EVALUATING COMMON STATISTICAL METHODS USED FOR SPECIES DISTRIBUTION MODELING OF TWO TREE SPECIES

Hou-Chang Chen[a], Nan-Jang Lo[b], Wei-I Chang[c], and Kai-Yi Huang[*d]

[a]Graduate student, Dept. of Forestry, Chung-Hsing University, Taiwan E-mail: zkchris@hotmail.com
[b]Specialist, EPMO, Chung-Hsing University, Taiwan E-mail: njl@dragon.nchu.edu.tw
[c]Chief, FP Sec., Forest Bureau, Council of Agriculture, Taipei 100, Taiwan E-mail: weii@forest.gov.tw
[d*]Professor, Dept. of Forestry, Chung-Hsing Univ., E-mail: kyhuang@dragon.nchu.edu.tw (corresponding author)
250 Kuo-Kuang Road, Taichung, Taiwan 402, Tel: +886-4-22854663; Fax: +886-4-22854663

KEY WORDS: Species distribution modeling (SDM), Ecological characteristics, Remote sensing, Spatial extrapolation, *Schima superba* (Chinese guger tree), *Rhododendron formosanum* (Formosan rhododendrons).

ABSTRACT: Various statistical techniques have been used to model species suitable environmental niche. This study performed a comprehensive assessment of statistical techniques and examined whether model predictions are associated with ecological characteristics of species. We chose two representative species: widespread species, Chinese guger tree (CGT) and semi-cluster species Formosan rhododendrons (FR) in the Huisun study area in Taiwan as modeling target. Eight algorithms, including decision tee (DT), discrimnant analysis (DA), logistic multiple regression (LMR), maximum entropy (MAXENT), DOMAIN, BIOCLIM, generalized linear models (GLM), maximum likelihood ML), were used in a GIS to incorporate topographic factors and vegetation index for species distribution modeling (SDM). Model performance was evaluated based on *kappa* index, Independent t-test, and Bonferroni multiple comparison. The results indicated that models had strikingly different predictive abilities among them. The overall modeling accuracies of FR species were higher than those of CGT species and ecological characteristics of species affected predictive performance of models. DT, MAXENT, and DOMAIN were the best among the eight models. More importantly, SDM models merely based on topographic variables and their sample distribution could not be applied on a larger spatial scale since the topographic attributes of Tong-Feng and Kuan-Dau watersheds in Huisun are different. Consequently, the predictions from models built based on samples only from one watershed could not be accurately extrapolated to another; however, SDM performance of FRs was still better than that of CGTs. Improving model performance would be to run models on an iterative basis, and future work would incorporate predictor variables with soil and microclimate into a model so that it can be improved in accuracy and applied at a larger region.

1. INTRODUCTION

Species distribution model (SDM) has been the core of spatial ecology since the latter half of the 20th century (Guisan and Zimmermann, 2000). It can provide a measure of a species' occupancy potential in areas not covered by biological surveys and consequently is becoming an essential tool to conservation planning and forest management (Elith *et al*., 2006). Technological innovation over the last few decades, especially in the fields of remote sensing (RS) and geographic information systems (GIS), greatly enhanced scientists' capacity to face this challenge by giving them the ability to describe patterns in nature over broader spatial scales and at a greater level of detail than ever before (Miller *et al*., 2004).

Advances in statistical techniques enhance the ability of researchers to tease apart complex relationships, while effectively incorporated of RS and GIS tools permit more accurate descriptions of spatial patterns and guide field surveys (Engler *et al*., 2004). A broad suite of algorithms has been used to predict the geographical distributions of species (Elith *et al*., 2006). Here, we implemented decision tree (Breiman *et al*., 1984), discriminant analysis (Johnson and Wichern, 2007), logistic multiple regression (Johnson and Wichern, 2007), maximum entropy (Phillips *et al*., 2006), BIOCLIM (Busby, 1991), DOMAIN (Carpenter *et al*., 1993), generalized linear model (Guisan *et al*., 2002), Maximum likelihood (Myung, 2003) to build models since they produced useful predictions in other studies. These models use known distribution records of species, as well as environmental predictors (e.g. elevation, soil, rainfall), to build statistical functions for interpolating species' distributions across the environmental space (Guisan and Zimmermann 2000). Models may also extrapolate species' distributions to sets of environmental conditions outside those used to build the models (Peterson, 2003).

However, many studies pointed out that it was difficult to accurately depicted spatial patterns for all species regardless of the techniques used. This suggests that model performance may also be influenced by species ecological characteristics, sample size, model selection**,** and predictor contribution (Araújo and Guisan, 2006). Determining how species' traits influence model performance is particularly important because environmental tolerance, range size, niche width**,** and niche preference could provide the means to predict which species are suitable for modeling (Stockwell and Peterson, 2002; Guisan *et al*., 2007; Hanspach *et al*., 2010). Therefore it must be interpreted carefully of species' occupancy potential in areas not covered by biological surveys. Generally, models for species with broad geographic ranges and high environmental tolerance tend to be less accurate than those for species with smaller geographic ranges and limited environmental tolerance (Guisan *et al*., 2007).

Data characteristics and species traits are expected to influence the accuracy with which species' distributions can be modeled and predicted. This study aimed at (1) testing how the different ecological traits affect the model prediction; (2) evaluating the ability of spatial extrapolation with models; (3) determining which model or models are most useful for predicting plant species distribution. We selected two representative types of tree species, one was widespread and the other was cluster for the study. *Schima superba* (Chinese guger tree, CGT) is a widespread tree species with high dispersal ability (Liu *et al*., 1994). *Rhododendron formosanum* (Formosan rhododendrons, FR) is a kind of semi-cluster tree species, and always forms pure forests (Liu *et al*., 1994). CGT and FR were chosen as targets because they are representative species in the Huisun study area in central Taiwan. Field surveys were conducted to collect their samples from the study area by a GPS with sub-meter accuracy. Further, GIS technique was used to overlay the layer of species with environmental variables. Two sampling schemes were created for model development and validation via different combinations of CGT and FR samples taken from Tong-Feng and Kuan-Dau watersheds in Huisun. We assessed eight modeling techniques and used "multi-modeling assessment approach". This included the application of a single model to data describing patterns at different spatial scales and the comparison of several models using a common dataset.

2. STUDY AREA

The study area is a rectangular area, which encompasses Huisun Forest Station with an irregular shape (7, 477 ha), and it has a total area of 17,136 ha. Huisun Forest Station is located in the 24˚2′–24˚5′ N latitude and 121˚3′–121˚7′ E longitude (Figure 1). This station is the property of National Chun-Hsing University. The entire study area ranges in elevation from 454 m to 2,418 m, and its climate is temperate and humid.

In addition, the study area has nourished many different plant species more than 1,100 and is a representative forest in central Taiwan. It comprises five watersheds, including two larger watersheds, Kuan-Dau at west and Tong-Feng at east. All of the CGT and FR samples were collected from the Tong-Feng and Kuan-Dau sites in Huisun by using a GPS.

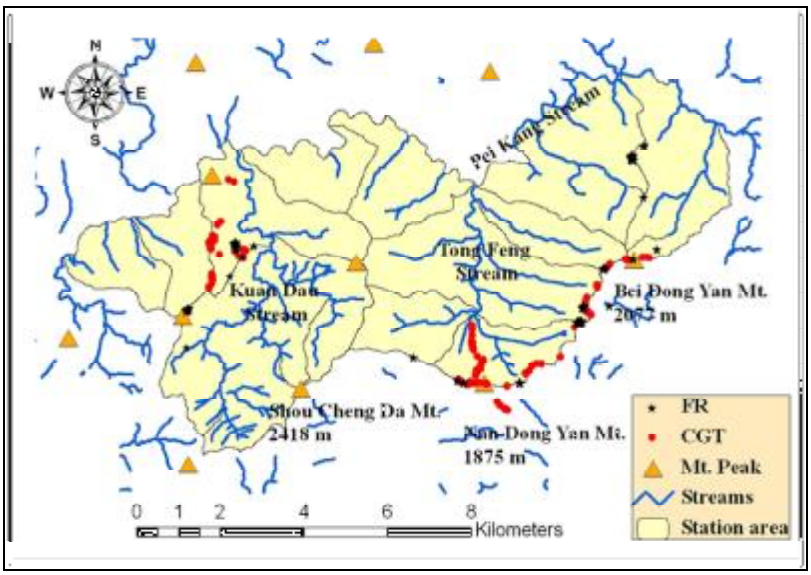

Figure 1: Location map of the study area.

3. MATERIALS AND METHODS

3.1 Species data and sampling design

We collected *in situ* CGT and FR samples by using a GPS linked with a 5-m expandable rod and a laser range finder, and then performed a post-processed differential correction that makes them have an accuracy of sub-meters. The dataset was eventually converted into ArcView shapefile format for later use. The sample sizes of the two tree species are shown in table 1.

Pseudo-absences (500 samples) were randomly generated from the study area for those models that required them (all except DOMAIN, BIOCLIM, MAXENT). Two sampling designs (SD) were created for model development and validation through different combinations of CGT and FR samples from aforementioned two sites (figure 1).

SD-1: we randomly selected two-thirds of Tong-Feng dataset for building "Tong-Feng base model" and the remaining one-third of that dataset for model validation.

SD-2: we used the same base model in SD-1 and only used samples taken from the Kuan-Dau watershed about 5 km away from Tong-Feng watershed to test the base model. Then we evaluated the spatial extrapolation ability of the eight models.

Table 1: The sample sizes of the two tree species.

Species	Tong-Feng watershed	Kuan-Dau watershed	Sample size
CGT	122	64	186
FR	118	61	179

To build reliable predictive models, sampling data points was repeated ten times for subsequent modeling. The GIS database used in the study was constructed by using ERDAS Imagine software module Layer Stack to overlay the layers of five environmental variables. CGT and FR sample layers were overlaid with five data layers, respectively, and those pixels of the five layers lying at the same position with tree sample pixels were clipped out.

3.2 Environmental variables

We collected digital elevation model (DEM) of 40 m resolution, orthophoto base maps (1:10,000), and two-date SPOT images. DEM was acquired from the Aerial Survey Office, Forestry Bureau of the Council of Agriculture, Taiwan. To meet the requirements of the study, the DEM was interpolated and resampled into 5 × 5 m grid size, geo-referenced to the coordinate system, TWD67 (Taiwan Datum, spheroid: GRS67) and Transverse Mercator map projection over two-degree zone with the central meridian 121°E.

The two-date SPOT-5 images were acquired from Center for Space and Remote Sensing Research, National Central University (CSRSR, NCU), Taiwan (© SPOT Image Copyright 2004 and 2005, CSRSR, NCU). System calibration and geometric correction with level 2B were performed on the images, and then they were rectified to the TWD67 Transverse Mercator map projection and resampled to 5 m resolution to be consistent with the layers from DEM. We chose the two-date SPOT-5 images (07/10/2004 and 11/11/2005) since they have the best quality with the amount of clouds less than 10%.

Elevation, slope, and aspect were generated from DEM by ERDAS Imagine software module. The ridges and valleys in the study area were used together with DEM to generate terrain position layer. The main ridges and valleys over the study area were directly interpreted from the orthophoto base maps; these lines were then digitized to establish the data layer by using ARC/INFO software for later use. The data layer in a vector format was then converted into a new data layer in a raster format by ERDAS Imagine software module, and then combined with DEM to generate terrain position layer (Skidmore, 1990). Vegetation index was derived from the two-date SPOT images based on the concepts stated in Hoffer (1978), and the index derived from the ratio of difference of NIR and MIR is expressed as followed:

$$\frac{NIR_{autumn} - MIR_{autumn}}{NIR_{summer} - MIR_{summer}} \quad (1)$$

3.3 Model development

Eight predictive techniques were used to build the species distribution models (table 2). They were: (1) DT: decision tree; (2) DA: discriminant analysis; (3) LMR: logistic multiple regression; (4) MAXENT: based on maximum entropy and Bayesian probability; (5) BIOCLIM; (6) DOMAIN; (7) GLM: generalized linear model; (8) ML: Maximum likelihood.

Table 2: Modeling technologies implemented in this paper.

Model	Core concept	Data type	Software	Source
DT	Classification tree	PA	SPSS	(Breiman *et al*., 1984)
DA	Discriminant function	PA	SPSS	(Johnson and Wichern, 2007)
LMR	Non-linear probability model	PA	SPSS	(Johnson and Wichern, 2007)
MAXENT	Entropy value and Bayesian probability	PE	MAXENT (free software)	(Phillips *et al*., 2006)
DOMAIN	Continuous point-to-point similarity metrix	PO	ModEco (free software)	(Carpenter *et al*., 1993)
BIOCLIM	Envelope model	PO	ModEco (free software)	(Busby, 1991)
GLM	Non-linear regression model	PO	ModEco (free software)	(Guisan *et al*., 2002)
ML	Maximize the likelihood function	PA	ModEco (free software)	(Myung, 2003)

* PO, only presence data used; PE, presence compared against the entire region; PA, presence and absences used.
**For these analyses, we randomly selected 500 pseudo-absences from each region.

Decision tree (DT) builds the rule by recursive binary partitioning into regions that are increasingly homogeneous with respect to a class variable. The homogeneous regions are called nodes. At each step in fitting a classification tree, an optimization is carried out to select a node, a predictor variable, and a cut-off or group of codes that result in the most homogenous subgroups for the data, as measured by the Gini index (Breiman *et al*., 1984). This criterion could set the optimum tree as a trade-off between goodness of fit on training data and size of the tree. Such a classification tree is said to be full grown, and the final regions are called terminal node. Terminal nodes are assigned a final outcome based on group membership of the majority of observations (Breiman *et al*., 1984; De'ath and Fabricius, 2000).

Discriminant analysis (DA) is a classification technique, which discriminates among k classes (objects) based on a set of independent or predictor variables. The objectives of DA are to find linear composites of n independent variables that maximize among-groups to within-groups variability. Then test if the group centroids of the k dependent classes are different. Further determine which of the n independent variables contribute significantly to class discrimination. Finally, assign unclassified or "new" observations to one of k classes (Johnson and Wichern, 2007).

Logistic multiple regression (LMR) can be bi- or multinomial which the observed outcome can have only two possible types (e.g., "presence" vs. "absence" or "yes" vs. "no"). Generally, the outcome is coded as "0" and "1" in binary logistic regression as it leads to the most straightforward interpretation. Where given P is the probability potential habitat of the tree species at a particular position, Y_h is dependent variable, x_1, …, x_n are a set of independent predictor (biophysical) variables, and B_0, . . ., B_n are logistic coefficients. The output probability values range from 0 to 1, with 0 indicating absence (a 0 percent probability of the target habitat) and 1 indicating presence (a 100 percent probability). The default threshold of 0.5 implies that probabilities above 0.5 are target habitat and below 0.5 are non-target habitat (Johnson and Wichern, 2007).

Maximum entropy (MAXENT) is a promising new method for recent publication (Phillips *et al*., 2006). MAXENT finds a probability distribution, defined over the study area and satisfies a set of constraints derived from the occurrence data. Each constraint requires that the expected value of an environmental variable (or function thereof) must be within a confidence interval of its empirical mean (the mean over the presences). Among distributions that satisfy the constraints, MAXENT chooses the one that maximizes entropy, i.e., is the closest to uniform condition, as any other choice would represent constraints on the distribution that are not justified by the data (Phillips *et al*., 2006; Phillips *et al*., 2008).

DOMAIN uses a point-to-point similarity metric (based on the Gower distance) to assign a value of habitat suitability to each potential site based on its proximity in the environmental space to the closest (most similar) occurrence location (Carpenter *et al*., 1993). A threshold value of suitability can then be selected to determine the boundaries of the ecological niche. Note that in contrast to all previous methods, environmental envelopes defined by DOMAIN are not necessarily continuous in the environmental space.

BIOCLIM defines the ecological niche of a species as the bounding hyper-box that encloses all the records of the species in the core-climate (Busby, 1991). Thus, it creates a rectilinear envelope in the environmental space, defined by the most extreme records (minimum and maximum) of the species on each environmental variable. To reduce the sensitivity of model predictions to outliers, the species records are sorted along each variable and only the records that lie within a certain percentile range of the data are used for model construction. In this study we applied a percentile range of 95% (disregarding 2.5% of the values on each side) (Carpenter *et al*., 1993).

Generalized linear model (GLM) is a generalization of general linear models. General class of linear models is made up of three components: random, systematic and link function. Random component identifies response variable E(Y) and its probability distribution. Systematic component identifies a set of predictor variables (X_1,...,X_k). Link function identifies a function of the mean that is a linear function g(μ) of the predictor variables. By using a logit link function that transforms the scale of the response variable, being able to relax the distribution and constancy of variances assumptions that are commonly required by traditional linear models (McCullagh and Nedler, 1989). Consequently, the GLM model is particularly suitable for predicting species distributions, and has been proven to be successful in various ecological applications (Guisan *et al*., 2002).

Maximum likelihood (ML) is assuming that the heights are normally distributed with some unknown mean and variance, the mean and variance can be estimated with ML while only knowing the heights of some samples of the overall population. ML would accomplish this by taking the mean and variance as parameters and finding particular parametric values that make the observed results the most probable. In general, for a fixed set of data and underlying statistical model, the method of ML selects values of the model parameters that produce a distribution that gives the observed data the greatest probability (i.e., parameters that maximize the likelihood function). Maximum-likelihood estimation gives a unified approach to estimation, which is well-defined in the case of the normal distribution and many other problems (Myung, 2003; Jensen, 2005).

3.4 Model Evaluation

Predictions of each model were compared to the validation data set to form a confusion matrix (Fielding and Bell, 1997), from which Cohen's *Kappa* was calculated. The *kappa* value range from -1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random (Cohen, 1960; Lillesand *et al*., 2008).

The aims of this study were to achieve the following specific objectives: (1) testing how the different ecological traits affect the model prediction; (2) evaluating the ability of spatial extrapolation with models; (3) determining which model or models are most useful for modeling distribution of a plant species. Since the same measurement of model performance was calculated ten times (different modeling procedures) on each species, a repeated measures hierarchical design was used for testing difference in model performance.

Differences between overall performance of each modeling procedure, assuming no interaction between model procedure and species type, were tested using the independent t-test that was used to compare for difference between two type species. Bonferroni correction tests were performed for post-hoc multiple comparisons of the models. Finally, we used split-sample validation for evaluating methods run under the different samplings. The first one (training set) was used to build model; the other one (test set) was used to validate the model.

4. RESULTS AND DISCUSSION

At beginning, we integrated field survey data and overlaid five environmental variables including four topographic factors and vegetation index derived from SPOT-5 satellite images. The total number of model runs made for this study was 2 species × 2 sampling designs × 5 predictors × 8 modeling algorithms × 10 replication, which results in 1,600 spatial predictions of species distributions. Owing to a very large amount of calculations, we reduced the predictor-variable dimension to improve calculating efficiency. The software module for each of the eight algorithms can evaluate the relative importance of five predictor variables with three predictive models for predicting the potential habitat of two species, as a reference for choosing the most effective variables. Overall, elevation, slope and terrain position had the highest relative importance for CGT and FR, respectively. Hence, we used the three variables to build predictive models.

4.1 Analyses of species traits

The line graph shows (figure 2) the relation between variances of techniques in model accuracy across species and variance of species across techniques. Line pattern in this diagram indicates that model accuracies were

influenced by species traits and modeling algorithms. The variance in model accuracy across species was greater than that across techniques. This result responded to those in previous studies (Thuiller, 2003). Then we explored whether individual species' characteristics could explain variation in model performance, we pooled eight model performances and obtained average values. According to the independent t-test, the overall modeling accuracies of FR species were higher than those of CGT species (mean *kappa*: FR (0.66) > CGT (0.51)) and the ecological characteristics of the two species substantially affected the predictive performance of models ($t = -9.42$, $p < 0.001$).

We found that CGT species has a broad and dispersed distribution, while FR species has a specialized, narrow, and clustery distribution, usually forming a pure forest. Specifically, FR species was hard to compete with many other species in a good environment due to its slow growth rate, but it can grow in a poor environment with thin, acidic, and infertile soils where most species are hard to grow. Consequently, the ecological characteristics of species can affect modeling accuracy, and species with a widespread distribution (or broad ecological amplitude) like CGT are generally more difficult for modeling than species with a clustery distribution like FR. Our result was consistent with the viewpoints asserted by Guisan *et al*. (2007) and Stockwell and Peterson (2002).

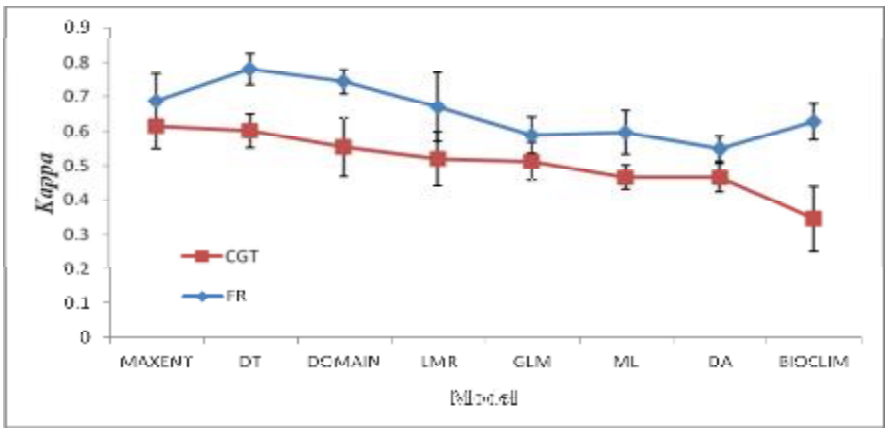

Figure 2: Performance of species for each technique based on their mean *kappa* ± one standard deviation.

4.2 Model evaluation and comparison

The figure 3 shows of performance of modeling techniques compared to the best prediction (pooled two species) by their mean *kappa* ± one standard deviation. The patterns clearly showed that model accuracy varied with species. The top three algorithms were DT (0.69), MAXENT (0.65), DOMAIN (0.65), followed by LMR (0.63), GLM (0.55), ML (0.53), all outperforming DA (0.51) and BIOCLIM (0.49). The diagram also indicated that DT, MAXENT, and DOMAIN models had relatively small standard deviations. BIOCLIM model had the highest standard deviation with respect to mean *kappa* and the poorest model performance.

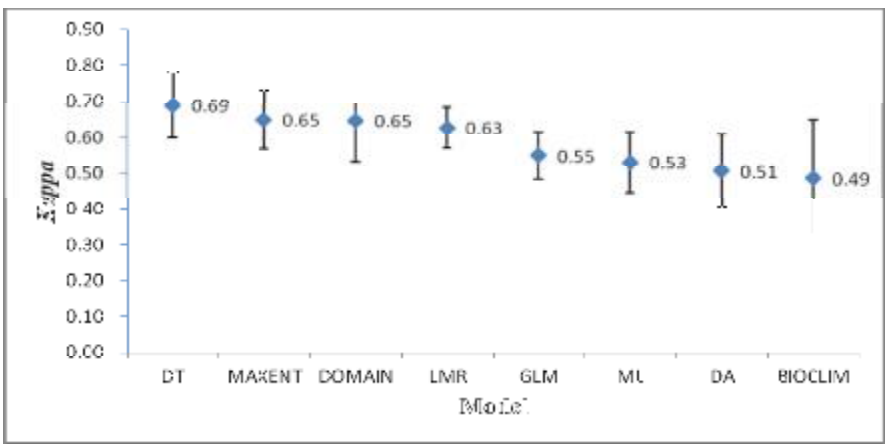

Figure 3: Performance of modeling techniques compared to the best prediction (pooled two species) based on their mean *kappa* ± one standard deviation.

Moreover, we integrated the multiple comparisons of predictive performance (*kappa*) of the eight modeling algorithms (table 3). We note that the profile methods (i.e. BIOCLIM) did not perform very well in predictions, and the model performance was significantly lower than those of other models ($p < 0.001$). DT, MAXENT, DOMAIN and LMR belong to the same level of accuracy (non-significant in Bonferroni test $p > 0.05$), followed by GLM and ML model, and all of the previous models were much better than BIOCLIM and DA models.

Table 4 indicated that predictions of DT, MAXENT and DOMAIN models generated high potential areas of CGTs and FRs, and substantially reduced the area of field survey to less than about 10% (1,714 ha) of the entire study area (17,136 ha) for CGTs, while these three models considerably reduced the area of field survey to less than 5% (857 ha) of the entire study area for FRs. These three models were better suited for predicting the tree's potential habitat since they save both cost and labor in the future investigation. We suggest that these models may be useful for conservation purpose and management tool.

Table 3: Multiple comparisons of predictive performance (*kappa* values) of the eight modeling techniques.

	DT	DA	LMR	MAXENT	DOMAIN	BIOCLIM	GLM	ML
DT	-							
DA	**	-						
LMR	n.s.	n.s.	-					
MAXENT	n.s.	**	n.s	-				
DOMAIN	n.s.	**	n.s	n.s.	-			
BIOCLIM	***	n.s.	***	***	***	-		
GLM	**	n.s.	n.s	n.s.	n.s.	n.s.	-	
ML	***	n.s.	n.s	n.s.	*	n.s.	n.s.	-

*Bonferroni correction for multiple testing; * $P < 0.05$; ** $P < 0.01$*** $P < 0.001$; n.s., $P > 0.05$.

Table 4: The distribution statistics of models predicting the potential habitat of the two species.

Species	Predictive area %	DT	DA	LMR	MAXENT	DOMAIN	BIOCLIM	GLM	ML
CGT	Habitat	3%	17%	4%	4%	9%	38%	4%	14%
	Non-habitat	97%	83%	96%	96%	91%	62%	96%	86%
	total	100%	100%	100%	100%	100%	100%	100%	100%
FR	Habitat	2%	23%	3%	2%	5%	13%	7%	7%
	Non-habitat	98%	77%	97%	98%	95%	87%	93%	93%
	Total	100%	100%	100%	100%	100%	100%	100%	100%

*Total area = 17,136 ha

4.3 Evaluating the ability of spatial extrapolation with models

Next, assess the spatial extrapolation ability of these models (see table 5). We created two sampling designs with different spatial scales via different combinations of CGT and FR samples taken from Tong-Feng and Kuan-Dau watersheds in Huisun. According to the SD-1 in table 5(b), we extended the prediction from one area to another and assessed the robustness of underlying relationships. As shown in table 5(c) (SD-2), the mean *kappa* values of these models declined sharply as the eight models were tested by independent samples from Kuan-Dau sites.

Each model had a similar change by mean *Kappa* values, and we found two level effect of change. First, models had different predicting ability and species ecological traits may affect model performance of prediction. Second, different sampling designs with different sample space distribution also affect model prediction. We take the map predicted by MAXENT as an example and enlarged a local area with red dashed line box in Kuan-Dau sites to show the difficulty of spatial extrapolation (figure 4). The figure clearly explained that "Tong-Feng base models" failed to pass validation by Kuan-Dau test samples despite passing validation by Tong-Feng test samples.

Topographic variables are frequently used to build predictive models because they are most easily measured in the field and have a good relation with observed species patterns in a small spatial scale. Again, such variables usually replace a combination of different resources and direct gradients (e.g. climate, rainfall) in a simple way (Guisan *et al*., 1999). However, the model performed poorly on spatial extrapolation from Tong-Feng to Kuan-Dau because the topographic attributes of the two watersheds are quite different from each other. Therefore, the models developed from topographic variables can only be applied within a limited geographical extent without

significant error.

Table 5: Comparisons of model performance for eight algorithms evaluated by the test dataset for three treatments.

(a) species	CGT			FR		
(b) SD-1	**Mean *kappa***	**Rank**	Standard deviation	**Mean *kappa***	**Rank base**	Standard deviation
DT	**0.60**	**2**	**0.05**	**0.78**	**1**	**0.05**
DA	0.47	6	0.04	0.55	8	0.04
LMR	0.52	4	0.08	0.67	4	0.10
MAXENT	**0.61**	**1**	**0.06**	**0.69**	**3**	**0.08**
DOMAIN	**0.55**	**3**	**0.09**	**0.74**	**2**	**0.04**
BIOCLIM	0.35	8	0.09	0.63	5	0.05
GLM	0.51	5	0.05	0.59	9	0.05
ML	0.47	6	0.04	0.60	6	0.06
(c) SD-2	**Mean *kappa***	**Rank**	Standard deviation	**Mean *kappa***	**Rank base**	Standard deviation
DT	0.01	6	0.03	0.00	8	0.00
DA	**0.28**	**1**	**0.05**	0.37	5	0.02
LMR	0.01	7	0.02	**0.43**	**3**	**0.15**
MAXENT	0.00	8	0.00	0.03	7	0.06
DOMAIN	0.03	4	0.04	0.18	6	0.08
BIOCLIM	**0.12**	**2**	**0.10**	0.41	4	0.04
GLM	0.01	5	0.02	**0.45**	**2**	**0.07**
ML	**0.04**	**3**	**0.00**	**0.47**	**1**	**0.04**
Overall	0.51	2	0.10	**0.66**	1	0.10

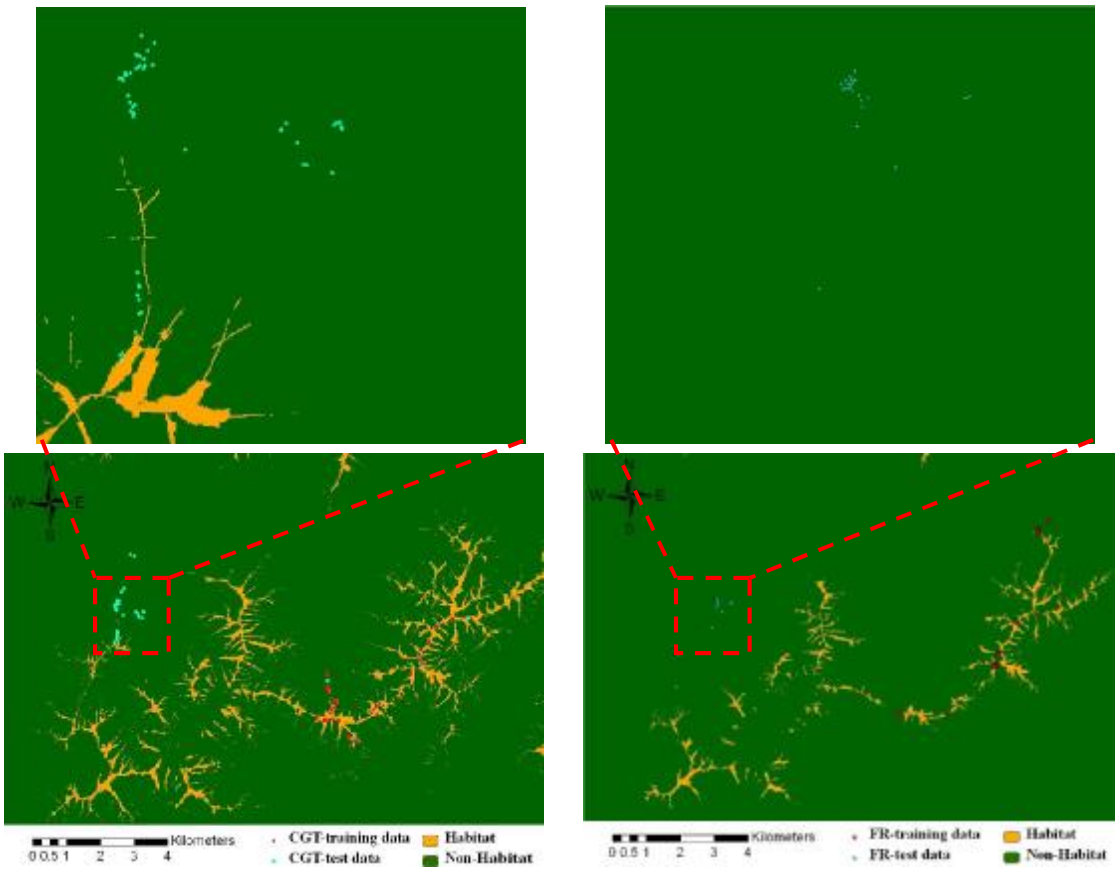

Figure 4: The lower maps are the potential habitat of CGT and FR species in Huisun generated from MAXENT and the upper ones are enlarged from a local area with red dashed line box in Kuan-Dau watershed.

5. CONCLUSIONS

Modeling algorithms and species traits with which species' distributions can be modeled and predicted are expected to influence the accuracy. The study illustrated a broad comparative exploration of species ecological characteristics with different organisms and processes responding to their environments, and the ways that these responses vary geographically. The results indicated that models built based on different algorithms or for different species had strikingly different predictive abilities among them. Moreover, the variance in model accuracy across species was greater than that across techniques. The overall modeling accuracies of FR species were higher than those of CGT species and ecological characteristics of species affected predictive performance of models. DT, MAXENT, and DOMAIN were the three best among the eight models.

More importantly, SDM models merely based on topographic variables and sample distributions corresponding to them could not be applied on a larger spatial scale since the topographic attributes of Tong-Feng and Kuan-Dau watersheds in Huisun are different. Consequently, the predictions from models built based on samples only from one watershed could not be accurately extrapolated to another; however, SDM performance of FRs was still better than that of CGTs. Improving model performance would be to run models on an iterative basis, and future work would incorporate predictor variables with soil and microclimate into a model so that it can be improved in accuracy and applied at a larger region.

6. REFERENCES

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

Breiman, L., J. H. Friedman, R. A. Olsen and C. G. Stone, 1984. Classification and Regression Trees. Chapman and Hall, New York, USA. 357 pp.

Busby, J. R., 1991. BIOCLIM-a BIOCLIMate analysis and prediction system. Nature conservation: cost effective biological surveys and data analysis (ed. by C.R. Margules and M.P. Austin), pp. 64-68. CSIRO, Canberra, ACT, Australia.

Carpenter, G., A. N. Gillison and J. Winter, 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation, 2, 667-680.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 (1): 37-46.

De'ath, G. and K. E. Fabricius, 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81 (11), pp. 3178-3192.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, 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. Overton, M. Mcc., A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz and N. E. Zimmermann, 2006. Novel methods improve prediction of species' distribution from occurrence data. Ecography, 29: 129-151.

Engler, R., A. Guisan and L. Rechsteiner. 2004. An imporved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Applied Ecology, 41: 263-274.

Fielding, A. H. and J. F. Bell, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence model. Environmental Conservation, 24, 38-49.

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

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, 615-630.

Guisan, A., S. B. Weiss and A. D. Weiss, 1999. GLM versus CCA spatial modeling of plant species distribution. Plant Ecology, 143, 107-122.

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

Hanspach, J., I. Kuhn, S. Pompe and S. Klotz, 2010. Predictive performance of plant species distribution models depends on species traits. Perspectives in Plant Ecology, Evolution and Systematics, 12, 219-225.

Hoffer, R., 1978. Biological and physical considerations in applying computer-aided analysis techniques to remote sensor data. In : Remote sensing : the quantitative approach, Swain P. H. & Davis S. M. Ed., McGraw-Hill pp 227-289.

Jensen, J. R., 2005. Introductory digital image processing-a remote sensing perspective, 3rd ed., Pearson Education,

Inc, New Jersey.
Johnson, R. A. and D. W. Wichern, 2007. Applied Multivariate Statistical Analysis, Prentice-Hall, Inc.
Lillesand, T. M., R. W. Kiefer and J. W. Chipman, 2008. Remote sensing and image interpretation. 5th Edition. John Wiley & Sons. Inc., New York.
Liu, Y. C., F. Y. Lu and C. H. Ou, 1994. Trees of Taiwan. Taichung, Taiwan: College of Agriculture, National Chung-Shing University, pp. 440, 451.
McCullagh, P. and J. A. Nelder, 1989. Generalized linear model. 2nd ed. Chapman and Hall, London. 512 pp.
Miller, J., R. G. Turner, E. A. H. Smithwick, C. L. Dent and E. H. Stanley, 2004. Spatial extrapolation: the species of predicting ecological patterns and processes. BioScience, 54 (4): 310-320.
Myung, I. J., 2003. Tutorial on maximum likelihood estimation. Mathematical Psychology, 47, 90-100.
Peterson, A. T., 2003. Predicting the geography of species' invasions vias ecological niche modelling. Quarterly Review of Biology, 78:419-433.
Phillips S. J., R. P, Anderson and R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions. Ecological Modeling, 190: 231-259.
Phillips, S. J. and M. Dudík, 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31:161-175.
Skidmore, A. K., 1990. Terrain position as mapped from a grided digital elevation model. Geographical Information Systems. 4 (1): 33-49.
Stockwell, D. R. B. and A. T. Peterson, 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling, 148, 1-13.
Thuiller, W., 2003. BIOMOD-Optimizing predictions of species distributions and projecting potential future shifts under global change. Global change biology, 9: 1353-1362.