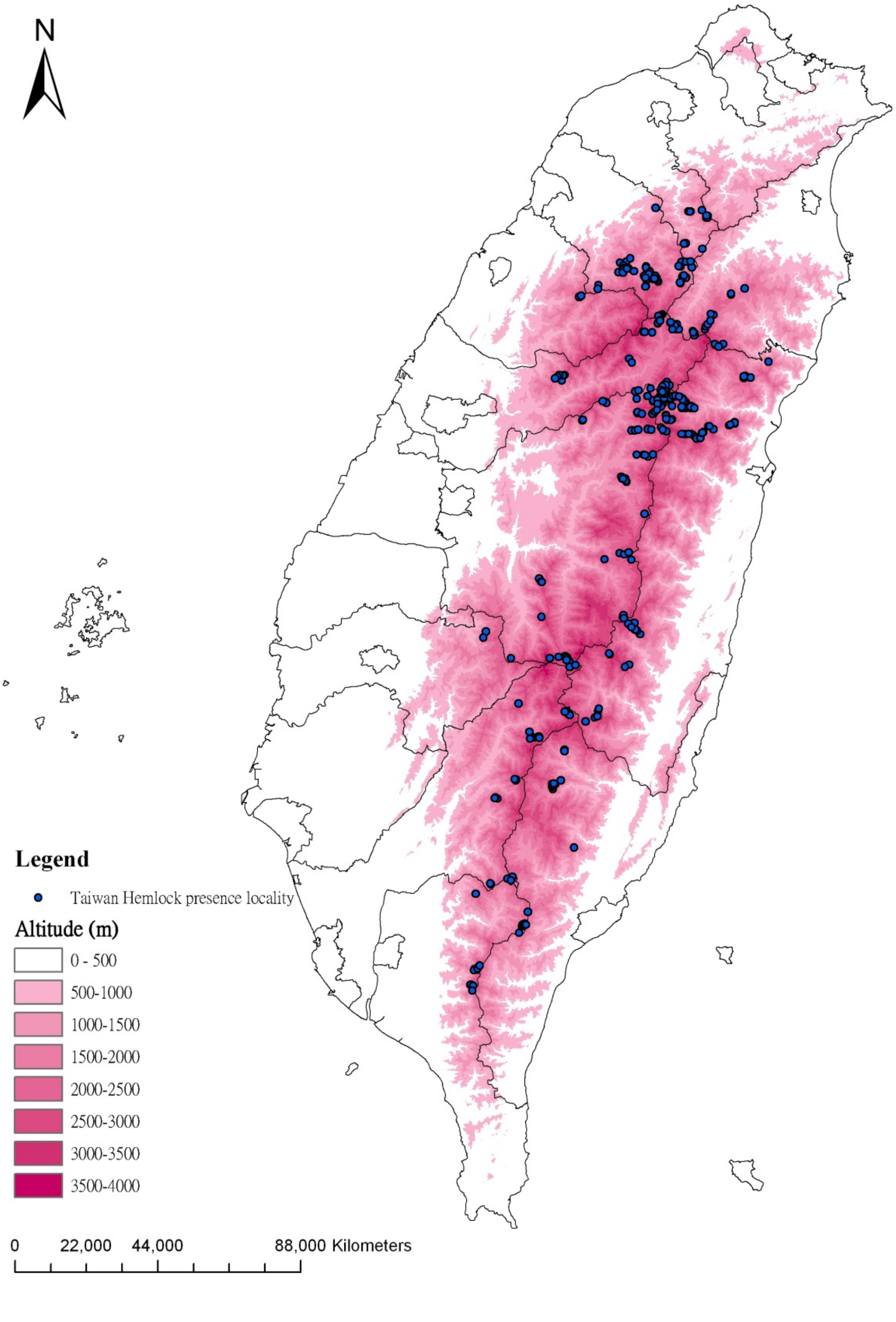


Figure 6. 408 occurrence of Taiwan Hemlock from National Vegetation Diversity Inventory and Mapping Project (NVDIMP)

Table 5. Elevation range of Taiwan Hemlock from National Vegetation Diversity Inventory and Mapping Project (NVDIMP) and Third Forest Resource and Land-Use Inventory (TFRLUI) data sets

Elevation (m)	TFRLUI	NVDIMP
Minimum value	600	1400
Maximum value	3300	3300
Mean value	2400	2500
Standard Deviation	400	400

3.3.2 Environmental Layers

Altitude and latitude are the two main environmental factors that affect the species distribution (Su, 1987; Guisan *et al.*, 1998; Yen, 2007). Other variable such as annual precipitation is not as important for determine species distribution in Taiwan as altitude and latitude, because Taiwan Island receives abundant precipitation all year round, thus precipitation is not a limited factor to vegetation distribution in Taiwan (Kuo，1978; Yen, 2007). Different latitude and altitude lead to different radiation absorption and energy (heat) store. Thus temperature is the major limiting factor to vegetation distribution (Su, 1992). In this study warmth index (WI) layer calculates from temperature layers of Liang (2004) estimated by linear equation model of Taiwan. Lang's data obtained from weather stations of Central Weather Bureau and Water Resource Agency since 1990 to 2002. Warmth Index is a proxy of annual sum of monthly average temperature, which is greater than 5 ℃ (Kira, 1948).

$$\mathrm{WI} = \sum (T_m - 5) \quad ; \quad T_m > 5\ ℃$$

Where WI is the abbreviation of Warmth Index.

Table 6 shows abbreviation and description of 16 environmental layers conducted by Resource Investigation and Analysis Laboratory (RIAL), School of Forestry and Resource Conservation, National Taiwan University.

Table 6. List of environmental variables (revised by Lindsay, 2005)

Variable	Names	Description	Unit	Reference	Software
ASP	Aspect	Direction of maximum downward gradient	Degrees	Zevenbergen and Thorne, 1987	ArcGis
CUR	Tangential Curvature	Tangential Curvature Curvature in an inclined plane (Mit´aˇsov´a and Hofierka, 1993)	Deg./m	Mit´aˇsov´a and Hofierka, 1993	TAS
ELE	Elevation	Elevation derived from DTM	M	Liang, 2006	ArcGis
PLA	Plan	Plan curvature Along-slope curvature	Deg./m	Gallant and Wilson, 2000	TAS
PRCSP	Spring Precipitation	Average precipitation of Mar. to May.	Mm	Liang, 2004	ArcGis
PRCSR	Summer Precipitation	Average precipitation of Jun. to Aug.	Mm	Liang, 2004	ArcGis
PRCAU	Autumn Precipitation	Average precipitation of Sep. to Nov.	Mm	Liang, 2004	ArcGis
PRCWT	Winter Precipitation	Average precipitation of Dec. to Feb.	Mm	Liang, 2004	ArcGis
PRCME	Mean of ann. Precipitation	Average precipitation of a year	Mm	Liang, 2004	ArcGis
PRCSU	Sum of month Precipitation	Summation of every monthly precipitation	Mm	Liang, 2004	ArcGis
PRO	Profile	Profile curvature Down slope curvature (Zevenbergen and Thorne, 1987)	Deg./m	Moore *et al.*, 1993	TAS

Table 6. List of environmental variables (revised by Lindsay, 2005) (Cont.)

Variable	Names	Description	Unit	Reference	Software
SLP	Slope	Slope gradient (Zevenbergen and Thorne, 1987)	Degrees	Zevenbergen and Thorne, 1987	ArcGis
STH	Southness	Southness = 180 - | aspect - 180 |	Unitless	Chang *et al.*, 2004	ArcGis
SVF	Sky View Factor	Sky View Factor (SVF), represents an estimation of the visible area of the sky from a ground viewpoint	Unitless	Steyn, 1980; Oke, 1981	SkyRatio
WI	Warmth Index	Sum of monthly mean temperatures greater than 5 ℃ from Jan to Dec	℃	Chiou *et al.*, 2004	ArcGis
WST	Westness	westness = | 180 - | aspect - 270 | |	Unitless	Chang *et al.*, 2004	ArcGis

The spatial location of each environmental layers are recorded by using the 2-degree Transverse Mercator projection coordinate system (TM2), including latitude/longitude (TMX and TMY), climate variables (WI and PCP), digital terrain model (DTM), and its topographic and radiation derivate models (Chang *et al.*, 2004) provided by Forestry Bureau, Council of Agriculture, Taiwan (R.O.C.). Table 5 shows a list of environmental layers. Topographic information derived from DTM includes aspect (ASP), slop (SLP), curvature (CUR), profile curvature (PRO), and plan curvature (PLA). Southness index (STH), westness index (WST) are derived form ASP (Chang *et al.*, 2004). Solar radiation derived from DTM includes sky view factor (SVF). Climate variable includes warmth index (WI), precipitation (PRC) and its derivate including spring precipitation (PRCSP, from Mar. to May), summer precipitation (PRCSR, from Jun. to Aug), autumn precipitation (PRCAU, from Sep. to Nov.), winter precipitation (PRCW, from Dec. to Feb.), annual mean precipitation (PRCME), and sum of annual precipitation (PRCSU) are investigated by Liang (2004).

3.3.3 Vegetation Analysis

Using dominant species for representing forest subtype, firstly calculate the relative dominance (RDo) of 212 Taiwan Hemlock presence localities from TFRLUI to emphasize the importance of dominance tree and to lower the disturbance of rare or small species. RDo is calculated by following:

$$Do = \sum_{i=1}^{n} BA_i \,, \quad RDo_j = \frac{Do_j}{\sum_{j=1}^{n} Do_j}$$

Where Do represents dominance, BA represents basal area of the tree, Do_j represents dominance for j_{th} species, and RDo_j represents relative dominance for j_{th} species.

To group sample plots, cluster analysis supported by PC-ORD 5.0 statistical software (McCune, B. and M.J. Mefford., 1999 PC-ORD. Multivariate Analysis of Ecological Data. MjM Software, Gleneden Beach, Oregon, USA) is used in classifying Taiwan Hemlock presence data, and then uses Euclidean method for distance measure and Ward's method for group linkage method of cluster analysis. Detrended correspondence analysis (DCA) and principal component analysis (PCA) ordination are used for representing relationships between classified groups and environmental gradients.

3.4 Environmental Factor Analysis for SDMs Performance

3.4.1 Avoidance of Multicollinearity

Multicollinearity refers to highly correlation among two or more explanatory variables and leads to over estimate of least square estimation and enlarges variance of the estimation, thus, inference might be misleading (Lin and Chen, 2005). Although two highly correlated predictor variables can both appear non-significant, each would explain a significant proportion of the deviance if considered separately (Guisan *et al.*, 2002). To avoid such case, correlation analysis is chosen for distinguishing variables that are highly multicollinearity by comparing their correlation coefficients. If absolute correlation coefficient is larger than 0.1 and smaller than 0.3, there is a small correlation between two variables; if absolute correlation coefficient is larger than 0.3 and smaller than 0.5, there is a medium correlation between two variables; if absolute correlation coefficient is larger than 0.5 and smaller than 0.1, there is a large correlation between two variables (Cohen, 1988). Thus, the variables with high correlation will be removed to 1 variable to retain. Algorithm of correlation coefficient is analyzed by R foundation for ecological computing (version 2.6.2.).

3.4.2 Attributes of Environmental variables

Descriptive statistic provides a basic sense and structure of the data. Presence and (N = 212) absence localities (N = 3784) from TFRLUI are compared by extracting values from environmental layers without high correlation to each other. Descriptive statistic is calculated by `pastecs` package in R foundation for ecological computing (version 2.6.2.). Basic statistic includes the number of values, the number of missing values, the minimal value, the maximal value, the range, and the sum of all non-missing values; the descriptive statistic includes the median, the mean, the standard error on the mean, the confidence interval of the mean, the variance, the standard deviation and the variation coefficient defined as the standard deviation divided by the mean. Normal Q-Q plot method is used to see if the data are normal distributed and then two-tailed t-test and 95 % confidence interval are used to test if the two data sets are significant the difference. Finally use histograms to demonstrate the attribute of each extracted environmental variable.

3.4.3 Try and Error Approach

Try each of 16 environmental layers to build SDM and compare results. Each of 16 environmental layers has its own attributes and influences the model performance separately. To identify how each environmental layer affects SDM results, single environmental layer prediction is able to provide each contribution of environmental layer to the model performance. On the other hand, to identify interaction of environmental variable effects, all 16 environmental layers for modeling contrasting the single input is the other way for testing model performance.

3.4.4 PCA Approach

Try and error approach shows how does single and all environmental layers contribute the model performance and is lack of statistical test for species in relation to environmental layers. Statistical approaches used in this study are principal component analysis (PCA) supported by PC-ORD 5.0. Those approaches mainly focus on finding the most variant axis to environmental layers and suppose the environmental variable of the axis is the most explainable for the species occurs. The first 3 components of PCA are able to explain the most variance of the data and thus are selected for model building.

3.4.5 Data Mining Approach

Except statistical approaches, one of machine learning technique, data mining, is also introduced in this study. Data mining is the search for new, valuable, and nontrivial information in environmental layers. Two of the data mining approaches are adopted here: (i) classification and regression tree (CART) and (ii) conditional inference tree (CIT). CART model uses the package `tree` (Ripley, 1996) in R foundation for ecological computing (2.6.2.). CIT model uses the package `party` (Hothorn *et al.*, 2006) in R foundation for ecological computing (2.6.2.). The extracted environmental variables that are selected by those two methods for splitting the nodes of the tree are used for further model building because those selected variables are able to reduce the variance of each split groups and to distinguish each vegetation type by using influential environmental variables.

3.5 *Predicting Species Distribution*

3.5.1 *Model Building*

According to the two data sets (NVDIMP and TFRLUI) are independently surveyed, TFRLUI data set uses for training (build the model) because the data of TFRLUI are systematic sampled and suitable for model building without autocorrelation and NVDIMP data set uses for testing (evaluate the model performance). Thus, there is no need to partition data. And besides, the number of data from NVDIMP (408) is greater than the number of data from TFRLUI (212) and it represents using more localities to test the model performance for more precise evaluation. This study used maximum entropy (MAXENT) technique (Phillips *et al.*, 2004) for SDM development. Since occurrence data and environmental layers are available, input those data with appropriate format to the models (appropriate data format of SDMs is listed in Table 7.) and set any parameter if needed. Detailed model setting for each SDM will describe in next two sections.

Table 7. Data format of SDM

Model techniques	Type of data	Input format	Output format	software	URL
Maximum entropy	Occurrence data	Csv	Asc	Maxent	http://www.cs.princeton.edu/~schapire/maxent/
	environmental layers	Asc	Mxe Grd		

Maxent program for maximum entropy modeling predicts species geographical distributions and is firstly introduced by Phillips *et al*. (2006). The probability distribution of maximum entropy, which is the concept of Maxent, is the distribution closest to uniform distribution or most spread out. Maxent evaluate a target probability distribution through looking for the distribution for maximum entropy (Phillips *et al*., 2006). The incomplete information for the target is represent by a set of constrains which is influential to Maxent. A set of real value (also called feature, observed value) is used as the available information for the target distribution. Constrains are obtained by matching expected value of each feature with empirical average, an average value of a set of sample localities derived from the target distribution.

The tutorial of Maxent stated that while running procedure, the gain calculated by Maxent is closely related to deviance, a measure of goodness of fit used in GAM and GLM and starts at 0 and increases towards an asymptote. The gain is defined as:

$$gain = \overline{\log(p)} - k$$

Where $\overline{\log(p)}$ the average log probability of the presence localities and k is a constant that makes the uniform distribution have zero gain.

Finally, the gain indicates how closely the model is closed around the presence localities; for example, if the gain is 1.5, it means that the average likelihood of the presence samples is exp(1.5) ≈ 4.5 times higher than that of a random background pixel. Note that Maxent isn't directly calculating "probability of occurrence" and raw values are an exponential function of the environmental variables, however, logistic format transforms it into probability of presence and sets to default option.

3.5.2 How Map Resolution and Environmental Variables Affect Model Performance

Higher resolution (smaller grid size) of the predicted background can reflect the more detail of topographical variables than climatic variables but is time consuming due to a very large data size for model estimation. Lower resolution (larger grid size) of background, on the other hand, is much faster when calculating but lacks of or reduces detail information of the meso-environmental variables. A grid of lower resolution may contains more than 1 presence or absence localities and causes misleading when presence and absence localities in the same environmental grid. To test the effect of background resolution, three resolutions are selected for modeling, 40×40 m ,100×100 m, and 1000×1000 m respectively, for the following analysis. Environmental variable combination of 7 different methods mentioned in section 3.4 will be compared to each other to find the most explainable combination of environmental variables by comparing their AUC values.

3.5.3 Vegetation and Species Units for Model Input

Assume species unit of Taiwan Hemlock is not as homogeneity as vegetation unit which may be suitable for difference environmental condition. Therefore, use all

Taiwan Hemlock presence localities and classified Taiwan Hemlock vegetation type subunit separately for model input to compare how difference between vegetation and species units relates to the environmental variable selected by Maxent and then to produce a potential vegetation map of Taiwan Hemlock with sub vegetation unit within it.

3.6 Model Evaluation

Although unbiased estimate of a model's predictive performance is evaluating with independent data collected from sites other than those used to train the model (Pearce and Ferrier, 2000), splitting the data into two partition, one for training and the other for test, is the alternative way for the model assessment while a independent testing data set is not available. Model performance is then tested at fixed specifically thresholds (threshold-dependent) and across all thresholds (threshold-independent) methods.

3.6.1 Threshold Independent AUC

ROC analysis provides the whole information that each threshold contributes a pair of sensitivity (absence of omission error) and 1 – specificity (commission error) and represents the trade-off for both values. Only AUC measurement for the performance of SDM is invariable to the prevalence (proportion of presence to sum of presence and absence) (Pearce and Ferrier, 2000; Rase and Steege, 2007). The larger the AUC estimated, the higher the sensitivity rate and the lower the 1-specificity rate happened. An AUC value of 1 represents an ideally diagnostic test because it means the value of both sensitivity and specificity are also 1 (i.e. no either omission or commission error). An AUC value of 0.5 indicates high omission and commission errors and random prediction. (Cantor *et al.*, 1999; Rase and Steege, 2007).

In this study, AUC value is calculated by `PresenceAbsence` package in R (2.6.2) with 408 Taiwan Hemlock presence localities from NVDIMP and 3784 Taiwan Hemlock absence localities from TFRLUI respectively. Phillips *et al.* (2006) suggested a sufficiently large sample of pseudo-absence is needed, typically 1000 to 10000, to reasonably represent the environmental variable restricted by the geographical area. Those pseudo-absences, however, results in a low prevalence value because the pseudo-absences are much larger than presences. A major drawback of using pseudo-absence is changing the perfect fit value of AUC is 1 - a/2 instead of 1, where a is the geographical area substituting to a species' true distribution (Phillips *et al.*, 2006; Rase and Steege, 2007). Therefore, 3784 inventory pseudo-absence are large enough to evaluate the model performance.

3.6.2 Threshold Dependent Confusion Matrix

Outputs of SDMs are continuous probability distribution layers of MAXENT. Use the test samples to complete confusion matrix and calculate the sensitivity, specificity of each model output after selecting a specific threshold.

Kappa statistic value is calculated by equation of table 3 from confusion matrix and relative to the accuracy that may have resulted by only chance. It ranges from negative one to positive one which indicates flawless agreement between observations and predictions and the value less or equal to 0 indicates no better performance than random classification (Tsoar *et al.*, 2007). If Kappa value ranges from 0 to 0.4, it means low strength of predicted accuracy; from 0.4 to 0.75, represents a good predicted accuracy; from 0.75 to 1, motioned above with perfect predicted accuracy (Landis and Koch, 1977; Tsao, 2007).

3.6.3 Null Model for Significant test

A null model is a model based on randomizations of the ecological data or random sampling from a know area (Gotelli and McGill, 2006; Rase and Steege, 2007). To ensure prediction of SDM is based on environmental layer to survey plots and not randomly selects from spatial localities, firstly estimates the AUC value of a SDM and a null model is established by repeating 999 times randomly selecting points equal to the number of the input occurrences from background area to estimate the AUC value and to generate a randomly AUC distribution on a histogram and compare with their output of AUC results and test if the null model is true. Using one-tailed 95 % CI for the null probability distribution of the randomly generated AUC values to test the significance as conventional statistic does (Rase and Steege, 2007). If there is a significant difference between output's AUC value and null model's AUC value, surveyed occurrences of *T. chinensis* are not randomly appeared by chance but are relative to environmental variables. In this study, 14 null models are generated for 7 projection areas which are the area of Taiwan Island and 6 geo-climatic regions multiplying 2 grid resolutions.

3.7 Potential Nature Vegetation Mapping

3.7.1 Specific Threshold to Presence

Muñoz and Felicísimo (2004) noted that the objective of the study determines the final threshold for PVM, considers how relative importance of false positives and false negatives error rates affects the PNM, and the decision made independent of model accuracy results. Muñoz and Felicísimo (2004) concluded that if the goal of the study is to identify localities where a species occurrence can be predicted with a great certainty, the false positive error rates should be minimized; conversely, if the purpose is conservation of the a species, the threshold must be chosen to minimize false negative error rates. The ultimate objective of PVM is usually a map of vegetation occurrence thus requiring a specific threshold to be selected to determine which probability range will be considered present (Gilmer, 2007).

A threshold that determines which predictions are considered absent/present have to be identified for most classification accuracy methods (Gilmer, 2007); however, threshold selection is subjective and can be selected based on several methods: (i) threshold = 0.5, (ii) sensitivity = specificity, (iii) maximizes (sensitivity+specificity)/2, (iv) maximizes Kappa, (v) maximizes percent correctly classified (PCC), (vi) predicted prevalence=observed prevalence, (vii) threshold=observed prevalence, (viii) mean predicted probability, and (ix) minimizes distance between ROC plot and up-left corner (0,1) (Cantor *et al.*, 1999; Manel *et al.* 2001; Wilson *et al.*, 2004) and those methods are available in `PresenceAbsence` package of R.

Although many studies use 0.5 as the threshold for considering present or absent, this value is somewhat arbitrary and can result in unacceptably low model sensitivity when the target species is rare (Fielding and Bell, 1997; Manel *et al*., 1999; Miller, 2005). Miller (2005) and Gilmer (2007) selected the threshold near the point that sensitivity and specificity cross, with an emphasis on ensuring sensitivity is relatively high. Miller and Franklin (2002) used maximize CPP due to sensitivity and specificity did not cross when plotted on a 0 to 1 scale, placing more importance on CPP, as is usually the case in vegetation mapping. Tsao (2007) and Tsoar *et al.,* (2007) selected the threshold with maximize kappa statistic. Prates-Clark *et al.* (2008) threshold selection was calculated based on the number of the probability of species' occurrence are 30, 40, 50, and 75 % and this method provided information which threshold made the greatest contribution to the model, and also indicated which model should be remained as the best potential predicting species distribution model. In this study, Taiwan Hemlock is not endangered species but a dominance tree in the alpine ecosystem in Taiwan. Therefore, threshold with max-Kappa is selected for higher predicted accuracy.

3.7.2 Model Combination and PNV Mapping Criteria

Prates-Clark *et al.* (2008) described the ideally and accurately predictive model for each tree species' potential geographical distribution based on (i) the lowest omission and commission rates, (ii) highest AUC value, (iii) higher percentage of predicted probability of species' presence localities, and (iv) a set of predictor variables biologically meaningful to summarize the ecological niche of its species. Predicted map generates by GIS platform and for this study ArcGIS 9.2 is used.

Input the raster of probability layers generated by SDMs for model combination. There are two kinds of strategies for combination used in this study. One of which is to combine the individual probability maps of sub-units of vegetation from vegetation analysis to generate potential vegetation map of Taiwan Hemlock sub-unit vegetation. The other combines the vegetation based sub-units and Taiwan Hemlock species based unit probability map into the potential vegetation map of Taiwan Hemlock. The method of combining vegetation sub-unit maps of Taiwan Hemlock is determined by the pixels with the greatest probability value of which vegetation sub-units. For example, if there are three vegetation sub-units probability maps generated by a SDM, overlapping the three probability maps and each pixel would have three probability values. If the highest value of probability is from sub-unit vegetation type 1 and then this pixel is determined to the potential area that sub-unit vegetation type 1 would occur and so on. The method of the second combination is to overlap the Taiwan Hemlock species probability map with all the sub-unit vegetation maps of Taiwan Hemlock and to demonstrate the potential distribution of species Taiwan Hemlock including the detail information of spatial distribution area of Taiwan Hemlock vegetation sub-units.

Chapter 4: Results

4.1 Vegetation Classification of Taiwan Hemlock Presence

There are 16 samples of outliner excluded at final classification due to the sample with too low relative dominance of Taiwan Hemlock or some rare or unique species occurred in the sample which might produce meaningless groups with very few samples or mislead the classified group. Classification of cluster analysis (Figure 7 and 8) by PC-ORD divides 196 plots into 6 sub groups and each group represents different species dominance companying with Taiwan Hemlock:

V_1: Vegetation type 1 represents Taiwan Hemlock-Taiwan Cypress (*Tsuga chinensis-Chamaecyparis formosensis*) dominance vegetation type.

V_2: Vegetation type 2 represents Taiwan Hemlock (*Tsuga chinensis*) dominance vegetation type.

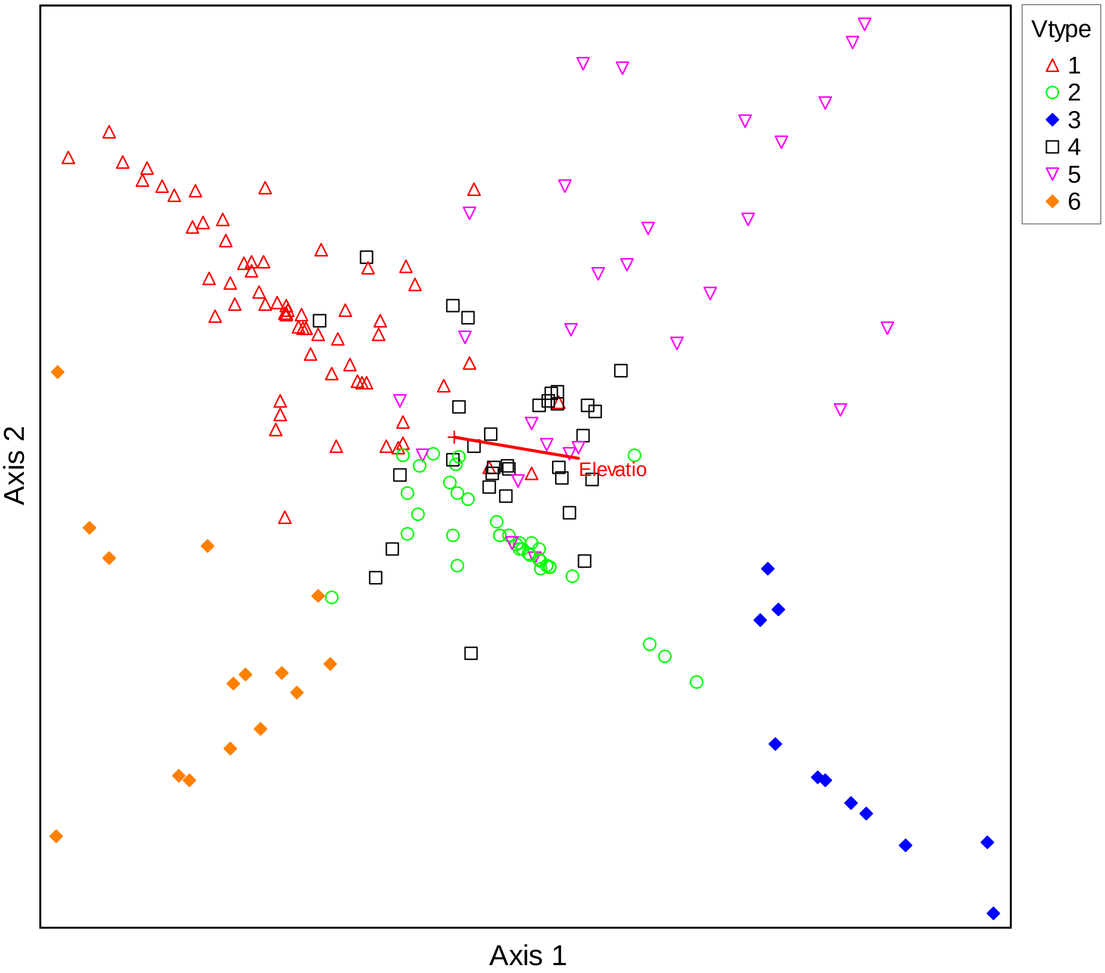
V_3: Vegetation type 3 represents Taiwan Hemlock-Taiwan Fir (*Tsuga chinensis-Abies kawakamii*) dominance vegetation type.

V_4: Vegetation type 4 represents Taiwan Hemlock-Taiwan Red Pine (*Tsuga chinensis-Pinus taiwanensis*) dominance vegetation type.

V_5: Vegetation type 5 represents Taiwan Hemlock-Taiwan Spruce (*Tsuga chinensis-Picea morrisonicola*) dominance vegetation type.

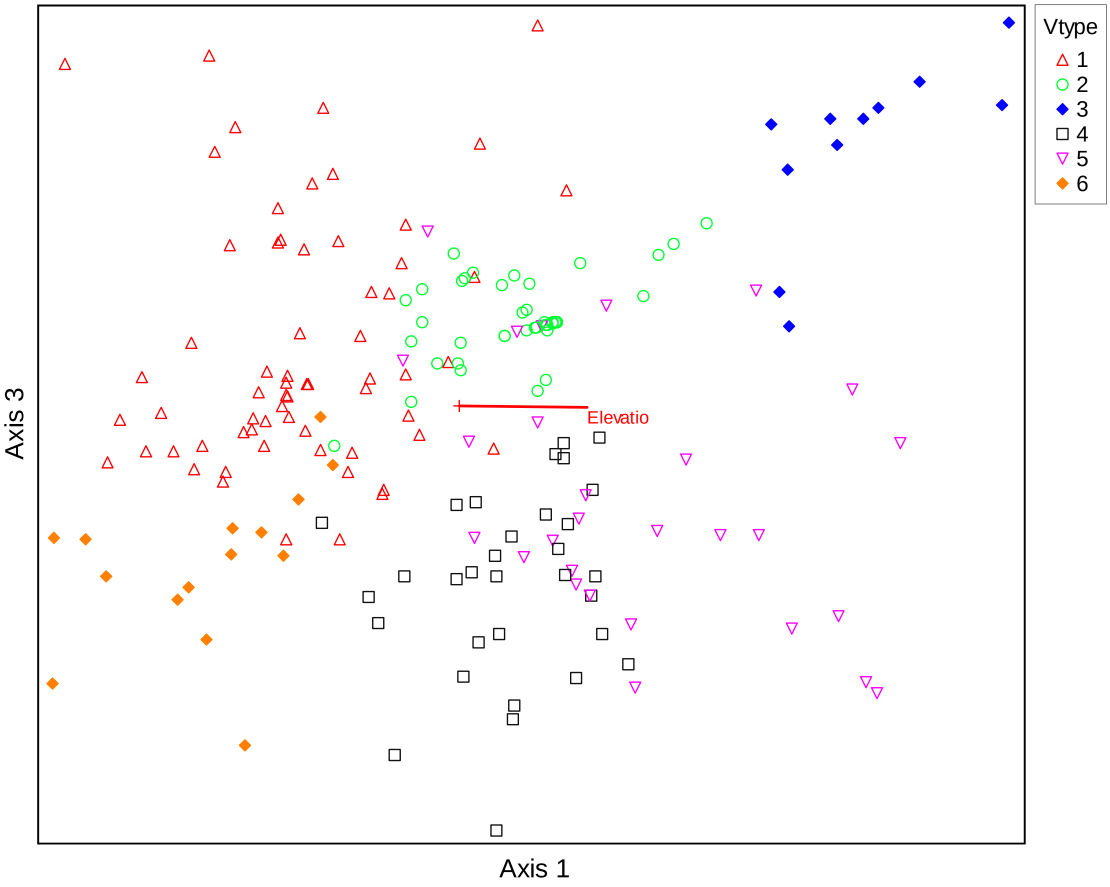
V_6: Vegetation type 6 represents Taiwan Hemlock-Taiwan Yellow Cypress (*Tsuga chinensis-Chamaecyparis obtusa* Var. *formosana*) dominance vegetation type.

Table 8 shows the correlation between environmental variables and DCA axis.



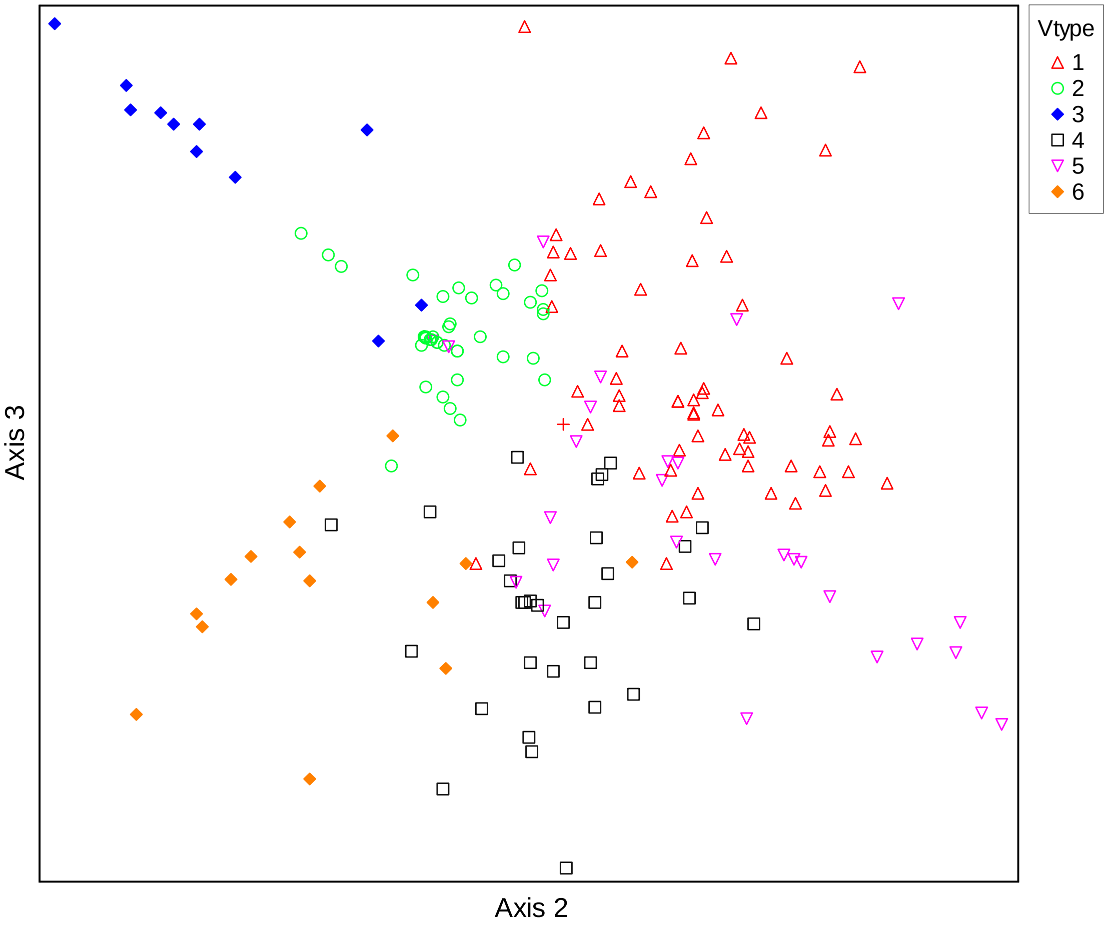
(a)

Figure 7. DCA odination of 6 sub groups of Taiwan Hemlock: (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axsis 2 and 3. Vtype 1 represents Taiwan Hemlock-Taiwan Cypress (*Tsuga chinensis-Chamaecyparis formosensis*) dominance vegetation type; Vtype 2 represents Taiwan Hemlock (*Tsuga chinensis*) dominance vegetation type; Vtype 3 represents Taiwan Hemlock-Taiwan Fir (*Tsuga chinensis-Abies kawakamii*) dominance vegetation type; Vtype 4 represents Taiwan Hemlock-Taiwan Red Pine (*Tsuga chinensis-Pinus taiwanensis*) dominance vegetation type; Vtype 5 represents Taiwan Hemlock-Taiwan Spruce (*Tsuga chinensis-Picea morrisonicola*) dominance vegetation type; Vtype 6 represents Taiwan Hemlock-Taiwan Yellow Cypress (*Tsuga chinensis-Chamaecyparis obtusa* Var. *formosana*) dominance vegetation type.



(b)

Figure 7. DCA odination of 6 sub groups of Taiwan Hemlock: (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axsis 2 and 3. Vtype 1 represents Taiwan Hemlock-Taiwan Cypress (*Tsuga chinensis-Chamaecyparis formosensis*) dominance vegetation type; Vtype 2 represents Taiwan Hemlock (*Tsuga chinensis*) dominance vegetation type; Vtype 3 represents Taiwan Hemlock-Taiwan Fir (*Tsuga chinensis-Abies kawakamii*) dominance vegetation type; Vtype 4 represents Taiwan Hemlock-Taiwan Red Pine (*Tsuga chinensis-Pinus taiwanensis*) dominance vegetation type; Vtype 5 represents Taiwan Hemlock-Taiwan Spruce (*Tsuga chinensis-Picea morrisonicola*) dominance vegetation type; Vtype 6 represents Taiwan Hemlock-Taiwan Yellow Cypress (*Tsuga chinensis-Chamaecyparis obtusa* Var. *formosana*) dominance vegetation type.



(c)

Figure 7. DCA odination of 6 sub groups of Taiwan Hemlock: (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axsis 2 and 3. Vtype 1 represents Taiwan Hemlock-Taiwan Cypress (*Tsuga chinensis-Chamaecyparis formosensis*) dominance vegetation type; Vtype 2 represents Taiwan Hemlock (*Tsuga chinensis*) dominance vegetation type; Vtype 3 represents Taiwan Hemlock-Taiwan Fir (*Tsuga chinensis-Abies kawakamii*) dominance vegetation type; Vtype 4 represents Taiwan Hemlock-Taiwan Red Pine (*Tsuga chinensis-Pinus taiwanensis*) dominance vegetation type; Vtype 5 represents Taiwan Hemlock-Taiwan Spruce (*Tsuga chinensis-Picea morrisonicola*) dominance vegetation type; Vtype 6 represents Taiwan Hemlock-Taiwan Yellow Cypress (*Tsuga chinensis-Chamaecyparis obtusa* Var. *formosana*) dominance vegetation type.

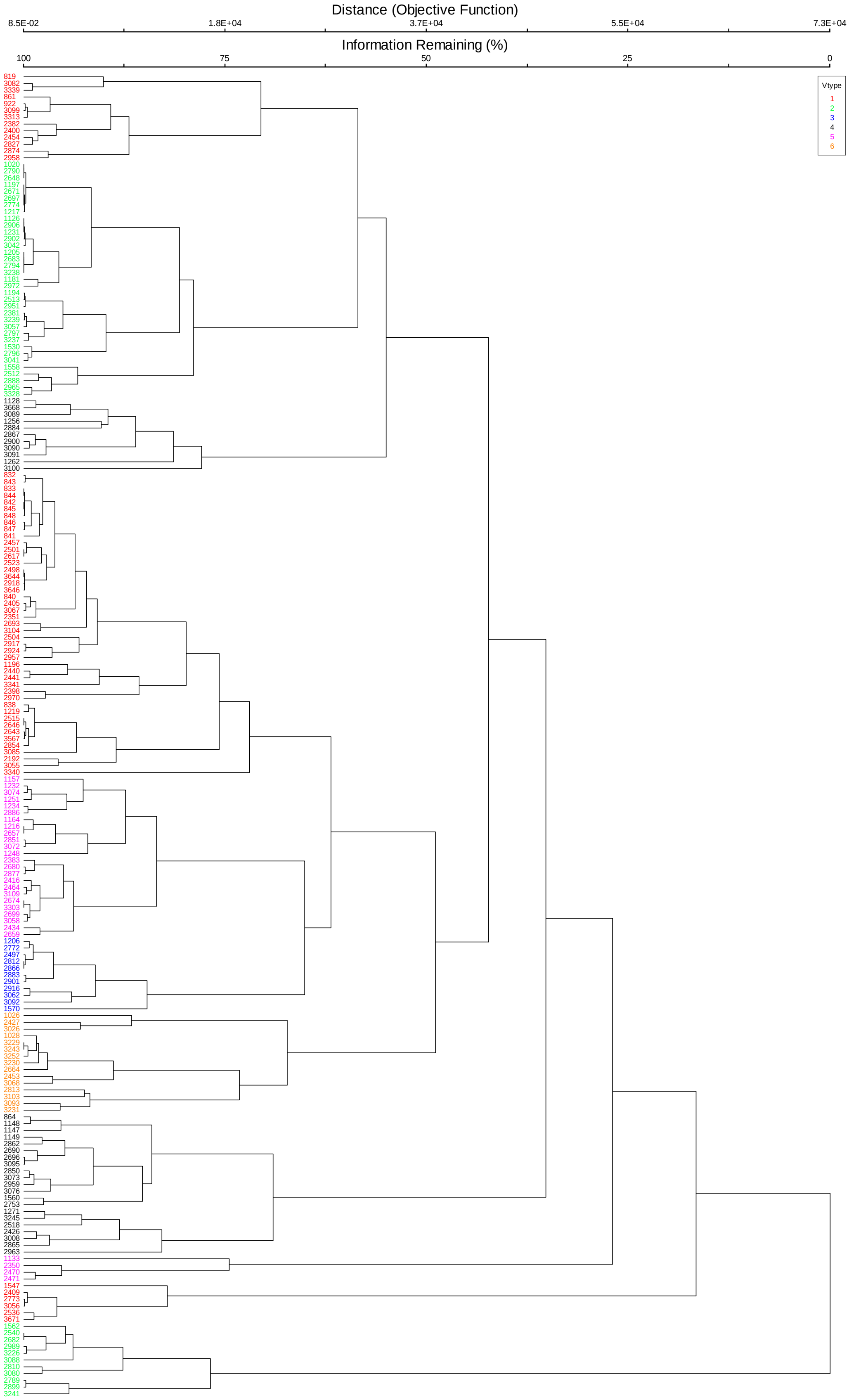


Figure 8. Cluster analysis dendrogram of 6 sub groups of Taiwan Hemlock.

Table 8. Pearson and Kendall correlation between surveyed environmental gradients and DCA and PCA axes. (N= 196)

Env.	DCA-1	DCA-2	DCA-3	PC1	PC2	PC3
Elevation	0.52	-0.22	-0.06	-0.54	0.58	0.08
Slope	0.03	-0.04	0.08	-0.51	-0.51	-0.68

Note: DCA-1 to DCA-3 represents axis 1 to axis 3; PC1 to PC3 represents component 1 to component 3 of PCA. Env.: environmental variable

Elevation gradient has high correlation with axis 1 (correlation coefficient is 0.52) and relative small correlation with axis 2 (correlation coefficient is -0.22). Axis 3, however, shows little correlation with both environmental gradient Elevation and Slope (correlation coefficient is -0.06 and 0.08 respectively).

Table 9. Pearson and Kendall correlation between 16 extracted environmental gradients and DCA and PCA axes. (N= 196)

Env.	DCA-1	DCA-2	DCA-3	PC1	PC2	PC3
ASP	-0.09	-0.04	0.00	-0.08	0.13	-0.35
CUR	-0.01	0.03	-0.06	-0.86	-0.39	0.17
PLA	-0.04	0.04	-0.02	-0.78	-0.31	0.14
PRCAU	-0.18	0.03	-0.12	-0.12	0.09	-0.47
PRCME	-0.13	0.07	0.03	-0.36	0.83	0.34
PRCSP	0.11	-0.10	0.09	0.08	0.31	0.77
PRCSU	-0.13	0.07	0.03	-0.36	0.83	0.34
PRCSR	-0.13	0.18	0.03	-0.52	0.71	0.08
PRCWT	0.11	-0.17	0.08	0.41	-0.04	0.48
PRO	-0.04	-0.01	0.08	0.74	0.39	-0.16
SLP	-0.11	-0.01	0.10	0.15	0.20	-0.14
STH	0.00	0.00	-0.20	-0.04	-0.10	0.08

Table 9. Pearson and Kendall correlation between 16 extracted environmental gradients and DCA and PCA axes. (N= 196) (Cont.)

Env.	DCA-1	DCA-2	DCA-3	PC1	PC2	PC3
SVF	0.11	-0.02	-0.11	-0.66	-0.47	0.25
ELE	0.52	-0.22	-0.07	0.31	-0.28	0.68
WST	-0.06	-0.02	0.02	-0.10	0.02	-0.36

Note: ASP: Aspect; CUR: Tangential Curvature, PLA: Plan curvature; PRCAU: Autumn Precipitation (from Sep. to Nov.); PRCME: Mean of annual Precipitation; PRCSP: Spring Precipitation (from Mar. to May); PRCSU: Sum of annual Precipitation; PRCSR: Summer Precipitation (from Jun. to Aug.); PRCWT: Winter Precipitation (from Dec. to Feb.); PRO: Profile curvature; SLP: Slope; STH: Southness Index; SVF: Sky View Factor; ELE: Elevation; WI: Warmth Index; WST: Westness Index.

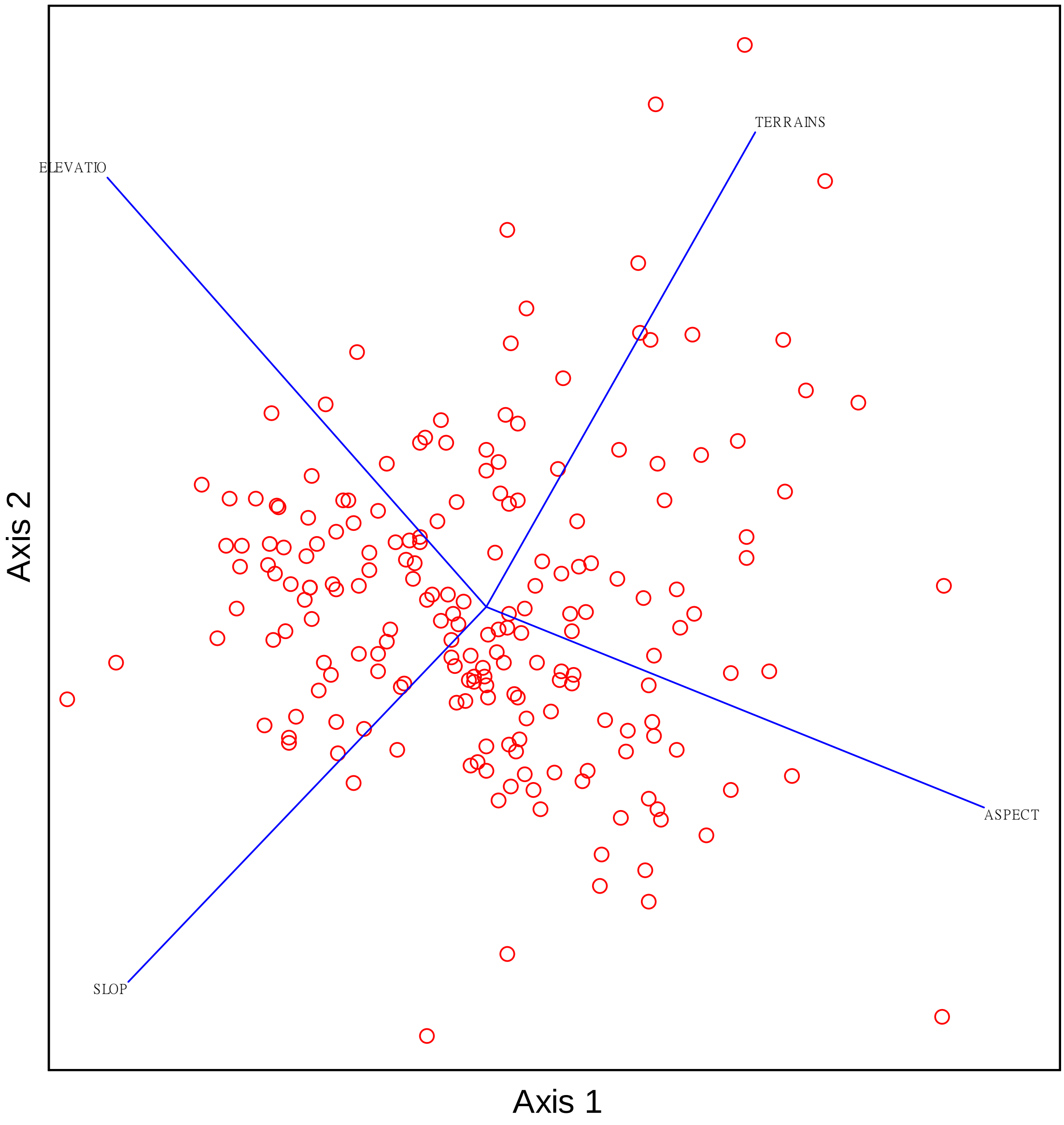
There is similar trend while examining correlation with extracted environmental variables from RIAL and each axis (Table 9). Except elevation gradient, axis 1 is slightly correlated with precipitation, slope, and sky view factor variable (correlation coefficients are all about 0.1 for each variable). Axis 2 is also slightly correlated with precipitation in spring, summer, and winter (correlation coefficients are all about 0.1). Axis 3 is slightly correlated with autumn precipitation, sky view factor and slope variable (correlation coefficients are all about 0.1).

PCA for the relationship between 4 surveyed and 16 extracted environmental variables and 196 Taiwan Hemlock presence localities was showed in Figure 8 and 9, and Eigen value of each component was listed in Table 10 and 11. In 4 surveyed environmental variables case, the first 3 components of PCA explained 79 % variance. On the other hand, the first 5 component of 16 extracted environmental variables PCA explained 75% variance. This result indicates that whether 4 or 16 environmental variables had many similar variance trends (see Figure 9 and 10) and spread radically on the principal component axes and leaded to average each component of variance (i.e. each component almost equally explained the variance of variables). This situation might cause PCA unable to find which direction varied most and leaded to the low explanation of the former components. The following correlation analysis is used to reduce similar variables and avoid of multicollinearity.

Table 10. Variance extracted first 10 axes of PCA from 4 surveyed environmental variables.

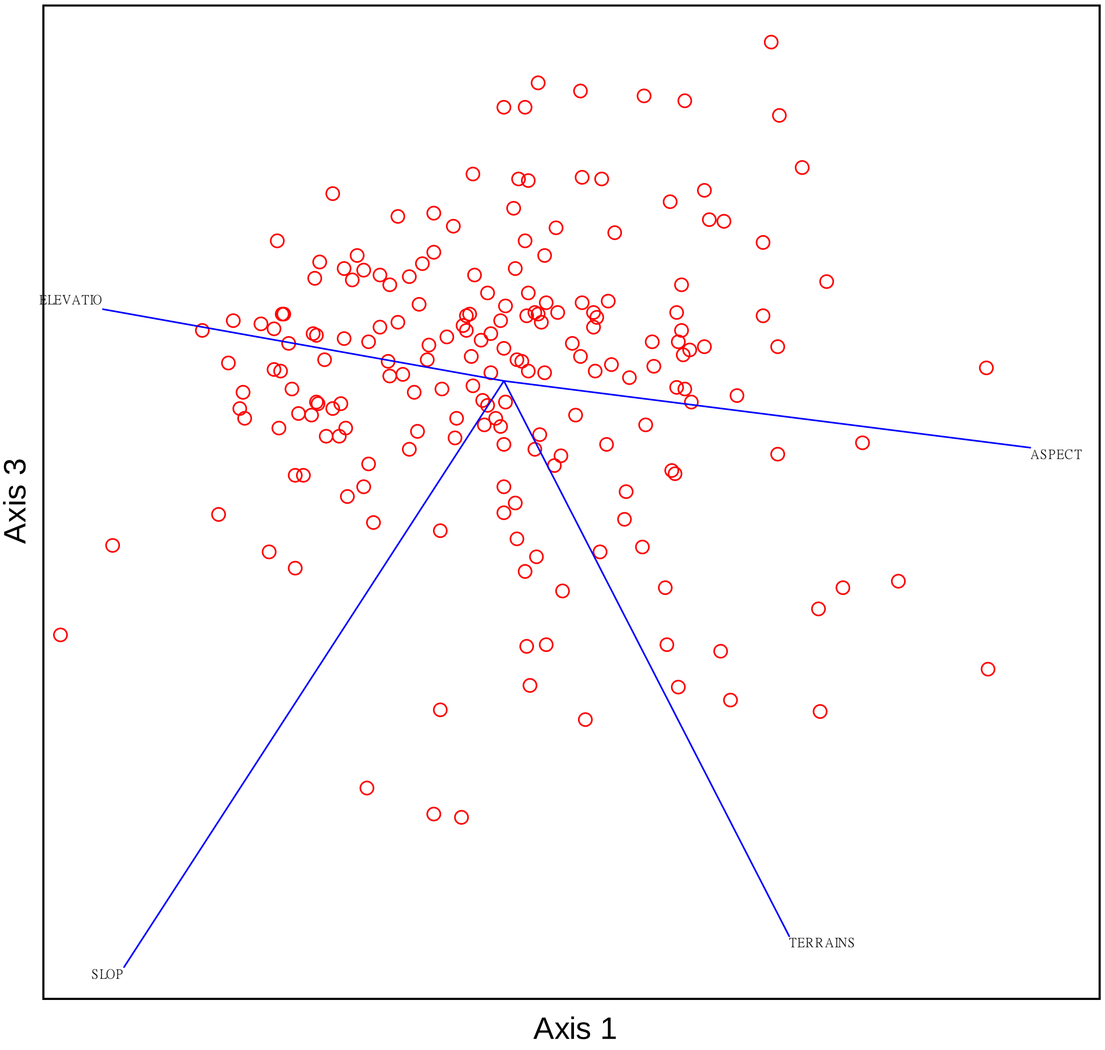
Axis	Eigen value	% of Variance	Cum.% of Var.	Eigen value
1	1.2	29.7	29.7	2.1
2	1.1	27.2	56.8	1.1
3	0.9	22.3	79.1	0.6
4	0.8	20.9	100.0	0.3

Note: Cum.% of Var.: cumulative percentage of variance; %: percentage



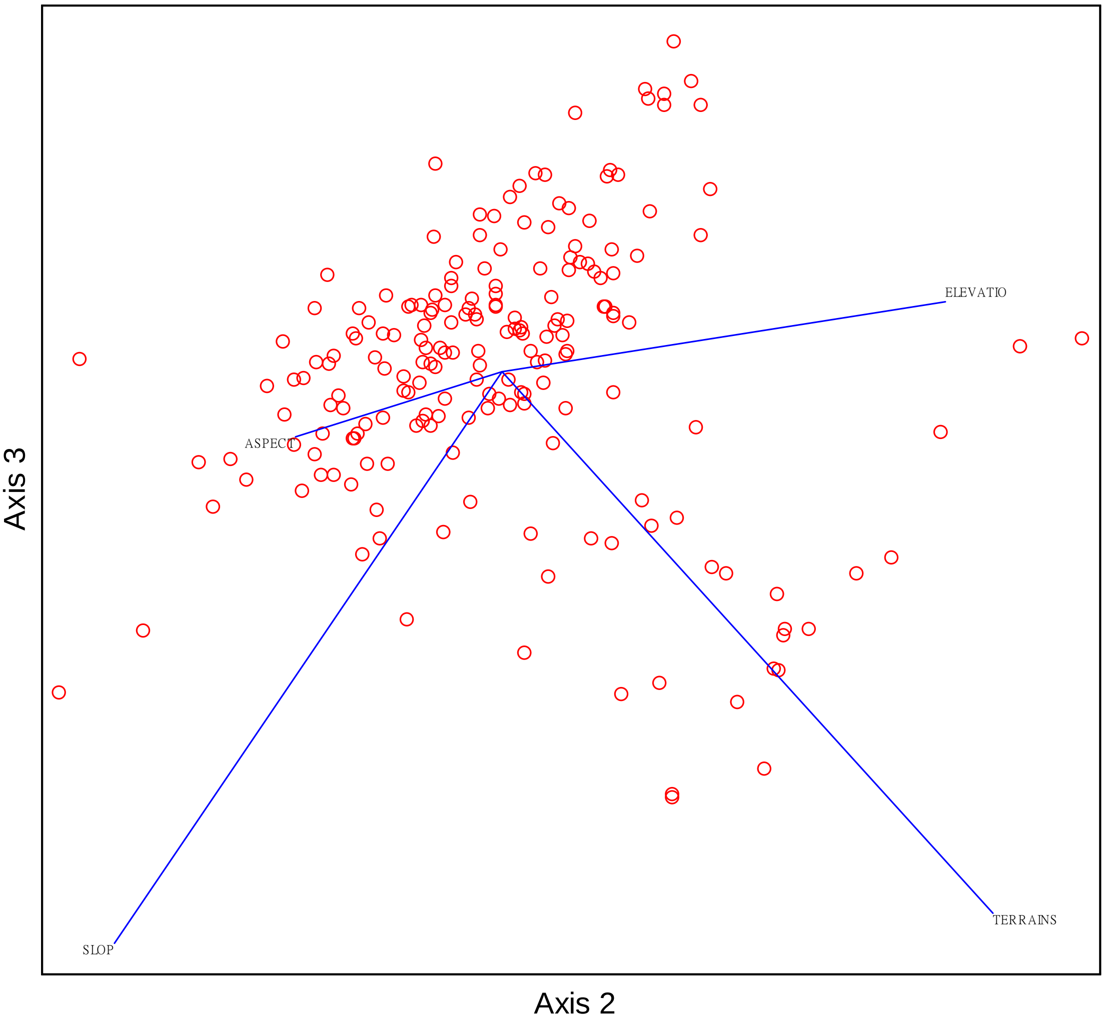
(a)

Figure 9. PCA ordination of Taiwan Hemlock with 4 surveyed environtment variables including ASP, CUR, ELE, PLA, PRCSP, PRCSR, PRCAU, PRCWT, PRCME, PRCSU, PRO, SLP, STH, SVF, WI, and WST. (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.



(b)

Figure 9. PCA ordination of Taiwan Hemlock with 4 surveyed environtment variables including ASP, CUR, ELE, PLA, PRCSP, PRCSR, PRCAU, PRCWT, PRCME, PRCSU, PRO, SLP, STH, SVF, WI, and WST. (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.



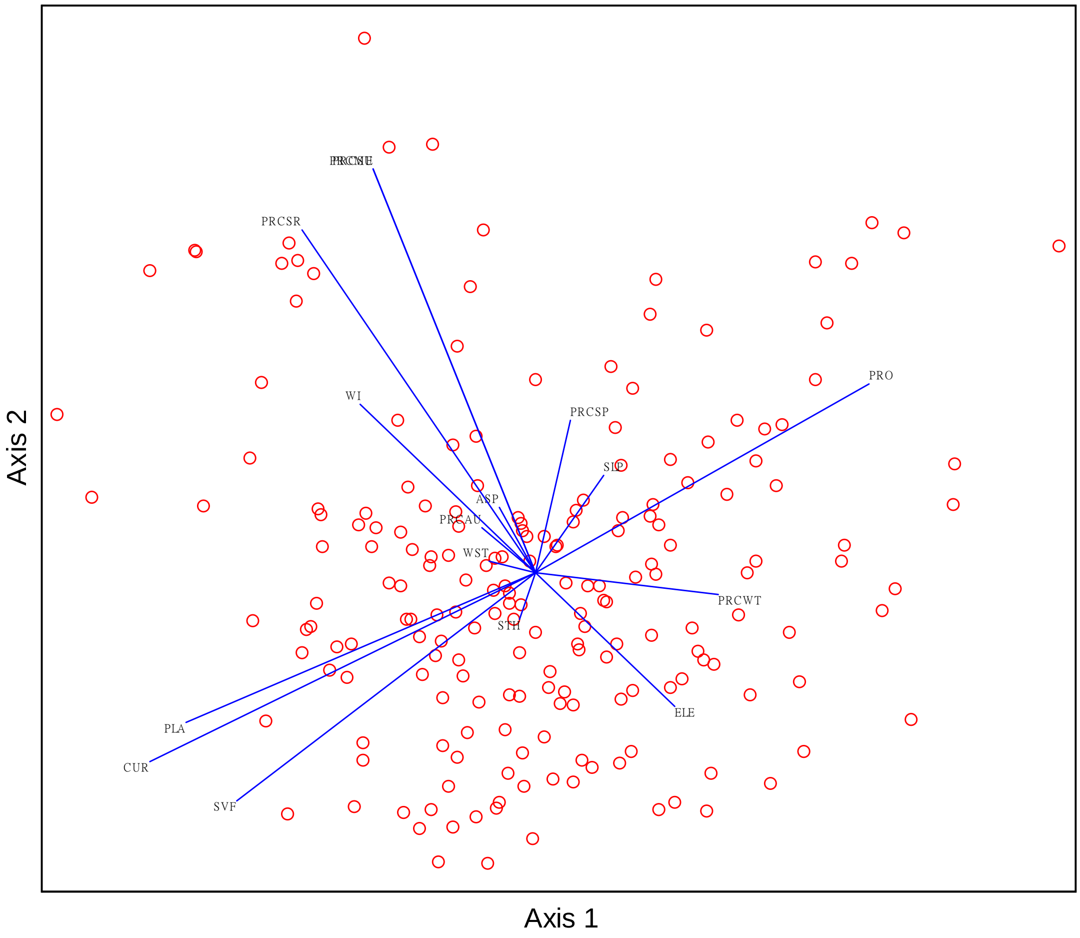
(c)

Figure 9. PCA ordination of Taiwan Hemlock with 4 surveyed environtment variables including ASP, CUR, ELE, PLA, PRCSP, PRCSR, PRCAU, PRCWT, PRCME, PRCSU, PRO, SLP, STH, SVF, WI, and WST. (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.

Table 11. Variance extracted first 10 components of PCA from 16 extracted environmental variables.

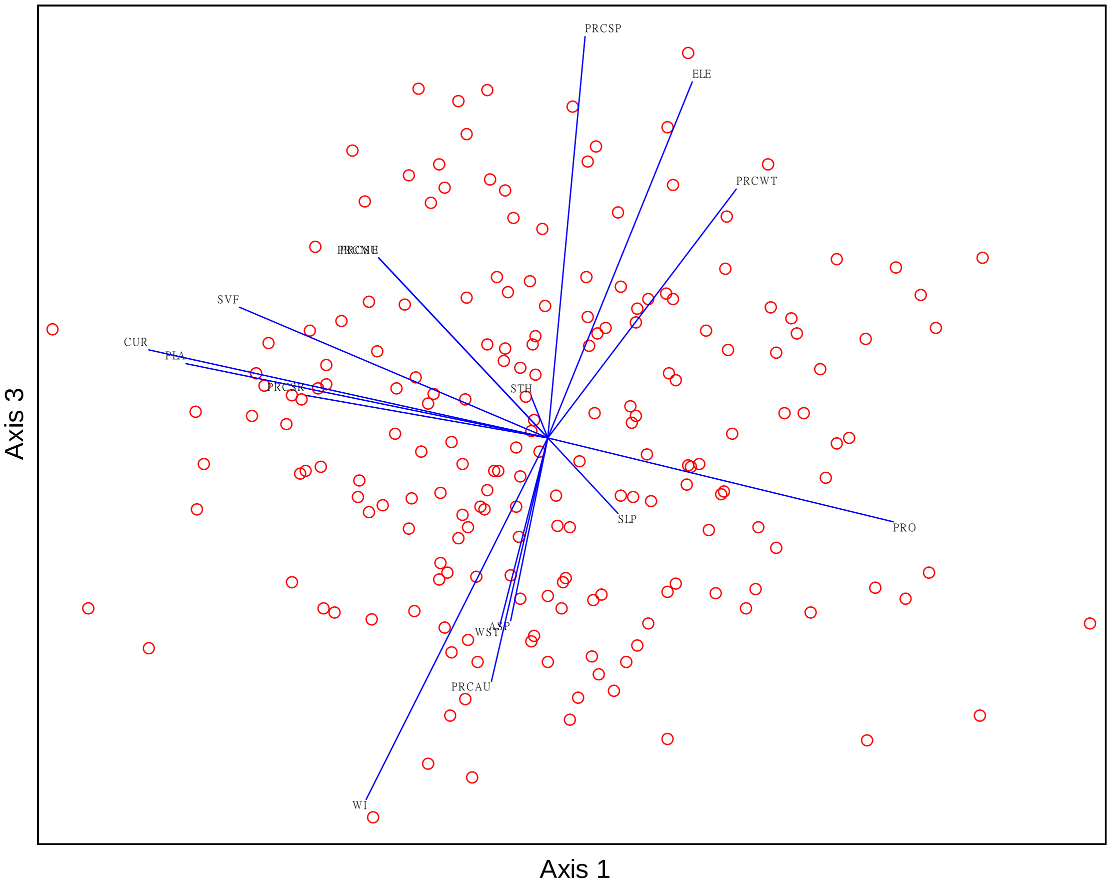
Axis	Eigen value	% of Variance	Cum.% of Var.	Eigen value
1	3.3	20.8	20.8	3.4
2	2.9	18.0	38.8	2.4
3	2.6	16.5	55.2	1.9
4	1.7	10.5	65.7	1.5
5	1.6	10.1	75.8	1.3
6	1.1	6.9	82.7	1.1
7	1.0	6.5	89.2	0.9
8	0.8	4.8	93.9	0.8
9	0.4	2.7	96.6	0.7
10	0.3	1.7	98.3	0.6

Note: Cum.% of Var.: cumulative percentage of variance; %: percentage



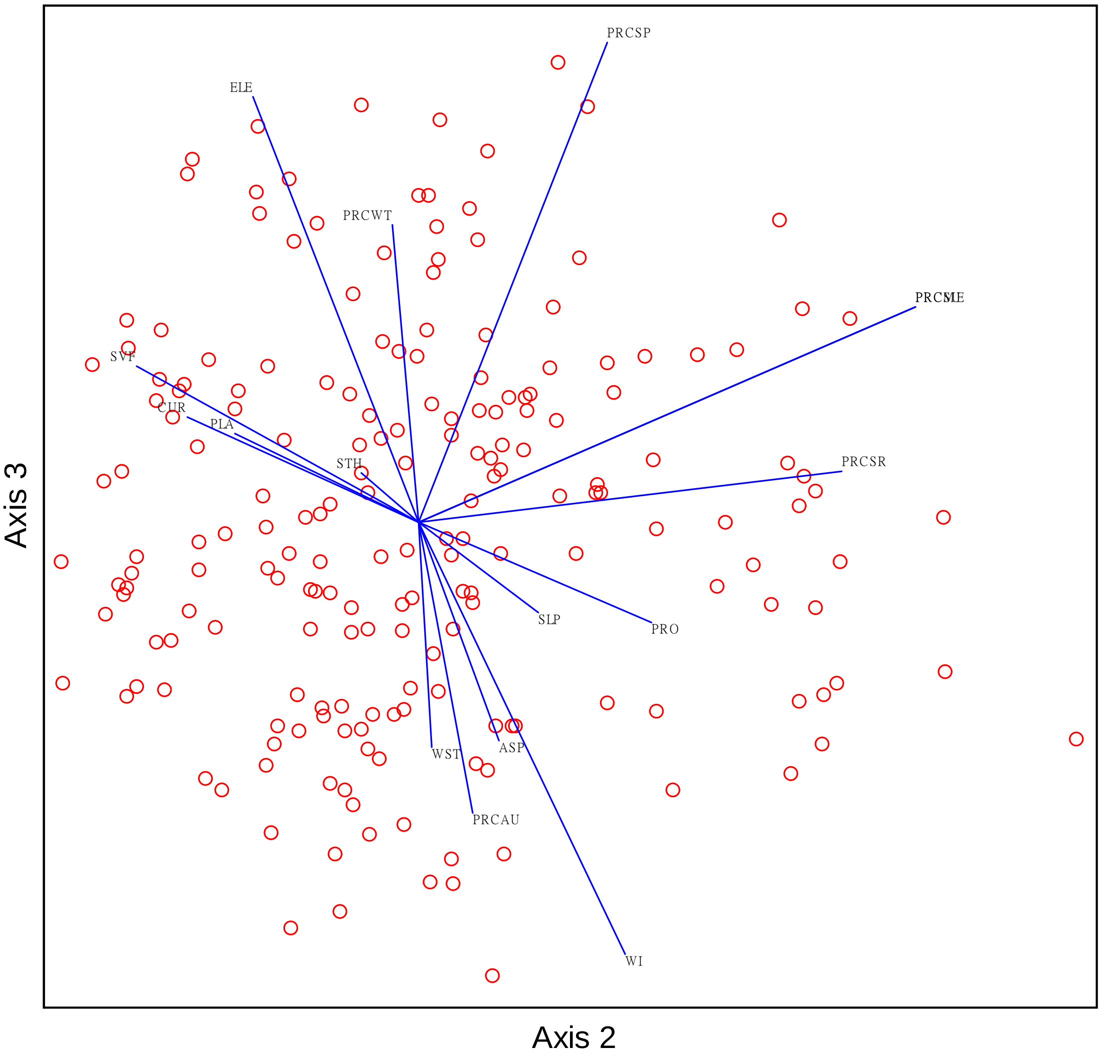
(a)

Figure 10. PCA ordination of Taiwan Hemlock with 16 extracted environtment variables including ASP, CUR, ELE, PLA, PRCSP, PRCSR, PRCAU, PRCWT, PRCME, PRCSU, PRO, SLP, STH, SVF, WI, and WST. (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.



(b)

Figure 10. PCA ordination of Taiwan Hemlock with 16 extracted environtment variables including ASP, CUR, ELE, PLA, PRCSP, PRCSR, PRCAU, PRCWT, PRCME, PRCSU, PRO, SLP, STH, SVF, WI, and WST. (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.



(c)

Figure 10. PCA ordination of Taiwan Hemlock with 16 extracted environtment variables including ASP, CUR, ELE, PLA, PRCSP, PRCSR, PRCAU, PRCWT, PRCME, PRCSU, PRO, SLP, STH, SVF, WI, and WST. (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.

4.2 Environmental Layers Analysis

4.2.1 Correlation Analysis of Environmental Variables

Table 12 demonstrated the results of correlation among 16 environmental variables. To avoiding multicollinearity, the pair with correlation coefficient larger than 0.7 is considered as high correlation to each other and combined in a group to leave 1 variable for building model. The highly correlative variable groups are listed as follow:

(i) ASP, WST

(ii) CUR, PLA, PRO

(iii) PRCAU, PRCSP, PRCW

(iv) PRCME, PRCSU, PRCSR

(v) ELE, WI

For ecological consideration, choose the correlative variables which influence the plant's distribution most. First group, westness index indicates the gradient strength to east and west which is easier to understand than quantitative aspect variable. For example, north aspect both includes 315° to 360° and 0° to 45° and that may cause the greatest and smallest values the same aspect. Second, CUR is selected because CUR represents the curvature of the topographic of Taiwan Island and high value of it indicates convex and low value of it indicates concave. PLA and PRO are similar variables with CUR and only differ at down slope or up slope direction of curvature. PRCW is selected in third group because of difference of dryness and wetness in southern and northern Taiwan in winter. In the fourth group, PRCSU is selected because the summation of the total precipitation represents the maximum intensity of the variable to reflect the extremely climate condition. In the last group, although temperature can be limiting factor to plant growth, ELE is selected because it was directly measured while WI is generate by secondary estimation and might contain more uncertainties than ELE. After correlation analysis, 8 environmental variables are remained: CUR, PRCSU, PRCW, SLP, STH, SVF, WST, and ELE.

Table 12. Correlation analysis for extracted environmental variable of 196 localities of Taiwan Hemlock.

	ASP	CUR	PLA	PRCAU	PRCME	PRCSP	PRCSU	PRCSR	PRCWT	PRO	SLP	STH	SVF	ELE	WST	WI
ASP	1															
CUR	-0.02	1														
PLA	-0.02	0.91	1													
PRCAU	0.03	-0.04	-0.01	1												
PRCME	0.03	0.03	0.07	0.16	1											
PRCSP	-0.05	0.00	0.02	-0.67	0.42	1										
PRCSU	0.03	0.03	0.07	0.16	1.00	0.42	1									
PRCSR	0.07	0.11	0.10	-0.05	0.76	0.14	0.76	1								
PRCWT	-0.06	-0.12	-0.08	-0.15	0.04	0.55	0.04	-0.54	1							
PRO	0.01	-0.86	-0.57	0.08	0.02	0.01	0.02	-0.09	0.14	1						
SLP	0.01	-0.04	-0.05	-0.07	0.00	0.05	0.00	0.00	0.06	0.03	1					
STH	0.07	0.00	0.03	0.03	0.00	-0.03	0.00	-0.01	0.01	0.04	-0.25	1				
SVF	-0.06	0.66	0.55	0.03	-0.02	-0.08	-0.02	0.08	-0.22	-0.63	-0.46	0.17	1			
ELE	-0.17	-0.10	-0.13	-0.17	-0.05	0.23	-0.05	-0.15	0.19	0.05	-0.12	0.09	0.17	1		
WST	0.72	0.04	0.06	0.04	-0.05	-0.10	-0.05	0.01	-0.11	0.01	0.05	0.04	0.01	-0.13	1	
WI	0.17	0.12	0.14	0.21	0.13	-0.28	0.13	0.29	-0.32	-0.07	0.12	-0.09	-0.14	-0.98	0.11	1

4.2.2 Attributes of Environmental Variables

Descriptive statistic provides a basic sense and structure of the data. Taiwan Hemlock presence data set (N = 196) and absence localities (N = 3784) are compared by extracting values from 8 selected environmental layers. Missing values are removed from the extracted absence localities and remains 3770 absence localities instead. Both basic and descriptive statistics are listed in Table 13 and 14.

Table 13. Basic and descriptive statistic of presence localities (N = 196)

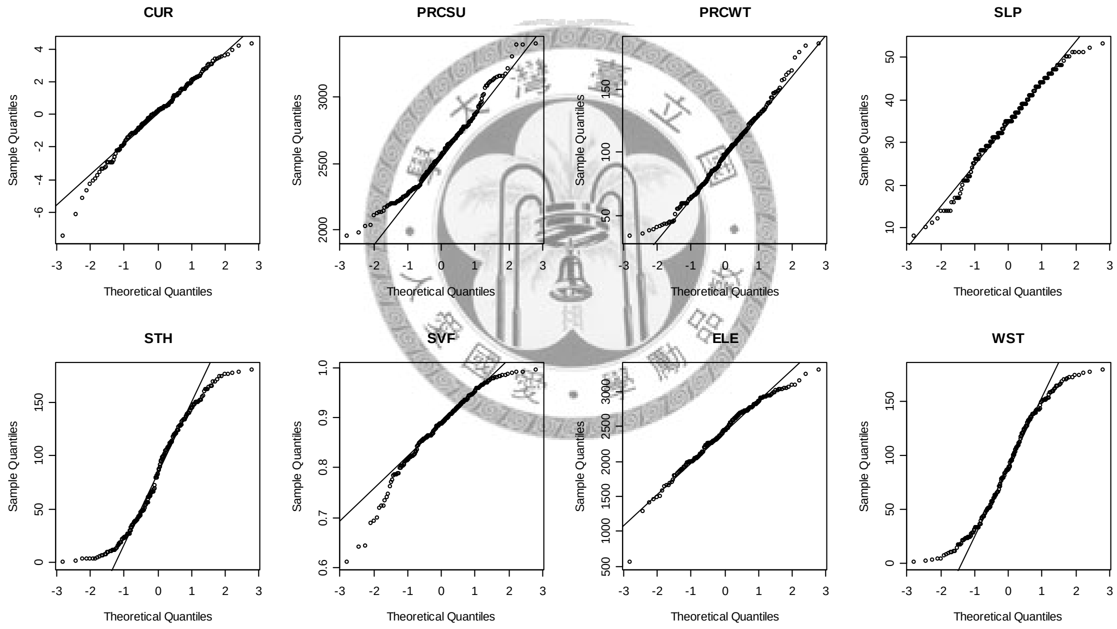
	CUR	PRCSU	PRCWT	SLP	STH	SVF	ELE	WST
min	-7.50	1949	33	8	1	0.61	559	2
max	4.31	3396	185	53	180	1.00	3315	179
range	11.81	1447	152	45	179	0.38	2756	177
median	0.13	2550	97	35	82	0.89	2440	87
mean	0.03	2577	97	34	83	0.88	2416	89
std.dev	2.00	302	32	10	52	0.07	430	50

Table 14. Basic and descriptive statistic of absence localities (N = 3770)

	CUR	PRCSU	PRCWT	SLP	STH	SVF	ELE	WST
min	-7.56	1132	13	0	0	0.45	0	0
max	10.25	5489	649	67	180	1.00	3777	180
range	17.81	4357	636	67	180	0.55	3777	180
median	0.00	2300	60	19	85	0.96	375	91
mean	0.05	2365	80	19	84	0.93	685	93
std.dev	1.41	673	71	16	56	0.07	767	50

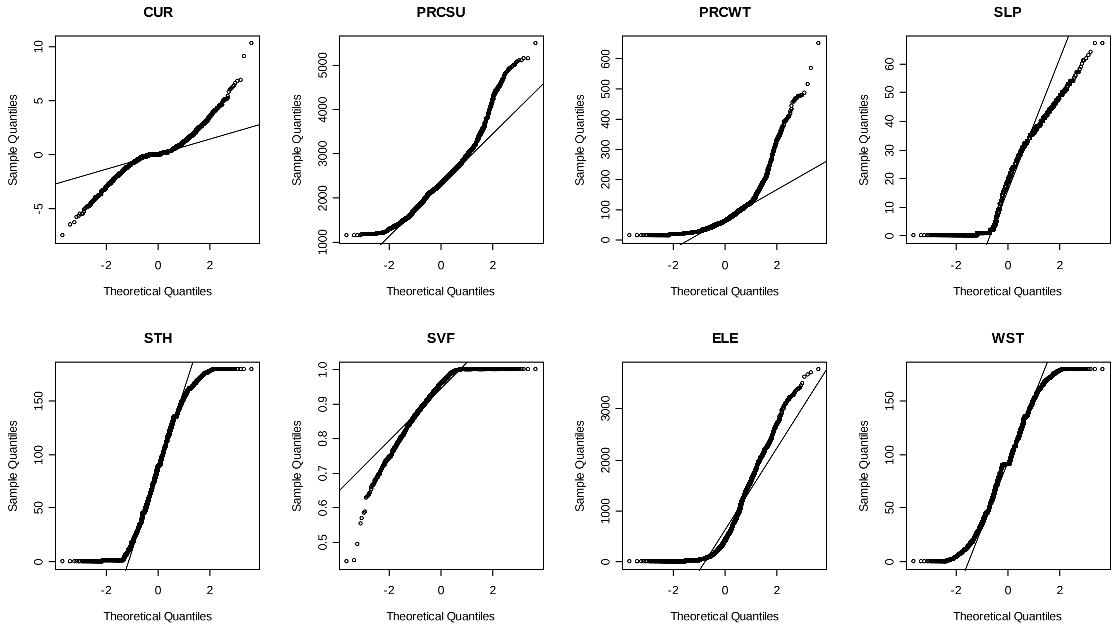
Before t-test, normal Q-Q plot method showed in Figure 11 represented the data is almost normal distributed at Taiwan Hemlock presence data sets but not at Taiwan

Hemlock absence data sets. Two-tailed t-test and 95 % confidence interval are used to test if the two data sets are significantly difference (Table 15.). The null hypothesis of the t-test is that true difference in means is equal to 0 and alternative hypothesis is that true difference in means is not equal to 0. Only STH and SVF are not significantly difference between absence and presence localities. On the other hand, the rest 5 environmental variables are significant difference between absence and presence localities. Figure 12 shows the visual sense of the difference in environmental variables between absence and presence by histograms with relative frequency at y axis.



(a)

Figure 11.Normal Q-Q plot of 8 extracted environmental variables (a) Taiwan Hemlock presence localities, (b) Taiwan Hemlock absence localities.



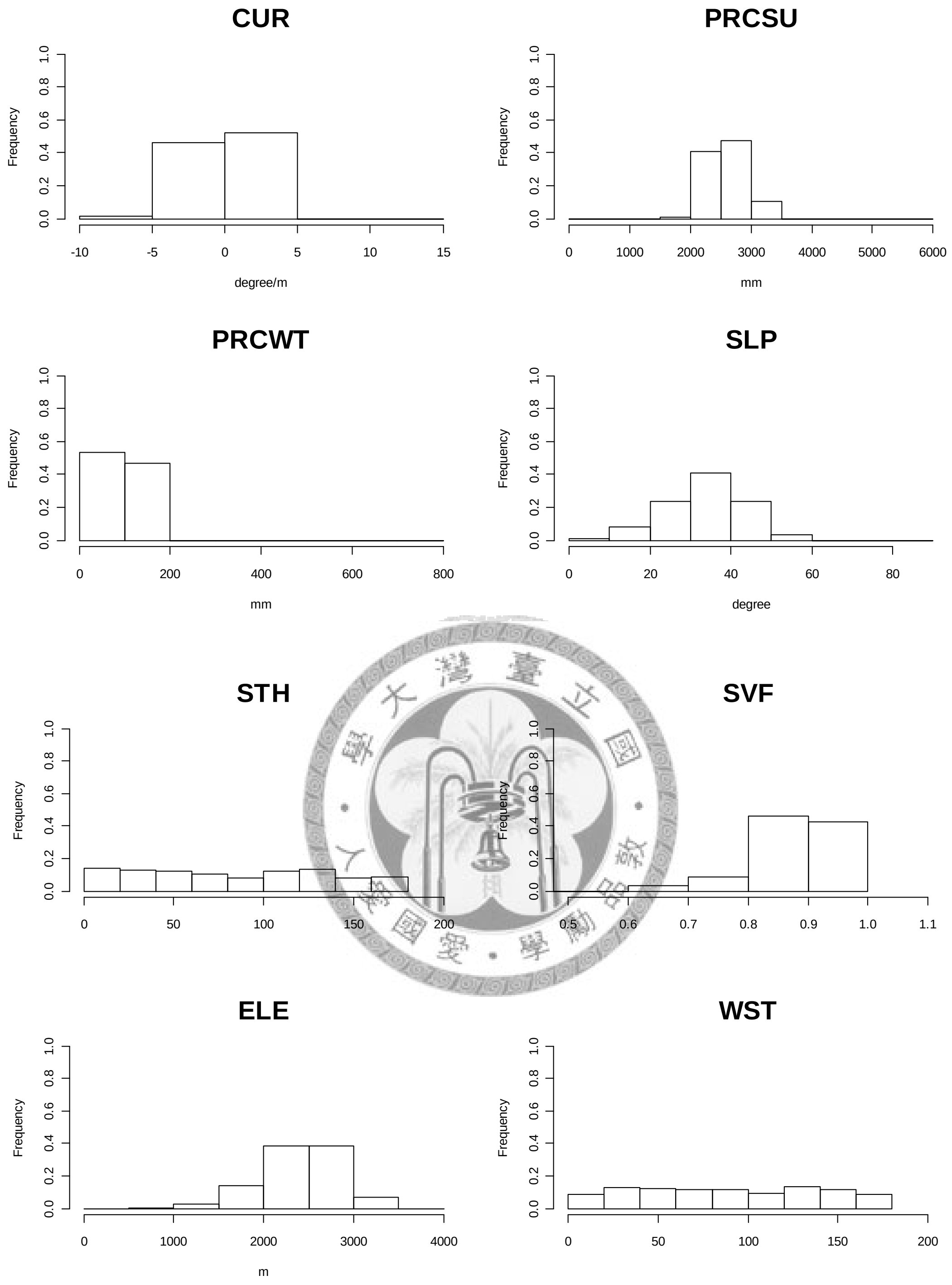
(b)

Figure 11.Normal Q-Q plot of 8 extracted environmental variables (a) Taiwan Hemlock presence localities, (b) Taiwan Hemlock absence localities.

Table 15. t-test of absence and presence localities

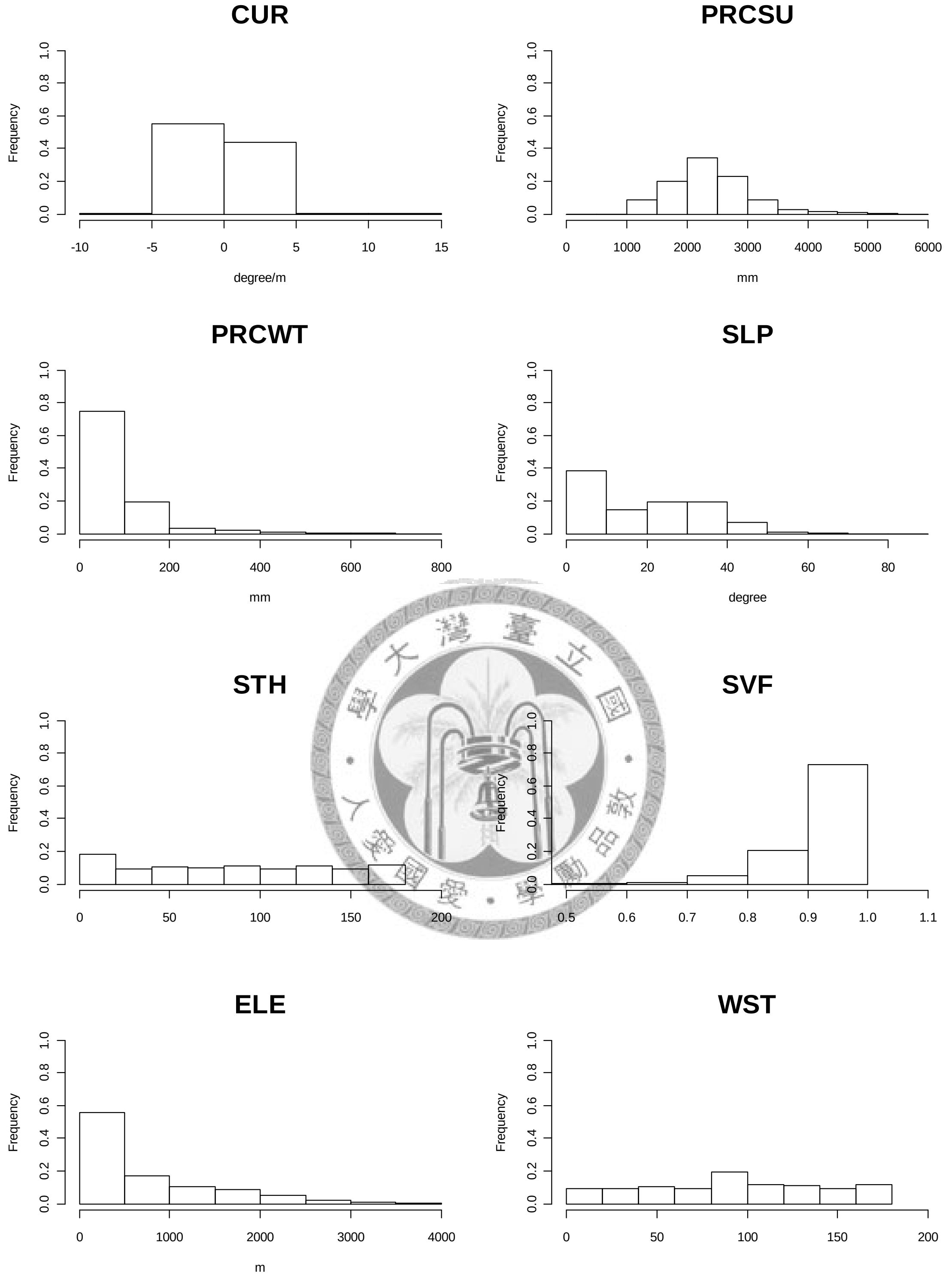
	CUR	PRCSU	PRCW	SLP	STH	SVF	ELE	WST
t	0.1	-8.8	-6.5	-21.1	0.3	10.3	-52.2	1.0
p-value	0.92	< 0.001	< 0.001	< 0.001	0.8	< 0.001	< 0.001	0.32
		**	**	**		**	**	

Note: **: significant



(a)

Figure 12. Histogram of 8 extracted environemtal variable between absence and presence localities. Y axis represents the relative frequency of counts of sample. (a) Presence data set. (b) Absence data set.

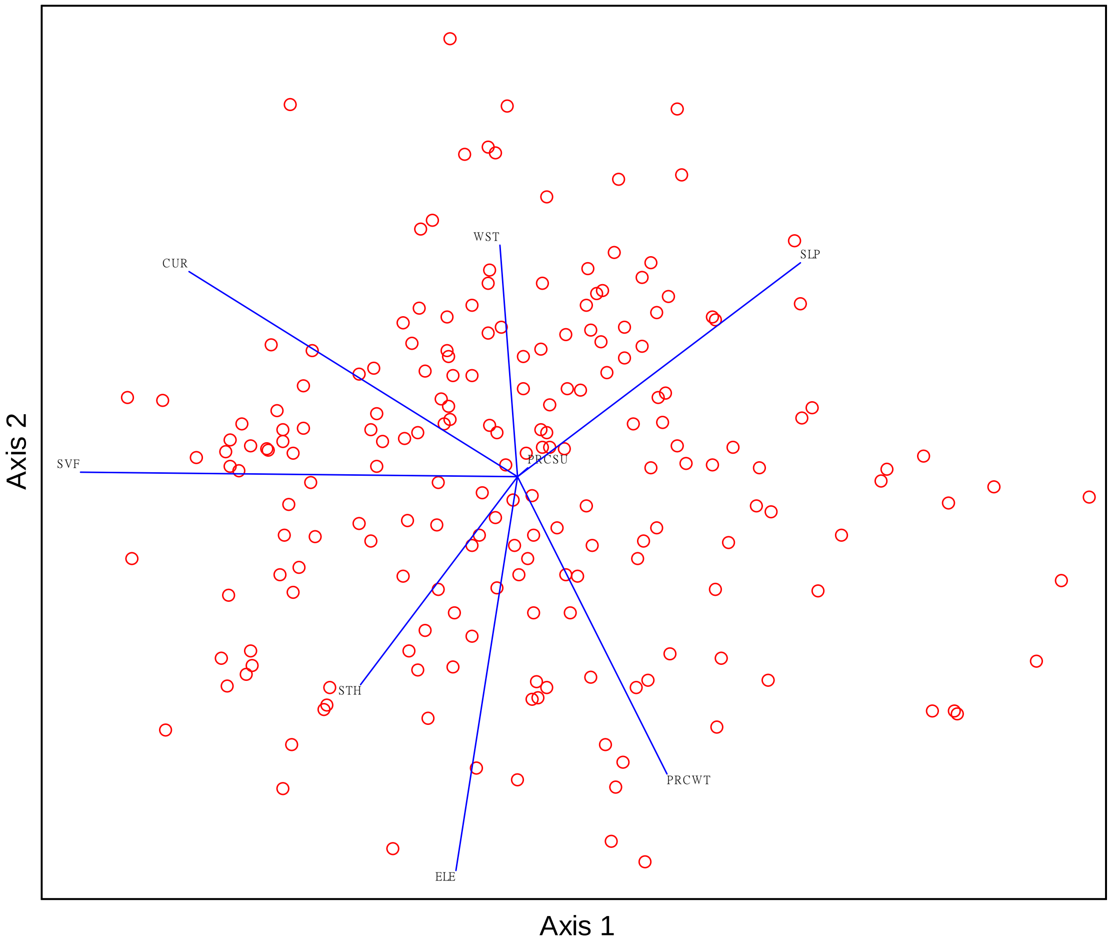


(b)

Figure 12. Histogram of 8 extracted environemtal variable between absence and presence localities. Y axis represents the relative frequency of counts of sample. (a) Presence data set. (b) Absence data set.

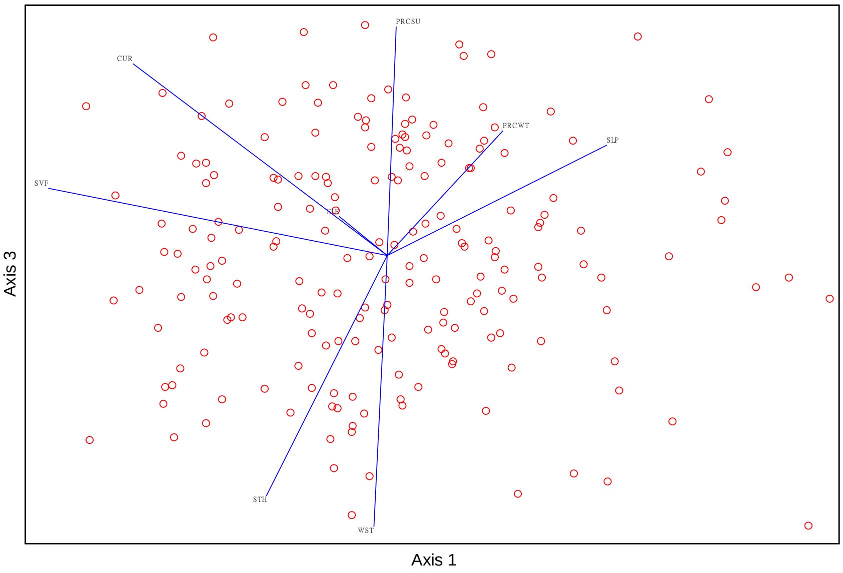
4.2.3 PCA Approaches

Plots of two PCA analysis for all occurrences (N = 196) is showed in Figure 13 and revealing similar trend whether in axis 1, 2, or 3.

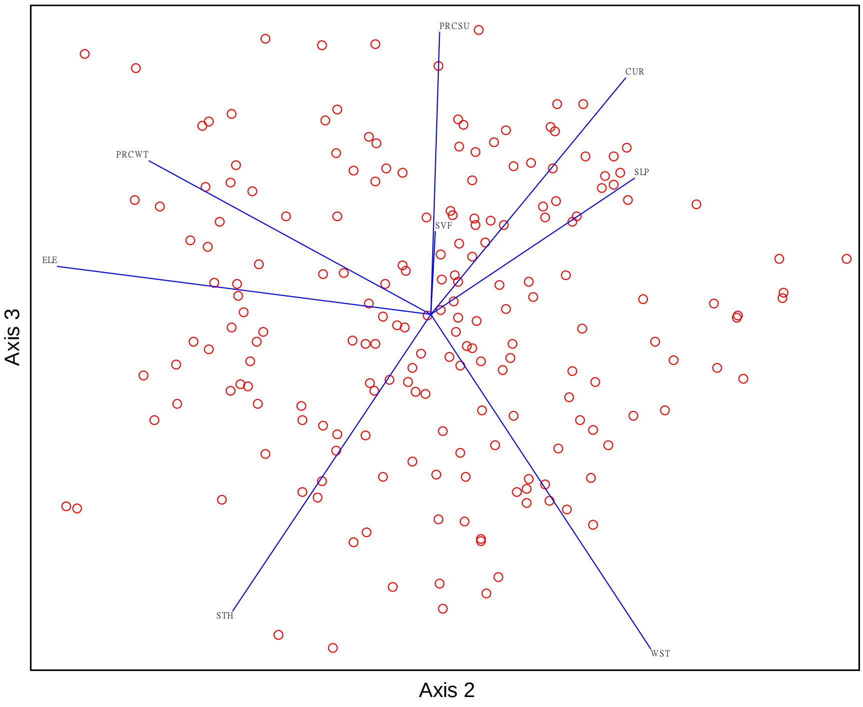


(a)

Figure 13. PCA ordination of Taiwan Hemlock and 8 extracted environtment variables including CUR, PRCSU, PRCWT, SLP, STH, SVF, ELE and WST. (N=196) (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.



(b)



(c)

Figure 13. PCA ordination of Taiwan Hemlock and 8 extracted environtment variables including CUR, PRCSU, PRCWT, SLP, STH, SVF, ELE and WST. (N=196) (a) Axis 1 and 2, (b) Axis 1 and 3, and (c) Axis 2 and 3.

The variance explained by the component 1 to 3 are 24%, 17%, and 13% respectively for all occurrences data and cumulative percentage of variance explained by the first three components are 24%, 41% and 55% respectively (Table 16). First 6 eigenvectors, listed in Table 17 and each scaled to its standard deviation (SD), sometimes called V vectors, and when applied to PCA of a correlation matrix, are the same as the correlation coefficient between scores for occurrence data and the environmental variables. First component is highly related to the three environmental variables, CUR, SLP, and SVF, second component is highly related to the PRCWT and ELE variables, and the third component is highly related to the STH and WST variable respectively (Table 17).

Table 16. Percentage of variance and cumulative percentage of variance from extracted 8 components of PCA.

Axis	Eigen value	% of Variance	Cum.% of Var.	Eigen value
1	2.0	24.4	24.4	2.7
2	1.4	17.3	41.8	1.7
3	1.1	13.7	55.5	1.2
4	1.0	12.6	68.1	0.9
5	0.9	11.3	79.4	0.6
6	0.8	9.7	89.0	0.4
7	0.7	8.7	97.8	0.3
8	0.2	2.2	100.0	0.1

Note: Cum.% of Var.: cumulative percentage of variance; %: percentage

Table 17. First 6 eigenvectors each scaled to its standard deviation of PCA.

	Eigenvector					
Env.	1	2	3	4	5	6
CUR	-0.70	0.37	0.40	0.14	-0.28	-0.23
PRCSU	0.02	0.02	0.48	-0.83	-0.11	0.28
PRCWT	0.32	-0.54	0.26	0.03	-0.54	-0.34
SLP	0.60	0.39	0.23	0.18	-0.28	-0.14
STH	-0.33	-0.37	-0.50	-0.41	-0.09	-0.43
SVF	-0.93	0.01	0.14	0.11	-0.07	0.05
ELE	-0.13	-0.71	0.08	0.30	-0.21	0.47
WST	-0.04	0.42	-0.56	-0.08	-0.62	0.31

4.2.4 Data Mining Approach: CART

11 of 16 environmental variables, ELE, PRCSU, PRO, ASP, PRCAU, PRCSR, WI, PRCSP, WST, SVF, and STH, are selected by the classification tree to split each node of the tree (see Figure 14). First node is split by ELE variable indicates ELE is able to distinguish the two groups divided it. In other word, ELE is the most explainable variable from the 16 environmental variables due to the two groups split by ELE having more homogeneity than original one. And besides, the second splitting variable using by classification tree is PRCSU and WI and the third is PRCSR and PRO in the same node. On the left hand side of the tree is mainly consist of V_1 and a few V_2 and V_5, on the middle of the tree is mainly consist of V_3 and V_5, and on the right hand side is mostly V_2 and mixed with V_3 and V_4. Total number of terminal nodes is 24 and residual mean deviance is 1.74. Misclassification error rate of the six vegetation type is 32 %. Figure 15 and 16 showed how each split reduces the variance and the misclassifying rate of split groups.

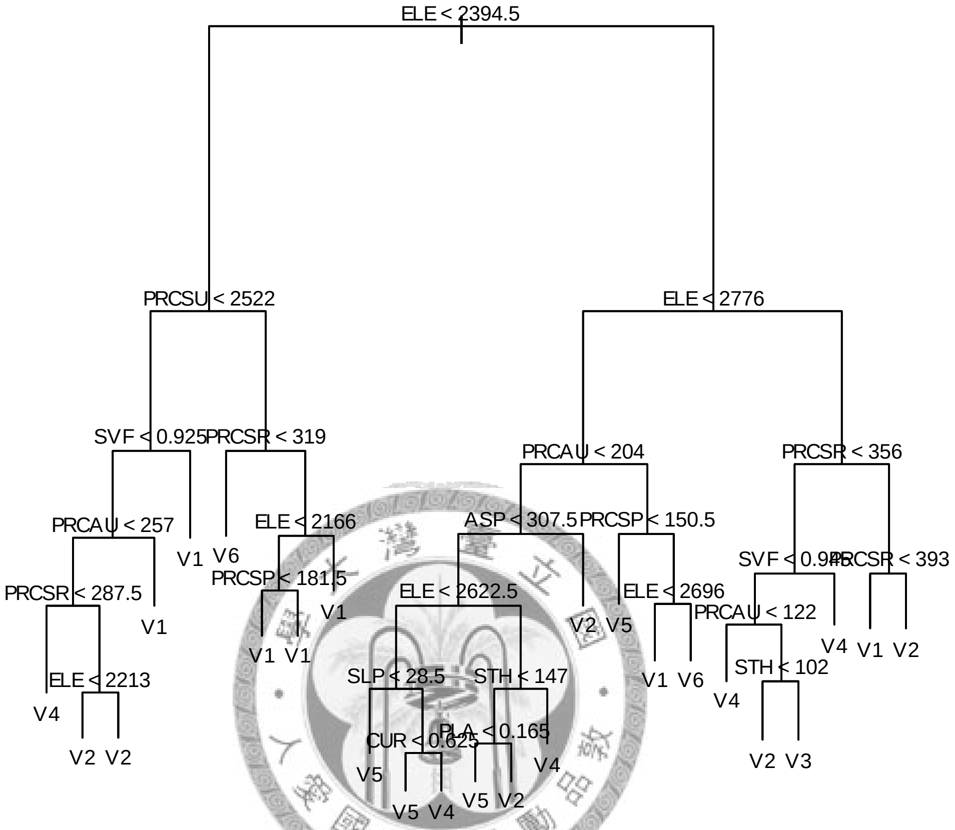


Figure 14. Tree plot of classification and regression tree (CART) analysis with 6 groups of 196 Taiwan Hemlock presence localities and 16 extracted environmental variables from RIAL.

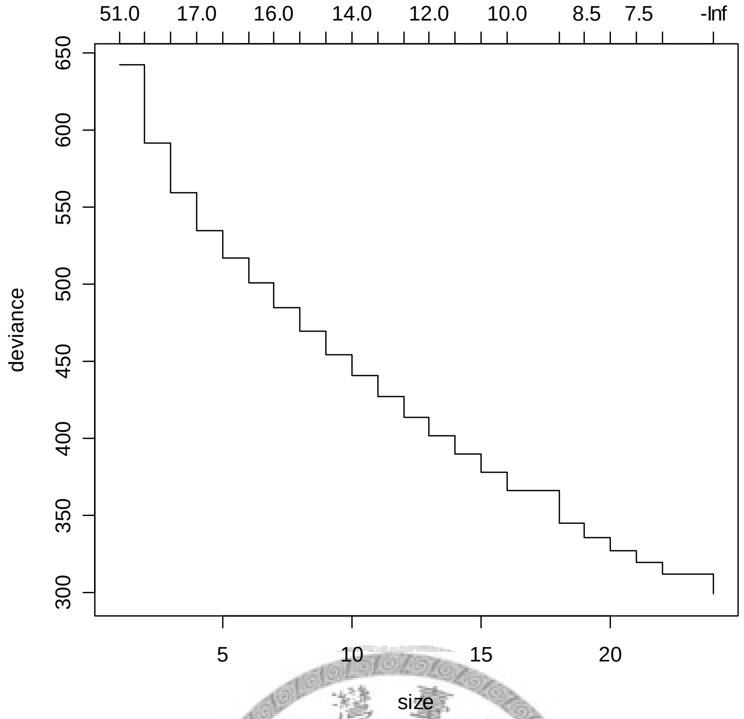


Figure 15. Prune tree of CART analysis from Figure 13 shows how each split affects the deviance.

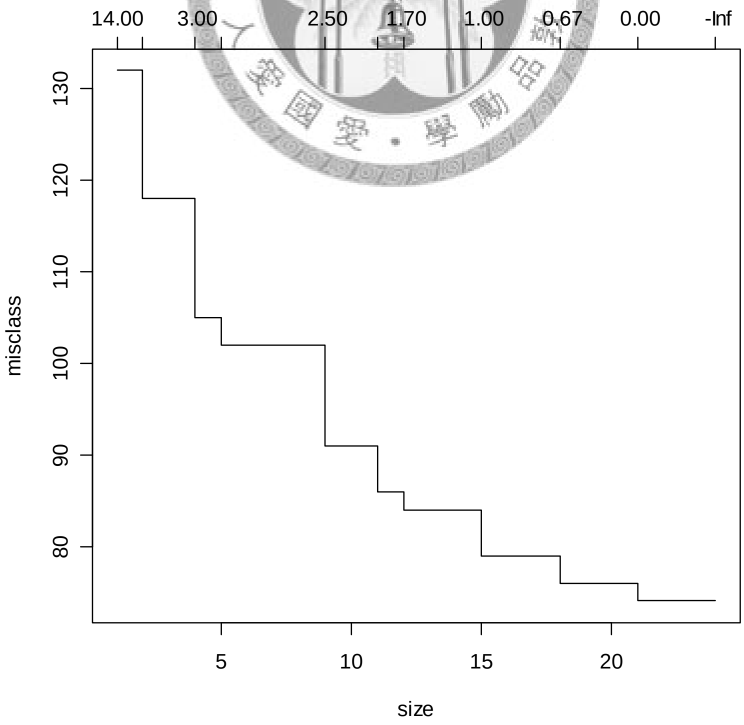


Figure 16. Missclassification of tree plot from Figure 13 shows how each split reduces misclassification number.

Figure 17 showed the result of combining V_1 and V_6 into V_1 due to similar habitat of both vegetation types found in field surveys. The combination resulted in lowering misclassifying rate from 35% into 27%. If considering V_4 for succession type and taking it as V_1, the misclassification of each group was reduced into 23%. The details of summary of CART analysis for 6 groups and 16 extracted environmental variables are listed in Table 18. First three splitting variables of Figure 16 are the same as the results of Figure 13. On the left hand side of the tree is mainly consist of V_1 and a few V_2 and V_5, on the middle of it is consist of V_2 and V_5, and on the left hand side is V_3 and a few V_2 and V_5. Total number of terminal nodes is 24 and residual mean deviance is 1.5.

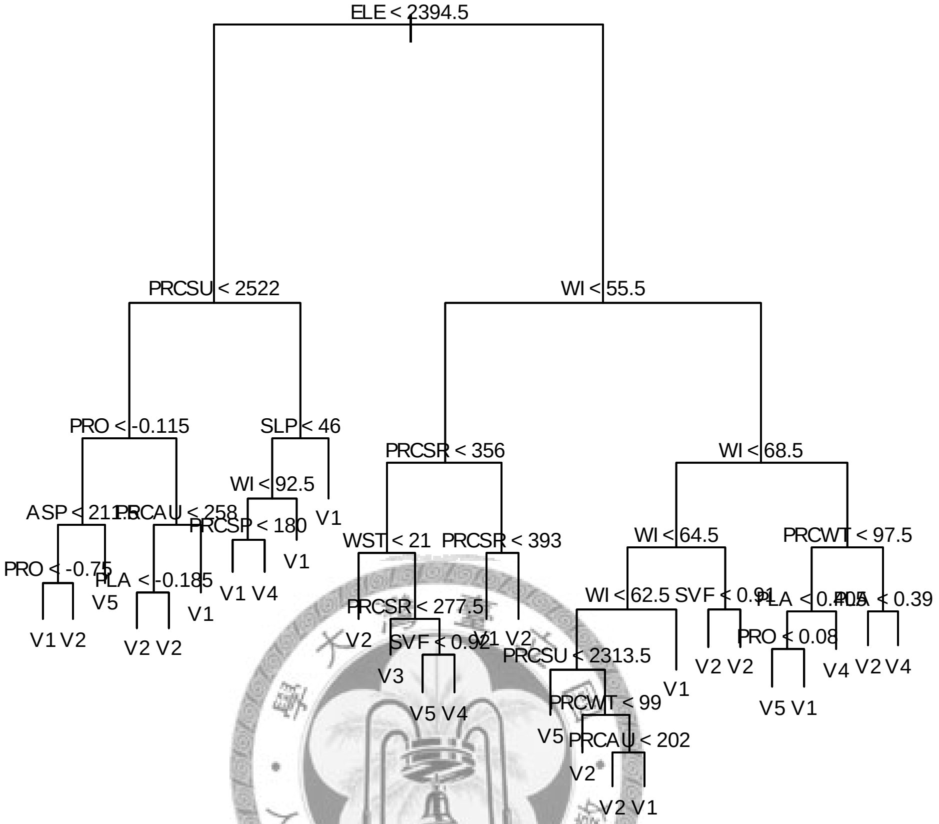


Figure 17. Tree plot of classification and regression tree (CART) analysis with 5 groups (combining V_1 and V_6 into V_1) of 196 Taiwan Hemlock presence localities 16 extracted environmental variables from RIAL.

The numbers of classified plot are greater than the original number of V_2, V_5 and V_6 vegetation type. In V_1, V_3, and V_4 situation, however, the numbers of classified plot are lesser than the origin number.

Table 18. Summary of CART analysis with 6 groups of 196 Taiwan Hemlock Presence localities 16 extracted environmental variables from RIAL.

	V_1	V_2	V_3	V_4	V_5	V_6
Number of original plot	64	46	11	32	28	15
Number of classified plot	61	60	4	27	32	16
Missing plot	18	9	7	21	6	7
Missing rate (%)	28	20	64	66	21	47
Total missing rate (%)			35			

4.2.5 Data Mining Approach: CIT

Only 1 of the 16 environmental variables, WI, was actually used to construct the conditional inference tree and represent the significant difference between the both groups split by the environmental variable on the node and the originally undivided group (Figure 18). Total numbers of the terminal nodes is 3.

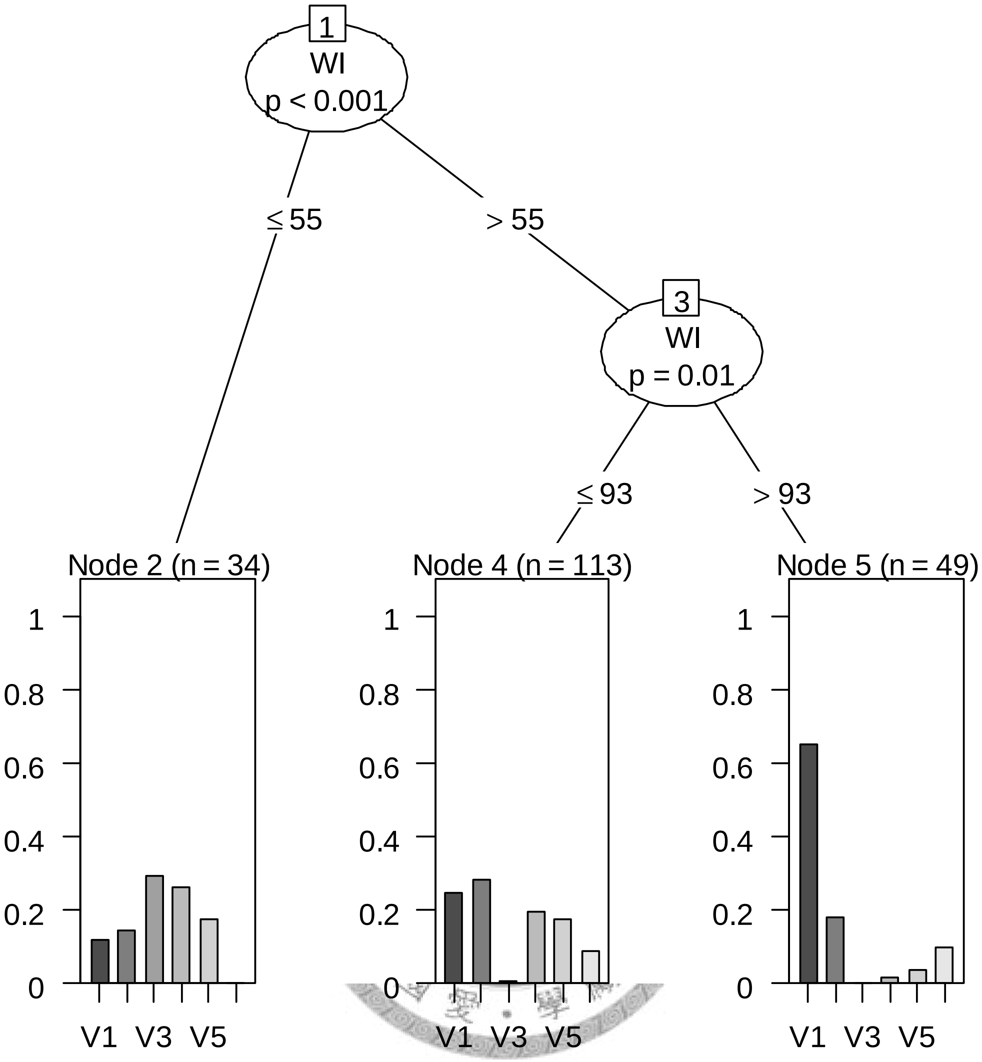


Figure 18. Tree plot of CTT analysis with 6 groups of 196 Taiwan Hemlock presence localities 16 extracted environmental variables from RIAL.

On the right hand side of CIT in Figure 17 is mainly consist of V_1 and mixed V_2 and V_6, on the middle of the node is mainly consist of V_2 mixed with V_1 , V_4 and $V_{5,}$ and on the left hand side is V_3 and companied with $V_{2,}$ V_4 and V_5. The results indicated that warmth index gradient could distinguish Taiwan Hemlock presence localities into 3 three groups: (i)Taiwan Hemlock- Taiwan Cypress group, (ii)Taiwan Hemlock mixed with conifers groups, and (iii) Taiwan Hemlock-Taiwan Fir group. The details of summary of CART analysis for 6 groups and 16 extracted environmental variables are listed in Table 19. The result of CIT showed relative high misclassification rate comparing to the results of CART analysis.

V_4 to V_6 are not classified by the CIT analysis and gains 100 % missing rate which influencing the total misclassification rate in comparing to the misclassification rate of occurrences data without V_4 to V_6 vegetation type or V_4 to V_6 are able to be distinguished by 16 extracted environmental variables. And besides, V_4 and V_5 were mainly predicted to V_2 and V_3 and that indicated they shared similar environmental gradient with V_2 to V_3; V_6 was mainly predicted to V_1 and V_2 and still implied its environmental gradient was similar with V_1 and V_2. The numbers of classified plot are greater than the original number of V_2 and V_3 vegetation type. In V_1 situation, however, the number of classified plot is lesser than the origin number.

Table 19. Summary of CIT analysis with 6 groups of 196 Taiwan Hemlock Presence localities 16 extracted environmental variables from RIAL.

	V_1	V_2	V_3	V_4	V_5	V_6
Number of original plot	64	46	11	32	28	15
Number of classified plot	49	113	34	0	0	0
Missing plot	32	14	1	32	28	15
Missing rate (%)	50	30	9	100	100	100
Total missing rate (%)			62			

16 environmental variables are reduced to 8 by correlation analysis and then partially selected by each statistical and data mining methods summarized in Table 20. PC1 selected 3 environmental variables, CUR, SLP, and SVF, PC2 selected 2 environmental variables, PRCWT and ELE, PC3 selected 2 environmental variables,

STH and WST, CART selects almost all environmental variables except CUR, PLA, PRCME, PRCWT and SLP, and CIT selects only 1 environmental variable, WI. In addition to the methods mentioned above, try and error approach (i.e. selecting all and individual environmental variables for model building) is also considered while building the Maxent distribution model.

Table 20. Summary of all approaches selecting influential environmental variables to distribution of Taiwan Hemlock. Character "V" means the variable is selected by the method.

	CA	PC1	PC2	PC3	CART	CIT	ALL
ASP					V		V
CUR	V	V					V
PLA							V
PRCAU					V		V
PRCME							V
PRCSP					V		V
PRCSU	V				V		V
PRCSR					V		V
PRCWT	V		V				V
PRO					V		V
SLP	V	V					V
STH	V			V	V		V
SVF	V	V			V		V
ELE	V		V		V		V
WST	V			V	V		V
WI					V	V	V

Note: V means the environmental variable was selected by each method; CA means correlation analysis; PC1 represents environmental variables with high correlation with component 1 of PCA.PC2 and PC3 represent the same meaning with PC1; CART represents classification and regression tree; CIT means conditional inference tree; ALL represents all 16 extracted environmental variables.

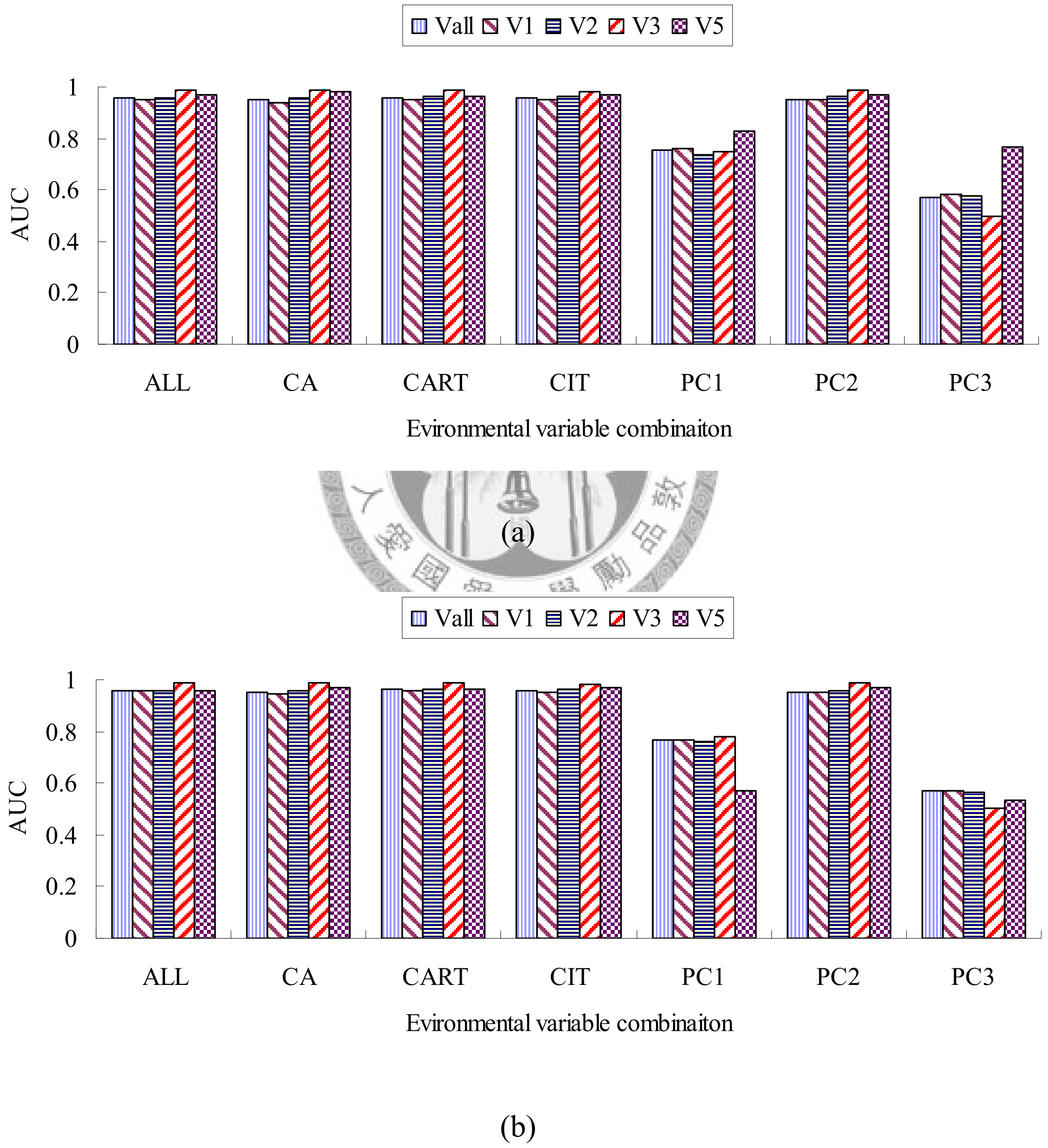
Therefore Total 23 combination of environmental selection for SDM building including 7 environmental combinations in Table 20 (CA, PC1, PC2, PC3, CART, CIT, and ALL) and each of 16 extracted environmental variables.

4.3 SDM Outputs and AUC

4.3.1 Resolution, Presence Unit, Environmental Variable Selection and AUC

Vegetation unit is considered a more homogeneous unit and stable in succession stage than all species occurrence localities which may contain mixed plant compositions and more variant in data structure. Figure 19 shows how AUC of V_1, V_2, V_3, V_5, and V_{all} differ from each other. Due to V_1 and V_6 usually appear in similar environment and nearby on DCA plot, they were combined into V_1 only. In V_4 situation, pine species are not only considered as succession species and it is reasonable to explain the pure stand of Taiwan Hemlock with some disturbance and companies with pine species in nature, but also near to V_2 on DCA plot. So V_2 and V_4 are combined together into V_2 vegetation type. Multiple Behrens-Fisher tests by Package of npmc in R for 3 units, including resolution, input locality, and environmental variable combination, reveal effects of different respects. Three resolutions did not differ from each other significantly (all p-values > 0.86). For input locality perspective, only V_1 and V_5 are significant different with p-value equal to 0.048 and the rest combination did not differ significantly among each other (Table 21). There was no significant difference among

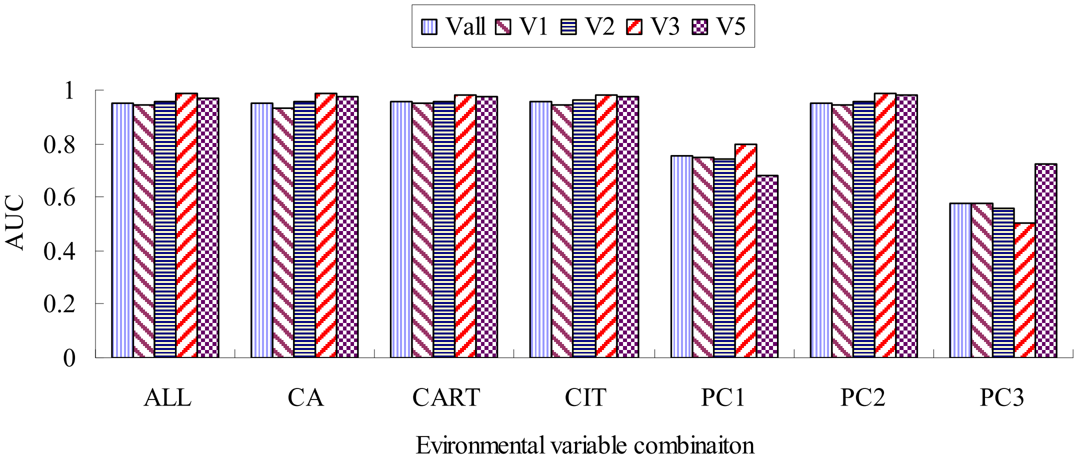
ALL, CA, CART, CIT and PC2 (all p-values > 0.99) and PC1 and PC3 are not only significantly differ to the former 5 combination but also significantly differ to each other (p-value < 0.001). Thus 3 rank of environmental combination estimated by npmc package are ALL, CA, CART, CIT and PC2 for rank 1, PC1 for rank 2 and PC3 for rank 3 respectively. The p-value of inter-group is almost 0 and that indicated each group of rank differ significantly.



(a)

(b)

Figure 19. Differences on AUC among 7 environmental variable combinations,ALL, CA, CART, CIT, PC1, PC2, PC3, and 5 type of occurrence localies, V_1, V_2, V_3, V_5, and V_{all} at 3 different kinds of map resulotion: (a) 40 × 40 m, (b) 100 × 100 m, and (c) 1 × 1 km.



(c)

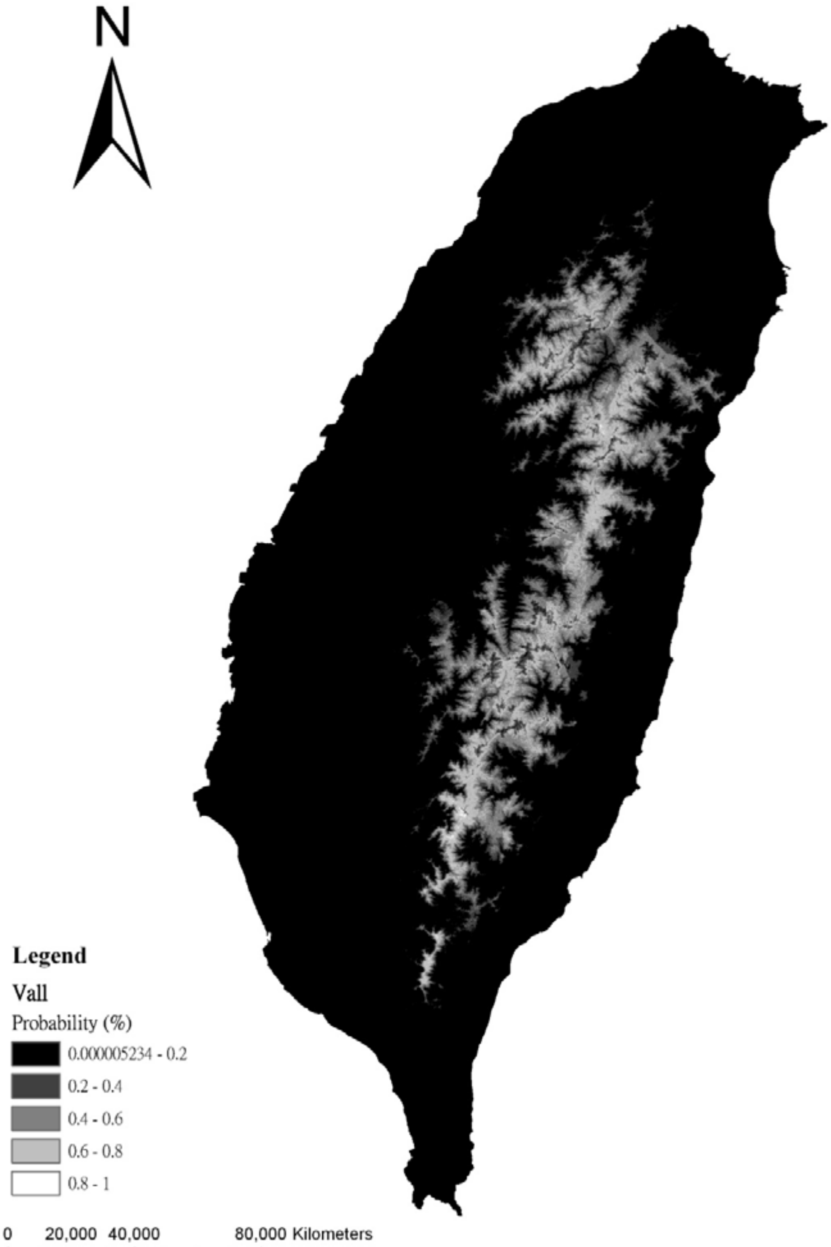
Figure 19. Differences on AUC among 7 environmental variable combinations,ALL, CA, CART, CIT, PC1, PC2, PC3, and 5 type of occurrence localies, V_1, V_2, V_3, V_5, and V_{all} at 3 different kinds of map resulotion: (a) 40 × 40 m, (b) 100 × 100 m, and (c) 1 × 1 km.

After comparing AUC values of different situations, environmental variable combination by CIT method was the highest AUC value among the 7 methods and CART method is the second high of AUC value. Maxent generates a series of statistical analysis and predicts the distribution probability of target species in logistic format (by default). Although CIT method of environmental variable selection performed best, it covered almost all alpine area of Taiwan and that is usually covered with Taiwan Fir pure stand or alpine grassland and not necessarily suitable for Taiwan Hemlock, and thus CART might be the best prediction for distribution of Taiwan Hemlock. Figure 20 shows 5 spatial predictions for Taiwan Hemlock (V_1, V_2, V_3, V_5, and V_{all}) in logistic format from Maxent by using CART method of environmental variable combination.

Table 21. Results of the multiple Behrens-Fisher tests for 4 vegetation types, V_1, V_2, V_3, V_5, and Vall Taiwan Hemlock localities Vall by 2-sided p-value.

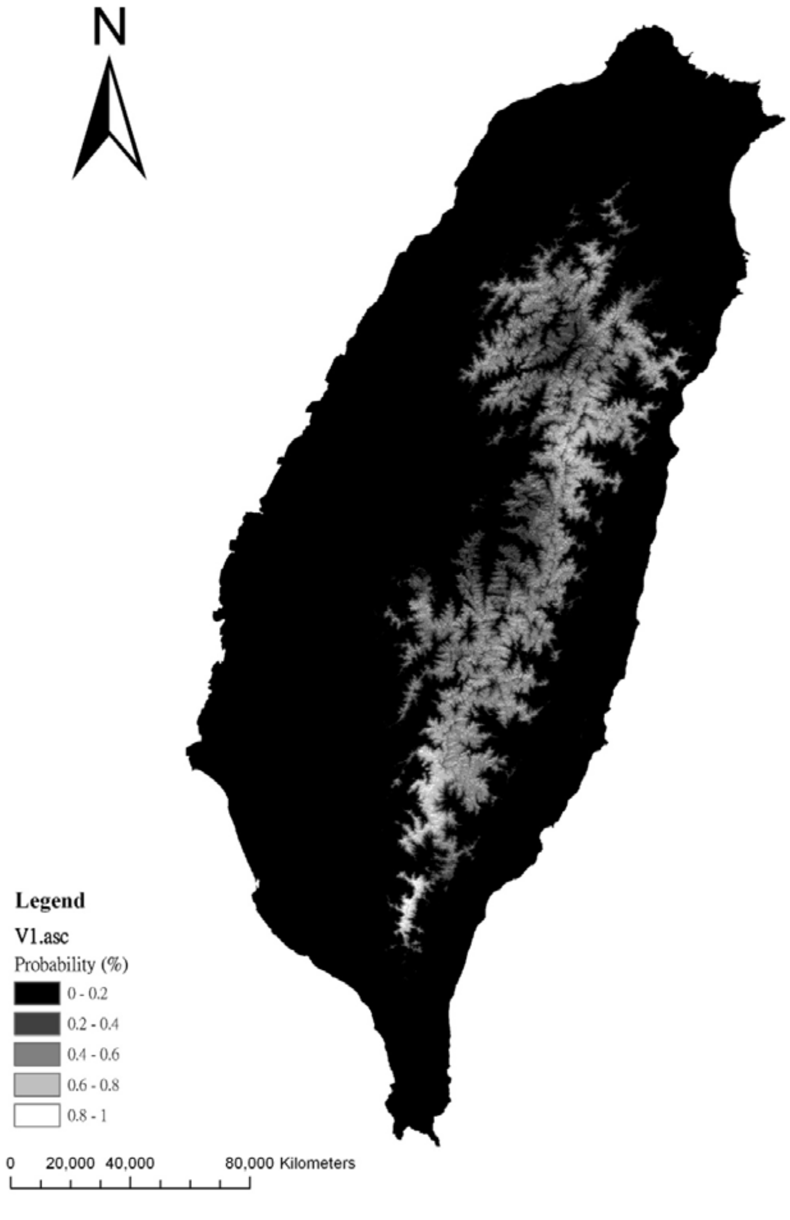
Compare vegetation types	2-sided p-value	
V_1- V_2	0.094	
V_1- V_3	0.066	
V_1- V_5	0.048	**
V_1- V_{all}	0.530	
V_2- V_3	0.055	
V_2- V_5	0.123	
V_2- V_{all}	0.151	
V_3- V_5	0.071	
V_3- V_{all}	0.065	
V_5- V_{all}	0.056	

Note: ** represent significant with p-value less than 0.05



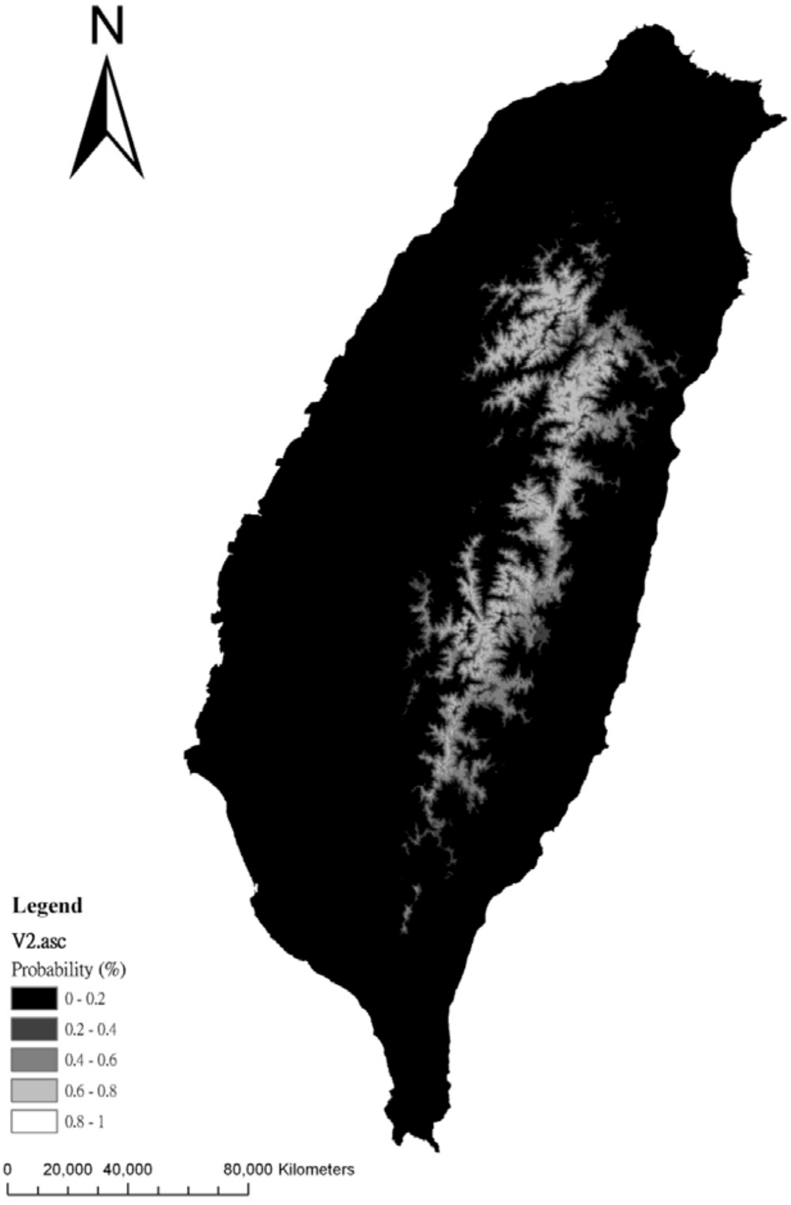
(a)

Figure 20. Probability pictures of Maxent model which uses environmental variable combination by CART method. (a) All occurrences data V_{all} vegetation type. (b) Taiwan Hemlock Taiwan Yellow Cypress and Taiwan Cypress dominance V_1 vegetation type. (c) Taiwan Hemlock mixed with pine species dominance V_2 vegetation type. (d) Taiwan Hemlock and Taiwan Fir dominance V_3 vegetation type. (e) Taiwan Hemlock and Taiwan Spruce dominance V_5 vegetation type. (40 × 40 m in resolution)



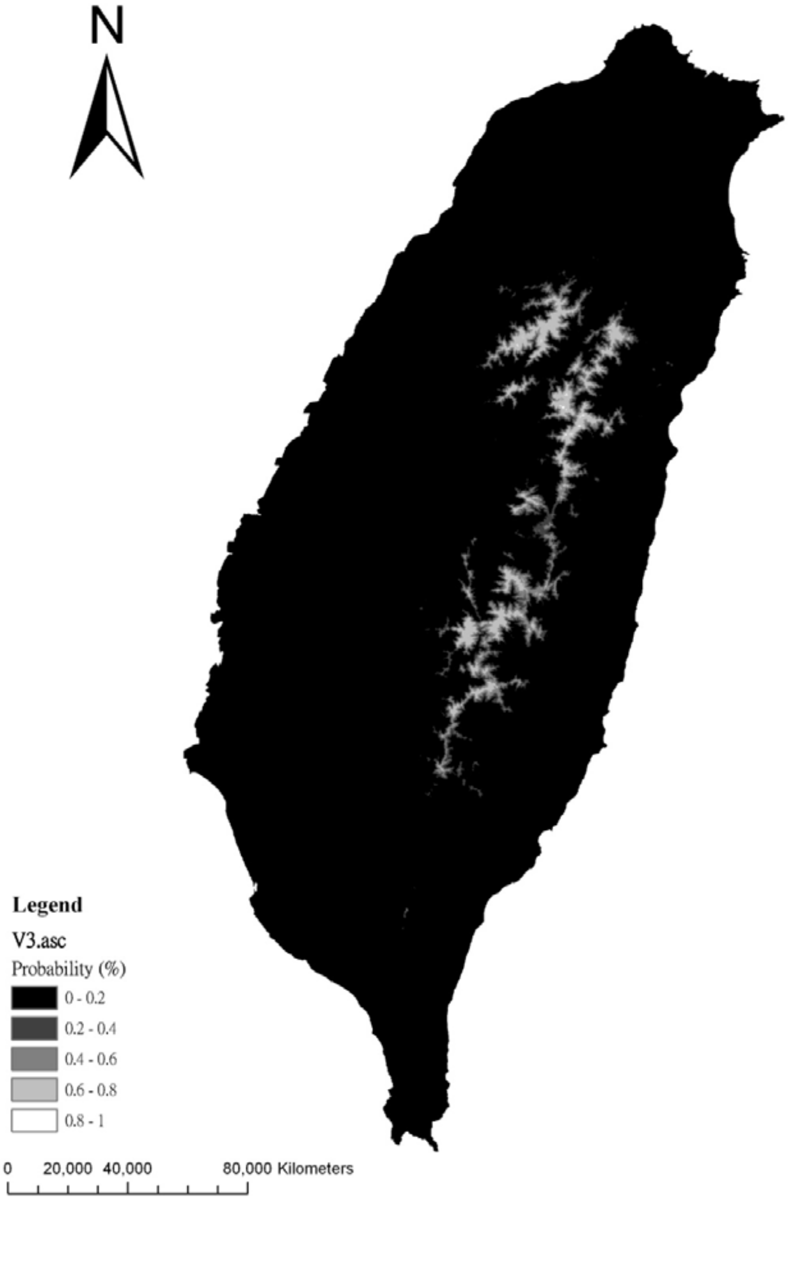
(b)

Figure 20. Probability pictures of Maxent model which uses environmental variable combination by CART method. (a) All occurrences data V_{all} vegetation type. (b) Taiwan Hemlock Taiwan Yellow Cypress and Taiwan Cypress dominance V_1 vegetation type. (c) Taiwan Hemlock mixed with pine species dominance V_2 vegetation type. (d) Taiwan Hemlock and Taiwan Fir dominance V_3 vegetation type. (e) Taiwan Hemlock and Taiwan Spruce dominance V_5 vegetation type. (40 × 40 m in resolution)



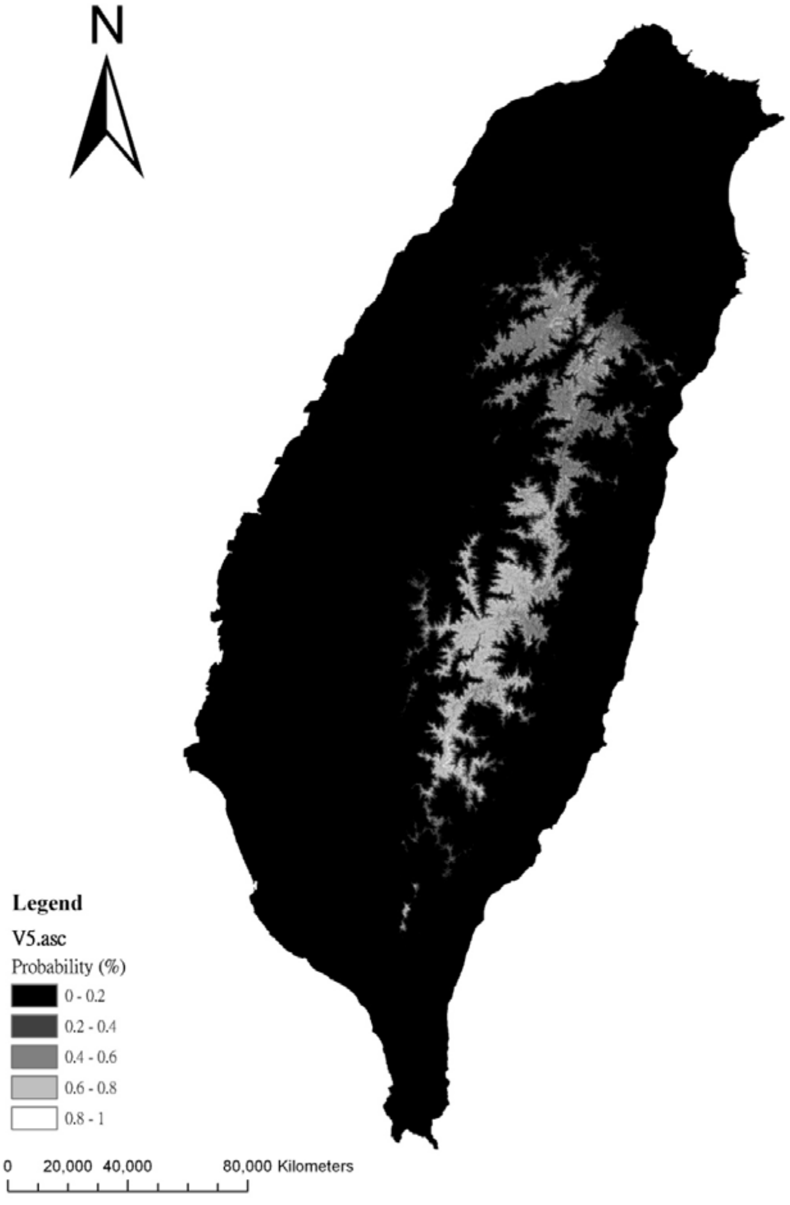
(c)

Figure 20. Probability pictures of Maxent model which uses environmental variable combination by CART method. (a) All occurrences data V_{all} vegetation type. (b) Taiwan Hemlock Taiwan Yellow Cypress and Taiwan Cypress dominance V_1 vegetation type. (c) Taiwan Hemlock mixed with pine species dominance V_2 vegetation type. (d) Taiwan Hemlock and Taiwan Fir dominance V_3 vegetation type. (e) Taiwan Hemlock and Taiwan Spruce dominance V_5 vegetation type. (40 × 40 m in resolution)



(d)

Figure 20. Probability pictures of Maxent model which uses environmental variable combination by CART method. (a) All occurrences data V_{all} vegetation type. (b) Taiwan Hemlock Taiwan Yellow Cypress and Taiwan Cypress dominance V_1 vegetation type. (c) Taiwan Hemlock mixed with pine species dominance V_2 vegetation type. (d) Taiwan Hemlock and Taiwan Fir dominance V_3 vegetation type. (e) Taiwan Hemlock and Taiwan Spruce dominance V_5 vegetation type. (40 × 40 m in resolution)



(e)

Figure 20. Probability pictures of Maxent model which uses environmental variable combination by CART method. (a) All occurrences data V_{all} vegetation type. (b) Taiwan Hemlock Taiwan Yellow Cypress and Taiwan Cypress dominance V_1 vegetation type. (c) Taiwan Hemlock mixed with pine species dominance V_2 vegetation type. (d) Taiwan Hemlock and Taiwan Fir dominance V_3 vegetation type. (e) Taiwan Hemlock and Taiwan Spruce dominance V_5 vegetation type. (40 × 40 m in resolution)

Reponses curves show how each environmental variable affected the Maxent prediction (Figure 21; 40 × 40 m in resolution) on all occurrences data V_{all} vegetation type. The y-axis was predicted probability of suitable conditions, given by the logistic method, with each variable set to their average value over the set of presence localities. The response curve of ASP was gradually the same high from 0 to 361 degree. The response of ELE was peaked from 2000 to 3100 m. The response of PRCAU, PRCSP, and PRCSR were different in amount of precipitation, but the response of total sum of annual precipitation PRCSU was range from 2400 to 3200 mm. The response curves of PRO, STH, SVF, and WST showed no trend of peak for predicted probability. The response curve of WI was higher ranging from 40 to 80 ℃.

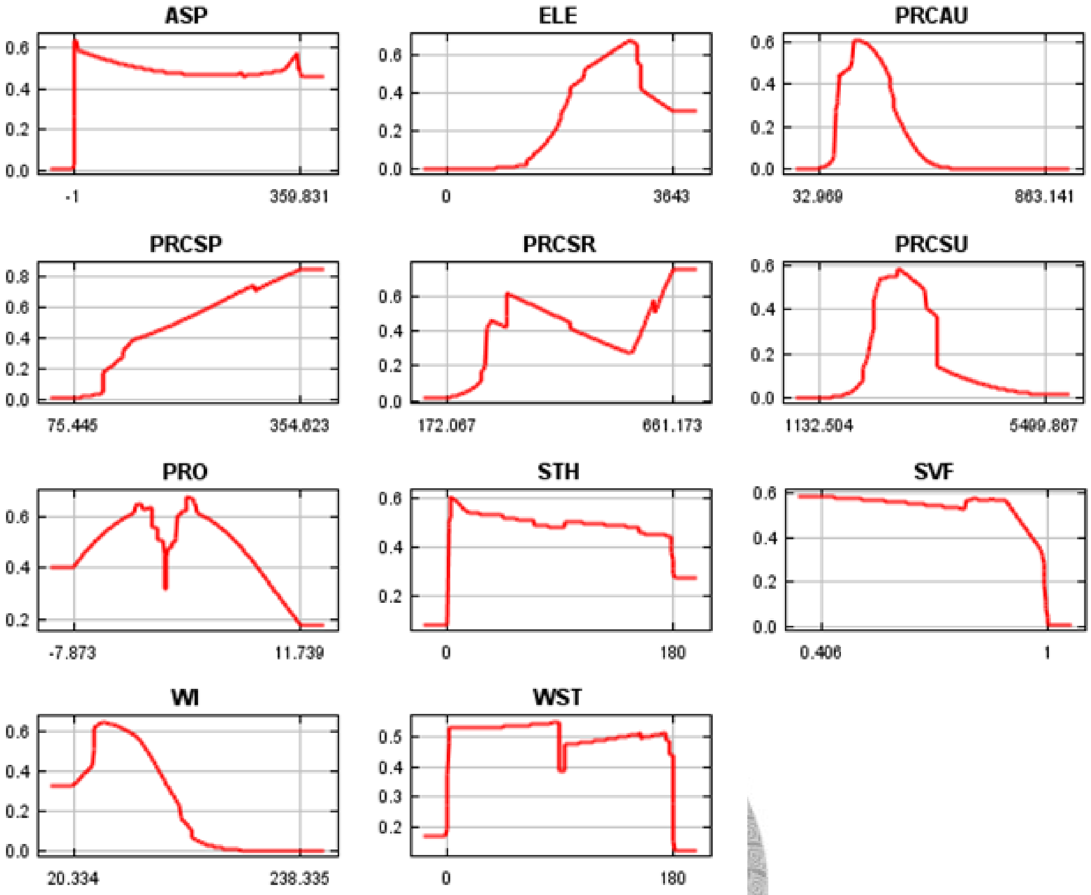


Figure 21. 11 Responese Curves of environmental variables for Maxent prediction of All occurrences data V_{all} vegetation type. X-axis is the range of environmental variable value and Y-axis is the logistic output of probability of presence (40 × 40 m in resolution).

In summary, V_{all} is suitable for elevation range from 2000 to 3100 m, sum of annual precipitation range from 2400 to 3200 mm, and warmth index ranging from 40 to 80 ℃. Table 22 gives a heuristic estimate of relative contributions of the environmental variables to the Maxent model. The variable contributions should be interpreted with caution when the predictor variables are correlated. WI and ELE had most contribution for modeling and the precipitation variables, the rest environmental variables were specific contribution for model to different vegetation type.

Table 22. Contributions of the environmental variables to the Maxent model with V_{all}, V_1, V_2 and V_3

Rank	V_{all}	Percent contribution	V_1	Percent contribution	V_2	Percent contribution	V_3	Percent contribution	V_3	Percent contribution
1	WI	50.9	ELE	89.4	ELE	88.1	WI	51.3	ELE	91.3
2	ELE	43.7	PRCSR	2.3	WI	9.5	ELE	47.3	WI	4.6
3	PRCAU	1.6	STH	1.8	PRCAU	1.6	PRCSP	1.4	PRCAU	1.1
4	PRCSR	0.7	WI	1.8	ASP	0.2	PRCSR	0	SVF	0.9
5	PRCSU	0.7	PRCAU	1.6	PRCSP	0.2	PRCSU	0	ASP	0.9
6	STH	0.6	WST	1	PRO	0.2	PRCAU	0	WST	0.4
7	ASP	0.6	PRCSU	0.9	STH	0.1	WST	0	PRCSP	0.4
8	SVF	0.5	SVF	0.4	PRCSR	0.1	SVF	0	PRCSU	0.2
9	WST	0.5	PRO	0.2	PRCSU	0	STH	0	PRCSR	0.2
10	PRO	0.2	PRCSP	0.2	SVF	0	PRO	0	STH	0.1
11	PRCSP	0.1	ASP	0.2	WST	0	ASP	0	PRO	0

4.4 SDM Assessment with Threshold and Null Model

4.4.1 Threshold to Presence

Table 23 lists 8 methods for determination of threshold to presence. Although threshold selection is depending on the purpose of the study, MaxKappa is chosen for more accurately binary predicted map with presence and absence values. Noted that, threshold of MaxKappa method are actually lower in geo-climatic regions than the whole Taiwan Island.

Table 23. The 8 methods of threshold selection produced by `PresenceAbsence` package of R

Method	V_{all}	V_1	V_2	V_3	V_5
Sens=Spec	0.27	0.31	0.19	0.17	0.11
MaxSens+Spec	0.18	0.24	0.09	0.07	0.11
MaxKappa	0.40	0.50	0.32	0.34	0.70
MaxPCC	0.50	0.78	0.62	0.90	0.90
PredPrev=Obs	0.47	0.56	0.50	0.64	0.77
ObsPrev	0.10	0.04	0.05	0.01	0.00
MeanProb	0.10	0.09	0.07	0.03	0.04
MinROCdist	0.19	0.27	0.16	0.16	0.11

Note: Sens=Spec: sensitivity=specificity; MaxSens+Spec: maximizes; (sensitivity+specificity)/2; MaxKappa: maximizes Kappa; MaxPCC: maximizes PCC (percent correctly classified); PredPrev=Obs: predicted prevalence=observed prevalence; ObsPrev: threshold=observed prevalence; MeanProb: mean predicted probability; MinROCdist: minimizes distance between ROC plot and (0,1)

4.4.2 Threshold Dependent Indices

After a specific threshold is selected, confusion matrix can derive many indices to calculate model performance. Table 24 and 25 compared a specific threshold is chosen based on the MaxKappa occurs and outperformed simply using a half probability with a 0.5 value for deciding a threshold. Noted that, AUC is threshold independent and therefore will not be changed by different threshold is selected. When threshold = 0.5, there is lower value of sensitivity for all models and while threshold changed to MaxKappa, sensitivity is raised but did not drop the value specificity too many. Thus applying a specific threshold is necessary when implement to the specific aim.

Table 24. Threshold dependent indices for each occurrence unit in threshold = 0.5

Unit	threshold	PCC	Sensitivity	specificity	Kappa	AUC
V_{all}	0.5	0.93	0.55	0.97	0.55	0.96
V_1	0.5	0.94	0.57	0.96	0.41	0.95
V_2	0.5	0.95	0.51	0.97	0.48	0.96
V_3	0.5	0.98	0.61	0.98	0.41	0.98
V_5	0.5	0.97	0.50	0.97	0.03	0.96

Note: PCC: percent correctly classified; AUC: area under ROC curve

Table 25. Threshold dependent indices for each occurrence unit in threshold = MaxKappa

Unit	threshold= MaxKappa	PCC	Sensitivity	specificity	Kappa	AUC
V_{all}	0.4	0.92	0.74	0.94	0.61	0.96
V_1	0.5	0.94	0.57	0.96	0.41	0.95
V_2	0.32	0.93	0.80	0.94	0.52	0.96
V_3	0.34	0.98	0.89	0.98	0.45	0.98
V_5	0.7	0.99	0.25	0.99	0.08	0.96

Note: PCC: percent correctly classified; AUC: area under ROC curve

4.4.3 Null Model for Significant test

To test if the model's algorithm could succeed to analysis the relationship between species' occurrences and environmental variables and predict the species'' spatial distribution rather than predicting by chance, a null model was generated by randomly repeating sampling for 999 times of background cells for the same number of the presence data set to get the random AUC's distribution (Figure 22). By adding the predicted AUC values to the null model to see if the model result is high governed by chance.

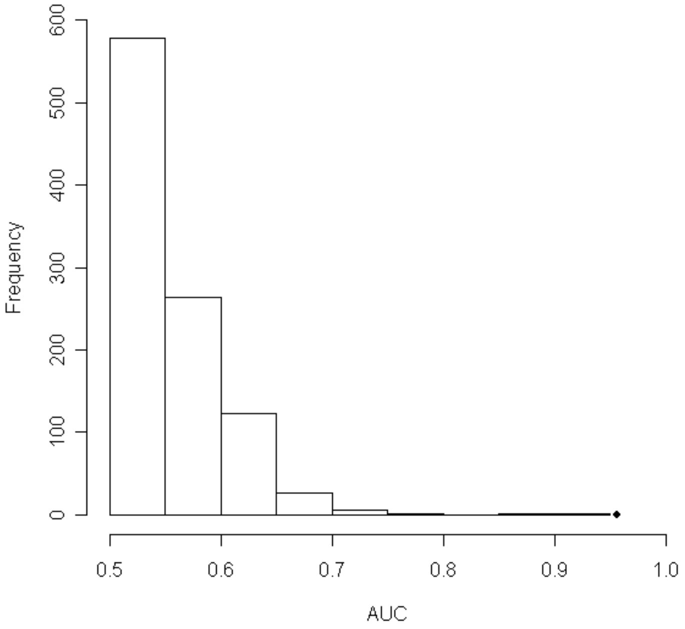


Figure 22. Null model test for V_{all} data set. The black dot on the bottom-right is the AUC value generated by the predictied model Maxent.

4.5 PNV Mapping Criteria and PVM for Taiwan Hemlock

To map the potential vegetation map of Taiwan Hemlock, 5 probability maps were considered, V_1, V_2, V_3, V_5 and V_{all} respectively from 40 × 40 m size in raster files. The potential vegetation maps of Taiwan Hemlock are as following. The composition of the Taiwan Hemlock species map of V_5 and Taiwan Hemlock sub-units vegetation map of $V_{1,}$ V_2, V_3, and V_5 was generated from each binary map split by threshold and then summed together (Figure 23). Finally, in order to evaluate potential predicted area, all occurrence data of Taiwan Hemlock from NVDIMP (N = 408) and absence data of Taiwan Hemlock from TFRILU (N= 3770) were used to calculate the accuracy of the model predicted presence.

The confusion matrix of each potential predicted map was listed from Table 26 to 30. Table 31 listed indices derived from confusion matrix (from Table 28 to 32). All Taiwan Hemlock species presence data set V_{all} predicted the most widely spatial range of Taiwan Hemlock with a high value of sensitivity (0.77) and specificity (0.94) and with the highest the Kappa statistic. That meant 77 % of Taiwan Hemlock presence localities were successfully predicted presence and 94% of Taiwan Hemlock absence localities were also successfully predicted absence. The sub-units of Taiwan Hemlock vegetation type were all predicted too constrain area and thus decline the sensitivity for each predicted map. The best sensitivity was at V_2 vegetation map and the best specificity appeared at V_2, V_3, and V_5 (with specificity 0.99). Because the predicted range of sub-unit vegetation types were so limited that there was a very high rate of predicting absence successfully and leaded to a high value of specificity. The smallest predicted area of the maps is V_5 because it had only 4 occurrence samples of the sub-unit vegetation type. V_{all} map was also had the highest value of positive predicted power (PPP = 0.57) because it predicted more true positive localities than other 4 vegetation maps. On the other hand, 5 vegetation maps predicted relative the same true negative localities and gained with high value of negative predicted power (NPP).

Table 26. Confusion matrix of potential natural vegetation map of Taiwan Hemlock by V_{all} with threshold equals to MaxKappa.

		actual	
		presence	absence
predicted	presence	313	235
	absence	95	3535

Table 27. Confusion matrix of potential natural vegetation map of Taiwan Hemlock by V_1 with threshold equals to MaxKappa.

		actual	
		presence	absence
predicted	presence	47	202
	absence	85	3817

Table 28. Confusion matrix of potential natural vegetation map of Taiwan Hemlock by V_2 with threshold equals to MaxKappa.

		actual	
		presence	absence
predicted	presence	53	105
	absence	148	3872

Table 29. Confusion matrix of potential natural vegetation map of Taiwan Hemlock by V_3 with threshold equals to MaxKappa.

		actual	
		presence	absence
predicted	presence	17	61
	absence	27	4073

Table 30. Confusion matrix of potential natural vegetation map of Taiwan Hemlock by V_5 with threshold equals to MaxKappa.

		actual	
		presence	absence
predicted	presence	1	30
	absence	3	4144

Table 31. Indices derived from confusion matrix of potential natural vegetation map of Taiwan Hemlock with threshold equals to MaxKappa.

Type	Sensitivity	Specificity	PPP	NPP	Odds-ratio	Kappa	Predicted Area (km^2)
V_{all}	0.77	0.94	0.57	0.97	49.56	0.61	3780
V_1	0.26	0.97	0.34	0.96	13.21	0.26	1800
V_2	0.39	0.99	0.22	0.99	42.04	0.27	960
V_3	0.25	0.99	0.03	1	46.04	0.06	360
V_5	0.25	0.99	0.03	1	46.04	0.06	240

Note: PPP: Positive Predicted Power; NPP: Negative Predicted Power

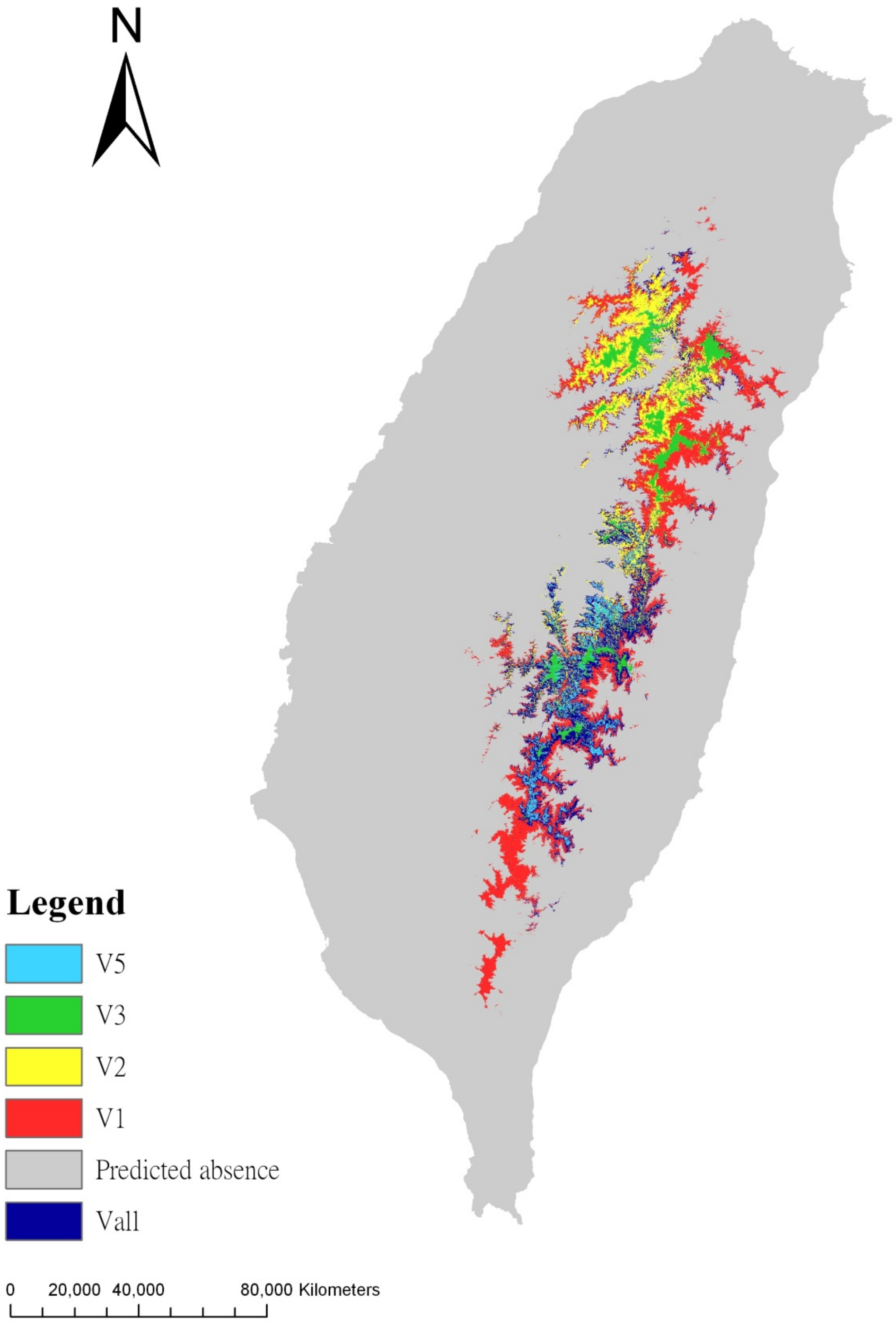


Figure 23. Potential vegetation map of Taiwan Hemlock with threshold equals to MaxKappa and 4 sub-units of Taiwan Hemlock vegetation types (V_1, V_2, V_3, and V_5).

Chapter 5: Discussion

5.1 Vegetation Analysis

Vegetation analysis grouped Taiwan Hemlock presence localities into 6 main composition of vegetation type. Taiwan Hemlock species centered on the DCA plot and displayed some attributes that match the field experience. First, Taiwan Hemlock and Taiwan Fir dominance V_3 vegetation type is far away from the rest 5 groups and indicated that this vegetation type would not mix with the vegetation type of the farthest distance groups V_1 and V_6 which represented Taiwan Hemlock and Taiwan Cypress dominance and Taiwan Hemlock and Taiwan Yellow Cypress dominance vegetation type respectively. And beside, V_3 group on the DCA plot gathered like a line and it indicated the pure stand composition of Taiwan Hemlock-Taiwan Fir vegetation type and this result could indirectly supported by Chen (1995) introduced 3 kinds of vegetation type of Taiwan Fir classified by Sen (1937) companied with Taiwan Hemlock tree species. On the other hand, V_1 might mix with some near groups on the DCA plots like V_2, V_4 and V_5 and revealed that they were at similar elevation range because the first 2 axes was relative to the elevation gradient. V_4 was also concentrated on the center of the DCA plot like V_2 did and that was a interesting result that indicating they were also stay at similar elevation but not specifically close to any other vegetation type. One reasonable explanation is V_4 belongs to V_2 but suffered some disturbance and the pioneer pine species gathered in. The results of Maxent model was also predicted well of distribution of V_2 vegetation type.

5.2 Analysis of Species-Environment Relationship

Environmental variables are extracted from GIS environmental grid layers to field surveys points. This extraction could cause uncertainty because of little difference when synthesize and transform points to layer models from different ordinations. The results show that each approach selects its own environmental variable group which is the most variant to Taiwan Hemlock localities. Correlation analysis avoids the problem of co linearity and reduces the numbers of highly relative predictive variables. This approach, however, does not provide the relationship between species and environmental layers.

The results of PCA reveal traditionally statistical analysis could not handle the non-linear distribution of so many environmental variables, whereas some environmental variables are chosen to represent the most variant environmental variables to Taiwan Hemlock. Component 1 and 2 of PCA can explain about 50% variation of the data. The ordination of the data for PCA shows the cluster-like ordination of data points and means none specific variable spread out in the axis 1, 2, and 3. Moreover, the first component of PCA represents the most variant axis of the data but is not necessarily relative to or limiting species' distribution. The second component of PCA (composite of temperature and precipitation factors) had a higher contribution to the distribution of Taiwan Hemlock than component 1.

The results of CART and CIT showed two different ways of environmental variable selection. CART chooses elevation for the early splits and almost all other topographical and climatic variables for further splits (except slope, curvature and plan curvature, precipitation of winter and month mean precipitation). CIT on the other hand, split by elevation only and was the most important environmental variable to the distribution of Taiwan Hemlock because the variable split by CIT separates two groups with significant difference. This result indicates that CART has the ability to distinguish as detailed homogeneous groups as if sufficient variables are given. CIT, however, has the ability to choose the variable that split node with more homogenous offspring groups significantly different with mother group.

5.3 Environmental Variables to Taiwan Hemlock and Model Assessment

Kappa statistic yields similar results to the AUC (Guisan *et al.*, 2007) as well as the result of this study. According to Swets (1988), AUC values greater than 0.9 are considered with high accuracy, in range from 0.7 to 0.9 are considered as useful, and lower than 0.7 though of poorly. In this study, the first rank group of environmental combination for AUC evaluation included the ALL, CA, CART, CIT, and PC2 and elevation variable was always contained by each environmental variable combination. It indicates that elevation variable is the most important environmental variable to distribution of Taiwan Hemlock in this study. The only elevation variable chosen by CIT method reached AUC value to 0.96 and it was vary well performance of the predicted. However, the result of inputting environmental variables chosen by CIT might over predict at the alpine area because of lacking of test samples so the AUC is still high in this situation. CART, on the other hand, avoid this problem of predicting Taiwan Hemlock presence over all alpine area. The first environmental variable splits of

CART was also elevation and that split reduced the most of the variance of the offspring groups and it implied the same signal that elevation is the most important environmental variable for splitting Taiwan Hemlock species into sub unit vegetation type. The environmental variable far from the root of the tree is minor to contribute for reducing the variance (i.e. relatively lower contribution to the predicted model). This situation was also proved by the contribution of environmental variable to the Maxent model (Table 24). In this study, therefore, precipitation is the second important environmental variable for Taiwan Hemlock. The topographic variables were not as important as the former 2 variables for Taiwan Hemlock distribution. The maximum AUC of the environmental variable combination is CIT and then ALL, CART, CA, PC2, PC1 and then PC3 orderly. CIT combination used the fewest environmental variables to achieve the best model performance but only can conclude CIT combination is the best fit for Maxent modeling and had the ability to discriminate between a suitable environmental condition and a random absence rather than suitable and unsuitable conditions (Hernandez *et al.*, 2006). Although many studies used of Kappa statistic for measuring model performance, Kappa is a threshold dependent statistic that calculated the proportion of correctly classified units (A+D) in confusion matrix after accounting for the probability of chance agreement. As threshold is larger tends to decrease commission error and increase omission error (Fielding and Bell, 1997) and affects the Kappa statistic. AUC measured with full information provided by all possible thresholds (Pearce and Ferrier, 2000) and more informative than Kappa statistic in a 0.5 threshold. Hernandez *et al.* (2006) described that multiple evaluation measures were suitable for evaluating the model performance with presence only-data.

Maxent performed well if suitable environmental variables were puts into it. CART and CIT successfully analyzed and chose the most effective environmental variable to the distribution of Taiwan Hemlock and this approach is useful while too many irrelative environmental variables are available.

Lobo *et al.* (2008) described five drawbacks of AUC assessment including (i) AUC is discriminant assessment and ignores predictive probability values and goodness-of-fit of the model, (ii) it summaries the test performance which one would rarely operate, (iii) it weights both omission and commission errors equally, (iv) it does not offer information about the spatial distribution of model errors, and (v) well predicted absences and the AUC scores area influenced by the total extent. This study used the equal number of presence and absence localities while testing the model to avoid the fifth point stated by Lobo *et al*. They also concluded that AUC provides information about the generalist or restricted distribution of a target species along with specific environmental variables in the study area, but does not provide information about the good performance of the model (Lobo *et al*., 2008) and that means model uncertainties will not be considered by AUC. One purpose of this study is to predict the spatial range of Taiwan Hemlock and AUC provides a good discriminant between presence and absence localities.

5.4 Vegetation and Species Based Units and Map Resolution

Vegetation units derived form all occurrence data of Taiwan Hemlock is a special case to the sample size reduction. Hernandez *et al.* (2006) concluded Maxent was more capable to model and produced useful results while data was incomplete or sample size is as small as 25 or even lower (smallest 5 samples in their study) and model accuracy was better to the species with small range in geographic distribution and limited environmental tolerance. The sample size effect on vegetation (smaller sample size) and species (total sample size) of this study demonstrated the similar results as the Hernandez *et al.* (2006). Pearson *et al.* (2007) described vary low sample sizes (as low as five records) to a fixed Maxent probability value of 10 to significantly recover all known presences. In the same context, a lower Maxent value was useful in revealing uncertain but potentially important distributional ranges. V_3 with the smallest sample size resulted in that perfect accurate in AUC measure (valued to 0.98) compared to the rest (AUC of V_1, V_2, V_3 and V_{all} are 0.96, 0.95, 0.96, and 0.96 respectively).

As motioned above, resolution environmental variables and background pixels is really influential to the model performance? The result of this study does not support the statement that resolution of environmental variable and background cells is influential to the distribution of Taiwan Hemlock in 40, 100 and 1000 m resolution.

5.5 Combination of Models for Predicting Vegetation Map

Although combining models reaches out the optimization of each model algorithm and reduces model based uncertainties (Clemen, 1989; Gilmer, 2008), some risk of combining models with less accuracies predictions are needed to consider. In this study, 4 vegetation sub-unit based models with different sample size smaller than the whole occurrence data of Taiwan Hemlock classified by cluster analysis had a higher accurate prediction, would it be the effects of small sample to the AUC (Hernandez *et al.*, 2006). Because the V_1, V_2, V_3 and V_5 are partial samples of V_{all} and the good performance of each sub-unit model predicted probability value higher at surrounding grid cells.

Maxent was not yet clear how significant were the differences between various probability distribution values (Phillips *et al.*, 2006); as such, the difficult task to the user of selecting the appropriate threshold, below which the model may loose predictive power and become too general (Pape and Gaubert, 2007). Pearson *et al.* (2007) addressed the threshold issue of small samples available for ENM. The acceptable threshold value depends on the purpose: if the interest is in observing general distributional patterns, then a 'free' threshold is suitable (i.e. over-predicting is informative). When conservation applications, however, are of principal interest, a 'conservative' threshold is more adequate (i.e. over-predicting is not suitable) (Pape and Gaubert, 2007). Although in this study selected the threshold by the threshold that gave the maximum of Kappa statistic, too many absence localities limited the threshold to be more conservative to presence localities because if a liberal threshold was selected, too many absence localities would be predicted presence incorrectly. And the threshold affected the potential maps of each sub-unit vegetation type and limited the predicted area of each map resulting in a conservative prediction of potential vegetation map of Taiwan Hemlock. Thus, lower sensitivity were appeared in each map evaluation but specificity performed well for evaluating each vegetation map.

Confusion matrix of each vegetation map was high at specificity because too many of absence localities to be correctly predicted. Some problem with so many absence data because if too many absence data, there was easy to predicted all localities absence and would get not bad model performance. Another question is the realized niche we observed is not necessarily the whole environment that suitable for the target species and maybe geographical limitation leaded the suitable environment without the target species. Therefore, too many absence localities will constrain the potential area and decline the model performance. On the other hand, if too few absence localities were used, it would come up with a liberal threshold selection and over-predict the map for more information (Pape and Gaubert, 2007).

Why predicted area is so important because of the low prevalence of the data (i.e. absence localities are much greater than presence). An error rate of wider predicted range might cause in many commission errors than a limited predicted range which might cause in fewer omission error. The reason is too many absence localities can cause more error rate where a relative few presence localities is not able to reach. For conservative perspective, if the vegetation map of Taiwan Hemlock generated in this study is able to distinguish 77% of all Taiwan Hemlock presence localities, the predicted area is conservatively suitable for Taiwan Hemlock. On the other hand, the predicted area for Taiwan Hemlock sub-unit vegetation map is more conservatively constrained or limited each vegetation sub-unit of Taiwan Hemlock.

Chapter 6: Conclusion

Taiwan Hemlock concentrates in range from 2000 to 3100 m in elevation, 2000 to 3100 mm in annual precipitation, and 25 to 40 degrees in slope and consists of 4 main vegetation type, Taiwan Hemlock-Taiwan Cypress-Taiwan Yellow Cypress dominance, Taiwan Hemlock-Pine species dominance, Taiwan Hemlock-Taiwan Fir dominance, and Taiwan Hemlock-Taiwan Spruce dominance vegetation types. The classification is response to the environmental variables mainly by elevation and warmth index which is highly relative with elevation. Whether CART, CIT or Maxent method chose the elevation variable for representing the characteristics of distribution of Taiwan Hemlock. Other variables used in this study were minor to affect the distribution of Taiwan Hemlock.

The analysis of species distribution and environmental relationships reveals the extrinsic effects on species' distribution. None of the best model is defined as the universal tools for predicting species distribution, however, the attempt to analysis those relationships gives the implication of how the species reacts to any environmental disturbance and where does the species can escape from this impact of changes. Clemen (1989) concluded model combination as:

"Combining forecasts has been shown to be practical, economical and useful. Underlying theory has been developed, and many empirical tests have demonstrated the value of composite forecasting. We no longer need to justify this methodology. We do need to find ways to make the implementation of the technique easy and efficient."

The analysis of the relationship between species and environmental variables is quantified by many new approaches rather than using just a traditional statistical analysis. Machine learning methods jump over the traditional method due to the multi-consideration of the algorithm with modern computer intensive ability to analyze more precisely prediction. Many tasks of SDM discussed in this study such as how to select appropriate environmental variables, if more homogeneous samples affects the selection of environmental variables?, and resolution of the environmental layers and the cell space to be predicted. However, restricted to the unavailable data quality, the accuracy of the SDM still needs to be revised by actually examined by a specific experiment on id the potential environmental condition really suitable for the target species or the potential predicted area is really suitable for planting or growing of the target species? Although nowadays the predicted modeling is widely spread the model needs more experiments by further study to support the PVM.

Maxent performed well if suitable environmental variables were puts into it. CART and CIT successfully analyzed and chose the most effective environmental variable to the distribution of Taiwan Hemlock and this approach is useful while too many irrelevant environmental variables are available.

Combining model approach makes the SDM and its relevant model such as ENM more flexible to apply in a specific purpose. However, it still need further study for completing it and encourages the recent scientists to have the foundation to establish new combination approaches, as in this study pays efforts on changing combination target from model techniques to species and vegetation units, which is followed the plant community conception. How can ecological theory help the model performance of accuracy is still needs further study, however, this research gives an initial implication and hopes for more interesting ideas.

For conservation management, further alpine ecological researches are needed in Taiwan to adapt the climate change impact. A physical based model is the possible approach to improve the cons in the statistic models, i.e. the parameters are still robust under the climate change?

References

Anderson, R.P., A.T. Peterson, and M. Gomez-Laverde (2002b) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *OIKOS*, 98: 3–16.

Anderson, R.P., and E. Mart´ınez-Meyer (2004) Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (*Heteromys*) of Ecuador. *Biololical Conservation*, 116: 167–179.

Anderson, R.P., D. Lewc, and A.T. Peterson (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, 162: 211–232.

Anderson, R.P., M. Gomez-Laverde, and A.T. Peterson (2002a) Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, 11: 131–141.

Araújo, M.B., and A. Guisan (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33: 1677–1688.

Austin, M.P., A.O. Nicholls and C.R. Margules (1990) Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species, *Ecological Monographs*, 60: 161–177.

Bates, J.M. and C.W.J. Granger (1969) The combination of forecasts. *Operational Research Quarterly*, 20: 451–468.

Beven, K.J. (1997) Distributed Modelling in Hydrology: Applications of the TOPMODEL Concepts, Wiley, Chichester.

Beven, K.J., M.J. Kirkby (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24: 43–69.

Box, E.O. (1981) Macroclimate and plant formm: An introduction to predictive modeling in phytogeoraphy. *Volume 1 of Tasks of Vegetation Science*, The Hague: Junk Publishers, Pp. 132.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, 26: 123–140.

Breiman, L. (2001) Random forest. *Machine Learning*, 45: 5–32.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984) Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey, CA.

Brovkin, V. (2002) Climate-vegetation interaction. *Journal De Physique IV – Proceedings*, 12:10–57.

Cairns, D.M. (2001) A comparison of methods for predicting vegetation type. *Plant Ecology*, 156: 3–18.

Cantor, S.B., C.C. Sun, G. Tortolero-Luna, R. Richards-Kortum, and M. Follen (1999) A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. *Journal of Clinical Epidemiology*, 52(9): 885–892.

Cha, G.S. (1998) Estimation of changes in potential forest area under climate change. *Journal of Korea Forest Society*, 87(3): 358–365.

Chang, C.R., P.F. Lee, M.L. Bai, T.T. Lin (2004) Predicting the geographical distribution of plant communities in complex terrain - a case study in Fushan Experimental Forest, northeastern Taiwan. *Ecography*, 27(5): 577–588.

Chang, H.J. (1999) Vegetation analysis of long-term sites of Taiwan Fir at Ho-Huan mountain. *Project of Taroko National Park*, Construction and Plaining Agency Ministry of the Interioir, Pp. 64. (in Chinese)

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Society, Series A*, 158: 419–466.

Chen, Y.F. (1995) Series of Taiwan Fir reaserch (I): study of the history. *Annual Journal of Museum of Taiwan*. 38: 23–53.

Chen, Y.F. (1997) Research on the response of vegetation to climat change: modeling. *Progress in Geography*, 16(3): 24–28.

Chen, Y.F. (1999) China climate-vegetation model based on soil classification. *Progress in Nature Science*, 9(1): 54–60.

Chen, Y.F. (2004) Taiwan Hemlock zone in Taiwan (I). Avangard Publisher, Taipei,. Pp. 452. (in Chinese)

Chiou, C.R., C.F. Yu, and C.F. Li (2005) The planning and present situation of Taiwan Vegetation Information System. Proceeding of the Third Symposium of Vegetation Diversity in Taiwan, Forest Bureau, Council of Agriculture, Taipei, Taiwan, 202–215. (in Chinese)

Chiou, C.R., J.R. Lin and C.F. Li (2006) Analysis of distribution characteristics of Taiwan Hemlock communities. *Proceedings of Fourth Symposium of Vegetatoin Diversity in Taiwan Vegetation Mappong Series*, 8: 280–305. (in Chinese)

Chiou, C.R., Y.J. Lai, C.F. Li, and Y.C. Laing (2004) The application of GIS on the simulation of climate change impact on forest – a case study on Taiwan Cypress forest. Greater China GIS Conference and Exhibition 2004, Hong Kong GIS System Association. (in Chinese with English abstract)

Chiu, C.A., K.C. Lu, P.H. Lin, and M.C. Liao (2005) Mapping Holdridge’s life zones at Taiwan. *Academic Journal of Naitonal Park*, 15(1): 61–78.

Clemen, R. T. (1989) Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5: 559–583.

Cohen, J. (1988) Statistical power analysis for the behavioral sciences (2nd ed.) Hillsdale, Lawrence Erlbaum Associates, NJ.

Cramer, W.P. and R. Leemans (1993) Assessing impacts of climate change on vegetation using climate classification system. In: Lolomon A.M. and H.H. Shugart(eds.). Vegetation Dynamics and Global Change. Chapman and Hall, New York, Pp. 190–217.

Danijela, P.M. (2003) Predictive vegetation modeling for forest conservation and management in settled landscapes. A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy, Graduate Department of Faculty of Forestry, University of Toronto.

De'Ath, G., and K.E. Fabricius (2000) Classificaiton and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11): 3178–3192.

Della Pietra, S., V. Della Pietra, and J. Lafferty (1997) Inducing features of random fields. IEEE Trans. *Pattern Annual Machine Intellegence*, 19 (4): 1–13.

Dias, E., R.B. Elias, and V. Nunes (2004) Vegetation mapping and nature conservation: a case study in Terceira Island (Azores). *Biodiversity and Conservation*, 13: 1519–1539.

Dormann, C.F. (2007) Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, 8: 387–397.

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

Elith, J. and M.A. Burgman (2003) Habitat models for PVA. In: Population Viability in Plants. Conservation, Management and Modeling of Rare Plants, Springer-Verlag, New York, Pp 203–235.

Elith, J., C.H. Graham, R.P. Anderson, M. Dudk, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J.McC. Overton, A.T. Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberon, S. Williams, M.S. Wisz, and N.E. Zimmermann (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29: 129–151.

Fielding, A.H. and J.F. Bell (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1): 38–49.

Foley, J.A., S. Levis, I.C. Prentice, D. Pollard, and S.L. Thompsons (1998) Coupling dynamic models of climate and vegetation. *Global Change Biology*, 4: 561–579.

Forest Bureau (1995) The Third Forest Resource and Land-Use Inventory. Council of Agriculture, Taipei, Taiwan. (in Chinese)

Franklin, J. (1995) Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress Physical Geography*, 19: 474–499.

Franklin, J. (1998) Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*, 9: 733–748.

Franklin, J. (2002) Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral. *Applied Vegetation Science*, 5: 135–146.

Franklin, J. (2002) Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of vegetation science*, 9:733–748.

Friedman, J.H. (2001) Greedy function approximation: A gradient boosting machine. *Annual of Statistic*, 29: 1189–1232.

Fu, K.M. (2002) Vegetation Ecology of Dan-da Region. Department of Forestry, National Chung Hsing University, Pp. 137. (in Chinese)

Gibson, L., B. Barrett, and A. Burbidge (2007) Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. *Diversity and Distributions*, 13: 704–713.

Gilmer, B.F. (2007) Predictive Vegetation Models: A Comparison of Model Combination Approaches. Master Thesis, Department of Geology and Geography Morgantown, West Virginia University.

Gotelli, N.J. and B.J. McGill (2006) Null versus neutral models: what's the difference? *Ecography*, 29(5): 793–800.

Graham, C.H., and R.J. Hijmans (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, 15: 578–587.

Graham, C.H., J. Elith, R.J. Hijmans, A. Guisan, A.T. Peterson, B.A. Loiselle, and The Nceas Predicting Species Distributions Working Group (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45: 239–247.

Grinnell, J. (1917) Field tests of theories concerning distributional control. *The American. Naturalist*, 51: 115–128.

Grinnell, J. (1924) Geography and Evolution, *Ecology*, 5: 225–229.

Gu, S.L., T.T. Chen, C.W. Lai, J.R. Lin and Y.J. Hsia (2006) Predicting species spatial distributions in Li-Wu Watershed using generalized additive models. *Proceedings of Fourth Symposium of Vegetatoin Diversity in Taiwan Vegetation Mappong Series*, 8: 80–99. (in Chinese)

Guisan, A. and N.E. Zimmerman (2000) Predictgive habitat distribution models in ecology. *Ecological Modelling*, 135: 147–186.

Guisan, A., A. Lehmann, S. Ferrier, M.P. Austin, J. Mc. C. Overton, R. Aspinall, and T. Hastie (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, 43: 386–392.

Guisan, A., and W. Thuiller (2005) Predicting species distribution: offering more than simple habitat models. *Ecological Letters*, 8: 993–1009.

Guisan, A., C.H. Graham, J. Elith, F. Huettmann, and the NCEAS Species Distribution Modelling Group (2007) Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, 13: 332–340.

Guisan, A., J.P. Theurillat, and F. Kienast (1998) Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*, 9:65–74.

Guisan, A., N.E. Zimmermann, J. Elith, C.H. Graham, S. Phillips, And A.T. Peterson (2007) What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? Ecological Monographs, 77(4): 615–630.

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

Hardtle, W. (1995) On the theoretical concept of the potential vegetation and proposals for an up to date modification. *Folia Geobotanica et Phytotaxonomica*, 30(3): 263–276.

Hastie, T.J., and R.J. Tibshirani (1986) Generalized additive models. *Statistical Science*, 1: 297–318.

Hastie, T.J., and R.J. Tibshirani (1990) Generalized Additive Models. Chapman and Hall, New York.

Hengeveld, R. (1990) Dynamic biogeography. Cambridge University Press, Cambridge.

Hernandez P.A., C.H. Graham, L.L. Master, and D.L. Albert (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29: 773–785.

Holdridge, C.T. (1947) Determination of world formations from simple climatic data. *Science*, 105: 367–368.

Holdridge, L.R. (1967) Life Zone Ecology. San Jose, Costa Rica. Tropical Science Center, Pp. 54.

Hothorn, T., K. Hornik and A. Zeileis (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3): 651–674.

Hsu, H.H., and C.T. Chen (2002) Observed and projected climate change in Taiwan. *Meteorological Atmosphere Physics*, 79: 87–104.

Huang, T.C., H. Keng, W.C. Shieh, J.L. Tsai, C.F. Hsieh, J.M. Hu, C.F. Shen, K.C. Yang, and S.Y. Yang (1994) Flora of Taiwan. Vol. 1: 567–569.

Huberty, C.J. (1994) Applied discriminant analysis. Wiley Interscience, New York, Pp. 466.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22: 415–427.

Hutchinson, G.E. (1978) An introduction to population ecology, Yale University Press, New Haven.

IPCC (2001) IPCC third assessment report, summary of policy maker: the scientific basis, WGI: Scientific aspects of climate.

IPCC (2007) IPCC forth assessment report, summary of policy maker: the scientific basis, WGI: Climate Change 2007: The Physical Science Basis.

James, F.C. and C.E. McCulloch (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annual Review of Ecological System*, 21: 129–166.

Jansen, M.E.; W. Shmidt; V. Stuber; H. Wachter; C. Neader; M. Weckesser and F.J. Knauft (2002) Modeling of nature woodland communities in Harz mountains. Spatial modeling in forest ecology and management: a case study. Berlin, Springer, New York, 162-175.

Johnson, W.C. (1989) The role of Blue Jays (*Cyanocitta cristata*) in the postglacial dispersal of *Fagaceous* tree in eastern north-America. *Journal of Biogeography*, 16: 561–571.

Kellman, M.C. (1980) Plant geography. Methuen and Co. Ltd., London, Pp. 181.

Kessell, S.R. (1976) Gradient modeling: a nre approach to fire mofeling and wildness resource management. *Environmental Management*, 1:39–48.

Kira, T. (1948) On the altitudinal arrangement of climate zone in Japan – a contribution to the rational utilization in cool highlands. *Agricultural Science of the North Temperate Region*, 2: 143-173. (in Japanese)

Köppen, W. (1931) Grundriss der Klimakunde, Berlin: DeGruyter. Pp. 388.

Kriegler, B. (2007) Cost-sensitive stochastic gradient boosting within a quantitative regression framework. Ph. D. Dissertation, University of California Los Angeles.

Kuo, B.C. (1978) Relationship between distribution of forest and plants, and warmth index in Taiwan. Journal of Association of Chinese Argriculture, 105–113. (in Chinese)

Landis, J.R. and G.C. Koch (1977) The measurement of observer agreement for categorical data. *Biometrics,* 33: 159–74.

Lassueur, T., S. Joost, and F. Randin (2006) Very high resolution digital elevation models: Do they improve models of plant species distribution? *Ecological Modelling*, 198: 139–153.

Leniha, J.M. and R.P. Neilson (1993) A rule-based vegetation formation model for Canada. *Journal of Biogeography*, 20: 615–628.

Liang, Y.C. (2004) Studies on Zoning the Ecoregion at Domain and Division Levels in Taiwan. Master Thesis, School of Forestry and Resource Conservation, National Taiwan University. Pp. 122. (in Chinese)

Lim, B. K., Petereson, A.T., Engstrom, M.D. (2002) Robustness of ecological niche modeling algorithms for mammals in Guyana. *Biodiversity and Conservation*, 11: 1237–1246.

Lin, H.L. and Z.S. Chen (2005) Statistics: approaches and applications (II). Yeh Yeh Book Gallery, Pp. 630. (in Chinese)

Lindsay, J.B. (2005) The Terrain Analysis System: a tool for hydro-geomorphic applications. *Hydrological Process*, 19: 1123–1130.

Liu, J.C. (2003) Study of plant forms of important habitats of wild animals at Ci-Lan mountain (II). Conservation Serie, Forestry Bureau, Council of Argriculture Executice Yuan, Pp. 97. (in Chinese)

Liu, J.Y. and Y.H. Tseng (1999) Study of vegetation ecology at Sha Li Shian River watershed in Yushan National Park. *Report of National Park*, 9(1): 11–31. (in Chinese)

Liu, T.S. (1962) A phytogeographic sketch on the forest flora of Taiwan. *Acta Phytotaxonomica et Geobotanica*, 20:149–157.

Lobo, J.M., A. Jimenez-Valverde, and R. Real (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17: 145–151.

Lowell, K.E. (1991) Utilizing discriminant function analysis with a geographical information system to model ecological succession spatiality. *International Journal of Geographical Information System*, 5:175–191.

Luckman, B. (1990) Mountain areas and global change: a view of Canadian Rockies. *Mountain Research and Development*, 10(2): 183–195.

Mackey, B.G., and D.B. Lindenmayer (2001) Towards a hierarchical framework for modelling the spatial distribution of animals. *Journal of Biogeography*, 28: 1147–1166.

Manel, S., H.C. Williams, and S.J. Ormerod (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38: 921–931.

Manel, S., J.M. Dias, S.J. Ormerod (1999) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, 120: 337–347.

Miller, J. (2005) Incorporating spatial dependence in predictive vegetation models: Residual interpolation methods. *Professional Geographer*, 57(2): 169–184.

Miller, J. and J. Franklin (2002) Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling*, 157: 227–247.

Miller, J., J. Franklin, R. Aspinall (2007) Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, 202(3-4): 225–242.

Mit´aˇsov´a, H., J. Hofierka (1993) Interpolation by regularized spline with tension: II. Applications to terrain modeling and surface geometry analysis. *Mathematical Geology*, 25(6): 657–669.

Miyawaki, A. (1988) Restoration of urban green environments based on the theories of vegetation ecology. *Ecological Engineering*, 11(1-4): 157–165.

Miyawaki, A. and K. Fujiwara (1988) vegetation mapping in Japan. Pp. 427–442 in A.W. Kulcher and I.S. Zonneveld, eds. Vegetation Mapping. Kulwer Academic Publisher, Pp. 635.

Miyawaki, A., K. Fujiwara, and S. Okuda (1987) The status of nature and re-creation of green environment in Japan Pp. 357–376 in A.Miyawaki; A. Bogenrider; S. Okuda and J. White, eds. Vegetation Ecology and Creation of New Environment, Tokyo.

Moisen, G. G. and T. S. Frescino (2002) Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, 157: 209–225.

Moore, I.D., R.B. Grayson, and A.R. Ladson (1991) Digital terrain modeling: a review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, 5: 3–30.

Moore, I.D., T.W. Norton, and J.E. Williams (1993) Modelling environmental heterogeneity in forested landscapes. Journal of Hydrology, 150: 717–747.

Muñoz, J. and A. Felicísimo (2004) Comparison of statistical methods commonly used in predictive modeling. *Journal of Vegetation Science*, 15: 285–292.

Murphy, A.H., and R.L. Winkler (1987) A general framework for forecast verification. *Monthly Weather Review*, 115: 1330–1338.

Murphy, A.H., and R.L. Winkler (1992) Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7: 435–455.

Oke, T.R. (1981) Canyon geometry and the nocturnal urban heat island: comparison of scale model and field observations. *Journal of Climatology*, 1(1-4): 237–254.

Olmeda, I. and E. Fernández. (1997) Hybrid classifiers for financial multicriteria decision making: the case of bankruptcy prediction. *Computational Economics*, 10: 317–335.

Ou, C.H, J.C. Liu, C.C. Wang, M.C. Chang, C.A. Chiu and C.Y. Tseng (1994) Study of vegetation ecoloy of Shuan-Kuei Lake nature preserve. Conservation Serie, Forestry Bureau, Council of Argriculture Executice Yuan, Pp 107. (in Chinese)

Pape, M., and P. Gaubert (2007) Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions*, 13: 890–902.

Pearce, J., and S. Ferrier (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological modeling*, 133: 225–245.

Pearson, R.G. and T.P. Dawson (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, 12: 361–371.

Peterson, A.T., J. Soberon, V. Sanchez-Cordero (1999) Conservatism of ecological niches in evolutionary time. *Science*, 285: 1265–1267.

Peterson, A.T., L.G. Ball, and K.P. Cohoon (2002) Predicting distributions of Mexican birds using ecological niche modelling methods. *Ibis*, 144: E27–E32.

Peterson, A.T., M. Papes, and J. Soberon (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, 213: 63–72.

Peterson, A.T.and K.C. Cohoon (1999) Sensitivity of distributional prediction algorithms to geographic data completeness, *Ecological Modeling*. 117: 159–164.

Phillips, S.J. (2008) Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson *et al*. (2007). *Ecography*, 31: 272–278.

Phillips, S.J. *et al*. 2005. Maxent software for species distribution modeling. <http://www.cs.princeton.edu/~schapire/maxent/>.

Phillips, S.J., and M. Dudı´k (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31: 161–175.

Phillips, S.J., Dudik, M., Schapire, R.E. (2004) A maximum entropy approach to species distribution modelling. In: Proceedings ofthe Twenty-first Century International Conference on Machine Learning. ACM Press, New York, Pp. 655–662.

Phillips, S.J., R.P. Andersonb, and R. E. Schapired, (2006) Maximum entropy modeling of species geographic distributions, *Ecological Modelling*, 190: 231-259.

Prates-Clark, C.D.C., S.S. Saatchib, and D. Agostid (2008) Predicting geographical distribution models of high-value timber trees in the Amazon Basin using remotely sensed data. *Ecological Modelling*, 211: 309–323.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecological Letters*, 3: 349–361.

Raes, N., and H. ter Steege (2007) A null-model for significance testing of presence-only species distribution models. *Ecography*, 30: 727–736.

Reid, D. J. (1968) Combining three estimates of gross domestic product. *Economica*, 35: 431–444.

Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Rodríguez, J.P., L. Brotons, J. Bustamante and J. Seoane (2007) The application of predictive modelling of species distribution to biodiversity conservation. *Diversity and Distributions*, 13: 243–251.

Scott, J.M., M. Murray, R.G. Wright, B. Csuti, P. Morgan, and R.L. Pressey (2001) Representation of natural vegetation in protected areas: capturing the geographic range. *Biodiversity and Conservation*, 10: 1297–1301.

Scott, J.M., P.J. Heglund, J.B. Haufler, M. Morrison, M.G. Raphael, and W.B.Wall (2002). Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, CA.

See L. and R.J. Abrahart (2001) Multi-modal data fusion for hydrological forecasting. *Computers and Geosciences*, 27: 987–994.

Segurado, P., M.B. Aeaujo, and W.E. Kunin (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, 43: 433–444.

Seibert, P. and M. Conrad-Brauner (1995) Concept, mapping and application of the potential natural vegetation taking the PNV-map of the lower Inn-Valley as an example. *Tuexenia*, 15:25–43.

Sen, B.Y. (1937) Study of plant communities of *Abies kawakamii* near sub-alpine vellege (I); (II). *Animals and Plants*, 6(9): 46–52.

Song, G.Z.M., C.T. Lin, C.R. Chiou, and Y.C. Lu (2007) Comparing three species distribution models - Applied in Tsuga chinensis distribution in Taiwan. *Proceedings of the Fifth Symposium of Vegetaion Diversity in Taiwan, Vegetation Mapping Series*, 9: 49–65.

Steyn, D.G. (1980) The calculation of view factors from fisheye-lens photographs. *Atmosphere-Ocean*, 18(3): 254–258.

Stockwell, D.R.B. (2006) Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecological Modelling*, 192: 188–196.

Stockwell, D.R.B., and A.T. Peterson (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148: 1–13.

Strasser, H. and C. Weber (1999) On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, 8: 220–250.

Su, H.J. (1984a) Studies on the climate and vegetation types of the natural forests in Taiwan (I). Analysis of the variation in climatic factors. *Quarterly Journal of Chinese Forestry*, 17(3): 1–14.

Su, H.J. (1984b) Studies on the climate and vegetation types of the natural forests in Taiwan (II). Altitudinal vegetaion zones in relation to temperature gradient. *Quarterly Journal of Chinese Forestry*, 17(4): 57–73.

Su, H.J. (1985) Studies on the climate and vegetation types of the nature forests in Taiwan (III): a scheme of geographical climatic resgions. *Quarterly Journal of Chinese Forestry*, 18(3): 33–44.

Su, H.J. (1987) Forest habitat factor and its quantititative estimation. *Quarterly Jounal of Chinese Forestry*, 20(1): 1–14。(in Chinese)

Su, H.J. (1988) Study of vegetation ecology at nature preserve of *Juniperus sqyanata* var. *morrisonicola* in Hsuehshan. *Conservation Serie*, Forestry Bureau, Council of Argriculture Executice Yuan, Pp 123. (in Chinese)

Su, H.J. (1991) Study of vegetation ecology at nature preserve of conifer and deciduous forest in Pei Da Wu Mountain (II): estimation of representativeness of nature preserve of vegetation analysis. Study of Vegetation Ecology of Nature Preserve of

Nature Forest of Taiwan, Forestry Bureau, Council of Argriculture Executice Yuan, Pp. 141. (in Chinese)

Su, H.J. (1992) Vegetation of Taiwan: altitudinal vegetation zone and geographical climate regions. Institute of botany, Academia Sinica Monograph series, *Academia Sinica, Taiwan*, 11: 39-53.

Swets, J.A. (1988) Measuring the accuracy of diagnostic system. *Science*, 240: 1285–1293.

Taiwan Government Information Office. (2008) http://www.gio.gov.tw/

Thomas, C.D., A. Cameron, R.E. Green, M. Bakkenes, L.J. Beaumont, Y.C. Collingham, B.F.N. Erasmus,M.F. de Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A.S. van Jaarsveld, G.F. Midgley, L. Miles, M.A. Ortega-Huerta, A.T. Peterson, O.L. Phillips, S.E. Williams (2004) Extinction risk from climate change. *Nature*, 427: 145–148.

Thuiller, W., M. Araujo, and S. Lavorel (2003) Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, 14: 669–680.

Tsao, L.S. (2007) Using generalized additive models to establish the relationships between distribuition ranges and climatic factors for six conifer species of Taiwan. Master Thesis, School of Forestry and Resource Conservation, National Taiwan University. Pp. 76.

Tsoar, A., O. Allouche, O. Steinitz, D. Rotem, and R. Kadmon (2007) A comparative evaluation of presenceonly methods for modelling species distribution. *Diversity and Distributions*, 13: 397–405.

Tuhkanen, S., (1980) Climatic Parameters and Indices in Plant Geography. Almqvist and Wiksell International, Sweden, Pp. 110.

Urban, D.L., G. B. Bonan, T.M. Smith, and H.H. Shugart (1991) Spatial applications of gap models. *Forest Ecology and Management*, 42: 95–110.

Vayssières, M. P., R. E. Plant, and B.H. Allen-Diaz (2000) Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, 11: 679–694.

Walter, H. (2002) Walter's vegetation of the Earth: The Ecological Systems of the Geo-Biosphere. 4th, Completely Revised and Enlarged Edition. Springer-Verlag, Berlin, Pp. 527.

Walther, G.R. (2004) Plants in a warmer world. Perspective in Plant Ecology, *Evolution and Systematics*, 6: 169-185.

Wang, Y.S., B.Y. Xie, F.H. Wan, Q.M. Xiao, and L.Y. Dai (2007) The potential geographic distribution of *Radopholus similis* in China. *Agricultural Sciences in China*, 6(12): 1444–1449.

Wang, Y.S., B.Y. Xie, F.H. Wang, Q.M. Xiao, and L.G. Dai (2007) Application of ROC curve analysis in evaluating the performance of alien species' potential distribution models. *Biodiversity Science*, 15(4): 365–372.

Whittaker, R.H. (1975) Classification of Plant Communities. Wiksell International, Sweden, Pp. 110.

Wilson, J.D., (1984) Determining a TOPEX score. *Scottish Forestry*, 38: 251–256.

Wilson, K.A., M.I. Westphal, H.P. Possingham, and J. Elith (2004) Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, 22(1): 99–112.

Yang, Y.L. (1997) A study on the application of potential vegetation to planting design-as case study syn-show mountain of Taipei. Master Thesis, Department of Horticulture, National Taiwan University.

Yen, S.M. (2007) Modeling species distributions of three coniferous forest type in Taiwan. Master Thesis, Department of Geography, College of Science, National Taiwan University, Pp. 96.

Yen, S.M., C.R. Chiou, K.C. Chang, and J.R. Lin (2007) Development and evaluation of Taiwan Hemlock distribution model in Taiwan. Quarterly Jounal of Chinese Forestry, in publishing.

Zevenbergen, L.W., C.R. Thorne (1987) Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12: 47–56.