Proceedings of International Conference on Biodiversity: IBD2019 (2019); 156 - 162

Exploring the suitability map of wild banana (*Musa serpentina* Swangpol & Somana) in Thailand using species distribution models with the limited occurrence data

Thanayut Changruenngam and Jantrararuk Tovaranonte*
School of Science, Mae Fah Luang University, Muang District, Chiang Rai, Thailand
*Corresponding author e-mail: jantrararuk@mfu.ac.th

Abstract: Species distribution model (SDMs) is one of the powerful tools to predict the suitability map to address the ecology and conservation approaches. However, the limited number of occurrence data has been a problem in model performance. Here, the three algorithms (MaxEnt, generalized linear model: GLM, and random Forest: RF) were selected to project the suitable map with eight occurrences of *Musa serpentina* Swangpol & Somana wild banana which endemic to the west of northern and central Thailand and eleven environmental variables. Due to a limitation of occurrence data, the fuzzy logic had been applied to GLM and RF model to enhance the occurrence data. MaxEnt is presence-only data, it could not use the fuzzy logic to create pseudo absence data. The results showed that three climatic variables had affected all three models especially precipitation of warmest quarter (BIO8) and precipitation of coldest quarter (BI019). All algorithms could predict the suitable map well with high AUC values (> 0.9). GLM had the highest performance with AUC value of 0.991.

Keywords: MaxEnt, Random forest, GLM, Fuzzy logic, *Musa serpentine*

Introduction

Species distribution models (SDMs) play an important role for conservation of endangered species. It can predict future results under climate change and generate species distribution models under geospatially explicit layers of abiotic or biotic data which defines the ecological requirements of species under study (Franklin, 2010). SDMs consist of 2 categories according to the characteristics of data. First, presence data-only algorithms such as genetic algorithm for ruleset prediction (GARP) and maximum entropy (MaxEnt) (Phillips, 2008). MaxEnt is commonly shown accurate prediction capabilities of these models. Second, binary presence/absence data algorithms such as classical modelling approach: the generalized linear model (GLM) and modern method: random forest (RF) and boosted regression trees (BRT). From previous studies, consideration of the actual values of the predictions which emphasize more clearly found that MaxEnt, BRT and GLM performed well, followed by RF then GARP (Elith & Graham, 2009). The applications of SDMs are widely available, such as predicting distribution of disease vectors. (Khatchikian et al., 2011). The other applications include conservation of the suitable habitat of species (Tovaranonte et al., 2015) and the evaluation of species vulnerability to climate change (Trisurat, Shrestha, & Kjelgren, 2011).

Musa serpentina Swangpol & Somana ('Nakkharat' or 'Naga') is a rare wild banana in *Musaceae*. It is endemic to the west of northern and central Thailand such as Tak, Mae Hong Son and Kanchanaburi Provinces. It commonly grows in the open mixed deciduous forest by stream banks or low slopes by roadsides. The altitude of habitat is approximately 240-570 meters above sea level. Due to a small seed set and drastic fragmentation, its conservation status is considered as endangered species based on IUCN Red List Categories and Criteria: Version 3.1 (Criteria D in Section V; IUCN, 2001)(Swangpol & Somana, 2011).

SDMs should provide conservation practitioners with the estimated spatial distributions of species requiring attention. These species are often rare and have a small sample size, posing challenges for creating accurate species distribution models. The studies indicated that the number of known occurrences had greatly affected the accuracy of SDMs whereas MaxEnt always had the highest accurate predictive ability (Hernandez, Graham, Master, & Albert, 2006). Wisz et al. (2008) compared AUC form 14 SDMs and found MaxEnt and GLM had higher performance. GLM gave the highest AUC compared to RF and MaxEnt in different species distribution (Elith & Graham, 2009). However, RF algorithm can also generate better species predictive distributions for a habitat prediction of small sample size. while MaxEnt had lower performance (Mi, Huettmann, Guo, Han, & Wen, 2017). Since the sample size of *Musa serpentina* is small. Hence, the suitability map of Musa serpentine should be predicted by MaxEnt, RF and GLM algorithm**.**

GLM and RF are algorithms for binary presence/absence data. The methods of pseudo-absence selection influence a performance of species distribution models which is binary presence/absence data (Mary S. Wisz & Guisan, 2009). The pseudo-absence data was established by using opposite condition with environment of *Musa serpentina*. Generally, the boundaries of suitable environment cannot be clearly defined. Therefore fuzzy logic is a tool for increasing flexibility of boundaries of suitable environment. It is used to handle the concept of partial truth where the true values may range between completely true and completely false that is these values of variables

are any real numbers in closed interval 0 to 1 (0 = false, 1= truth). These are a feature of the membership function of fuzzy set.

Here, this study aimed to compare the predicted performance between GLM and RF which developed by Fuzzy pseudo-absence data with MaxEnt for the small sample size condition. And there is establish the suitability map for the conservation of the habitat of *Musa serpentina*.

Materials and Methods

Study area and the occurrence data

Thailand is located between 5°45′ and 20°30′N and 97°30′ and 105°45′E. The total area is 511,731 square kilometers or around 200,000 square miles. There are eight localities of *Musa serpentina* were obtained from published data (Swangpol & Somana, 2011), herbarium data and unpublished data

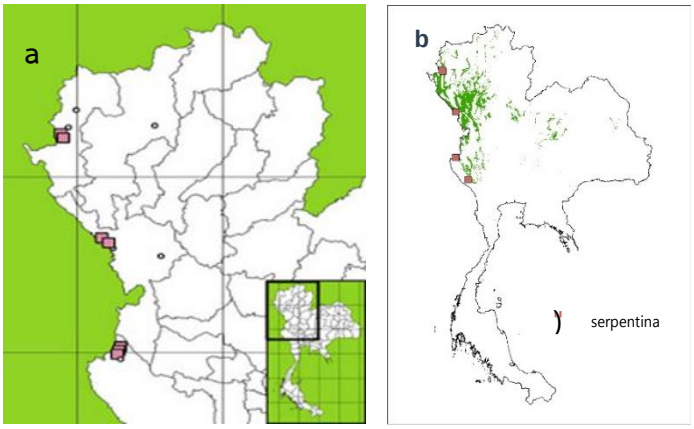

Figure 1. (a). Distribution map of *Musa serpentina* in the west side of northern and central regions of Thailand (Swangpol & Somana, 2011) and **(**b). suitability map of *Musa serpentina* overlaid by the area of Band 4 and 5 of three models

Environmental variables

Twenty environmental variables were used at 30 seconds ($1km^2$) resolution. Twenty climate variables were downloaded from the WorldClim website (http://www.worldclim.org, 11th March 2017, Hijmans et al., 2005). These variables are BIO1 (annual mean temperature), BIO2 (mean diurnal range), BIO3 (Isothermality), BIO4 (temperature seasonality), BIO5(Max temperature of warmest month), BIO6 (min temperature of coldest Month), BIO7 (temperature annual range), BIO8 (mean temperature of wettest quarter), BIO9 (mean temperature of driest quarter), BIO10 (mean temperature of warmest quarter), BIO11 (mean temperature of coldest quarter), BIO12 (annual precipitation) BIO13 (precipitation of wettest month), BIO14 (precipitation of driest month), BIO15 (precipitation seasonality), BIO16 (precipitation of wettest quarter), BIO17 (precipitation of driest quarter), BIO18 (precipitation of warmest quarter), BIO19 (precipitation of coldest quarter) and Alt (altitude).

Fuzzy Logic

Fuzzy Logic is a computer system that works by fuzzy set invented by Zadeh (Zadeh, 1965). It is used to handle the concept of partial truth where the truth values may range between completely true and completely false (0 = false, 1= truth). These are a feature of the membership function of fuzzy set. It was created by the degree of truth as an extension of valuation. The membership functions were used as input and output variables of the rule-based system (IF-THEN rules) derived from knowledge. Fuzzy logic was used to create pseudo-absence data for identifying suitable habitat of *Musa serpentine.* After that, areas with probability level of less 0.6 as area for 1,000 randomly sampled was used as pseudo-absence data or Fuzzy pseudo-absence data.

Fuzzy pseudo-absence data

A suitable habitat of *Musa serpentina* is in highly disturbed habitats of open mixed deciduous forest by stream banks or low slopes by roadsides and altitude 240–570 meters (Swangpol & Somana, 2011). The open mixed deciduous forest often found at height of 50-800 meters, rainfall range from 1,200-1,600 mm per year and not found in the South of Thailand (Marod, Pinyo, Duengkae, & Hiroshi, 2010). The presence data and absence data were defined as 1 and 0 respectively. A membership function of fuzzy logic was created with Triangular function and the rule-based system as follows.

Input

Altitude (0 -1): low (less 300 m), appropriate (200 - 600 m), high (more than 500)

Rainfall (0-1): low (less 1300 mm), appropriate (1,100 - 1,700 mm), high (more than 1500 mm)

Output

Minimum of the Fuzzy membership from the input (0 - 1)

Rule-based system

1. If altitude = low then 0
2. If altitude = High and then 0
3. If altitude = appropriate and rainfall = low then 1
4. If altitude = appropriate and rainfall = appropriate then 1
5. If altitude = appropriate and rainfall = high then 0

The map outputs were overlaid type "and" are shown in Figure 2. We used areas with a level of less 0.6 as an area for 1,000 randomly sampled to use as pseudo-absence data (Morgane, Frederic, Jiguet, Cecile Helene, & Wilfried, 2013).

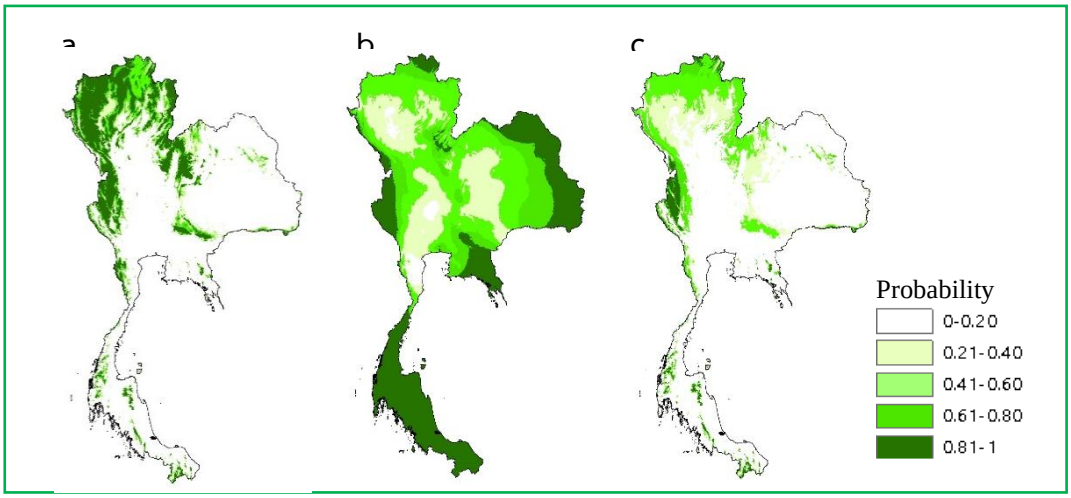

Figure 2. The fuzzy map in a fuzzy logic system to create pseudo-absence data from (a) altitude data, (b) rainfall data and (c) The overlaid map

Modelling and evaluation methods

The program named MaxEnt version. 3.2.1 was used to project eight samples and 1000 background samples were created randomly. The function random forest was used to build an ensemble of classification trees. The model generated 500 trees from both binary presence data and Fuzzy pseudo-absence data. We used function GLM to create all possible subsets of models with the options for each variable from presence data and Fuzzy pseudo-absence data. Area Under the ROC Curve (AUC) was used to evaluate the performance of the models (Liu, White, & Newell, 2011). The AUC is the current best practice for evaluating the success of a model for binary presence/absence data (Rushton et.al, 2004). Therefore, the value of AUC is an efficient tool for testing the performance of SDMs. For example, in 2013, the value of AUC was used to evaluate the distribution model of the particle to find the most effective model (Tovaranonte et al., 2015). The criteria for the AUC values that are frequently seen are 0.90-1.00: excellent, 0.80-0.90: good, 0.70-0.80: fair, 0.60-0.70: bad and 0.50-0.60: failed (Swets, 2015). In addition, each model different in prediction which can be observed from the Pearson product-moment correlation coefficient and percentage of duplicate area. Summarizing the modelling and evaluation methods are shown in Figure 3.

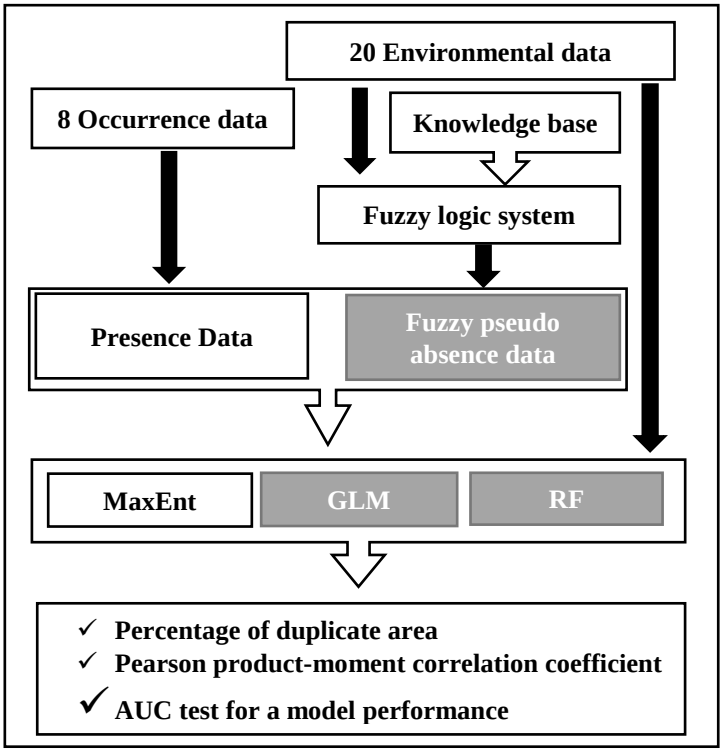

Figure 3. Flow chart summarizing the modelling and evaluation methods.

Result & Discussion

The suitability map of Musa serpentine

The evaluation of habitat prediction from the sample areas revealed the generalizability of the model predictions. This method was used to evaluate that occurrence data matched prophecy of the distribution area, as a spatial assessment of model performance. It is a visual method to show the transferability of models from sampled to interesting areas. The suitability map for *Musa serpentina* in Thailand was developed in three models (MaxEnt, GLM and RF) as shown in Figure 4. Those maps were classified by the criteria for the probability values are 0 - 0.20: Band 1, 0.21-0.40: Band 2, 0.41-0.60: Band 3, 0.61-0.80: Band 4, and 0.81 - 1: Band 5. Each Band was shown in green according to the intensity of the shades. Hence, the dark green color indicated that the most suitable locations were in the western part of Thailand.

The area predicted by MaxEnt model most appeared in Band 3 about 41% of Thailand. However, for Band 4 and 5 where probability more than 0.6 had the sum area about 8%, spread in the west and the lower North of Thailand. GLM model provided the most area predicted at Band 1 where 64% of Thailand after that the area continued to decrease to 1% in Band 5. For RF model had no area of prediction in Band 1 but it had the most at Band 2 (about 48%) and had the predicted area in Band 4 and 5 higher than the area predicted by MaxEnt and GLM. Results of the predicted area of each model are shown in Table 1.

Band 4 and 5 of each model represented to the habitat of *Musa serpentina* and had a total area of 172,988 km^2. While the overlapping areas of three models had 30,942 km^2 are shown in the green area (Figure 1b). The three suitability maps predicted by MaxEnt, GLM and RF had high performance. However, the overlapping of these maps represented the suitability map of *Musa serpentina* did likewise. Hence, the overlay maps cover all 8 samples are shown in Figure 1a. Therefore, this area is the suitability map that we recommend to be the conservation of the habitat for *Musa serpentina*. Band 5 represented more specific residential of *Musa serpentina* where the suitability maps from the three models had a total area of 46,281 km^2. However, the area with the highest probability presence was developed by overlapping of Band 5 in each model. It had 450 km^2 spread in the North and West of Thailand especially in Mae Hong Son Province (Figure 5d).

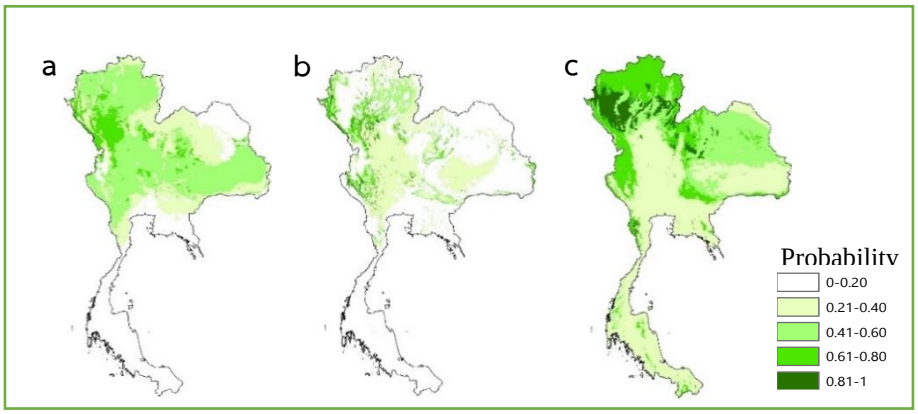

Figure 4. Suitability map of *Musa serpentina* in Thailand created from (a) MAXENT, (b) Random Forest and (c) GLM.

Species distribution models

MaxEnt, GLM and RF had high performance for creating the suitability map of *Musa serpentina*. Those maps can provide conservation practitioners to estimate the accurate spatial distributions of *Musa serpentina*. However, each algorithm had estimated in difference size of the area in the map. Therefore, the percentage of duplicate area, Pearson product-moment correlation coefficient and AUC from each model were compared.

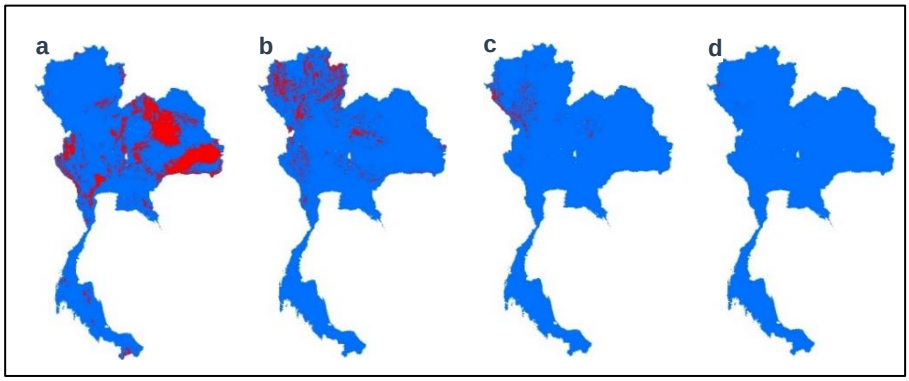

Figure 5. The same area predicted from MaxEnt, RF and GLM (a) Band 2, (b) Band 3**,** (c) Band 4, (d) Band 5 Percentage of the duplicate area represented the prediction area of each algorithm in the same band. However, the predicted size of areas from three models were different. Band 1 had no duplicate area as the RF's map had a probability of not less than 0.2. Range 0.21 - 0.40 (Band 2) had duplicated area for 52% which is close to GLM's predicted area (Figure 5a) and Band 3 had 27% of duplicated area (Figure 5b). Band 4 and Band 5 were used to define the suitability habitat because of the probability > 0.6. However, the predicted area of each model is different. Therefore, predicted for the same area of Band 4 found 13 % (Figure 5c). Band 5 is the model's predicted area that has the highest probability (>0.80), represented to the suitability map of *Musa serpentina* where all models shown a duplicate area of 9 %.

Table 1. Percentage of predicting area by MAXENT, GLM, RF of Musa serpentine in Thailand.

Band	Probability	Model's predicted area of *Musa serpentine* in Thailand (%)			Percentage of duplicate area (%)
		MAXENT	GLM	RF	
1	0.00 – 0.20	28	64	0	0
2	0.21 – 0.40	24	24	48	52
3	0.41 – 0.60	41	6	25	27
4	0.61 – 0.80	7	4	21	13
5	0.81 – 1.00	0.06	1	7	9

The r-value of MaxEnt and GLM had a strong positive relationship (0.564) but the relationship between RF with the remaining two models had a moderate positive relationship (Table 2). However, the reliability of the correlation coefficient depends on hypothesis test of the significance of the correlation coefficient where H_0: $\rho = 0$, H_1: $\rho \neq 0$ represented no significant correlation exists and a significant correlation exists. Testing the significance of the correlation coefficient of every pair of model's predicted maps had p-value ≤ 0.001 (Table 2). Therefore, there is sufficient evidence to conclude that there is a significant linear positive relationship between

these prediction maps. In other words, MaxEnt GLM and RF established the suitability map of *Musa serpentina* are in the same direction.

Table 2. the Correlation Coefficient of Model's predicted maps.

Model	MAXENT		GLM		RF	
	r	p-value	r	p-value	r	p-value
MAXENT	1					
GLM	0.664	0.00*	1			
RF	0.446	0.00*	0.478	0.00*	1	

*P-Value < 0.05

The Area under ROC curve (AUC) was usually used to evaluate model accuracy. The prediction performance of three models were similar. The three models had relatively high AUC values. Our results showed that the AUC values from GLM were highest (0.991), followed by RF (0.989) and MaxEnt (0.931), respectively. Three variables included precipitation of warmest quarter, mean temperature of wettest quarter and precipitation of coldest quarter had affected MaxEnt models especially precipitation and maximum temperature of warmest month, precipitation of warmest quarter and precipitation seasonality which were the top three of the variables affected GLM model. For RF model was the most affected from variables precipitation of wettest month, precipitation of coldest quarter and precipitation seasonality. It can be seen that precipitation of warmest quarter (BIO18) and precipitation of coldest quarter (BIO19) were the two variables that had affected all models. In addition, both variables represented characteristic of the habitat of *Musa serpentina* with open mixed deciduous forest.

Conclusions

The prediction of the suitability map of *Musa serpentina* found that the most suitable locations were in the North and West part of Thailand. However, the areas with the highest probability of presenting were in Mae Hong Son Province cover the area of 450 square kilometers

MaxEnt is acceptable as a high-performance model for the prediction of the species distribution with small sample size. Fuzzy pseudo-absence data can develop the algorithm of GLM and Random Forest as efficient as MaxEnt at the small sample size. Although, the predicted area of each model had some differences whereas PMCC test can represent that each algorithm established the suitability map of *Musa serpentina* in the same direction. In conclusion, all three algorithms are highly efficient tools of species distribution models for limited occurrence data

Acknowledgement

We would like to thank our colleagues at the Computational Science Department, School of Science, Mae Fah Luang University for fruitful interaction and encouragement of the first author. We would like to thank Chiang Rai Rajabhat University for supporting the ArcGIS software and Mae Fah Luang University for funding this research.

References

Elith, J., & Graham, C. H. 2009. Do they? How do they? WHY do they differ? on finding reasons for differing performances of species distribution models. *Ecography*, *32*(1), 66–77. https://doi.org/10.1111/j.1600-0587.2008.05505.x.

Franklin, J. 2010. Spatial Inference and Prediction. In Cambridge (Ed.), *Mapping Species Distributions* (Vol. 141). https://doi.org/10.2140/agt.2016.16.1403.

Hernandez, P. A., Graham, C., Master, L. L., & Albert, D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*(5), 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276.

IUCN. 2001. IUCN Red List Categories and Criteria: Version 3.1. IUCN Species Survival Commission. In *IUCN*. Gland, Switzerland and Cambridge, UK..

Khatchikian, C., Sangermano, F., Kendell, D., & Livdahl, T. 2011. Evaluation of species distribution model algorithms for fine-scale container-breeding mosquito risk prediction. *Medical and Veterinary Entomology*, *25*(3), 268–275. https://doi.org/10.1111/j.1365-2915.2010.00935.x.

Liu, C., White, M., & Newell, G. 2011. Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, *34*(2), 232–243. https://doi.org/10.1111/j.1600-0587.2010.06354.x.

Marod, D., Pinyo, P., Duengkae, P., & Hiroshi, T. 2010. The role of wild banana (Musa acuminata Colla) on

wildlife diversity in mixed deciduous forest, Kanchanaburi Province, Western Thailand. *Kasetsart Journal - Natural Science*, *44*(1), 35–43.

Morgane, B.-M., Frederic, Jiguet, C., Cecile Helene, A., & Wilfried, T. 2013. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution 2012*, *3*, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Phillips, S. 2008. A Brief Tutorial on Maxent. *AT&T Research*, 1–38. https://doi.org/10.4016/33172.01.

RUSHTON, S. . P., ORMEROD, S. J., & KERBY, G. 2004. New paradigms for modeling species distributions? *Journal of Applied Ecology*, *41*(April 2004), 193–200. https://doi.org/10.1111/j.0021-8901.2004.00903.x.

Swangpol, S., & Somana, J. 2011. Musa serpentina. *THAI FOREST BULLETIN (BOTANY)*, *39*, 31–36.

Swets, J. A. 2015. Measuring the Accuracy of Diagnostic. *American Association for the Advancement of Science*, *240*(4857), 1285–1293. Retrieved from http://www.jstor.org/stable/1701052?seq=1&cid=pdf-reference#references_tab_contents.

Tovaranonte, J., Blach-Overgaard, A., Pongsattayapipat, R., Svenning, J. C., & Barfod, A. S. 2015. Distribution and diversity of palms in a tropical biodiversity hotspot (Thailand) assessed by species distribution modeling. *Nordic Journal of Botany*, *33*(2), 214–224. https://doi.org/10.1111/j.1756-1051.2013.00217.x.

Trisurat, Y., Shrestha, R. P., & Kjelgren, R. 2011. Plant species vulnerability to climate change in Peninsular Thailand. *Applied Geography*, *31*(3), 1106–1114. https://doi.org/10.1016/j.apgeog.2011.02.007.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., … Zimmermann, N. E. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x.