Zhonglin XU1*, Chuanyan ZHAO2, Zhaodong FENG 1, 3

1 MOE Key Laboratory of Western China’s Environmental Systems and College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China, *e-mail: galinwa@gmail.com (corresponding author)

2 MOE Key Laboratory of Arid and Grassland Agroecology, Lanzhou University, Lanzhou 730000, China

3 College of Resource and Environmental Science, Xinjiang University, Urumqi, 830046, China

SPECIES DISTRIBUTION MODELS TO ESTIMATE THE DEFORESTED AREA OF PICEA CRASSIFOLIA IN ARID REGION RECENTLY PROTECTED: QILIAN MTS. NATIONAL NATURAL RESERVE (CHINA)

ABSTRACT: The earth is now facing the land degradation due to human disturbance, natural habitats were converted to rural and agricultural areas in order to fulfill the increasing demand of human population. The deforestation of Picea crassifolia (Qinghai spruce) forest at Qilian Mts is an example of such disturbance. P. crassifolia is an ecologically and hydrologically important plant species in the northwestern arid area of China. However, the forests have been intensively and extensively deforested. In order to restore the human-disturbed ecosystems, the spatial distribution of P. crassifolia needs to be delineated. This study employed Genetic Algorithm for Rule-set Prediction model (GARP) and Maximum entropy model (Maxent) and four environmental variables (mean temperature of the warmest quarter, precipitation of the wettest quarter, annual solar radiation, topographic wetness index) to predict the potential distribution of P. crassifolia in Qilian Mts. Genetic Algorithm for Rule-set Prediction model (GARP) produces a model of species niches in geographic space based on heterogeneous rule-sets. Maximum entropy model (Maxent) focuses on fitting a probability distribution for occurrence based on the idea that the best explanation to unknown phenomena will maximize the entropy of the probability distribution, subject to the appropriate constraints. The environmental variables were spatially interpolated throughout the entire study area. We used sensitivity-specificity sum maximum approach to select the threshold value. The projected niche space for the mean temperature of the warmest quarter is between 8.5 and 18.1°C; the space for the precipitation of the wettest quarter is between 149 and 245 mm; the space for annual solar radiation is 118–1100×10^3 wh m^-2 and the space for topographic wetness index is between –0.4 and 5.1. The results show that both GARP and Maxent’s models produce acceptable predictions, but the overall comparison shows that GARP prediction is better than Maxent’s; the comparison between the observed distribution and the predicted distribution suggests that 61% (2869 km^2) of P. crassifolia forests have been deforested.

KEY WORDS: restoration, potential distribution, species distribution modeling, Genetic Algorithm for Rule-set Prediction model, Maximum entropy model, deforested area

1. INTRODUCTION

Generally, each drainage basin contains an upper mountainous part that, with the coverage of vegetation, plays an important role in the water cycle of the soil–plant–atmosphere continuum (SPAC), as well as provides habitat for various life forms. Although the upper mountainous part provides important functions/services, it is influenced by natural (such as fires and insect outbreaks, etc) and anthro-
pogenic (deforestation, agriculture, mining, road construction, reservoir construction, etc) disturbances. Such disturbances may result in changes to forest ecosystem structure and functioning (Zou et al. 1995, Wilk et al. 2001, Bala et al. 2007, Coe et al. 2009). The vast arid area of northwestern China is distinctively characterized by independent inland drainage basins (Tang 1985), and the *P. crassifolia* forests in the Qilian Mts. have been well demonstrated to be extremely important in maintaining the stability of the ecosystems, regulating the hydrological cycles, and even sustaining the socioeconomic systems in the associated drainage basins (Liu and Wang 2001, Dang et al. 2006, Zhao et al. 2003). But the forests have been subjected to deforestation since the Han Dynasty (~200 AD) when the Han Chinese started a nearly permanent occupation of the Hexi Corridor, the lowlands at the eastern front of Qilian Mt. The deforestation has accelerated since the Ming Dynasty (~1400 AD) and the most intensive and extensive deforestation occurred between 1950s and 1980s (Wu 2000, Wang et al. 2002). Consequently, the ecological systems have been recently deteriorating at an alarming rate and the associated hydrological cycles have been severely disrupted so that the ecological degradation has been further accelerated and the socioeconomic systems have been threatened.

In order to restore the human-disturbed ecosystems, maximize the hydrological potentials, and sustain the socioeconomic development in the Qilian Mts. and the associated drainage basins, the Qilian Mts National Nature Reserve was established at 1986 and some reforestation and afforestation practices were conducted within the reserve. The present study attempts to spatially depict the deforested area of the *P. crassifolia* forests that account for as much as ~75% of the total forested areas in the Mountains. This spatial depiction of deforested area of *P. crassifolia* in the reserve is urgently needed for providing the theoretic foundation to guide the governmental ongoing efforts in reforestation and afforestation. In this study, the main objectives are: 1) to predict the potential spatial distribution of *P. crassifolia* by GARP model, GARP appears especially well-suited to reducing error in predicted distributions, both in the form of omission of real distributational areas and inclusion of areas not holding actual populations. 2) to compare the potential distribution with the present distribution of *P. crassifolia* for delineating deforested area of the species in Qilian Mts. area. We hope that the research result could provide some insights into the feasibility of restoring large-scale ecosystems and also of improving land-use planning and management.

Fig. 1. Location and Digital Elevation Model (DEM) of study area. The Qilian Mts. rises along the northeastern rim of the Qinghai-Tibet Plateau to a maximum elevation of 5434 m.
2. STUDY AREA

The Qilian Mountains National Nature Reserve (93.38–104.05°E, 36.03–40.53°N), is located in the northwestern China with elevations ranging from 1500 to 5400 m above sea level (Fig. 1) and covered 13,801 km². The climate is characterized by cold and dry winters and relatively warm and humid summers. The mean annual air temperature (MAT) is ~6 °C at the lowlands and is −10°C at the highest elevations. The mean annual precipitation (MAP) also varies with elevations with 88% of the precipitation occurring between May and September. It is only about 250 mm at lowlands and up to 700 mm in highest elevations. It should also be noted that both MAT and MAP have a decreasing trend from the southeast to the northwest.

3. METHODS

3.1. Data collection

The climate data sets used in this study were obtained from the Worldclim data set (http://www.worldclim.org.). The data are characterized by grid format with a resolution of 1 km. This data set includes annual trends (e.g., mean annual temperature and annual precipitation), seasonality (e.g., annual ranges in temperature and precipitation), and the extremes (e.g., temperatures of the coldest and warmest month, and precipitations of the wet and dry quarters). It should be pointed out that the Worldclim data are spatially interpolated dataset with a spatial resolution of 1 km and the interpolation was based on the geographic locations (i.e., longitudes and latitudes) and the elevations (Hijmans et al. 2005). We also used the second set of climate data from China Meteorological Data Sharing Service System to check the acceptability of the Worldclim data. The actual distribution of *P. crassifolia* (totally 2094 grid cells, used as presence samples), which was obtained from the Bureau of Qilian Mts. National Natural Reserve, was used for the training and testing of the two models. Through field surveys conducted in August of 2008 (Fig. 2), we obtained 93 presence sites and 33 absence sites of *P. crassifolia* for the validation of models.

3.2. Selection and spatialization of environmental variables

In this study, we used Maxent – a species distribution model which enables researchers to estimate species distribution probability by finding the probability distribution of maximum entropy, can also be used to select important variables for the distribution of species (Phillips et al. 2006). Maxent model was used to investigate the contribution of each one of the environmental variables to the species distribution (i.e., *P. crassifolia*). Since temperature and precipitation are the most important factors influencing the distribution of *P. crassifolia*, 19 temperature- and precipitation-related data layers (e.g., minimum temperature, maximum temperature, precipitation in the wettest quarter, etc.), downloaded from the Worldclim Database, were included for evaluating the importance of each one of the environmental variables in contributing to the species distribution. In addition, annual solar radiation (or net incoming short-wave radiation) and topographic wetness index were also included for evaluating the importance. The jackknife test (with the help of Maxent model, see Phillips et al. 2006) of these environmental variables reveals that four environmental variables are the most important variables in influence the distribution of *P. crassifolia* and they were therefore used as predictors to run GARP model. GARP detects non-random relationships between two sets of data: (a) georeferenced occurrence records of the species, and (b) a set of digital raster data layers representing environmental variables potentially relevant to determining the species’ geographic distribution at that particular scale of analysis.

The annual solar radiation (or net incoming short-wave radiation) can be estimated from measured sunshine hours according to the following empirical relationship (Shuttleworth 1992):

\[ R_n = (1 - \alpha)(0.25 + 0.5 \frac{n}{N})S_0 \]

(1)

Where \( n \) is the bright sunshine hours per day (h), \( N \) the total day length (h), \( \alpha \) the reflection coefficient and the recommended value is 0.16 (Shuttleworth 1992, Zhao et
al. 2006). The $S_0$ in Eq. (1), the extraterrestrial radiation, can be calculated using:

$$S_0 = 15.392\frac{d}{r} (\sin \varphi \sin \delta + \cos \varphi \cos \delta \sin t)$$

(2)

where

$$d = 1 + 0.033 \cos \left(\frac{2\pi}{365} \times J\right)$$

is the relative distance between the earth and the sun ($J$ the Julian day number), $t = \arccos(-\tan \varphi \tan \delta)$ is the sun-set hour angle (in radians), $\varphi$ the latitude, $\delta$ the solar declination (in radians) and $\delta = 0.4093 \sin \left(\frac{2\pi}{365} \times J - 1.405\right)$.

The topographic wetness index is determined by DEM-based topographic factors (Beven and Kirkby 1979, Dymond and John-son 2002) and expressed as the following:

$$SM_i = \ln\left(\frac{A}{\tan B}\right)$$

(3)

where $SM_i$ is topographic wetness index, $A$ the upslope contributing area per unit contour ($m^2m^{-1}$), and $B$ is the slope angle.

The environmental variables extracted from Worldclim data set are spatial data. Annual solar radiation and topographic wetness index are needed to be spatialized. The parameters in Eqs. (1), (2) and (3), such as $\varphi$, $A$, $B$ were calculated from DEM with a resolution 1 km. Then the spatial distribution of annual solar radiation and topographic wetness index are obtained from the Eqs. (1), (2) and (3) using ArcGIS software. It should be noted that we tested the acceptability of the Worldclim-extracted data layers by comparing the observed data at 27 stations within and around the study area with the Worldclim-extracted data for these 27 stations.

### 3.3. Introduction to GARP model

The fundamental ecological niche of a species can be defined as the set of ecological conditions that allow for its long-term survival. Whereas the realized niche is subset of the fundamental niche that the species actually occupies (Hutchinson 1957). The fundamental niche describes the climax stage when the species of interest reaches the equilibrium state with the environmental conditions. Theoretically speaking, the realized niche expresses the non-equilibrium state before the climax stage was reached (Schn-neider and Kay 1994). Due to intensive and extensive adverse human impacts on the eco-systems, a species’ fundamental niche may be much greater than the realized niche. In other words, the species should potentially distributes much more extensively than it is actually observed. A niche-based model represents an approximation of a species’ fundamental niche (or potential niche) in the environmental or climate space. The approximation is based on the assumption that the statistically and randomly sampled existing relationships
between the realized niche and the associated environmental variables can be projected to the geographic space (or entire study area) to spatially depict the fundamental niche.

The genetic algorithm, one of many machine learning methods, has been demonstrated to be quite promising for modeling the potential distribution of certain plant species or a group of ecologically associated species through the application of GARP (e.g., Stockwell and Peters 1999, Peterson et al. 2002, Anderson 2003, Elith et al. 2006, Stockman et al. 2006, Schmidt et al. 2008, Contreras et al. 2009). Due to three advantages, GARP have been widely used in species niche modeling. First, as an evolutionary-computing algorithm, the genetic algorithm can efficiently seek for the possible solutions in a complex highly-dimensional variable space. Second, as a nonparametric method, the genetic algorithm makes no assumptions about the underlying data distribution, which is often advantageous when analyzing ecological data. Third, because GARP uses a combination of multiple rules (e.g., logistical regression, bioclimatic envelope rules, etc.), it has a greater predictive ability than any single rule applied independently (Stockwell and Peters 1999). Detailed descriptions of GARP are provided by Stockwell and colleagues (e.g., Stockwell and Noble 1992, Stockwell and Peters 1999, Stockwell 1999).

In this study, two steps were conducted to model potential distribution of the species by the GARP: (1) running the model, and (2) determining the threshold value. After the Maxent-selected four environmental variables and 2094 presence-only samples were input into the model, we set GARP to perform 500 runs in our study. Due to randomness of its algorithm, each running will produce a different estimation of species’ distribution map, so 500 grid files with presence as 1 and absence as 0 were produced. 500 maps were summed to make a cumulative map. Grid values of the cumulative map range from 0 to 500. In order to create a final predictive map, we chose a threshold value (between 0 and 500) to exclude unsuitable cells in the cumulative map. The cumulative map, together with the total 126 (i.e., 93 presence and 33 absence) fields work collected samples, were used to choose the threshold value through constructing a presence-absence confusion matrix or error matrix (Fielding and Bell, 1997). The error matrix is composed of 4 elements denoted by a, b, c and d (Table 1). a expresses true positive (recorded present, predicted present), b is false positive (recorded present, predicted absent), c false negative (recorded present, predicted absent) and d true negative (recorded absent, predicted absent). a and d are correct classifications. b represents omission error and c commission error. Two more terms were introduced to define the threshold value: (1) sensitivity index to define the accuracy level of the presence-prediction in the observed presence cells and the index is calculated as a/(a+c), and (2) specificity index to define the accuracy level of the absence-prediction in the observed absence cells and the index is calculated as d/(b+d). The presence threshold is set when the sum of sensitivity and specificity indices reaches the maximum, namely sensitivity-specificity sum maximum approach (Jiménez-Valverde and Lobo 2006). Specifically, this approach gives us a threshold value of 415, meaning that 415 of total 500 “votes” are needed to assign “1” to a particular cell as predicted presence of P. crassifolia.

3.4. Assessment of the model performance

Several methods have been used to assess performance of models, and the widely used ones include overall classification rate, sensitivity, specificity, precision, Cohen’s kappa index, and area under receiver operating characteristic curve (AUC) (Liu et al. 2005). Among them, overall classification rate (also accuracy, Ac) and kappa index (discrimination, K) are easy to compute and also quite effective (Elith et al. 2006, Godown and Peterson 2000, Manel et al. 1999) and we thus adopt these two methods to evaluate the performance of GARP model used in this study. Overall classification rate and kappa index can be defined as follows using the confusion matrix in Table 1.

\[ Ac = \frac{(a+d)}{(a+b+c+d)} \]  \hspace{1cm} (4)

\[ K = \frac{[(a+d)-(a+c)(a+b)+(b+d)(c+d)])/n]}{[n-(a+c)(a+b)+(b+d)(c+d))/n]} \]  \hspace{1cm} (5)
4. RESULTS AND DISCUSSION

4.1. Environmental variables selected

As mentioned above, the Maxent-supported evaluation reveals that four environmental variables accounted for 91.4% of the species distribution as a sum of their relative contribution and they were therefore used as predictors to run GAPR model. The four variables are mean temperature of the warmest quarter (36.9%), precipitation of the wettest quarter (28.5%), annual solar radiation (24.2%), and topographic wetness index (12.8%). It was reported the distribution of *P. crassifolia* in Qilian Mts. is affected by temperature in July, annual precipitation and aspect (Zhao et al. 2010). Aspect was used for an appropriate surrogate of solar radiation (Grayson et al. 1997, Gómez-Plaza et al. 2001). Therefore environmental variables selected are consistent with other research results.

With validation at 27 stations, the Worldclim-extracted data are acceptable. The correlation coefficients are 0.63 for the mean temperature of the warmest quarter and 0.66 for the precipitation of the wettest quarter.

4.2. Predicted potential distribution

Based on GARP model and important variables selected in the study, the potential distribution of *P. crassifolia* in Qilian Mts. was estimated. In order to explore the spatial heterogeneity in the potential distribution of *P. crassifolia* resulted from the variations in elevations and in geographic locations (i.e., longitude and latitude) and to detect the extent of human disturbance, the study area (i.e., the Qilian Mts. within Gansu Province) was divided into four parts (A, B, C, and D in Fig. 3). As shown in Fig. 3, GARP model failed to predict any potential distribution of *P. crassifolia* in Part A. But, small patches of *P. crassifolia* were actually observed along a NW-SE strip of Part A (see Fig. 3). The failure may be attributed to two reasons. First, the presence samples from Part A are not

---

**Table 1. Confusion matrix used in distribution models of *P. crassifolia***

<table>
<thead>
<tr>
<th></th>
<th>Recorded present</th>
<th>Recorded absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted present</td>
<td>a (true positive)</td>
<td>b (false positive)</td>
</tr>
<tr>
<td>Predicted absent</td>
<td>c (false negative)</td>
<td>d (true negative)</td>
</tr>
</tbody>
</table>

Element *a* represents known distributional areas correctly predicted as present. *d* reflects regions where the species has not been found and correctly predicted as absence. Element *c* are pixels of known distribution but predicted as absence. *b* represents those areas the species is found to be absence but predicted as presence.
sufficient for training the model. Second, the spatial annual precipitation for the entire Part A is < 270 mm, being below the minimal threshold for *P. crassifolia* to grow (Zhao *et al.* 2006).

GARP model performed well at Part B (Fig. 3) as demonstrated by Ac and Kappa indexes which are 0.88 and 0.65, respectively. The high performance indices (especially K) further confirm the acceptability of the modeled results. Ac and Kappa index are as high as 0.93 and 0.71 at Part C. The accuracy (Ac) and Kappa index are 0.83 and 0.49 at Part D.

4.3. Estimated deforested area

Comparing the potential distribution of *P. crassifolia* with the present distribution of the species, we can obtain the deforested area (Fig. 3). We also calculated the ratio between the observed distribution and the predicted distribution of *P. crassifolia*. The ratios of the observed distribution to the predicted distribution are 51.4% in Part B, 19.1% in Part C, 44.4% in Part D. That is, about 50% of *P. crassifolia* has been deforested in Part B, nearly 80% of *P. crassifolia* forests has been deforested in Part C, over 50% of *P. crassifolia* forests has been deforested in Part D.

To sum up, the prediction shows that the most intensive deforestation occurred in Part C, the upper reach of the Heihe River basin. The forest land has been transformed to farmland in the southeastern part and grassland in the northwestern part (Zhao *et al.* 2010). About a half of the forests (i.e., *P. crassifolia*) has been deforested in Part B where most part of forest land has been changed into grassland. Over 50% of the forest land has been used as farmland in Part D, because of the combination of temperature and precipitation suitable to crop growth. The calculated overall ratio (0.39) of the observed distribution to the predicted distribution suggests that 61% (2869 km²) of *P. crassifolia* forests have been deforested.

4.4. Projected niche spaces

The ranges of the four environmental variables in the distribution area of *P. crassifolia* may differ from those in the whole study area. On Fig. 4, dark gray histogram represents variables of the localities where the species occurs, while the light gray histogram is computed for the Qilian Mountains National Nature Reserve. As shows in Fig. 4, the mean temperature in the warmest quarter is 14.7 °C in the study area, but mean of the variable is 12.5°C in the distribution area of the species (embraced by two shortest solid lines in A of Fig. 4). 68% of *P. crassifolia* potentially distributes at cells with mean temperature of the warmest quarter ranging from 10.7 to 15.1°C (range embraced by dashed lines in A of Fig. 4), and 95% of *P. crassifolia* potentially distributes at cells where temperature of the warmest quarter is ranging from 9.0 to 16.3°C. Precipitation of the wettest quarter has a mean value of 180 mm throughout the entire study area, while the mean of this variable is 191 mm in the distribution area of the species. The range embraced by dashed lines (i.e., 176–207 mm, B of Fig. 4) consists of 68% *P. crassifolia* potential distribution, and 95% of *P. crassifolia* distributes in the range embraced by solid lines (i.e., 163–221 mm, solid lines embraced range in B of Fig. 4) of precipitation of the wettest quarter. The annual mean solar radiation is 1413×10³ wh m⁻² in the study area, whereas the mean of the variable is 753×10³ wh m⁻² in the distribution area of the species. The values 510–870×10³ wh m⁻² consists the range within which 68% of *P. crassifolia* potentially distributes (range embraced by short solid lines in C of Fig 4), 95% of target species distributes at annual solar radiation ranging from 372 to 1083×10³ wh m⁻² (range embraced by dashed lines in C of Fig. 4). The mean topographic wetness index is 1.1 in the study area, but in the distribution area of the species, the mean of this variable is 1.9. The 68% of *P. crassifolia* potentially distributes at cells with topographic wetness index ranging from 0.8 to 3.3 (range embraced by dashed lines in D of Figure 4), and 95% of *P. crassifolia* potentially distributes at cells where topographic wetness index ranges from 0.3 to 4.7 (range embraced by solid lines in D of Fig. 4). It should be noted that the modeled distribution area of *P. crassifolia* is the area on which all four environmental variables (mean temperature of the warmest quarter, and precipitation of the wettest quarter, annual solar radiation, topographic wetness index) are suitable for *P. crassifolia*.
Several conclusions can be drawn from this study. First, GARP model produces acceptable predictions except for Part A in the study area. Second, the comparison between the observed distribution and the predicted distribution suggests that 61% (2869 km²) of *P. crassifolia* forests have been deforested, and restoration of the 61% *P. crassifolia* forests holds great potentials for maintaining the stability of the ecosystems, for regulating the hydrological cycles, and even for sustaining the socioeconomic systems. Third, the projected niche space for the first variable (i.e., mean temperature of warmest quarter) is between 8.5 and 18.1°C; the space for the second variable (i.e., precipitation of the wettest quarter) – between 149 and 245 mm; the space for the third variable (i.e., annual mean solar radiation) – between 118×10³ and 1100×10³ wh m⁻² and the space for the fourth variable (i.e., topographic wetness index) is between −0.4 and 5.1. Fourth, the accuracy of the predictions can be improved if higher resolution environmental variables become available. For example, the climate data layers can be improved through incorporating high-resolution DEM data (i.e., 30×30 m) into the spatial interpolation. The net radiation and the topographic wetness index can be greatly improved through incorporating high-resolution DEM data (i.e., 30×30 m) into the spatial interpolation. Fifth, the prediction accuracy needs to be improved with high-resolution data for providing practical guidelines to carry out the ongoing reforestation plans.

We concluded that models that predict species’ potential distributions by combining known occurrence records with digital layers of environmental variables have much potential for application in conservation science.
including predicting potential impacts of climate change, supporting conservation priorities and reserve selection, predicting the spread of invasive species and guiding field surveys to discover new species.

ACKNOWLEDGEMENTS: This project was supported by Nation Natural Science Foundation of China (No. 30770387, No. 40671067, No. 91025015); National Environmental Protection Commonwealth Project of China (NEPCP 200809098).

6. REFERENCES


Received after revision January 2012