Species distribution models for two subspecies of *Dodonaea viscosa* (Sapindaceae) in Indonesia

Arief Priyadi¹-²*

¹ “Eka Karya” Bali Botanic Garden, Tabanan, Indonesia
² Research Centre for Plant Conservation and Botanic Gardens, Bogor, Indonesia

*Corresponding author Email: arief.priyadi@lipi.go.id

Abstract. Maxent is one of the most popular software for Species Distribution Modelling because of its versatility and robustness, requires only species occurrence data and some environmental layers. The availability of default setting adds the usability, not surprisingly most reports were resulted from this option. However, cautions should be taken when modelling based on presence-only data; might be unavailable in default. This work aimed to test if running in personalized settings could result in better models. Here I run Maxent using data of *Dodonaea viscosa* subspecies *angustifolia* and *D. viscosa* subsp. *viscosa* (Sapindaceae) from GBIF and BIOCLIM for environmental layers, covering the Indonesian archipelago. The former subsp. is upland whereas the latter is coastal plant. Six models were built for each subsp. and their performance were further evaluated based on area under the ROC curve (AUC), prevalence, pixel counts of suitable area and visual observation of the final maps. Results showed all models of subsp. *angustifolia* were better compared to those of subsp. *viscosa*; proper measures applied before and during building might improve model performance. This work shows that Maxent may be more powerful when operated beyond its default setting. Users should be encouraged to experiment with more parameter values to produce the best model.

1. Introduction

Species distribution models (SDM) or also known as habitat suitability models and niche models is a method to correlate predictive environmental variables to the presence of species in corresponding habitats [1]. This method becomes popular with the recent publicly available environmental data, such as Worldclim [2, 3], and computer software [4-6]. Consequently, it has been applied in a broad range of studies: biogeography [7], invasion biology [8], and conservation biology [9, 10].

Maxent is one of popular software for individual SDM [1, 11] and developed to run based on presence only (PO) data [4, 5, 12]. Despite of being ease, precautions should be considered in using PO data to build SDM [11, 13]. Six determining variables were identified for better results: background data, feature classes, regularization multipliers, sampling bias, types of output, and evaluating models [11]. Additionally test data, overfitting evaluation, and varying program settings could be beneficial [7]. However those settings might be difficult to be applied in usual Maxent run by using its point-and-click interface [6] and requires more knowledge to run the program in batch mode/sample with data (SWD) format [13] which might become an arduous
task for general users. Alternatively, SDMToolbox can be used to help general Maxent users to prepare data, run, and evaluate results with ease [14, 15].

I asked whether modification of Maxent settings could result in models with different performance, which environmental data contributed most to corresponding SDM, and what was the easiest measure to determine the best model. Here two subspecies of Dodonaea Jacq. subsp. angustifolia (L.f.) J.G.West. and Dodonaea subsp. viscosa (L.f.) J.G.West. were used to generate SDM with publicly available environmental data. The best models were determined based on confusion matrix: Area under the ROC curve (AUC) and prevalence (the probability of occurrence across the study area) [16], also the SDM’s pixel counts of selected suitable area.

**Table 1.** The design of six species distribution models for Dodonaea viscosa subsp. angustifolia and subsp. viscosa.

| Model | Background Selection | Bioclim | Occurrence Points | Regularization Multiplier(s) |
|-------|----------------------|---------|-------------------|---------------------------- |
|       | No selection | MCP | Convex-Hull | All | Selected | All | Rarefied |
| SDM1  | ✓ | ✓ | ✓ | ✓ | ✓ | 1, 3, 5 |
| SDM2  | ✓ | ✓ | ✓ | ✓ | ✓ | 1, 3, 5 |
| Mxt1  | ✓ | ✓ | ✓ | ✓ | ✓ | 1 |
| Mxt2  | ✓ | ✓ | ✓ | ✓ | ✓ | 1 |
| Mxt3  | ✓ | ✓ | ✓ | ✓ | ✓ | 1 |
| Mxt4  | ✓ | ✓ | ✓ | ✓ | ✓ | 1 |

2. Materials and Methods

2.1. The study species

The genus Dodonaea is native to Australia, consists of 68 species [17]. The species in Malesian floristic regions includes D. angustifolia and D. viscosa (considered as two species) [18] but treated as sub spcies of D. viscosa subsp. angustifolia and subsp. viscosa [17]. The present taxonomic treatment is mainly based on leaf morphology [17, 19]. The formal classification of the genus is yet to be established because recent studies based on DNA sequence data showed that the morphological variation does not reflect geographical distribution [20] and current subspecies treatment does not agree with its history of evolution [21]. Here I follow the treatment of two sub species: D. viscosa subsp. angustifolia and subsp. viscosa. In South-East Asia, the habitat of subsp. viscosa is at 0–90 m and that of subsp. angustifolia is at 1200–3600 m [18]. Therefore, we have two subsp. with a contrast of habitat preference, the former is coastal and the latter is upland.

2.2. Occurrence points and Environmental layers

In total 155 occurrence points were browsed and downloaded from GBIF using categories: species name, with geographical coordinates, region Asia, and country Indonesia. Following duplicate removal, as many as 90 points were retained that consisted of 47 and 43 points subsp. angustifolia and subsp. viscosa. The data was formatted as .csv in three columns of species, longitude, and latitude (hddd.dddd°).

Nineteen BIOCLIM variables ver. 2.1 resolution of 30s [2] were downloaded from https://biogeo.ucdavis.edu/data/worldclim/v2.1/base/wc2.1_30s_bio.zip. The geotiff formats were converted to ascii and assigned to WGS84 projection using SDMToolbox 2.4 [14, 15], covering region between 94.5°E–142°E and 11.5°S–8°N.
2.3. Building of six models and their evaluation

Maxent version 3.4.1 and SDMToolbox version 2.4 were used in this study. I consider six key factors that affect the final distribution model [11] with emphasize on precautions described in SDMToolbox: background data and regularization multipliers [14, 15]. Additionally, removal of significantly correlated environmental layers (threshold=0.7) was applied [11] as well as rarifying of occurrence data [15]. The options applied for each model are shown in Table 1, and all models were run in 3 (three) replicates.

![Figure 1. Species distribution models: (a) D. viscosa subsp. angustifolia SDM1, (b) subsp. viscosa SDM1, (c) subsp. angustifolia Mxt3, and (d) subsp. viscosa Mxt 3. Habitat suitability scale ranged from 0 (unsuitable–blue) to 1 (highly suitable–red).](image)

Table 2. Percent contribution of each predictor to the corresponding model.

| Ssp.       | Models | Bio1 | Bio2 | Bio5 | Bio6 | Bio9 | Bio11 | Bio12 | Bio13 | Bio14 | Bio15 | Bio19 | Total |
|------------|--------|------|------|------|------|------|-------|-------|-------|-------|-------|-------|-------|
| **Subsp. angustifolia** |        |      |      |      |      |      |       |       |       |       |       |       |       |
| SDM1       | 64.6   | 3.3  | -    | -    | -    | 2.8  | 0     | -     | -     | 25.2  | -     | -     | 95.9  |
| SDM2       | 61.1   | 0.2  | -    | -    | -    | 0.1  | 1.7   | -     | 27.2  | -     | 90.3  |       |       |
| Mxt1       | 2.3    | 0.8  | 1.5  | 11.9 | 31.5 | 19.4 | 0.1   | 0     | 3.4   | 13.9  | 88.7  |       |       |
| Mxt2       | 72.5   | 1.3  | -    | -    | 0.1  | 0    | -     | 24    | -     | 97.9  |       |       |       |
| Mxt3       | 0      | 0.6  | 15.2 | 11.1 | 17.8 | 15.4 | 0     | 2.8   | 4.5   | 18.6  | 86    |       |       |
| Mxt4       | 72.4   | 0.6  | -    | -    | 0.1  | 0    | 24    | -     | 97.3  |       |       |       |       |
| **Subsp. viscosa** |        |      |      |      |      |      |       |       |       |       |       |       |       |
| SDM1       | 5.5    | 48.5 | -    | -    | -    | 11.4 | 13    | -     | 20.7  | -     | 99.1  |       |       |
| SDM2       | 4.9    | 45.7 | -    | -    | -    | 6    | 24.7  | -     | 11.5  | -     | 92.8  |       |       |
| Mxt1       | 0      | 29.1 | 2    | 0    | 0    | 0.4  | 2.4   | 7.5   | 17.1  | 6.4   | 1.5   | 66.4  |       |
| Mxt2       | 5      | 38.5 | -    | -    | -    | 9.4  | 20.1  | -     | 23.3  | -     | 96.3  |       |       |
| Mxt3       | 0      | 27   | 3.1  | 0    | 2.1  | 0    | 1.5   | 4.9   | 18.6  | 7.6   | 1     | 65.8  |       |
| Mxt4       | 8.2    | 34.6 | -    | -    | -    | 17.6 | 13.8  | -     | 22.4  | -     | 96.6  |       |       |

Bio1: annual mean temperature, Bio2: mean diurnal range, Bio5: max temperature of the warmest month, Bio6: min temperature of the coldest month, Bio9: mean temperature of the driest quarter, Bio11: mean temperature of the coldest quarter, Bio12: annual precipitation, Bio13: precipitation of the wettest month, Bio15: precipitation seasonality, Bio19: precipitation of the coldest quarter.
Table 3. Quantitative measures to evaluate model performance.

| Ssp.          | Models  | Test AUC | Prevalence | Pixel count |
|---------------|---------|----------|------------|-------------|
| Subsp. angustifolia | SDM1    | 0.941 a  | 0.105 b    | 3.0% a      |
|               | SDM2    | 0.851 a  | 0.153 a    | 2.9% a      |
|               | Mxt1    | 0.896 a  | 0.052 c    | 0.7% b      |
|               | Mxt2    | 0.895 a  | 0.062 c    | 0.8% b      |
|               | Mxt3    | 0.919 a  | 0.048 c    | 0.6% b      |
|               | Mxt4    | 0.899 a  | 0.060 c    | 0.8% b      |
| Average       | 0.900 p | 0.080 p  | 1.46% p    |
| Subsp. virgina | SDM1    | 0.717 a  | 0.323 bc   | 12.2% a     |
|               | SDM2    | 0.619 a  | 0.481 a    | 14.4% a     |
|               | Mxt1    | 0.688 a  | 0.322 bc   | 4.3% b      |
|               | Mxt2    | 0.692 a  | 0.387 b    | 5.2% b      |
|               | Mxt3    | 0.688 a  | 0.274 c    | 4.3% b      |
|               | Mxt4    | 0.701 a  | 0.319 c    | 4.7% b      |
| Average       | 0.684 q | 0.351 q  | 0.075 q    |

In the same column, values followed by same letter are not significantly different (Duncan test α=5%); total pixel counts of the study area 2,842,687.

Table 4. Pearson’s correlation between the quantitative measures.

|               | Test AUC | Prevalence | Pixel count |
|---------------|----------|------------|-------------|
| Test AUC      | 1        |            |             |
| Prevalence    | -0.82    | 1          |             |
| Pixel count   | -0.65    | 0.83       | 1           |

SDM1 and SDM2 were built using “Run MaxEnt: Spatial Jackknifing” available in SDMToolbox v2.4. Highly correlated environmental layers were removed, occurrence points were rared and regularization multipliers were varied. The difference was the method to select background point, SDM1 was built using minimum-convex polygon (MCP) and SDM2 local adaptive convex hull (Convex-Hull). The SDMToolbox’s outputs were a series of batch files as inputs for Maxent 2.4.1. This method resembles a description of running Maxent in batch mode [13]. In total 15 models were built for each subsp. by applying 3 (three) regularization multipliers combined with 5 (five) types of features: linear (L), hinge (H), linear quadratic (LQ), linear quadratic hinge (LQH), and linear quadratic hinge product threshold (LQHPF). The best model was chosen based on omission error rate and AUC.

Models of Mxt1-4 were built by Maxent’s point-and-click mode as described in its brief tutorial [6]. These models were set by default regularization multiplier and no background selection. The four were made by the combination of two levels of environmental layers (all and selected) and two levels of occurrence points (all and rared).

Three variables—AUC, prevalence, and pixel count of selected suitable areas—were used to compare the performance of 6 (six) models for 2 (two) Dodonaeae subsp. The values of test AUC and prevalence were extracted from a Maxent’s output “maxentResults.csv”. Pixel counts of areas with suitability index of 0.67-1 were determined using Reclassify tool in a GIS application. The final data matrix consisted of 36 rows (2 subsp, 6 models, and 3 replications) and 3 columns (AUC, prevalence, and pixel counts). The data was subjected to analysis of variance (Anova) continued by duncan’s multiple range test (Duncan Test) in PROC GLM available in SAS ver.
9.1 [22]. Pearson’s correlation matrix between variables were determined using Data Analysis add-ins in MS-Excel 2016.

3. Results and Discussions

3.1. Results

After rarefying of 47 and 43 the origin occurrence data, SDMToolbox maintained 37 and 40 points respectively. As many as 7 (seven) Bioclimatic variables (Bio1, 2, 4, 7, 12, 13, and 15) were determined as uncorrelated environmental layers. SDMToolbox assigned specific background layers for each subspecies and selection methods (MCP and Convex-Hull) in .asc format. Figure 1 shows 2 out of 6 models in this study for subsp. *angustifolia* and *viscosa*. These two represent two extremes of the model settings (Table 1). The models for subsp. *angustifolia* (Figure 1a, 1c) show suitable area of uplands whereas for subsp. *viscosa* (Figure 1b, 1d) show suitable area of lowlands. The area marked in red (highly suitable) is more prominent for SDM1 (Figure 1a, 1b) compared to those of Mxt3 (Figure 1c, 1d). Percent contribution of each variable to the corresponding SDM is shown in Table 2 (total range 65.8%-99.1%); Bio1, Bio2, and Bio15 contribute significantly to the models.

Three quantitative measures in Table 3 show the characteristics of modelled distributions of the two subspecies and 6 models for each subspecies. Subspecies *angustifolia* has higher AUC values and pixel counts compared with those of subspecies *viscosa*; the former has lower prevalence values compared with those of the latter.

The Pearson’s correlation values between the measures is shown in Table 4. Based on three measures in Table 3, six models in each subspecies can be grouped into SDMToolbox generated models (SDM1, 2) and Maxent direct run/point-and-click interface (Mxt1-4).

3.2. Discussions

I found that modifying Maxent’s settings resulted in different models with various performance. SDM1 and SDM2 tends to outperform Mxt1-4 for both subsp. *angustifolia* and subsp. *viscosa* (Figure1, Tabel 3). Bio1, Bio2, and Bio15 were three predictors with highest contribution to the corresponding model (Table 2). Bio1 and Bio2 corresponded to temperature whereas Bio15 corresponded to rainfall/precipitation (Table 2). The generated SDM agreed with real habitats of the study species (upland for subsp. *angustifolia* and lowland for subsp. *viscosa*), as shown in Figure 1 and Table 3. The correlation analysis showed that prevalence inversely proportional with test AUC and pixel count directly proportional with prevalence.

These results pinpoint that many parameters are available in Maxent settings and modifying one or more parameters would result in different model. This work only focusses on determining present potential habitats for two subspecies of *Dodonaea*. On the course of species range-shift, such present distribution models could be projected for past or future climate scenarios (e.g. [8, 23]). In such scenario with more model uncertainties of future climate predictions or past climate reconstructions, the construction of reliable present distribution models is a prerequisite. Therefore, utilization of available tools for data preparation to run Maxent such as SDMToolbox should be beneficial.

The results of this work showed that Bioclim variables related to temperature and precipitation contributed significantly to the SDM. This phenomenon is not surprising, because it is a matter of fact that distribution of floristic communities in the tropics is governed by those two factors (e.g. [24]). Related to temperature subsp. *angustifolia* is adapted to cooler climate (upland) whereas subsp. *viscosa* in warmer (lowland). This finding is in concordance with that is stated by [18] if subsp. *angustifolia* is adapted to cooler climate with less precipitation compared with that of subsp. *viscosa* that is adapted to warmer and more humid climate.
4. Conclusions
This work shows that varying maxent settings resulted different models with different performance, therefore users should be encouraged to explore beyond Maxent’s default settings to produce the best SDMs. The generated SDMs agree with real habitat preference of the study species.

Acknowledgments
I thank to Huiqin Yi (South China Botanical Garden, University of Chinese Academy of Sciences) for introducing Species Distribution Modelling in 2018 that lead to further exploration of its capabilities. Original species occurrence data is available from https://doi.org/10.15468/dl.gjnyqz.

References

[1] Ovaskainen O and Abrego N 2020 Joint Species Distribution Modelling With Applications in R (Singapore: Cambridge University Press)
[2] Fick S E and Hijmans R J 2017 WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas International Journal of Climatology 37 4302-15
[3] Hijmans R J, Cameron S E, Parra J L, Jones P G and Jarvis A 2005 Very high resolution interpolated climate surfaces for global land areas International Journal of Climatology 25 1965-78
[4] Phillips S J, Anderson R P and Schapire R E 2006 Maximum entropy modeling of species geographic distributions Ecological Modelling 190 231-59
[5] Phillips S J, Anderson R P, Dudík M, Schapire R E and Blair M E 2017 Opening the black box: an open-source release of Maxent Ecography 40 887-93
[6] Phillips S J 2017 A Brief Tutorial on Maxent.
[7] Radosavljevic A and Anderson R P 2014 Making better Maxent models of species distributions: complexity, overfitting and evaluation Journal of Biogeography 41 629-43
[8] Elith J, Kearney M and Phillips S 2010 The art of modelling range-shifting species Methods in Ecology and Evolution 1 330-42
[9] Raes N, Roos M C, Slik J W F, Van Loon E E and Steege H t 2009 Botanical richness and endemicity patterns of Borneo derived from species distribution models Ecography 32 180-92
[10] Elith J and Leathwick J 2009 Conservation prioritisation using species distribution modelling Spatial conservation prioritization: quantitative methods and computational tools. Oxford University Press, Oxford, UK 70-93
[11] Merow C, Smith M J and Silander Jr J A 2013 A practical guide to MaxEnt for modeling species’ distributions: what it does, and why inputs and settings matter Ecography 36 1058-69
[12] Phillips S J and Dudík M 2008 Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation Ecography 31 161-75
[13] Elith J, Phillips S J, Hastie T, Dudík M, Chee Y E and Yates C J 2011 A statistical explanation of MaxEnt for ecologists Diversity and Distributions 17 43-57
[14] Brown J L 2014 SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses Methods in Ecology and Evolution 5 694-700
[15] Brown J L, Bennett J R and French C M 2017 SDMtoolbox 2.0: the next generation Python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses PeerJ 5 e4095
[16] Fielding A H and Bell J F 1997 A review of methods for the assessment of prediction errors in conservation presence/absence models Environmental conservation 38-49
[17] West J 1984 A revision of Dodonaea Miller (Sapindaceae) in Australia Brunonia 7 1-194
[18] van Welzen P C 2001 *Plant Resources of South-East Asia: Medicinal and Poisonous Plants 2*, ed J v Valkenburg and N Bunyapraphatsara (Bogor, Indonesia: PROSEA Foundation) p 785

[19] West J G and Noble I R 1984 Analyses of Digitised Leaf Images of the Dodonaea viscosa Complex in Australia *Taxon* 33 595-613

[20] Harrington M G and Gadek P A 2009 A species well travelled – the Dodonaea viscosa (Sapindaceae) complex based on phylogenetic analyses of nuclear ribosomal ITS and ETSf sequences *Journal of Biogeography* 36 2313-23

[21] Christmas M J, Biffin E and Lowe A J 2018 Measuring genome-wide genetic variation to reassess subspecies classifications in Dodonaea viscosa (Sapindaceae) *Australian Journal of Botany* 66 287-97

[22] SAS 2004 *SAS/STAT® 9.1 User's Guide* (Cary, North Carolina: SAS Institute Inc.)

[23] Qin A, Liu B, Guo Q, Bussmann R W, Ma F, Jian Z, Xu G and Pei S 2017 Maxent modeling for predicting impacts of climate change on the potential distribution of *Thuja sutchuenensis* Franch., an extremely endangered conifer from southwestern China *Global Ecology and Conservation* 10 139-46

[24] Whitmore T C 1975 *Tropical Rain Forests of the Far East*: Clarendon Press, Oxford)