**EVALUATE THE PERFORMANCE OF ECOLOGICAL NICHE MODELING FOR THE HBITAT OF JAPANESE ELAEOCARPUS**

Po-Huai Shen (1), Bao-Hua Shao (2), Nan-Chang Lo (3),Kai-Yi Huang (4)

1 Dept. of Forestry, Chung-Hsing University (CHU), No. 145 Xingda Rd, Taichung 402, China Taipei

2 Nan-Tou Division Office, Forest Bureau, The Council of Agriculture, 456 Shihguan Rd, NanTou 542

**3** Experimental Forest Management Office, Chung-Hsing University (CHU), Taichung 402

4 Dept. of Forestry, Chung-Hsing University (CHU), Taichung 402

Email: jeffery870214@gmail.com; baobao357@gmail.com; njl@nchu.edu.tw; kyhuang@dragon.nchu.edu.tw

Abstract: Conservation management and planning often rely on results of ecological niche modeling (ENM) to assist in new population searching, identifying, and prioritizing important biodiversity areas. We use ENM as an accessory appliance in investigation to proceed spatial extrapolation. It not only solves the problem when searching areas are inaccessible but also decreases labor cost and time cost. The study chose *Elaeocarpus japonicas* (Japanese Elaeocarpus, JE) as target species because JE is a sort of pioneer tree species in second succession of forest ecology. Besides, it is a superior species in middle altitude of Taiwan which widely spread through mountain ridge area and has important implications to ecologists. It used decision tree (DT), random forest (RF), maximum entropy (MAXENT) and discriminant analysis (DA) to develop ecological niche models, which incorporate topographic variables, including elevation, slope, aspect, terrain position (TP), surface curvature (SC), profile curvature (PRC), plan curvature (PLC), and global solar radiation (GSR) in a GIS. Eventually, four terrain-related variables elevation, slope, TP, and GSR were incorporated into the models based on relative importance of all predictor variables. Model calibration and evaluation for these models were implemented efficiently in the ArcGIS and SPSS software and some software modules written by Python. In the model evaluation, RF significantly outperformed (*kappa* value 0.74) the others with the same result as indicated in previous studies. Followed by DT and MAXENT, DT was nearly on a level with MAXENT (*kappa* value 0.69 and 0.67, respectively). DA was the worst but still had reasonable performance (*kappa* value 0.59). The four models accurately predicted the spatial distribution of JEs in Huisun Experimental Forest Station (HEFS), and substantially reduced the distribution area to less than 10% of the entire study area. As a result, they were well suited for spatial distribution modeling of JETs. In the first round of ENM iterative process, they can prioritize either the field-survey areas where it is viable to collect fine spatial-resolution microclimatic, edaphic, or biotic data for refining predictions of potential habitat in later rounds of ENM or search areas for new population discovery under the conditions of limited funding and manpower. After all, it is much difficult to predict the spatial pattern of JE species accurately since it has wide-spread, scattered distribution. Hence, the follow-up study will attempt use high-resolution DEM generated from LiDAR to derive above-mentioned terrain-related variables so that the predictive accuracy of ENMs can be improved substantially.

Keywords: Japanese Elaeocarpus, ecological niche modeling (ENM), random forest (RF), decision tree (DT), maximum entropy (MAXENT).

1. **INTRODUCTION**

Along with the rapid growth of science and technology and the quality of life improved, environmental pollution and ecological issues gradually be taken seriously, due to the influence of human activity and interference in the process of economic development can't all avoid, how to maintain the stability of the ecosystem, reduce loss of biodiversity is vital. Complex biodiversity is crucial for ecosystem so researchers should explore the basic stratum starting from species to assess the state of biodiversity. Species spatial information is a must watch and discuss issues, which helps us predict the distribution of specific species and drive the underlying information about biodiversity. Recently, researchers have focus on specific ecological niche modeling (ENM) and the extended idea of species distribution modeling (SDM). In the past, the spatial information of species was mostly obtained from ecological survey. However, due to the constraints of environment and transportation, it was difficult to obtain large range of information quickly, and relevant ecological survey might cost considerable manpower and funds. Therefore, the hypothesis of estimation should be made from representative sample areas.

Predictive distribution maps generated from either ENM or SDM can offer general but useful information for biodiversity conservation, forest management, etc. (Guisan and Zimmermann, 2000; Soberón and Peterson, 2004; Araújo and Guisan, 2006). The basic idea of SDM or ENM is the concept of “ecological niche” proposed by Odum. Species can only reproduce under their relative suitable environment due to their physiological limits. They generally have different distributed range because “niche breadth” of each species may not be same (Odum, 1971). The species with a wider niche breadth adjusts to wider environmental conditions, and then result in wider distribution. Ecologists especially value SDM because its base information closely related to current distribution of species. The species’ distributions are decided by various factors, direct and indirect factors. In general, direct factors include climatic, edaphic and biotic factors. Direct factors lack of accurate data in a wide range. The data are usually generated by spatial interpolation and geographical statistics through collected from the established survey stations. The estimated situation is often quite different from the actual situation, especially in the mountainous areas of Taiwan with complex terrain. Therefore, species distribution models mostly adopt indirect factors. Indirect factors can be obtained by remote sensing technology than generated DEM and relatively accurate variables can be obtained even in steep mountain areas. Relevant terrain variables can be derived from DEM such as elevation, slope, and terrain position. Therefore, many prior studies also generated prediction based on topographic factors and provided quality prediction (Brown, 1994; Guisan *et al.*, 1998; Guisan *et al.*, 1999). Even there were also some studies trying to interpolate direct factors, their results indicated that the precision of derived factors layers is poor in mountainous areas (Prudhomme and Reed, 1999; Marquinez *et al.*, 2003) Hence, in the area with complex terrain (e.g. Taiwan), introducing DEM to a GIS deriving physiographical factors may offer useful information for ENM or SDM. Most studies can’t obtain all of useful factors (i.e. direct and indirect factors) for model parameterization at one time; therefore, an iterative modeling process is needed (Williams *et al.*, 2009). First, model must be fitted by available but accurate factors than if new factors are quite accurate, modeler should iterate modeling process by adding these new factors.

Nowadays, as the evolution of technology emerge, many ecologists have combined 3S technology to establish a species distribution model which extracts the point data concurrent with related factors’ value to produce the potential distribution map. It not only predicts the fertile or potential distribution of species, but also establishes model rules for spatial extrapolation.

Generally, researchers would adopt various modeling methods to provide objective result for comparisons when executing the prediction manipulations. They can compare the distribution maps to rank the performance of different models. Breiman *et al.* (1984) and Thuiller (2003) applied decision tree (DT) method along with GIS to predict distribution maps. Maximum entropy (MAXENT)is a prospective tool in many domains in ecology field. MAXENT doesn’t suffer the statistical assumption and limitation, and it can use only a fewer number of data and incomplete information to perform robust predictions (Phillips *et al.*, 2006; Kumer and Stohlgren, 2009). The advantages of MAXENT modeling are very indispensable in ecological related field because it is unnecessary to collect abundant and representative point data in field survey. Guisan *et al.* (2007) used 10 method predicting distribution maps of tree species in Swiss and provided quality predictions. Wisz *et al.* (2008) manipulated SDM in broad scale (four continent) with 12 methods. Nevertheless, some novel method is applied by modelers for SDM, for instance random forest (RF) (Breiman, 2001; Elith *et al.*, 2006; Culter *et al.*, 2007; Hanberry *et al.*, 2012).

Another important aspect of SDM is applying different measures to reasonably assess model performance. Hernadez *et al.* (2006) emphasized that assessing model performance is needed. Owing to each measure has different consideration, in this study, we used: (1) *kappa*, (2) *Matthews’s correlation coefficient* (MCC), and (3) F1-score statistic, to evaluate model performance. Cohen’s *kappa* is commonly used in image classification, but also used in recent ecological modeling study (Pearson *et al.*, 2006; Guisan *et al.*, 2007; Wize *et al.*, 2008; Williams *et al.*, 2009).

Several studies have been published to demonstrate that date for static species has lower uncertainty than mobile species (Guisan *et al.*, 2007; Wisz *et al.*, 2008; Santos *et al.*, 2010). The SDM built for static species, therefore, have more useful and accurate prediction than that built for mobile species. For this reason, the focal tree species of this study is a kind of evergreen tree species, *Elaeocarpus japonicas* (Japanese Elaeocarpus tree, JET). JET’s leaf will turn red and peel in autumn and drought winter. It distributes in guangdong, sichuan, Japan, Taiwan area which produces in low altitude of 2200 m forest. It often mixed with other hardwood. Being as one of the pioneer tree species, it also like sunny environment. It needs less demand on environmental water and grow in the open areas surrounding the ridge. Luo Nanzhang (1922) 's survey at Dongfeng Creek also indicated that it was mainly distributed in shallow areas with direct sunlight and relatively dry soil. It has a large tolerance range for many factors, so its ecological amplitude is wide and has no important limiting factors, so it is widely distributed. JET is a kind of pioneer tree species in second succession, and therefore it plays a crucial role in ecosystem.

This study aimed at applying 3S (GIS, GPS and RS) technology to derive elevation, slope, aspect, terrain position, surface curvature (SC), profile curvature (PRC), plan curvature (PLC), and global solar radiation (GSR) from DEM, using these eight environmental layers to build predictive models. In this study, we adopted four methods (DT, DA, RF and MAXENT) and three measures (*kappa*, MCC and F1-score) for model performance assessment. DT, RF are closely related because RFis extended manipulation of DT, but implemented based on different considerations. MAXENT was reported with relative robust performance than most of the methods, this study used it as baseline for comparison. DA has relative low accuracy in these methods. Therefore, comparing model performance by multi measures was the main object of this study.

1. **STUDY AREA**

We chose the study area with rectangular shape (figure 1), which covers the Huisun Forest Station and has the total area of 18,164.3 ha. The analysis simulation was carried out according to this rectangular area. The Huisun Forest Station is situated in central Taiwan, situated within 24◦2´–24◦7´ N latitude and 120◦3´–121◦8´ E longitude. The station is the property of National Chung-Hsing University, and study area ranges in elevation from 454 m to 2,419 m, and its climate is temperate and humid. Hence, the study area has nourished a wide variety of plant species about 1,100 species and above. It is a typical forest in central Taiwan. It comprises five watersheds,includingtwo lager watersheds, Kuan-Dau at west and Tong-Feng at east. All of the JET samples were collected from the two watersheds by using a GPS.



Figure 1. Location map of the study area.

1. **METHODS AND METERIAL**

**3.1 Data Collection**

The collected data contained DEM with 4 resolutions including 1m,5m,20m,40m which is measured by lidar. The JET samples were acquired by field survey with Trimble PRO XR series GPS system. Furthermore, an expandable antenna rod with 5m in length and a laser ranging were adopted with GPS for enhancing the capacity of the system. Afterwards, we converted raw data into Shapefile format for analysis. A total of 224 samples of JET were collected in this study.

**3.2 Data Processing**

Elevation, slope, aspect, terrain position, surface curvature (SC), profile curvature (PRC), plan curvature (PLC) and global solar radiation (GSR) data layers were generated from different resolutions of DEM by using ArcGIS software. The ridges and valleys in the study area were used together with DEM to derive terrain position layer. JET samples obtained by a GPS were corrected by using post-processed differential correction and converted into ArcView shapefile format for later use.The layers of elevation, slope, aspect, terrain position, vegetation index, and JET sample data were overlaid by ArcGIS software.

**3.3 Presence and Absence Sample Data**

Species presence data are JET point sample collected by GPS in the field surveys and their concurrent environment factor value at the same locations. The ratio of absence to presence adopted in the study should be greater than a value of one to two. Species absence data were randomly selected from study area to avoid spatial autocorrelation (Pereira and Itami, 1991).

**3.4 Sampling Designs and Model Building**

Using test samples is an important issue of model evaluation. It helps us estimate the performance of modelers compared with reality world and provide strong evidences to demonstrate the prediction cam be reliable. Model validation includes split-sample validation, re-substitution validation and cross validation. In this study, we adopt split-sample validation process, and the original samples were divided to two parts, 70% for model building and 30% used as test sample for model evaluation. This method has the advantage of using untrained independent data to evaluate performance, which is more rigorous and less error than re-substitution method and is currently a better method for determining the spatial extrapolation ability of intelligent machines.

*3.5.1 Decision Tree (DT)*

DT (also called Classification and Regression Trees, CART) was developed by (Brieman etal., 1984), and Verbyla (1987) was first applied in ecological research. Due to the assumption that there was no normal distribution in statistics, this method was one of the non-parent number statistical methods (deng jiaku, 2004). DT could build classified rules from observations or some experiences (Guisan and Zimmermann, 2000). It makes predictions about the unknown data by analyzing the given known raw data and according to the classification rules. Leafnode refers to classification or prediction results (in this study, it is a habitat or a non-habitat), non-leaf nodes are judgment conditions (input variables), and the node at the top of the tree of judgment conditions at the beginning is called root node. The data were divided into two groups according to each property of the input data (environmental variables), so as to achieve the minimum difference between groups (the highest purity) and the maximum difference out of groups. This process was repeated to establish a tree structure. There is a unique path from the root node to each leaf node. When an unknown piece of data enters the decision tree, it will test the conditions one by one from the root node according to its corresponding classification rules, and finally get the predicted results (OBrien et al.,2005). DT calculation was performed in Python.

*3.5.2 Random Forest (RF)*

RF is a novel method for SDM, and it is recently applied in ecology and biogeography. This technique minimizes the bias of a single tree in DT (Breiman, 2001). RF built a number of simple trees by one of N partitions in original sample randomly. This process is iterated but with replacement from original sample, until the predefined tree number is reached. At the meanwhile, each tree is fitted with a number of m variables, here m << M (i.e. the number of variables in original sample set). Finally, the prediction of each cell (grain) is determined by voting of "forest (randomly built trees in RF)". The prediction value of each pixel is determined by the votes of this pixel (maxim value is the number of "trees" predefined in RF). RF calculation was performed in Python.

*3.5.4 Maximum Entropy (MAXENT)*

Maxent is a relatively new forecasting model that can make predictions or inferences with incomplete information which is based on the known information of target distribution. The principle is to explore the probability distribution (maximum degree of chaos (the most dispersed, or near uniform state), and then estimate the probability distribution of target species.Entropy represents the degree of chaos in the system, and can also be regarded as noise affecting the state of useful information in the system. Therefore, when the distribution of data reaches the maximum entropy, its theoretical distribution will be the closest to the actual distribution state. Its application in species distribution simulation includes many features and advantages and is considered to be applicable to all known species distribution pattern types.



Where: hinge feature;

λn: weight coefficient

Linear Predictor Normalizer: a constant for numerical stability

Z: constant of proportionality

*3.5.5 Discriminant analysis (DA)*

Discriminant analysis, also called differential analysis (DA) is a division of group technology, will be known in advance the category of observation values, select samples of classification effect, using the class variables (grouping variable (g) as a response variable (y) of the following type, multiple measuring the difference between the variables (discriminant variable) as explained variables (xi) of the following type, set up the difference function (discriminant function), the formula is as follows:

y = b0 + b1X1 + b2X2 +... + biXi +... + bnXn

y is the discriminant score, Xi is the discriminant variable, and Bi is the discriminant coefficient or weight. The discriminant function is used to appropriately classify the newly observed value (Lin Zhenyan, 2018). This study uses SPSS statistical software to generate linear discriminant automatically by iteration.

**3.6 Model Performance Assessment**

Various prior studies assess model performance by more than one measure (Hernandez *et al.*, 2006; Pearson *et al.*, 2006; Williams *et al.*, 2009;). In this study, we used two measures that are illustrated in the following subsections.

*3.6.1 Cohen’s kappa*

Cohen’s *kappa* agreement coefficient is extremely important to assess the agreement between predicted map and reference test data set. The *kappa* coefficient compares the marginal and diagonal value in matrix fairly owing to the calculation containing not only true predication (true positive and true negative) but also false prediction (false positive and false negative)( Rosenfield and Fitzpatrick-Lins, 1986; Paine and Kiser, 2003 ;Congalton, 1991). The original function of Cohen’s *kappa* (Cohen, 1960) is:

$$ Kappa = \left(\frac{ Pr(a)- Pr(e)}{1-Pr(e)}\right)$$

Where

*Pr(a)* = relative observed agreement among raters,

*Pe(e)* = hypothetical probability of chance agreement.

The aforementioned equation can be rewritten according to Table 2 as the following equation.

$$Kappa=\left(\frac{\left(a + b + c + d\right)\left(a + d\right)– \left(b + d\right)\left(c + d\right)- \left(a + b\right)\left(a + c\right)}{\left(a + b + c + d\right)^{2}– \left(b + d\right)\left(c + d\right)- \left(a + b\right)\left(a + c\right)}\right)$$

Table 1. Schema of an error matrix for assessment of model performance.

|  |  |
| --- | --- |
| Predicted | Observed  |
| Presence | Absence |
| Presence | a | b |
| Absence | c | d |

*3.6.2* *Matthews’s correlation coefficient (MCC)*

MCC is the correlation coefficient of binary classification, with a value between -1 and 1. 1 means all predictions are correct, 0 means random blind guesses, and -1 means all predictions are wrong, which is suitable for evaluating the accuracy of unbalanced samples.

*3.6.3 F1-score*

F1-score is the harmonic average of precision and recall, and the range is 0-1. The higher the value is, the higher the proportion of correct prediction will be, and it is widely used in various fields. However, the index is easily changed by the exchange of positive and negative sample labels, and it is questioned that the number of TN samples is not considered conceptually (Chicco and Jurman, 2020).

Table 2. three indices evaluating the accuracy of the model

|  |  |  |
| --- | --- | --- |
| Item | Index | Formulation |
| (a) | *Kappa* | $$\frac{Total×\left(TP+TN\right)-[\left(W×Y\right)+\left(X×Z\right)]}{Total^{2}-[\left(W×Y\right)+\left(X×Z\right)]}$$ |
| (b) | MCC | $$\frac{TP×TN-FP×FN}{\sqrt{W×X×Y×Z}}$$ |
| (c) | F1 | $$\frac{2}{\frac{1}{precision}+\frac{1}{recall}}$$ |

note1：*precision* = *TP* / *W*； note 2：*recall* = *TP* /*Y*； note 3：*Total* = *Y*+*Z* = *W*+*X*

**4. RESULTS AND DISCUSSION**

We calculated the statistics of eight environmental factors corresponding to the entire study area and all of the JET samples, and compared the difference including resolutions in statistics between them, as shown in Table 3-5. From 5m resolution JET sample data (Table 3) the elevation range of JET samples (648–1,740 m) were within the suitable distribution range, from low elevation to 1,700 m above sea level. The means of slope statistics were 42°. The mean slope of all JET samples is obviously higher than that of the entire study area; consequently, this result is due to the trait of JET. JETs prefer to grow near ridges on the convex slope areas with unclosed canopy structure, where they are illuminated by abundant solar radiation. This trait could also be demonstrated by the mean value of terrain position statistics. In the past survey, it was found that JET was often found in mid-altitude areas, and the soil was mostly shallow and dry or there was a podzolic horizon in the soil layer, and the accompanying plant height in the areas where it appeared was relatively low. In table 3, it was unexpectedly the value of mean of GSR for JET samples is lower than entire study area. This result may affect variable selection before model building.

Table 3. Statistics data of JET samples(5m)

|  |  |  |
| --- | --- | --- |
|  | Study Area | JET Samples |
| min | max | mean | std | min | max | mean | std |
| Elevation | 515.5 | 2372.1 | 1300 | 396.7 | 647.5 | 1740.4 | 1254.7 | 389 |
| Slope | 0.78 | 71.7 | 39.88 | 12.95 | 21.6 | 59.77 | 41.96 | 8.76 |
| Aspect | 1 | 8 | 4.88 | 2.33 | 1 | 8 | 4.37 | 2.63 |
| SC | -27.8 | 45.24 | -0.288 | 12.77 | -55.72 | 47.32 | -0.86 | 12.71 |
| PRC | -29.12 | 25.94 | -0.157 | 6.83 | -26.87 | 31 | -0.44 | 6.96 |
| PLC | 0.003 | 51.5 | 0.134 | 8.05 | -19 | 35.93 | 0.41 | 7.15 |
| TP | 0 | 1 | 0.48 | 0.28 | 0.01 | 0.99 | 0.54 | 0.26 |
| GSR(10,000) | 49 | 196 | 133 | 30 | 56 | 190 | 124 | 34 |

SC: surface curvature, PRC: profile curvature, PLC: plan curvature, TP: Terrain Position, GSR: global solar radiation

Table 4. Statistics data of JET samples(20m)

|  |  |  |
| --- | --- | --- |
|  | Study Area | JET Samples |
| min | max | mean | std | min | max | mean | std |
| Elevation | 514 | 2377 | 1292 | 402.4 | 650 | 1745 | 1285 | 375.6 |
| Slope | 5.27 | 57.51 | 36.53 | 10.74 | 16.9 | 57.7 | 38.58 | 7.4 |
| Aspect | 1 | 8 | 4.78 | 2.2 | 1 | 8 | 4.6 | 2.71 |
| SC | -9.75 | 5.75 | 0.019 | 2.733 | -8.08 | 26.28 | 1.90 | 6.03 |
| PRC | -6.46 | 4.783 | 0.018 | 1.634 | -15.3 | 6.1 | -0.8 | 3.4 |
| PLC | -4.44 | 6.8 | -0.001 | 1.68 | -8.16 | 11.73 | 1.15 | 3.68 |
| TP | 0 | 0.99 | 0.48 | 0.28 | 0.01 | 0.99 | 0.55 | 0.26 |
| GSR(10,000) | 71 | 196 | 140 | 28 | 73 | 192 | 132 | 30 |

Table 5. Statistics data of JET samples(40m)

|  |  |  |
| --- | --- | --- |
|  | Study Area | JET Samples |
| min | max | mean | std | min | max | mean | std |
| Elevation | 496.5 | 2291 | 1313 | 368 | 650.75 | 1740 | 1289.9 | 369 |
| Slope | 5.3 | 54.6 | 35.5 | 10.1 | 12 | 52.7 | 37.2 | 6.7 |
| Aspect | 1 | 8 | 4.8 | 2.4 | 1 | 8 | 4.6 | 2.7 |
| SC | -5.45 | 3.11 | 0.034 | 1.45 | -3.84 | 3.08 | 0.12 | 1.36 |
| PRC | -3.15 | 2.73 | 0.026 | 0.98 | -2.51 | 2.63 | 0.02 | 0.97 |
| PLC | -2.24 | 2.3 | -0.01 | 0.66 | -1.75 | 1.34 | -0.1 | 0.57 |
| TP | 0.01 | 0.99 | 0.56 | 0.26 | 0.01 | 0.98 | 0.48 | 0.27 |
| GSR(10,000) | 82 | 184 | 142 | 24 | 80 | 193 | 136 | 28 |

From figure 2-4, it is shown under the observation of different resolutions that model is better in simple appearance. The larger the parameter value of depth, the more complex the model structure becomes, that is, the more classification rules. Therefore, from the point of view when the depth of the model is smaller, the more concise it is. It can be set at the turning point of the curve which considered to be the turning point when the curve rises to the level. It would be a more appropriate selection. For the result, we select 5, 22 and 4 with the depth, the samples split and the samples leaf. Before building SDM, variable selection is needed to make process effectively and reduce noise. In this study, we performed best variables combination of JET (table 7). It shows that the variable combinations with the highest predicted performance were selected from four models and three resolutions. We use three assessments to evaluate the highest predicted performance of variable combinations. The result is actually similar and reveal coherent. Elevation and Terrain Position appears in all fields. They are the two most important variables which effect a lot to the model performance. Aspect is one of the key components at RF. SC, PRC and PLC have quite crucial influence in high resolutions. GSR is relative important in low resolutions.



Figure 2. selecting depth with DT



Figure 3. selecting samples split with DT



Figure 4. selecting samples leaf with DT

Table 6. Important variables combination of JET

|  |
| --- |
|  5 m 20 m 40m  |
| DA | [E, PRC, TP] | [E, TP] | [E, TP, GSR] |
| DT | [E, SC, PLC, TP] | [E, PRC, TP, GSR] | [E, S, TP, GSR] |
| RF | [E, A, PLC, TP] | [E, S, SC, PRC, TP, GSR] | [E, A, TP] |
| Maxent | [E, SC, TP] | [E, SC, PRC, TP, GSR] | [E, S, TP, GSR] |  |

The rank of these four methods is (from high performance to low): RF, DT, Maxent and DA (table 10). In this study, RF is the best method to build SDM for JET especially in 5m. In F1-score, the accuracy of RF is close to DT and Maxent. Besides, model assesses by F1-score has the ideal number of other two measures in three resolutions.

As we mentioned before, the optimum method should keep the balance of performance and condensed habitat area. So far, it is very clear that RF approached this goal among these four methods. In applications of SDM, we recommended that RF is an optimal method. But DT and MAXENT still had robust performance, and they should be adopted for specific objects. For example, DT had the smallest habitat area which were relatively higher probability of presence. That implies that habitat area was predicted by DT are relatively optimal for new population discovering with specific species, as JET in this study. MAXENT had the largest habitat prediction but with reliable performance, therefore, MAXENT may be substantial for conservation management and reservation area delimitation.

Table 7. Result of model evaluation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test *Kappa* (%)** | **MCC (%)** | **F1(%)** |
| **DA** | **5m** | 42 | 42 | 72 |
| **20m** | 45 | 46 | 72 |
| **40m** | 36 | 36 | 68 |
| **DT** | **5m** | 59 | 60 | 79 |
| **20m** | 61 | 64 |  81 |
| **40m** | 56 | 56 | 77 |
| **RF** | **5m** | 72 | 72 | 86 |
| **20m** | 60 | 62 |  80 |
| **40m** | 56 | 56 | 78 |
| **MAX****ENT** | **5m** | 60 | 60 | 80 |
| **20m** | 60 | 63 | 82 |
| **40m** | 56 | 56 | 76 |

**5. CONCLUSION**

The four measures we adopted appeared constant trend, the best model is RF followed by DT, Maxent and DA. More importantly, the method related to DT were slightly better than MAXENT. Again, the result didn't imply that RF must be the best method for every application of SDM. But we proposed that except for DA other three methods should be adopted as candidates in other applications because all of these methods had quite robust performance. If SDM is used for specific application, we recommended DT is preferred for new population discovering, and MAXENT is more optimal for reservation area delimitation. On the other hand, the applications of SDM should use multi-measure approach for model elevation. If different measures appear similar trend, it is more objective for multi-method comparison.

Follow-up studies will attempt to extract spectral information associated with species from hyperspectral data and LIDAR DEM and use it as variables for model development so that models are applicable on a broader spatial scale. Environmental factors for SDM application are hard to be available at one time, especially that difficult to mapping (e.g. biotic factors). Therefore, SDM should be performed iteratively, and improve SDM steadily. This study is an elementary attempt to apply novel method in complex terrain and dense forest, such as that in Taiwan. This study used only indirect factors and build robust predictions. Therefore, we will seek other available direct and indirect factors (e.g. temperature, biotic interactions, soil type, etc) to improve our present result.

Finally, we proposed four directions for further iterative manipulation. (1) Keep field survey to add new samples of JETs. (2) Improve the quality of available factor layers to distill subtle information for SDM. (3) Adding new factors, such as temperature, biotic interactions, soil type, etc., to accurately describe distribution and niche of JETs. (4) Finally, on the base of quick improvement of statistics, we have to test new methods that haven't applied for SDMs in Taiwan, such as support vector machine (SVM), multivariate adaptive regression splines (MARS), convolutional neural network (CNN)and genetic algorithm for rule-set prediction (GARP).

**6. REFERENCE**

Anderson, R. P. and A. Raza, 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimate of niche evolution: preliminary test with montance rodents (genus Nephelomys) in Venezuela. Journal of Biography, 37, pp. 1378-1393.

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

Bourg, N. A., W. J. Mcshea and D. E. Gill, 2005. Putting a CART before search: successful habitat prediction for a rare forest herb. Ecology, 86 (10), pp. 2793-2804.

Breiman, L., Friedman, J. H., Olshen, R. A., and C. J. Stone, 1984. Classification and regression trees. Chapmna and Hall, New York.

Breiman, L., 2001. Random Forests. Machine Learning, 45 (1), pp. 5-32.

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettman, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S., Richardson, K., Schachetti Pereira, R., Schapire, R. E., Sober\_on, J., Williams, S. E., Wisz, M. and N. E. Zimmermann, 2006. Novel methods improve predictions of species' distributions from occurrence data. Ecography, 29, pp. 129-151.

Engler, R and A. Guisan, 2009. MigClim: predicting plant distribution and dispersal in a changing climate. Diversity and Distribution, 15, pp. 590-601.

Giam, X., Bradshaw, C. J.A., Tan, H. T.W. and N. S. Sodhi, 2010. Future habitat loss and the conservation of plant biodiversity. Biological Conservation, 143, pp. 1594-1602.

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S. and A. T. Peterson, 2007. What matters for predicting the occurrence of trees: techniques, data, or species characteristics? Ecological Monographs, 77 (4), pp. 615-630.

Hernandez, A. P., 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, pp. 773-785.

Hanberry, B. B., He, H. S. and D. C. Dey, 2012. Sample size and model comparison metrics for species distribution models. Ecological Modelling, 227, pp. 29-33.

Jensen, J. R. 2005. Introductory digital image processing: a remote sensing perspective (3rd ed.). Pearson Education, Inc., New Jersey.

Kumar, S. and T. J. Stohlgren, 2009. MAXENT modeling for predicting suitable habitat forthreatened and endangered tree Canacomyrica monticola in New Caledonia. Journal of Ecology and Natural Environment, 1 (4), pp. 94-98.

Phillips, S. J., R. P. Anderson and R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190, pp. 231–259.

Santos, M. J., Greenberg, J. A. and S. L. Ustin, 2010. Using hyperspectral remote sensing to detect and quantify southeastern pine senescence effects in red-cockaded woodpecker (Picoides borealis) habitat. Remote Sensing of Environment, 114, pp. 1242-1250.

Williams, J. N., Seo C., Thorne J., Nelson J. K., Erwin S., O’Brien J. M. and M. W. Schwartz, 2009. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions, 15, pp. 565-576.

Wise, M. S., Hijmans, J. L., Peterson, A. T., Graham, C. H., Guisan, A. and NCEAS predicting Species Distribution Working Group, 2008. Effect of sample size on the performance of species distribution models. Diversity and Distribution, 14, pp. 763-773

Nan C.L., Wei I.C., Kai Y.H., 2011. Application of 3S and Multivariate Statistics to Predict the Potential Habitat of Elaeocarpus Japonicus and Castanopsis carlesii. Quarterly Journal of Forest Research, 33(3)