ASSESS TREE SPECIES TOLERANCE LIMITS ON THE ACCURACY OF PREDICTIVE HABITAT MODELS

Chin-Jin Kuo (1), Bao-Hua Shao (2), Nan-Chang Lo (3), Kai-Yi Huang (4)

[1] Dept. of Forestry, Chung-Hsing University, 145 Xingda Road., Taichung 402, China Taipei
[2] Nan-Tou Division Office, Forest Bureau, Council of Agriculture, 456 Shihguan Rd., NanTou 542
[3] Experimental Forest Management Office, Chung-Hsing University, 145 Xingda Rd., Taichung 402
[4] Dept. of Forestry, Chung-Hsing University, 145 Xingda Road., Taichung 402
Email: kuoking37@gmail.com; baobao357@gmail.com; njl@nchu.edu.tw; kyhuang@dragon.nchu.edu.tw

Keywords: tolerance limit, long-leaf chinkapin (*Castanopsis carlesii*), random forest (RF), maximum entropy (MAXENT), decision tree (DT)

Abstract: Prediction of species' potential habitat distribution has become a major research issue in applied ecology. To obtain accurate predictive results, it is necessary to scrutinize ecological factors in a predictive habitat model. This study aimed to assess whether the tolerance limits of different tree species would affect the accuracy of models for predicting their potential habitat. Long-leaf chinkapins (LCC) grow widely over mountainous areas in central Taiwan, but there is a minimum tolerance limit in elevation above which this species can only grow over there, and furthermore the species usually occurs on the flat and broad ridges for sufficient sunlight. By contrast, Japanese Elaeocarpus (JE), which has no limits of tolerance on any ecological trait, can grow widely in mountainous areas from low to medium elevation. Hence, JE species has a scattered distribution. The study attempted to predict the potential habitat of LCC and JE species in the Huisun Experimental Forest Station (HEFS) in central Taiwan. It used geospatial information system (GIS) to integrate the datasets for two species and environmental factors, including elevation, slope, aspect, terrain position, and surface curvature (SC), profile curvature (PRC), plan curvature (PLC), and global solar radiation (GSR). DEMs of three grid sizes (5, 20, and 40 m) were used to derive these terrain-related variables. The models incorporating these terrain-related variables were developed using decision tree (DT), random forest (RF), maximum entropy (MAXENT), and discriminant analysis (DA) to predict their potential habitats. The results show that the *kappa* values of RF, DT and MAXENT models with LCC are greater than that of DA, and the same results are with JE. More importantly, the accuracies of the four SDMs with LCC are much better than those with JE. It means that LCC has a minimum tolerance limit on elevation (i.e. 1,700 m), which plays a key role on the spatial distribution of LCC, thereby substantially raising the accuracy of predictive habitat models. The opposite is true for widespread, disperse species like JE, which does not have any particular

trait with small ecological amplitude. In addition, LCC trees usually prefer to occur in wide ridges with gentle slope for enough sunlight, and thus terrain position and slope are also important variables for predictive habitat models. The point may explain why the prediction of the species with a small ecological amplitude in a certain factor or some factors is much easier and more accurate than that of widespread species. To improve the accuracy of predictive models for a widespread species like JE, the follow-up study will attempt to use DEM with high spatial resolution (1m) derived from LiDAR to generate terrain-related variables.

1. INTRODUCTION

Species distribution model (SDM) is a useful tool for evaluating potential species distribution (de Oliveira *et al*., 2020). Recently, there are plenty of new powerful statistical techniques and GIS tools, causing the development of SDM rapidly increased in ecology (Guisan and Zimmermann, 2000). Choosing the fittest modeling method and variables to increase the accuracy of prediction becomes more and more important.

Ecological amplitude (EA) is the capability of a species to establish in various habitat lying along an environmental gradient (Varghese and Menon, 1999). By determining the range of conditions under which it persists in nature, the EA of a species may be most effectively established (Packham and Willis, 1976). Every tree species has different EAs, if one can find out the main factors that affect species and the limits of tolerance of tree species, it will be helpful for modeling potential habitat distribution.

Castanopsis carlesii (long-leaf chinkapins, LCC) is native to Hainan, Guangdong, Guangxi, Fujian and Taiwan (Lu *et al*., 2017). It is widely distributed in the mountainous area of central Taiwan with an altitude of about 1700 – 2100 m. LCC is an intolerant species, mostly distributed on wide flat ridges and on the platforms on both sides of the ridges. It is difficult to find this species from mountainside to valleys because these places are usually dark and humid. Even if it appears, its tree shape is far inferior to those growing on the ridge (Lo *et al*.,2008). *Elaeocarpus japonicus* (Japanese Elaeocarpus, JE) is distributed in forests from low-elevation to up to 2,000 meters across the Taiwan island. It is often mixed in broad-leaved forests in a scattered state. Lo (1992) found out that JE species is frequently distributed in shallow soil, direct sunlight and dry places.

The purpose of our research is to confirm whether the limits of tolerance of the species affect the accuracy of the model through establishing and comparing the species distribution models of LCC and JE. The research used digital elevation model (DEM) in three grid sizes (5, 20, and 40 m), which were acquired from the Satellite Survey Center, Dept of Land Administration, M.O.I. By using GIS software package, it generated the data layers for environmental factors,

including elevation, slope, aspect, terrain position (TP), surface curvature (SC), profile curvature (PRC), plan curvature (PLC), and global solar radiation (GSR) from DEM. Four models, decision tree (DT), random forest (RF), MAXENT, and discriminant analysis (DA), were developed to predict the potential habitat of the species.

2. STUDY AREA

The study area is located in the Huisun Experimental Forest Station (HEFS) in central Taiwan. Its geographic coordinates approximately fall within 121°1'–121°8' east longitude and 24°2'–24°6' north latitude. The total area of the station is 7,477 ha and divided into 19 forest classes. Altitude ranges from 454 m to 2,418 m, including the ecological environment from low altitude to medium-high altitude. There are about 1,100 kinds of plants in HEFS, which is a representative forest in central Taiwan (Lo *et al*., 2011).

3. MATERIALS AND METHODS

3.1 Field Data

The samples of LCC and JE are based on the data accumulated by the previous long-term ground surveys in our laboratory. These ground surveys were conducted to use Trimble Pro XR GPS with a 5 m telescopic extension antenna and a laser range finder.

There were 120 LCC and 224 JE samples collected in this study and the background samples are randomly sampled from other areas except the target sample point. The ratio of target sample to background sample is 1:1. This study compared the influence of DEM with different spatial resolutions on model performance. It used 5 m, 20 m, and 40 m grids as the unit to merge samples, which means samples that fall on the same cell were merged into a point. The sample size of each resolution is shown in Table 1, and sample ratio for model development and model validation is 7:3.

Table 1. Sample size of target tree species

Resolution	Long-leaf chinkapins	Japanese Elaeocarpus
5m	111	200
20m	82	132
40m	61	97

3.2 Environmental factors

In this study, ten environmental factors were selected. The layers for elevation, slope, aspect, surface curvature (SC), profile curvature (PRC), and plan curvature (PLC) were generated from DEM by ArcGIS software package. Global solar radiation (GSR) layer was calculated as the sum of direct and diffuse radiation. We used ArcGIS to output direct radiation raster and diffuse radiation raster, then calculated them together to get GSR. The calculation of terrain position (TP) is much more complicated than others. First, we digitized ridges and valleys from the DEMs, then calculated the Euclidean distance from each cell to the nearest ridge and valley line, and last determined the relative position ratio, the formula is shown as follows:

$$P_{ij} = \frac{PV}{(PV + PR)} \qquad (1)$$

where PV = Euclidean distance from point P to the nearest valley line
PR = Euclidean distance from point P to the nearest ridgeline
P = a validation point (cell)
P_{ij} = the relative position ratio of j row and i column

In this study, the ridgeline is the highest slope position, represented by "1", while the valley line is the lowest slope position, represented by "0". Also, we divided TP into three categories based on the complexity of the ridgeline and valley line. While mapping, ridges or valleys which seems larger will be classified in TP1. TP2 will add secondary ridgelines and valley lines in the map. The most intricate one is TP3, it contains almost every ridges and valleys which can be identified in the DEMs.

3.3 Model Development

This research used the machine learning module written in python programming language to develop SDMs, applying scikit-learn to create discriminant analysis (DA), decision tree (DT), and random forest (RF). To maximize the performance of the model, hyperparameter tuning played a key role. The hyperparameters of each model in this study, DA is the prior probability, both DT and RF are child-node, parent-node, and depth. In addition, this study used the Python programming language to construct a loop to exhaust the possibilities of all parameter combinations, and import the DA, RF and DT models one by one to find out which combination has the best *kappa* value.

3.3.1 Discriminant Analysis: Discriminant analysis is a technique to distinguish the differences between groups (Chen *et al.*, 2013). It has been used widely in many applications

(Ye *et al*., 2005). It selects the observed values of known categories in advance and choose samples which has classification effect, use grouping variable as the reaction variable, multiple measured discriminant variables as the explanatory variable to establish the discriminant function (Lo *et al*., 2011), the formula is as follows:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_i x_i + \cdots + b_n x_n \quad (2)$$

where y = discriminant score

x_i = discriminative variable

b_i = discriminant coefficient

Most of the hyperparameters of DA are preset by the scikit-learn module. The only hyperparametersp we tuned is prior probability, and set it as 0.5.

3.3.2 Random Forest: Random Forest (RF) is an ensemble classifier, which generates predictions by building multiple decision trees, and obtains the final prediction results through a majority vote (Shao, 2020). In RF algorithm, many decision trees are randomly created with "boot-strap samples" from the data set, and the final estimate is the average of all results from each tree (Yeşilkanat, 2020). The hyperparameters we choose and their setting values are the same as DT. The number of decision trees (n_estimators) is set to 500, and when cutting variables, the number of potential variable list (max_features) randomly sampled is also generally recommended to be set to the root of all variables.

3.3.3 Decision Tree: The principle of classification tree is to establish a dichotomy classification rule of one-to-multi-layer tree structure for the input original data, according to this rule to predict the unknown result data. The nodes are the points in a tree where a test is done on the attribute, and branches are test result that leads to another node (Vanfretti and Arava, 2020). There are three kinds of nodes: root node, internal node and leaf node. The root node is on the top, internal nodes are in-between and leaf nodes are assigned a final outcome based on group membership of the majority of observations.

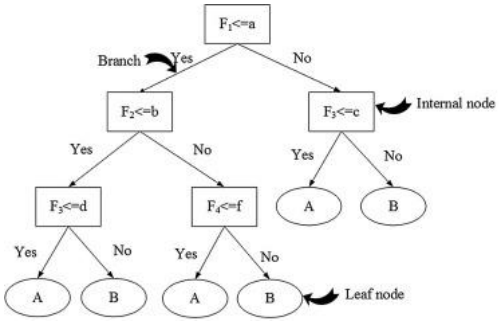

Figure 1. The diagram of the classified process of DT (Lan *et al.*, 2020)

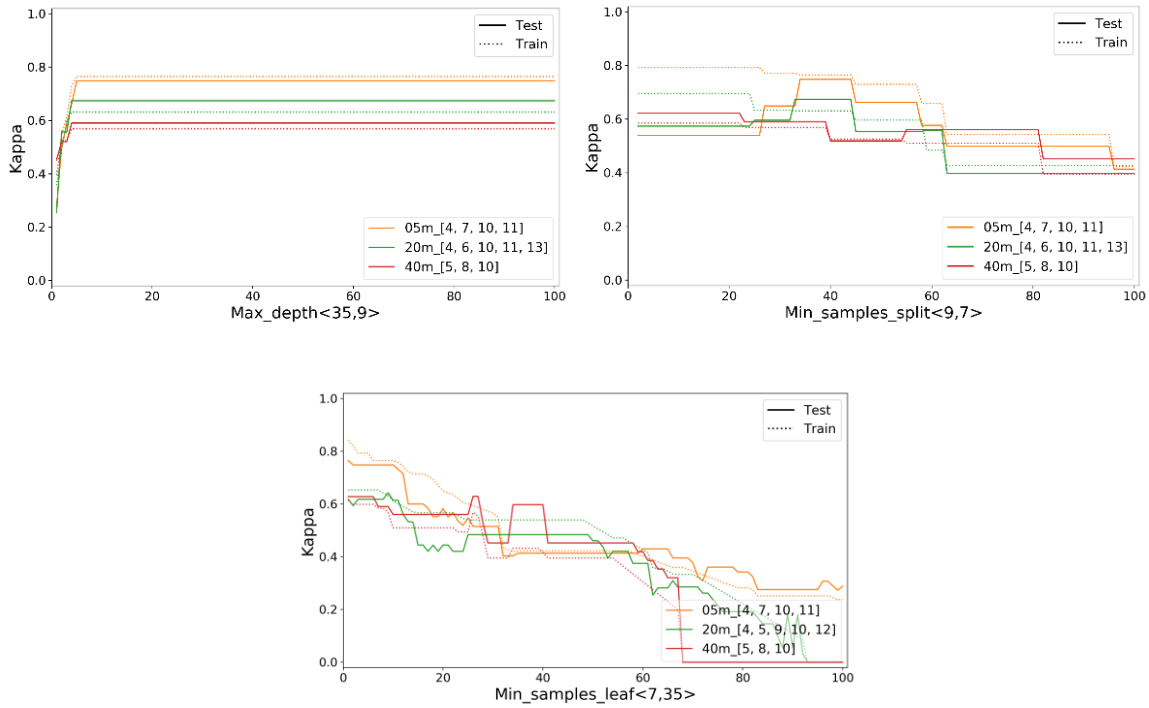

Figure.2 DT maximum depth**, parent node, and child node *kappa* value curve of Japanese Elaeocarpus** *each number present an environmental factor

3.3.4 Maximum Entropy: MAXENT software is freely available on the worldwide web (https://www.gbif.org/zh-tw/tool/81279/maxent). It is one of the most robust and advanced modeling approaches for presence-only data (Qin *et al.*, 2020). This method employs the maximum entropy algorithm and species occurrence to predict habitat distribution (Abolmaali *et al.*, 2018). It requires only presence data, and environmental information of the study area. Therefore, when there are not many training samples, it has an advantage (Phillips *et al.*, 2006).

3.4 Model Validation

The purpose of model validation is to analyze the prediction results and understand the reliability of the model. Accuracy assessment contains the *kappa* coefficient and Matthews correlation coefficient (MCC). The range of *kappa* value is -1 to 1, but it usually falls between 0 and 1. The higher the *kappa* value was, the more consistent it will be (Wang, 2012). MCC is the correlation coefficient of binary classification, the value is between -1 and 1, 1 means all predictions are correct, 0 means not better than random predictions, -1 means all predictions are wrong (Shao, 2020).

4. **RESULTS AND** DISCUSSION

According to the result of the optimal combination of parameter, we found out that elevation,

slope, and TP2 are the most important environmental factors that affect the performance of SDMs for two species . Table 2, 3, and 4 are the statistics of the environmental factors of two species and study area, it shows that the elevation range of JE was much greater than that of LCC, that were about 1630–2100 m and 650–1740 m, respectively. The average slope of JE was also steeper than that of LCC, and both TP1, TP2 shows that most LCC survive at ridges while JE can adapt to valleys or ridges but prefer to live on mountainside. For LCC, the average of SC and PRC pointed out that this species usually grows at convex surfaces, while the curvature of JE didn't provide useful information. The experience of the field survey in the past pointed out that LCC could only grow in elevation above about 1700 m (Lo *et al.*, 2011). It means that there are significant limits in elevation for LCC. As for JE, the limiting factors are fewer than those of LCC. According to table 3 and 4, we found out that the growth area of JE has no obvious ecological characteristics, its distribution is about 647 – 1,750 m, which almost covers the altitude range of the entire study area. This means that it has no distinct lower limit of tolerance at elevation, and the upper limit of tolerance is about 2,000 m, but still need to consider the latitude with more detailed investigation and analysis.

Table 2. The statistics of long-leaf chinkapins in study area (5m)

statistics	Elevation (m)	Aspect (°)	Slope (°)	SC (m^{-1})	PLC (m^{-1})	PRC (m^{-1})	TP1	TP2	TP3	GSR (Wh/m^2)
mean	1893	185	15	0.8	0.3	-0.4	1.0	1.0	0.3	1789310
min	1533	1	2	-24.7	-14.4	-10.8	0.4	0.8	0.2	1135050
max	2081	358	41	19.0	8.7	14.5	1.0	1.0	1.0	1996670

Table 3. The statistics of Japanese Elaeocarpus in study area (5m)

statistics	Elevation (m)	Aspect (°)	Slope (°)	SC (m^{-1})	PLC (m^{-1})	PRC (m^{-1})	TP1	TP2	TP3	GSR (Wh/m^2)
mean	1255	193	42	-0.9	-0.4	0.4	0.5	0.5	0.1	1244801
min	648	0.6	22	-56	-27	-19	0.0	0.0	0	560313
max	1740	359	60	47	31	36	1.0	1.0	1	1898924

Table 4. The statistics of study area (5m)

statistics	Elevation (m)	Aspect (°)	Slope (°)	SC (m^{-1})	PLC (m^{-1})	PRC (m^{-1})	TP1	TP2	TP3	GSR (Wh/m^2)
mean	1290	192	39	-0.4	-0.1	0.3	0.4	0.5	0.1	1348391
min	515	1	1	-75.4	-32.9	-29.1	0.0	0.0	0.0	491170
max	2372	360	72	45.2	25.9	51.5	1.0	1.0	1.0	1955534

The Python programming language had already found out which combination has the best *kappa*

values, we chose parameters which appears frequency and import to the DA, RF, MAXENT, and DT models. Table 5 shows the accuracies of models for the two species. SDMs that used 5 m DEM usually have higher accuracies then 20 m or 40 m. The possible reason is related to the reduction of the number of samples due to merge, resulting the stability of the model decrease. However, in JE's RF model, the accuracy of 40 m is better than that of 20 m, the reason for this result is still unclear. For JE the rank of these four method was (from high performance to low): RF, DT, MAXENT, and DA. In this study RF and DT both are suitable methods to build SDM. As for LCC, RF and DT also performed well, much better than MAXENT and DA. After comparing the models of the two species, it is no doubt that the model performance of LCC is significantly better than JE. The reason may be that JE is a species with a broad EA, causing it hard to correctly distinguish JE from background.

Table 5 model performance for predicting the potential habitats of two tree species

Model	Japanese Elaeocarpus						Long-leaf chinkapins					
	kappa			OA (%)			*kappa*			OA (%)		
	5m	20m	40m	5m	20m	40m	5m	20m	40m	5m	20m	40m
DA	0.50	0.56	0.46	75	78	73	0.95	0.91	0.89	97	94	93
DT	0.75	0.67	0.59	88	84	80	0.97	0.92	0.95	97	96	97
RF	0.73	0.58	0.70	87	79	85	0.98	0.96	0.95	99	98	97
MAXENT	0.65	0.57	0.51	88	78	75	0.95	0.92	0.90	97	95	94

5. CONCLUSIONS

The four methods we chose in this study, DT and RF both have higher accuracies than the remaining do. It points out that these two methods are fitted for predicting the habitat of JE and LCC, while MAXENT and DA only suitable for LCC.

Generally speaking, rare species have more limits factors than widespread species, grasping these key factors can improve the accuracy of SDM, this is why the former is relatively easier to simulate the spatial distribution of species. However, LCC isn't a rare species and is even widely distributed in the mountainous area. The reason why the accuracies of LCC's SDM is much higher than JE is due to the tolerance limits such as elevation and slope, which means that even if it is a widely distributed species, when discovered its growth limiting factors, we still can accurately predict its potential habitat, as in the case of this study.

Our study didn't find out the main factor which limits the growth of JE due to its broad EA. There are some ways to improve this situation, such as adding new factors (e.g. temperature, humidity, biotic interactions, soil type or other terrain-related variables), keep field survey to add new samples of JE, use DEM with high spatial resolution (1m) derived from LiDAR, and

test deep learning methods like convolutional neural network (CNN).

6. REFERENCES

Abolmaali, S. M. R., M. Tarkesh, and H. Bashari, 2018. MaxEnt modeling for predicting suitable habitats and identifying the effects of climate change on a threatened species, Daphne mucronata, in central Iran. Ecological Informatics, 43, pp. 116-123.

Chen, H. C., 2013. Evaluation of the Factors Affecting Predictive Performance of Species Distribution Models. Dept. of Forestry, Chung-Hsing University, master thesis.

de Oliveira, K., T. Araújo, A. Rotti, D. Mothé, F. Rivals, and L. S. Avilla ,2020. Fantastic beasts and what they ate: Revealing feeding habits and ecological niche of late Quaternary Macraucheniidae from South America. Quaternary Science Reviews, 231, pp. 106178.

Guisan, A., and N. E. Zimmermann, 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135 (2-3), pp. 147-186.

Lan, T., H. Hu, C. Jiang, G. Yang, and Z. Zhao, 2020. A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. Advances in Space Research, 65 (8), pp. 2052-2061.

Lo, N. C., 1992. Analysis on the vegetation of Tung-Feng-Hsi watershed in Hui-Sun experimental forest station. Dept. of Forestry, Chung-Hsing University, master thesis.

Lo, N. C., H. C. Shu, K. Y. Huang, 2008. Application of GIS and Logistic Multiple Regression (LMR) to Predict the Potential Habitat of Castanopsis carlesii. Quarterly Journal of Forest Research, 30 (1), pp. 29-43.

Lo, N. C., W. I. Chang, K. Y. Huang, 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), pp. 55-70.

Lu, F.Y., C. H. Ou, Y. H. Tseng, C. M. Wang, 2017. Trees of Taiwan. Chinese Yi-Chih Forest Plant Research Association, pp. 364.

Qin, A., K. Jin, M. E. Batsaikhan, J. Nyamjav, G. Li, J. Li, Y. Xue, G. Sun, L. Wu, T. Indree, Z. Shi, and W. Xiao, 2020. Predicting the current and future suitable habitats of the main dietary plants of the Gobi Bear using MaxEnt modeling. Global Ecology and Conservation 22: e01032.

Packham, J. R., and A. J. Willis, 1976. Aspects of the Ecological Amplitude of Two Woodland Herbs, Oxalis Acetosella L. and Galeobdolon Luteum Huds. The Journal of Ecology, 64 (2), pp. 485.

Phillips, S. J., R. P. Anderson, and R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190 (3), pp. 231-259.

Vanfretti, L., and V. S. N. Arava, 2020. Decision tree-based classification of multiple operating conditions for power system voltage stability assessment. International Journal of Electrical Power & Energy Systems 123, pp. 106251.

Varghese, A. O., and A. R. R. Menon, 1999. Ecological niches and amplitudes of rare, threatened and endemic trees of Peppara Wildlife Sanctuary. Current Science, 76 (9), pp. 1204-1208.

Wang, W. C., 2012 Using species distribution models in GIS to predict the spatial pattern of rare and endangered plant—Brainea insigni. Dept. of Forestry, Chung-Hsing University, master thesis.

Ye, J., R. Janardan, and Q. Li, 2005. Two-Dimensional Linear Discriminant Analysis. L. K. Saul, Y. Weiss, & L. Bottou (Eds.), Advances in Neural Information Processing Systems, 17, pp. 1569–1576.

Shao, B. H., 2020. The impact of the terrain-shelterbelt effects on species distribution modeling for tree species at various altitude. Dept. of Forestry, Chung-Hsing University, master thesis.

Yeşilkanat, C. M., 2020. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals, 140, pp. 110210.