Range Mgmt. & Agroforestry 36 (2) : 146-150, 2015
ISSN 0971-2070

Predictive distribution models for determination of optimal threshold of plant species in central Iran

Hossein PiriSahragard[1], Mohammad Ali ZareChahouki[2]* and H. Gholami[3]
[1]Range and Watershed department, University of Zabol, Iran
[2]Department of Rehabilitation of Arid and Mountainous Regions, University of Tehran, Iran
[3]University of Hormozgan, Iran
*Corresponding author e-mail: mazare@ut.ac.ir
Received: 30th October, 2014 Accepted: 5th November, 2015

Abstract
The aim of this study was determination of optimum threshold for predictive distribution models of plant species. For this purpose, vegetation sampling was carried out using random-systematic method. The plot size and sample sizes were determined using minimal area and statistical methods respectively. For sampling the soil at each habitat, eight holes was drilled and samples were taken from 0 to 30 and 30 to 80 cm depths. Plant distribution modeling was conducted using Logistic regression (LR), the Maximum entropy methods (MaxEnt) and Multi-layer perceptron of artificial neural networks (ANN). Threshold optimum was determined using sensitivity-specificity equal and maximum sensitivity approaches. Results indicate that in the LR model, *Seidlitzia rosmarinus* model was the poorest model (opp=0.3). However, the *Artemisia sieberi* model is the most accurate one (opp=0.7). The poorest and strongest of MaxEnt models were related to *Halocnemum strobilaceum* (opp=0.1) and *Seidlitzia rosmarinus* (opp=0.3). The poorest and most powerful models of ANN with 0.4 and 0.8 discrimination ability related to *Seidlitzia rosmarinus* and *Tamarix passerinoides* habitats respectively.

Keywords: Maximum sensitivity, Optimal threshold, Predictive distribution models, Sensitivity-specificity

Abbreviations: ANN: Artificial neural networks; **GLM:** Generalized linear model; **LR:** Logistic regression **MaxEnt:** Maximum entropy

Introduction

Predicting species distributions is becoming increasingly important since it is relevant to resource assessment, environmental conservation and biodiversity management (Fielding and Bell, 1997; Manel *et al*., 1999). Many modeling techniques such as generalized linear models (GLM), Maximum entropy (MaxEnt) and artificial neural networks (ANNs) have been used for this purpose (Guisan and Zimmermann, 2000; Moisen and Frescino, 2002; ZareChahouki *et al*., 2012). GLM is a logistic model, which uses a log it link to describe the relationship between the response and the linear sum of the predictor variables. MaxEnt is a maximum entropy based machine learning program that estimates the probability distribution for a species occurrence based on environmental constraints using only species presence data (Phillips *et al*., 2006). One of the robust rule-based modeling approaches which are used in bioclimatic envelope modeling is artificial neural networks techniques (Thuiller, 2003; Liu *et al*., 2005).

These modeling techniques often generate continuous maps that are useful for many conservation applications (Araujo *et al*., 2002; Wilson *et al*., 2005). A threshold or cut-off probability is needed to transform the probability or suitability data to presence/absence data. Besides that the determination of the threshold is needed when assessing model performance using the indices derived from the confusion matrix (Manel *et al*., 2001).

There are many approaches for determining thresholds, which fall into two categories: subjective and objective. A representative in the first category is taking 0.5 as the threshold, which is widely used in ecology (Manel *et al*., 1999; Stockwell and Peterson, 2002). In the objective approaches, thresholds are chosen to maximize the agreement between observed and predicted distributions. In these approaches, Kappa maximization approach is popular in ecology (Guisan *et al*., 2002; Liu *et al*., 2005).

The present study compare two different approaches including sensitivity-specificity equality and maximum sensitivity approaches for determining threshold with aim

to find the optimum threshold criteria for the plant distribution models.

Materials and Methods

***Study area*:** The study area is located in the central part of the Qom province in geographic coordinates area 50' 50° 30'' to 50' 54° 30'' E and 34 '59° 30'' to 35' 03° 30'' N. This region is located in west Qom city and covers an area of 3,000 hectares. The location of the study area in Iran and Qom province having minimum and maximum altitude of 796 and 1100 meters above sea level respectively is shown in figure 1.

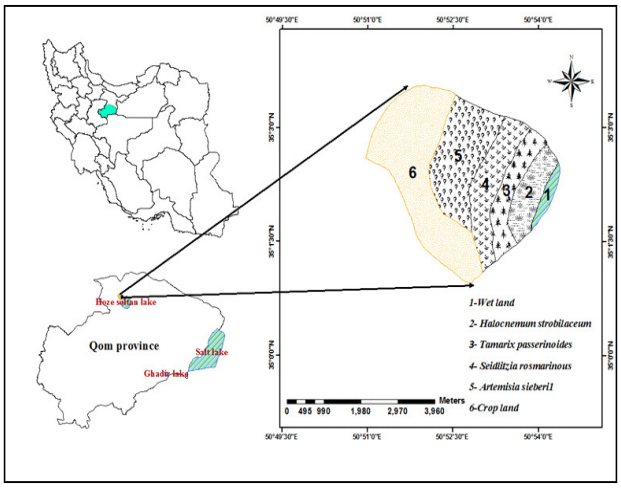

Fig 1. General location and vegetation types map of the study area

***Species and environmental data*:** After determination of homogeneous units using basic maps of the study area (digital elevation, aspect, slope and geology maps, scale 1:25000), vegetation sampling was carried out using random- systematic method in the plots which were established along four transect with 200-1000 m lengths. The plot size was determined using minimal area method from 2 to 25 m^2. Sample size was determined for 60 plots with respect to vegetation cover variations using statistical method. In order to sample the soil for each habitat eight holes was drilled and samples were taken from 0-30 and 30-80 depths. After sampling, soil characteristics consisting gravel per cent, texture, saturation moisture, available water, lime, gypsum, organic matter, acidity (pH), electrical conductivity (EC) and soluble solute (Na^{+}, Ca^{2+}, Mg^{2+}, K^{+}, Cl^{-}, $CO3^{2-}$, HCO^{3-} and $SO4^{2-}$) were measured by routine methods (Table 1). Using geostatistical and Kriging interpolation method with the same spatial resolution (pixel size 30*30 meters), soil digital layers were prepared. Arc GIS 9.3 and GS + Version fifth software were used for mapping soil properties.

Table 1. List of variables in the data set

Variable	Code	Unit	Mean± Standard deviation
Elevation	abs	M	790± 15
Slope	slope	%	5±0.2
Gravel	gr	%	9.85±1.86
Clay	caly	%	13.81±7.02
Silt	silt	%	32.34±10.16
Sand	sand	%	54.89±10.25
Saturation moisture	s m	%	38.34±5.09
Available water	A.W.	%	19.10±5.26
Gypsum	gy	%	3.31±1.34
Organic matter	OM.	%	0.57±0.27
Lime	Lime	%	7.01±0.48
pH (acidity)	pH	-	7.23±0.17
ECe	EC	ds/m	97.49±26.79
Sodium ion (Na^{+})	Na	meq/l	647.46±45.99
Potassium ion (K^{+})	K	meq/l	4.98± 0.58
Calcium ion (Ca^{2+})	Ca	meq/l	316.22± 11.46
Magnesium (Mg^{2+})	Mg	meq/l	99.88± 13.67
Chlorine (Cl^{-})	Cl	meq/l	831.75±45.61
Carbonate (CO_3^{2-})	Co	meq/l	1.4±0.068
Bicarbonate (HCO^{-}_3)	HCO_3^{-}	meq/l	9.35±1.82
Sulfate (SO_4^{2-})	SO_4	meq/l	235.46± 24.39

***Design of modeling process*:** Modeling vegetation was performed using LR, MaxEnt and ANN. In order to apply LR procedure, initial multi collinearity between variables was assessed and variables with variation inflation factors higher than 5 were eliminated (Cawsey *et al*., 2002). Then, relationships were extracted using SPSS version 18 and the model obtained was assessed using Hosmer and Lemshow statistic. In the MaxEnt procedure, we used the area under the receiver operating characteristic function (AUC) for evaluation of the discrimination ability (Fielding and Bell, 1997). For ANN modeling, the current study applies tangent sigmoid transfer function and Levenberg–Marquardt learning rule. The out-put layer neurons use of linear transfer functions and training network method is back propagation error.

***Model assessment indices and threshold determination approaches*:** Many indices can be used in the assessment of the predictions of species distributions, including sensitivity, specificity and Kohen's kappa statistic (Fielding and Bell, 1997; Liu *et al*., 2005). In this study, sensitivity-specificity equality and maximum sensitivity approaches were used that are popular in ecology (Cantor *et al*., 1999; Lehmann, 1998; Guisan *et al*., 1998).

Results and Discussion

Logistic regression statistics are given in Table 1 regarding determination of the coefficient and Hosmer–Lemeshow (HL) statistics. The obtained equations are significant at one percent level. Overall assessments of models were done and comparisons were made between ANN, MaxEnt models and LR in terms of classification accuracy. Results indicate that LR model for the *S. rosmarinus* alliance was the poorest model because none of the probability of presence values exceeded 0.3. However, the proposed model for *A. sieberi* is the most accurate model in order to classification of the presence or absence of species in this habitat (Sensitivity 0.7). Considering the 0.1 optimal threshold the poorest maximum entropy model is related to *H. strobilaceum* since discrimination ability of the model has not exceeded 0.5. The poorest and most powerful model of artificial neural network with 0.46 and 0.82 discrimination ability was related to *S. rosmarinus* and *T. passerinoides* habitats respectively (Table 2). An optimal threshold for each model, with an emphasis on ensuring sensitivity is relatively high, is provided using the ROC plot (Fig. 2). Determining the optimal threshold of plant species, can increase accuracy of predictive maps and validity of the results obtained from models. Although accuracy of prediction models is sensitive to threshold criteria applied in model derivation. (Freeman and Moisen, 2008).

Table 2. Logistic regression statistics for presence of plant species

Vegetation type	R^2	HL
H. strobilaceum (Ha. st)	0.840	1.00
T. passerinoides (Ta. pa)	0.727	0.89
S. rosmarinus (Se. ro)	0.815	0.99
A. sieberi (Ar. si)	0.771	0.97

Hosmer-Lemeshow (HL) statistics used to verify conformity between the observed and expected numbers of cases and higher values indicating greater conformity.

With due consideration to the key points, and importance of the kappa values in management decisions in relation to the management and modification of vegetation in rangelands, appropriate threshold cut-offs should be chosen in light of the intended use of the species distribution maps (Freeman and Moisen, 2008). In other words, the degree to which these errors are minimized depends on how the model will be used (Loiselle *et al*., 2003; Rondinini *et al*., 2006). Pirisahragard *et al.* (2015) determined the presence optimal threshold of plant species using sensitivity-specificity equal and concluded that modeling result can be used with greater confidence when the optimal threshold be determined.

In present study, due to the low prevalence values of plant species in some of the studied habitats, the sensitivity of the model in some habitat such as *H. strobilaceum*, *T. passerinoides* and *S. rosmarinus* is reduced. In confirmation of this finding, it has been reported that in cases where the purpose of the modeling is to identify all experimental sites which is suitable for certain species, then the best results were obtained from thresholds deliberately chosen so that the predicted prevalence equaled the observed prevalence. Hence it is necessary to develop a map with 99% sensitivity. In other words, the point is considered as the threshold at which the model sensitivity is maximum (Fielding and Bell, 1997; Miller and Franklin, 2002; Freeman and Moisen, 2008).

In cases where the two curves do not intersect, the high sensitivity of the model is more important and the level of probability which has maximum of sensitivity is considered as threshold (Fielding and Bell, 1997). Manel *et al.* (2001) examined a large set of species and concluded that results from a threshold which maximized the sensitivity–specificity sum were superior to results from a threshold of 0.5(Zweig and Campbell, 1993). Further, species with low prevalence or low model quality such as *H. strobilaceum* and *S. rosmarinus* were most sensitive to the choice of threshold and traditional default method such as 0.5 cutoff is unreliable, sometimes resulting in substantially lower kappa, with possible detrimental effects on a management decision. Thus, in these species the maximum sensitivity of the model was considered as the optimal threshold. This finding is consistent with studies of Fielding and Bell (1997), Miller (2005) and Freeman and Moisen (2008).

Conclusion

The modelers should take utmost care regarding the purpose of modelling, model quality and prevalence of the studied species. Optimal threshold of plant species should be determined using suitable objective methods based on the model quality species. The study demonstrated plant distribution modelling using different techniques. The performances were compared depending on quality and accuracy of the models and their applicability, which indicated that that *Artemisia sieberi* model was the most accurate model (opp=0.7). The poorest and most powerful models of ANN with 0.4 and 0.8 discrimination ability related to *Seidlitzia rosmarinus* and *Tamarix passerinoides* habitats respectively, whereas the poorest and strongest of MaxEnt models were related to *Halocnemum strobilaceum* (opp=0.1) and *seidlitzia rosmarinus* (opp=0.3).

Table 3. Optimum probability threshold and sensitivity/specificity for all models based on test data

Model	LR			MaxEnt			ANN		
Vegetation type	Optimum probability	Sensi-tivity	Speci-ficity	Optimum probability	Sensi-tivity	Speci-ficity	Optimum probability	Sensi-tivity	Speci-ficity
Ha. st	0.3	0.45	0.87	0.1	0.50	0.99	0.6	0.80	0.97
Ta. pa	0.3	0.43	0.87	0.2	0.53	0.96	0.8	0.82	0.96
Se. ro	0.3	0.31	0.88	0.3	0.57	0.95	0.3	0.46	0.97
Ar. si	0.7	0.95	0.79	0.3	0.56	0.96	0.4	0.58	0.87

Fig 2. Plots of sensitivity, specificity, and total accuracy for different probability thresholds in the test data. Optimum thresholds were selected to dichotomize probability of presence maps. opp= Optimum probability of presence.

References

Araujo, M. B., P. H. Williams and R. J. Fuller. 2002. Dynamics of extinction and the selection of nature reserves. *Proceedings of the Royal Society London-Biological Series* B269, pp. 1971-1980.

Cantor, S. B., C. C. Sun., G. Tortolero-Luna, R. Richards-Kortum and M. Follen. 1999. A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. *Journal of Clinical Epidemiology* 52: 885-892.

Cawsey, E. M., M. P. Austin and B. L. Baker. 2002. Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodiversity and Conservation* 11: 39-74.

Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/ absence models. *Environmental Conservation* 24: 38-49.

Freeman, E. A and G. G. Moisen. 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological modelling* 217:48-58.

Guisan, A., T. C. Edwards and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157:89-100.

Guisan, A., J. P. Theurillat and F. Kienast. 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetable Science* 9: 65-74.

Guisan, A and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147–186.

Lehmann, A. 1998. GIS modeling of submerged macrophyte distribution using generalized addition models. *Plant Ecology* 139: 113-124.

Liu, C., P. M. Berry, T. P. Dawson and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28:385-393.

Loiselle, B. A., C. A. Howell., C. H. Graham., J. M. Goerck., T. Brooks., K. G. Smith and P. H. Williams. 2003. Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology* 17: 1591-1600.

Manel, S., J. M. Dias., S. T. Buckton and S. J. Ormerod. 1999. Alternative methods for predicting species distributions: an illustration with Himalayan river birds. *Journal of Applied Ecology* 36: 734-747.

Manel, S., H.C. Williams and S. J. Ormerod. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38: 921–931.

Miller J. 2005. Incorporating spatial dependence in predictive vegetation models: Residual Interpolation Methods. *The Professional Geographer* 57: 169 -184.

Miller J. and J. Franklin. 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Journal of Ecological Modelling* 157: 227-247.

Moisen, G.G. and T. S. Frescino. 2002. Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling* 157: 209-225.

PiriSahragard, H. and M. A. Zare Chahouki. 2015. An evaluation of predictive habitat models performance of plant species in Hozesoltan rangelands of Qom province. *Ecological Modelling* 309-310: 64–71.

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

Rondinini, C., K. A. Wilson., L. Boitani., H. Grantham and P. Possingham. 2006. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters* 9: 1136–1145.

Stockwell, D. R. B. and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148: 1-13.

Thuiller, W. 2003. Optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology* 9: 1353-1362.

Wilson, K. A., M. I. Westphal., H. P. Possingham and J. Elith. 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation* 122: 99–112.

ZareChahouki, M. A. and L. KhalasiAhvazi. 2012. Predicting potential distributions of *Zygophyllum eurypterum* by three modeling techniques (ENFA, ANN and logistic) in North East of Semnan Iran *Range Management and Agroforestry* 2: 68-82.

Zweig, M.H. and G. Campbell. 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 39: 561-577.