



SPECIES DISTRIBUTION MODELS OF WILD TEA IN THE UPPER NORTH OF THAILAND

S.Nanglae* and R.Nilthong

School of Science, Mae Fah Luang University, 333 Moo 1, Tambon Tasud, Muang District, Chiang Rai, 57100, Thailand

*e-mail: Snanglae@gmail.com

Abstract

This study aims to find factors that affect the distribution of wild tea in the upper north of Thailand and build mathematical models that show the prediction of wild tea distribution. We obtained wild tea data from Tea Institute at Mae Fah Luang University. We used 3 climatic factors (rainfall, humidity and temperature) and 5 geographic factors (soil, slope, digital elevation model (DEM), distance from the main river and aspect) to build species distribution by generalized linear models (GLMs). We generated pseudo-absence points by randomly selecting outside the circle at difference radii from presence points and used the area under the curve (AUC) to statistically evaluate the model. The results showed that the radius size had a major effect on the predicted distribution area.

Keywords: Species Distribution Models (SDMs), Generalized Linear Models (GLMs), Area under the Curve (AUC).

Introduction

Wild tea is Assam tea (*Camillia sinensis* ver. *assamica*). It had origin from India and naturally distributes in the highland forest. In Thailand, the wild tea cultivation fields, known as “miang gardens”, have been maintained in the forests and recognized as a system of agroforestry. Pornchai Preechpanya [1] reported that many biodiversity features were found in miang gardens especially plants that are useful for health. Moreover, miang gardens are buffer zone to protect forest from invasion and disruption for agriculture.

This study aims to find factors that affect the distribution of wild tea in the upper north of Thailand and build mathematic model, known as “species distribution models (SDMs)”. The SDMs is an important tool for conservation and evolution study. It shows the relationship between species ranges and environmental parameters. The SDMs are categorized in two groups. First group, the methods that require presence-only data are called “profile techniques”, for example, a bioclimatic analysis and prediction system (BIOCLIM), a flexible modeling procedure for mapping potential distributions of plants and animals (DOMAIN) and ecological niche factors analysis (ENFA). Second group, the methods that require both presence and absence data are called “group discrimination techniques”, for example, generalized linear models (GLMs) and generalized additive models (GAMs). Both groups are based on statistical models. Elith et al [2] reported that when they compared the various SDMs, they found that group discrimination techniques tend to perform better than profile techniques. Thus, group discrimination techniques are increasingly used.

The main problem when we use group discrimination techniques is absence data are unavailable. So, many studies have used pseudo-absence data in place of real absence data [3]. There are several approaches for generating pseudo-absence data. It can be categorized in two main approaches. The first approach is randomly selecting the pseudo-absence points including background across all of the study area [4-5]. The second approach is selecting pseudo-absence points with two steps. First step is estimating suitability area by profile technique and the second step is selecting pseudo-absence points outside the suitability area [3, 5].

In this study, we chose group discrimination techniques for building SDMs with generalized linear models (GLMs). We assumed that the area in the circle is the suitable area for wild tea. So, the pseudo-absence points were randomly selected outside the circles at each radius (5 km, 10 km, 15 km, 20 km, 25 km, 30 km, 35 km, 40 km, 45 km, and 50 km) from presence points and then compared SDMs at each radius.

Methodology

Species and environmental data

We used 41 points of tea data with three climatic factors (rainfall, humidity and temperature) and five geographic factors (soil series, slope, digital elevation model (DEM), distance from the main river and aspect) in this study. The tea data were obtained from Tea Institute at Mae Fah Luang University. Rainfall data, humidity data; temperature data and river data were obtained from Remote Sensing and GIS at Asian Institute of Technology; DEM were obtained from Land Development Department and soil series data were obtained from Center for Information Technology Services at Mae Fah Luang University. We calculated slope and aspect from DEM and calculated distance from the main river from river data. The study area was Upper-North of Thailand. These areas covered eight provinces: Chiang Mai, Chiang Rai, Mae Hong Son, Nan, Phayao, Phrae, Lampang and Lamphun. Points of tea data and the study area are shown in Figure 1.

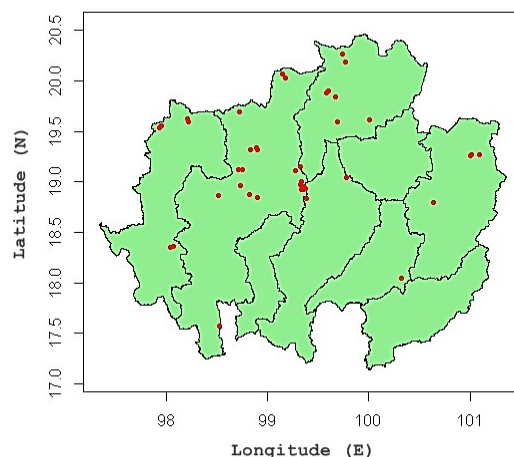


Figure 1 Points of tea data that distribute on Upper-North of Thailand

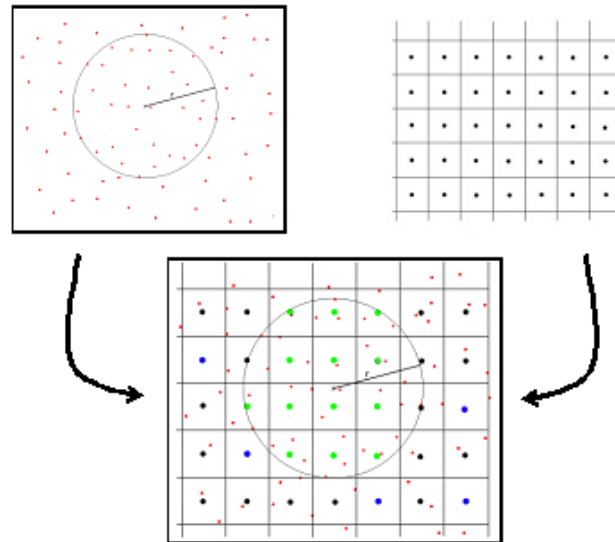


Figure 2 Generating pseudo-absence points: The center of circle is presence point. The red points are random points. The green points are the center points of each grid and also inside the circle. The blue points are the center points of the grid not containing the red point(s) and outside the circle. The black points are the center points of the grid containing the red point(s) and outside the circle called as *pseudo-absence points*.

Generating pseudo-absence points

We generated pseudo-absence points by randomly selected outside the circles at each radius (5 km, 10 km, 15 km, 20 km, 25 km, 30 km, 35 km, 40 km, 45 km, and 50 km) from presence points. From Figure 2, we were overlaying two maps with R programming. The first map, we drew circle around at each presence points with one radius. Then, we randomly selected 410 points on the map of study area (red points). The second map, we built map with 5 km × 5 km grid cells across of the study area with center point for each grid cell. When we overlaid two maps, the outside circle points (black points) were selected as pseudo-absence points. Although, the blue points were outside the circle but they were not selected as pseudo-absence points because the grid cells were not contained random points (red points).

Modeling experiment

Generalized linear models (GLMs) were introduced by Nelder and Wedderburn [10] in 1972. In GLMs, the functions in exponential family that are non-linear form are transformed to linear form. Estimation and inference are based on the theory of maximum likelihood estimation. We used logistic regression models that are one type of GLMs for binary (presence/absence) data to predict species distribution. Let $Y = 1$ or 0 denote the presence/absence of a species respectively and $p(x) = P(Y = 1 | X = x)$ be the probability that the species is present when $X = x$. The resulting presence-absence response curve is

$$p(x) = \frac{e^{g(u)}}{1 + e^{g(u)}}$$

where $g(u)$ is link function.

$$g(u) = \alpha + \beta^T x = \alpha + \sum_{j=1}^p \beta_j x_j$$

Results

This study, we randomly selected 20 times of pseudo-absence data set at each radius and used the area under the curve (AUC) to statistically evaluate each model. We ran GLMs and selected variables by stepwise method. In R programming, the criterion for selecting variables by stepwise was Akaike information criterion (AIC).

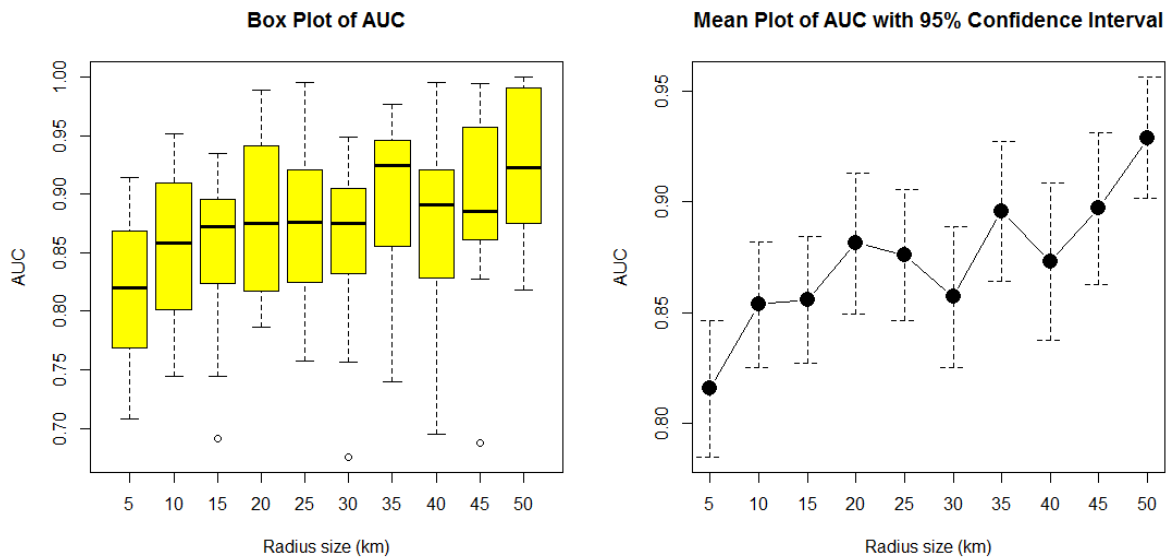


Figure 3 Box plot of AUC and mean plot of AUC at each radius

From Figure 3, we compared these models with AUC by box plot. The box plot showed the median value, the distribution and the outliers of AUC at each radius. The lines in the box indicated the median values of AUC. It showed that the median values of AUC at 5 to 10 km were in the period of increasing. The median values of AUC at 15 to 30 km were rather constant and the median values of AUC at 35 to 50 km were changing. The points outside the end of the vertical lines were outliers. The outliers of AUC at each radius were not considered. The mean plot showed the mean value of AUC and 95% confidence interval at each radius. The most of mean values were about 0.85 to 0.90. These models were considered excellent discrimination. After we deleted the outliers at each radius, we fitted the cubic curve for comparing AUC at each radius. We tried to find the suitability radius for generating pseudo-absence data set.

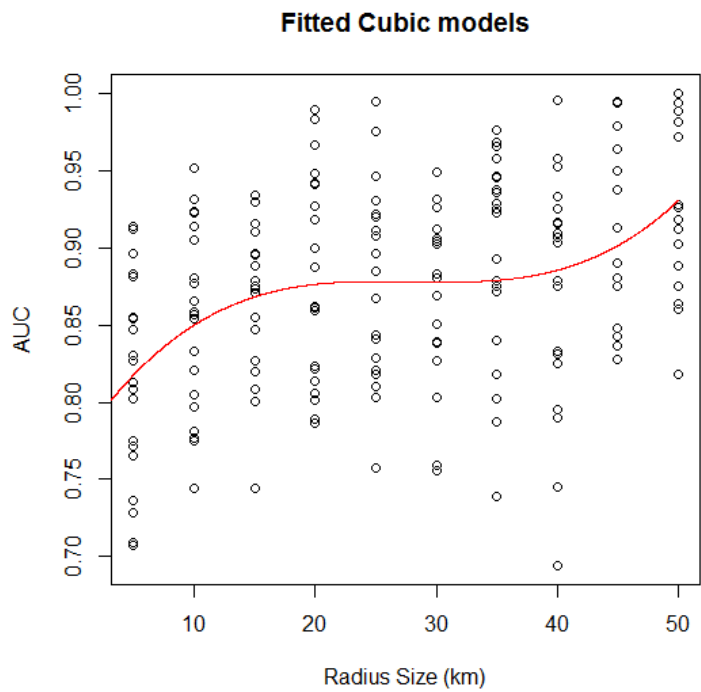


Figure 4 Curve fitting between AUC and radius size for cubic models

From Figure 4, it showed the cubic curve for comparing AUC at each radius. The range between 15 km up to 40 km was quite constant for AUC. Therefore, we selected smallest part of radius sizes 15 km and 20 km that still gave the excellent AUC to pursuit the suitability model.

Table 1 Statistic value of AUC at radius size 15 km and 20 km

Statistic value	Radius size	
	15 km	20 km
Mean	0.864565	0.881416
SD	0.048561	0.067682
CV	0.056168	0.076788
n	19	20

From Table 1, we randomly selected 20 times of pseudo-absence data set at each radius and the outliers of AUC were not considered. The mean values of AUC at each radius were closely. Although, the mean value at 20 km was more than that of at 15 km but the value of coefficient of variance (CV) at 20 km was more than that of at 15 km. It could interpret that the AUC at 20 km was varying more than AUC at 15 km so we couldn't say that the model at radius 20 km was better than the model at radius 15 km.

Then, we selected model that had AUC closed to the mean value to represent at each radius. We compared model and predictive maps.

Table 2. GLMs with stepwise method for radius size 15 km and 20 km

Factors	Radius Size = 15 km			Radius Size = 20 km		
	Coefficients	Std. Error	z value	Coefficients	Std. Error	z value
DEM	0.003243	0.000723	4.487 *	0.004121	0.00084	4.893 *
Rainfall	0.008095	0.001839	4.402 *	0.008751	0.00198	4.421 *
Humidity	-0.531300	0.142400	-3.73 *	-0.477800	0.13860	-3.447 *
Distance	0.000069	0.000037	1.882 .	0.000115	0.00004	2.766 *
Aspect				-0.016060	0.00635	-2.528 *
Intercept = 26.270000 AUC = 0.8674242			Intercept = 23.680000 AUC = 0.887931			

“ * ” is significance at 0.05 and “ . ” is significance at 0.10

Table 2 showed the factors that were selected to model by stepwise method. At 15 km, there were 4 factors that affected the model (DEM, Rainfall, Humidity and Distance). DEM was the strongest effect on the models. Subordinate factors were rainfall, humidity and distance, respectively. At 20 km, there were 5 factors that affected the model (DEM, Rainfall, Humidity, Distance and Aspect). DEM was the strongest effects like the model at 15 km. Subordinate factors were rainfall, humidity, distance and aspect, respectively.

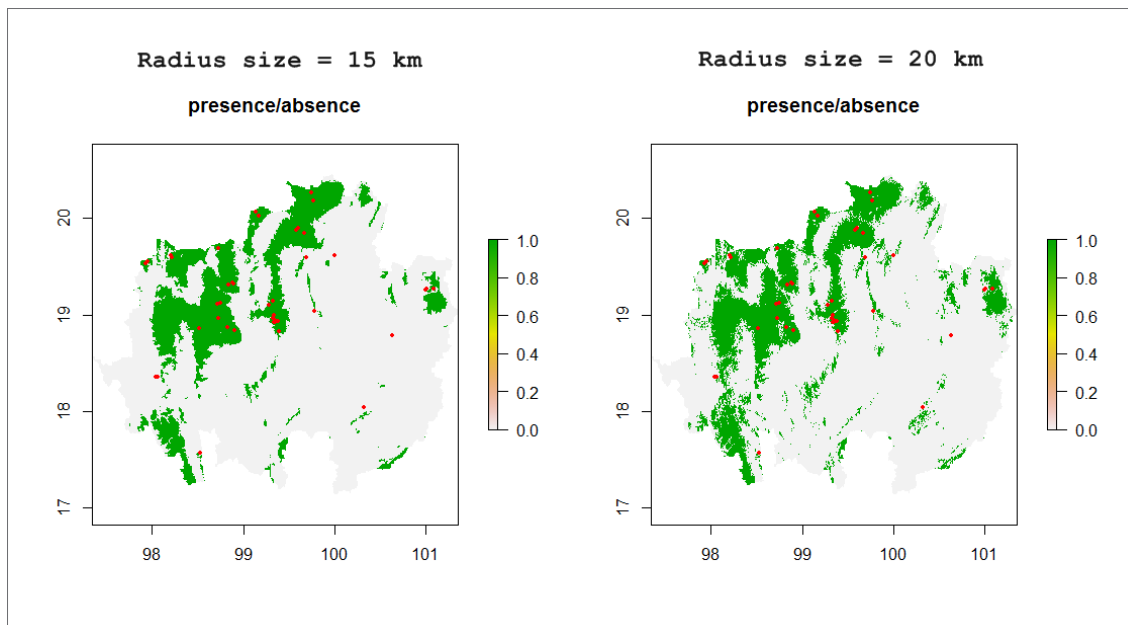
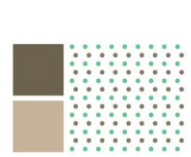


Figure 5 Estimation of presence area (Green area) and absence area (Gray area) from GLMs at radius equal 15 km and 20 km

From Figure 5 showed the estimation of presence and absence area, we compared two estimation maps with radius 15 km and 20 km and the red points on the maps were presence points. These figure showed that both maps estimated the presence area somewhat similarly but the map at radius 20 km seems to have more presence area than the map at radius 15 km and also AUC at radius 20 km was more than AUC at radius 15 km. Although, the map with larger presence area could not tell it was better than the other map because the AUC value at radius 20 km was more varying than AUC value at radius 15 km. So, we could not select the best model from this study but we could analyze the factors that affected the wild tea distribution. The next study, we would analyze factors with GLMs by choosing the other



method of generating pseudo-absence points and try to improve the model by reducing the variance of AUC at each radius.

Conclusion

Many studies demonstrated that the generating pseudo-absence points affected the resulting models [6]. In this study, we generated pseudo-absence points by randomly selecting outside the circle at each radius from presence points. It showed that species distribution models by GLMs depended on the radius size from presence points. If chosen radius size was too small (less than 15 km), the selected pseudo-absence points probably had the same geographical area or same climate whereas if radius size was too large (more than 35 km), surely the selected pseudo-absence points would have very different geographical area. Although our results were fairly effective, the random size would need to be tuned up by using other statistical techniques together with sufficient data for improving the model.

Acknowledgements

We also thank the Tea Institute and Center for Information Technology Services at Mae Fah Luang University, Remote Sensing and GIS at Asian Institute of Technology and Land Development Department for all data in this study.

References

1. Pornchai Preechapanya, *Indigenous Highland Agroforestry Systems of Northern Thailand*, Chiang Doa Watershed Research Station, Chiang Mai, Thailand, 2000.
2. Elith et al., *Novel method improve prediction of species' distributions from occurrence data*, *Ecography*, 2006, **29**, 129-151.
3. Mary S. Wisz and Antoine Guisan, *Do pseudo – absence selection strategies influence species distribution models and their predictions? An information – theoretic approach based on simulated data*, *BMC Ecology*, 2009.
4. Morgane Barbet-Massin, Frédéric Jiguet, Cécile Héléne Albert and Wilfried Thuiller, *Selecting pseudo – absences for species distribution models: how, where and how many?*, *Method in Ecology and Evolution*, 2012.
5. Jorge M. Lobo and Marcelo F. Tognelli, *Exploring the effects of quantity and location of pseudo – absences and sampling biases on the performance of distribution models with limited point occurrence data*, *Journal for Nature Conservation*, 2011, **19**, 1-7.
6. Jøgeir N. Stokkland, Rune Halvorsen, Bente Støa, *Species distribution modelling – Effect of design and sample size of pseudo-absence observations*, *Ecological Modelling*, 2011, **222**, 1800-1890.