ASSESSING GARP MODELING AND EFFECT OF PLANT SAMPLE POSITION ON PREDICTING SUITABLE HABITAT OF *Brainea insignis*

Wen-Chiao Wang[a], *Nan-Jang Lo*[b], *Wei-I Chang*[c], and *Kai-Yi Huang*[*d]

[a]*Graduate student, Dept. of Forestry, Chung-Hsing Univ., Taiwan E-mail: chiao87219@yahoo.com.tw*

[b]*Specialist, EPMO, Chung-Hsing Univ., Taiwan E-mail: njl@dragon.nchu.edu.tw*

[c]D*irector, Hsinchu forest district office, Forest Bureau, Hsinchu 300, Taiwan E-mail: weii@forest.gov.tw*

[d*]*Professor, same as with author*-a *E-mail: kyhuang@dragon.nchu.edu.tw (corresponding author)*

250 Kuo-Kuang Road, Taichung, Taiwan 402, Tel: +886-4-22854663; Fax: +886-4-22854663

KEY WORDS: Cycad-fern, Decision Tree (DT), Genetic Algorithm for Rule-set Prediction (GARP), Discriminant Analysis (DA), Geographic Information System (GIS), Remote Sensing (RS).

ABSTRACT: Forestry has begun to use 3S technologies (RS, GIS, and GPS) in routine inventory work and scientific research. Modeling ecological pattern of species needs to utilize the combination of 3S technologies and statistics, and it has become an important part in ecology. The study was intended to predict the suitable habitat of cycad-fern in the Huisun Forest Station by using multivariate statistics coupled with a GIS. The ecological pattern of the species was examined by overlaying the layer of cycad-fern samples collected with GPS on the layers of topographic variables and vegetation index derived from SPOT-5 images. We also combined easting and northing coordinates of grid cell with topographic variables to improve the accuracy of predictive models. Three models, decision tree (DT), Genetic Algorithm for Rule-set Prediction (GARP), and discriminant analysis (DA), were developed and validated. Accuracy assessment results indicated that the accuracies of DT with easting and northing added respectively were much greater than those of both GARP and DA with easting and northing added respectively, and GARP with easting and northing added respectively was also better than DA. More importantly, the accuracies of DT with easting and northing added respectively were greatly improved, those of DA with easting and/or northing added respectively were slightly improved, and the opposite was true with GARP. Easting was more effective than northing in improving model accuracy because of east-west distribution with cycad-ferns in the Kuan-Dau watershed of the Huisun area. However, easting and northing predictor variables were found to limit the ability of spatial extrapolation with models and make predictive results look more artificial. We shall attempt to incorporate predictor variables extracted from high spatial resolution, hyperspectral data into models to improve the ability of spatial extrapolation with models in a follow-up study.

1. INTRODUCTION

With computer technology and 3S technologies (remote sensing, geographic information system, and global positioning system) rapid advancement, forestry has begun to use these useful tools in routine inventory work and scientific research. Modeling spatial distribution of species needs to utilize the combination of these tools, and it has become popular in the field of ecology (Guisan and Thuiller, 2005). Applications in forest have included gaining the species of unknown distributional areas and undiscovered species, predicting species invasions, finding disease and supporting conservation planning (Bourg *et al.*, 2005; Pearson *et al.*, 2007; Asner *et al.,* 2008; Lelong *et. al.,* 2010).

Modeling species distribution allows us to understand the spatial distribution of species, and also identify the relationship between the species and environment. Indirect factors of environment variables (e.g. elevation, slope, aspect) are most easily measured in the field and are often used because of their good correlation with observed species patterns (Guisan and Zimmermann, 2000). Nowadays a variety of multivariate statistical methods have been used to model ecological niches and predict the geographical distributions of species, such as generalized additive models (GAM), genetic algorithm for rule-set prediction (GARP), discriminant analysis (DA), decision tree (DT) (Lowell, 1991; Guisan *et. al.,* 2002; Bourg *et al.,* 2005; Lisa *et al.*, 2008; Ke *et al.*, 2010).

In our study, we used three modeling techniques: GARP, DT and DA. We chose cycad-ferns as our target species, because they are a rare species and have limited distribution in Taiwan. The objective of the study was to predict the potential habitat of the species in the Huisun Forest Station in central Taiwan. Elevation, slope, aspect, terrain position, and vegetation indices derived from SPOT-5 images were accounted for in vegetation habitat evaluation and unknown site search. We also combined easting and northing coordinates of grid cell with topographic variables to improve the accuracy of predictive models and want to know the effect of position on predictive accuracy. The study developed the models that related known plant sites to habitat characteristics and extrapolated the plant’s potential sites in the study area. The study evaluated these models in terms of accuracy and implementation efficiency and determined the optimum for predicting the habitat of cycad-ferns.

2. STUDY AREA

The Huisun Forest Station is in central Taiwan (Figure1), situated within 24°2′–24°5′ N latitude and 121°3′–121°7′ E longitude. This station is the property of Chung-Hsing University, and has a total area of 7, 477 ha. The study area ranges in elevation from 454 m to 3, 419 m, and its climate is temperate and humid. Hence, the study area has nourished many different plant species and is a representative forest in central Taiwan. This study using samples from Sihwufongshan, Duhchuanling and Kuandaushan in Huisun, Sihwufongshan elevation from 680 m to 840 m, the highest elevation of Duhchuanling approximately 810 m, and Kuandaushan elevation approximately 760 m. According to the climate record of this study area, the annual mean warm is 21.0℃; the monthly mean warm highest is 30.6℃ in July, lowest is 20.5℃in January; mean annual precipitation 2453.5 mm, the average relative moisture is 85%.

Cycad-fern, *Blechnaceae* family, is only found in mountains in central Taiwan, such as Huisun Experimental Forest Station and Tong-Mao Mountain areas, and Huisun is the main area of growth. Because of its narrow distribution of limited ecological range, cycad-fern has been categorized as one of the rare species (Lu *et al*, 1986).

There were 202 cycad-fern samples collected from Sihwufongshan, Duhchuanling and Kuandaushan by GPS in this study, but a part of these samples remained after data integration because some cycad-ferns fall within the same pixels with others, respectively. The two-thirds of the dataset, 123 samples, were used for model development (training) and the remaining, one-third of the dataset, 59 samples were used for model validation (test).

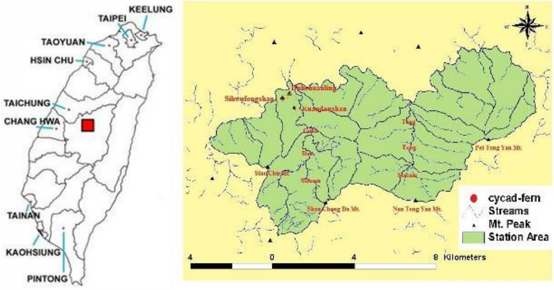

Figure1. Location map of study area.

3. MATERIALS AND METHODS

3.1 Data Collection

Digital elevation model (DEM) with grid size 5 × 5 m, orthophoto base maps (1:10,000), and nine-date SPOT images were collected. *In situ* data (cycad-fern samples) were also acquired by using a GPS linked with a laser range. Two-date SPOT images (07/10/2004 and 11/11/2005) were chosen because the two-date images have the best quality with the least amount of clouds among the nine-date SPOT images.

3.2 Data Processing

Slope, aspect, easting and northing data layers were generated from 5 × 5 m DEM. The ridges and valleys in the study area were used together with DEM to generate terrain position layer. The main ridges and valleys over the study area were directly interpreted from the contour lines shown on the orthophoto base maps; these lines were then digitized to establish the data layer of main ridges and valleys by using ARC/INFO software for later use. The data layer of main ridges and valleys in a vector format was converted into a new data layer in a raster format by ERDAS Imagine software, and then combined with DEM to generate terrain position layer (Skidmore, 1990). Vegetation indices were derived from the two-date SPOT-5 images, one in autumn, the other in summer, by using Spatial Modeler of ERDAS Imagine. Cycad-fern samples obtained by GPS were converted into ArcView shapefile format for later use.

3.3 Database Building

The GIS database used in the study was constructed by using ERDAS Imagine software module Layer Stack to overlay elevation, slope, aspect, terrain position, vegetation index, easting, and northing layers. The cycad-fern sample layer was overlaid with seven data layers, and those pixels of the seven layers lying at the same position with the cycad-fern pixels clipped out. To build statistical models, the sample data for both target groups (cycad-fern) and non-target groups (background) were taken from data layers by the random sampling to minimize spatial autocorrelation in the independent variables (Pereira and Itami, 1991). Because non-target sites (background) correspond to the vast majority of the study area, larger variation is expected in environmental characteristics for this group. The number of non-target pixels (sites) should be three times more than that of target pixels to increase the probability of acquiring a more representative sample of the habitat characteristics at non-target sites (Pereira and Itami, 1991; Sperduto and Congalton, 1996).

3.4 Model Development

The predictive models for selecting potential habitat of the trees were created using three statistical methods: (1) genetic algorithm for rule-set prediction, (2) decision trees, and (3) discriminant analysis. Model development and validation can be done by cross-validation (it is called split-sample validation). Split-sample validation can be implemented by dividing a dataset into two subsets, the first one (training data) typically comprising one-half to two-thirds of all data and the other (test data) comprising one-third to one-half of all data. The first one is used to build and test a model. The other one (an independent dataset) is just used to test the model, not used to build the model. Three models were implemented by using SPSS software package in this study.

3.4.1 Genetic Algorithm for Rule-set Prediction

Genetic algorithm for rule-set prediction has recently seen extensive use only in present studies. It seeks a collection of rules that together produce a binary prediction (Phillips *et al.*, 2006). GARP uses a set of point position records of species presence and a set of environmental layers that might limit the species' capabilities to survive. The model will use genetic algorithm to search heuristically for a good rule-set. There are four rules available currently in GARP software (DesktopGARP): atomic, logistic regression, bioclimatic envelope, and negated bioclimatic envelope rules, it uses the rules to search the correlation between species presence and absence and environmental variables for predicting suitable conditions for each pixel (Stockwell and Noble, 1992). GARP software is freely available on the worldwide web (http://www.nhm.ku.edu/desktopgarp/Download.html), named "DesktopGARP." It repeats times of statistical calculation based on runs set by user, and each of runs would generate a predictive distribution map. Part or all of maps generated from DesktopGARP would be overlaid, and then the integrative map could be generated according to the times of that each pixel was predicted as target. The preconditions of rules in GARP are simple conjunctive expressions:

$$V_1=v_1\&V_2>v_2\&\ldots\&V_m=(v_{m1},v_{m2}) \qquad (1)$$

where $v_1,v_2,...,v_m$ are values of the variables $V_1,V_2,\ldots V_M$. The variables 1 to m are a subset of the total number of variables, i.e. they are not repeated. The precondition selects a subset of data set. The conclusion is an assignment of a classification value to the selected subset.

3.4.2 Decision Tree

Decision trees are a sequential partitioning of the dataset in order to maximize differences on a dependent variable. Decision pathways originate from a starting node (root) that contains all observations and end at multiple nodes containing unique subsets of observations. Terminal nodes are assigned a final outcome based on group membership of the majority of observations (De'ath and Fabricius, 2000; Bourg *et al.*, 2005; O'Brien *et al.*, 2005). CART (Classification and Regression Trees) was used and it was implemented by using SPSS software package in this study. In theory there are several functions, but Gini splitting rule is most broadly used rule. If that the data set *S* contains *n* classes, *Gini* is defined as:

$$Gini(S)=1-\sum_{j=1}^{n} p_j^2 \qquad (2)$$

where P_j is the probability of j class in *S* dataset.

3.4.3 Discriminant Analysis

Discriminant analysis is a technique, which discriminates among *k* classes (objects) based on a set of independent or predictor variables. The objectives of DA are to (1) find linear composites of *n* independent variables which maximize among-groups to within-groups variability; (2) test if the group centroids of the *k* dependent classes are different; (3) determine which of the *n* independent variables contribute significantly to class discrimination; and (4) assign unclassified or "new" observations to one of *k* classes (Lowell, 1991). DA was implemented by using SPSS software package in the study. The variates for a discriminant analysis, also known as the discriminant function takes the following form:

$$Y_{jk}=\alpha+\beta_1X_{1k}+\beta_2X_{2k}+\ldots+\beta_nX_{nk} \qquad (3)$$

where

Y_{jk} = discriminant *Y* score of discriminant function *j* for object (class) *k*

α = intercept

β_i = discriminant weight for independent variable *i*

X_{ik} = independent variable *i* object (class) *k*

3.5 Model Validation

Model validation can be done by split-sample validation, as mentioned previously. For each model, predict the response of the remaining data, and calculate the error from the predictions and the observed values (De'ath and Fabricius, 2000). We also used overall accuracy and *kappa* coefficient to assess models, because overall accuracy only include the data along the major diagonal and excludes the errors of omission and commission, *kappa* incorporates the non-diagonal elements of the error matrix as a product of the row and column marginal (Lillesand *et al*., 2008).

4. RESULTS AND DISCUSSION

As compared the statistics of five predictor variables for the entire study area and cycad-fern sites, it was found that cycad-ferns had preference for elevation ranging from 728 to 916 m; slope was less than 69°, the mean slope was about 32°. The mean terrain position was 6, median was 7, and this indicates that the area may receive enough sunlight for cycad-fern growth.

Table 1 shows the accuracies of three predictive models with different combinations of predictor variables for predicting the potential habitat of cycad-ferns. By comparing the accuracies of the C1 and C3, vegetation index was found not able to improve the model performance significantly. This is because the spectral resolution and spatial resolution of SPOT imagery are not enough to discriminate cycad-fern from other plants. The same results were found with comparison between C1 and C2; that is, aspect variable could not improve the model performance significantly. Furthermore, we used elevation, slope and terrain position as our base models of predictor variables, and added easting and northing to know how they affected model prediction (table 2). After adding easting and northing, overall and *kappa* also increased, in DT model *kappa* increased by 0.13 (from 0.84 to 0.97) and GARP *kappa* increased by 0.069 (0.87 to 0.94), DA model *kappa* increased by 0.01 (from 0.62 to 0.63). Easting was more effective than northing in improving model accuracy because of east-west distribution with cycad-ferns in the Kuan-Dau watershed of the Huisun area. However, since the positions of predictive habitat were clustered around field locations, this gave rise to the possibility of exaggerating validation statistics.

Table 1 Accuracies of three predictive models with different combinations of predictor variables for predicting the potential habitat of cycad-ferns.

Class		GARP			DT			DA		
		C1	C2	C3	C1	C2	C3	C1	C2	C3
Training	Non-habitat (%)	95	94	94	99	98	98	84	83	85
	Habitat (%)	99	89	96	92	94	94	98	98	97
	Overall (%)	95	93	95	98	97	97	86	86	87
	Kappa	.87	.79	.85	.92	.92	.90	.68	.67	.79
Test	Non-habitat (%)	90	91	91	96	96	96	81	81	80
	Habitat (%)	98	92	97	93	93	93	98	98	98
	Overall (%)	91	91	92	96	95	95	84	83	82
	Kappa	.77	.75	.79	.84	.82	.84	.62	.61	.76

C1: Elevation, Slope, TP; C2: Elevation, Slope, Aspect, TP; C3: Elevation, Slope, TP, Vegetation Index; TP: Terrain Position

Table 2 Accuracies of three predictive models with added easting and northing for predicting the potential habitat of cycad-ferns.

Class		GARP			DT			DA		
		C4	C5	C6	C4	C5	C6	C4	C5	C6
Training	Overall (%)	98	99	93	99	99	99	87	87	87
	Kappa (%)	.94	.97	.82	.98	.96	.98	.67	.66	.68
Test	Overall (%)	96	97	90	99	98	98	85	86	85
	Kappa (%)	.87	.92	.73	.97	.94	.93	.63	.65	.61

C4: Elevation, Slope, TP, easting, northing; C5: Elevation, Slope, TP, easting; C6: Elevation, Slope, TP, northing; TP: Terrain Position

Table 3 Distribution statistics of predictive maps generated from two models with three variable cases (elevation, slope, and terrain position).

Class	GARP		DT	
	Area (ha)	%	Area (ha)	%
Habitat	998	5.8	248	1.4
Non-Habitat	16138	94.2	16,888	98.6
Sum	17136	100.0	17136	100.0

We also used elevation, slope and terrain position to predict cycad-fern's potential habitat. Because *kappa* and overall accuracy of DT and GARP model were much higher than those of DA, we just show the potential habitat of GARP and DT models for a comparison purpose. DT greatly reduced the area of potential habitat to less than 2% of the entire study area. GARP also reduced the area of potential habitat of 6% of the entire study area (table 3). After adding easting and northing of predictor variables, GARP reduced the area of potential habitat to less than 3%, and added easting had more influence than did added northing. DT had the same results as GARP (table 4 and 5). However, added easting and northing could limit the potential habitat to a small range around the positions of field samples (figure 2); it would become a limiting factor at a large scale. Thus, it would need more samples to predict the potential habitat at a large scale when easting and northing of predictor variables are used.

Table 4 Distribution statistics of predictive maps generated from GARP model with added easting and northing of predictor variables.

Class	C4		C5		C6	
	Area (ha)	%	Area (ha)	%	Area (ha)	%
Habitat	537	3.1	273	1.6	1,534	9.0
Non-Habitat	16,599	96.9	16,863	98.4	15,602	91.0
Sum	17136	100.0	17136	100.0	17136	100.0

C4: Elevation, Slope, TP, easting, northing; C5: Elevation, Slope, TP, easting; C6: Elevation, Slope, TP, northing; TP: Terrain Position

Table 5 Distribution statistics of predictive maps generated from DT model with added easting and northing of predictor variables.

Class	C4		C5		C6	
	Area (ha)	%	Area (ha)	%	Area (ha)	%
Habitat	107	0.6	204	1.2	252	1.5
Non-Habitat	17029	99.4	16932	98.8	16884	98.5
Sum	17136	100.0	17136	100.0	17136	100.0

C4: elevation, slope, TP, easting, northing; C5: elevation, slope, TP, easting; C6: elevation, slope, TP, northing; TP: terrain position

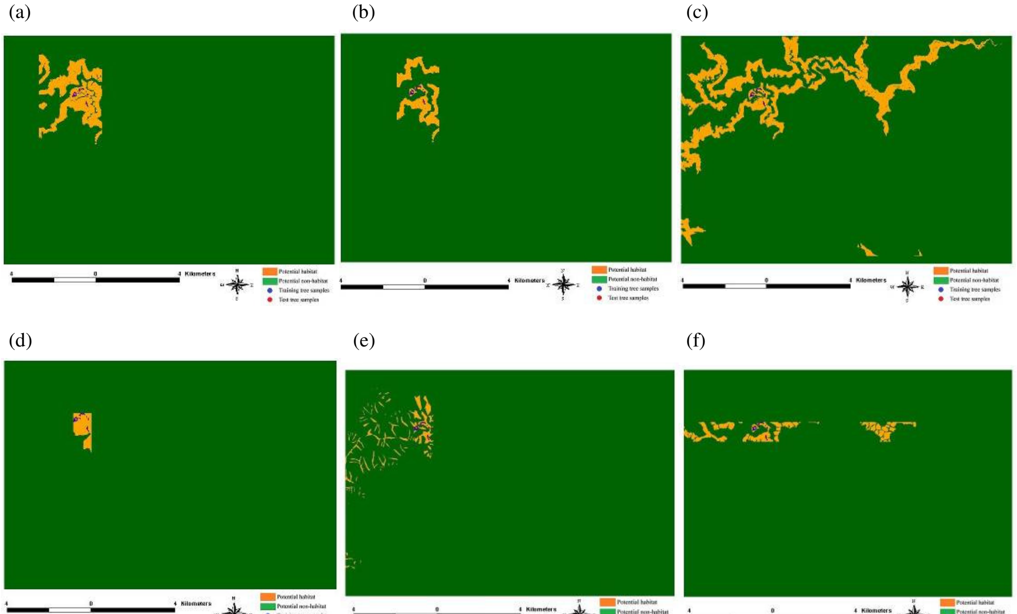

Fig. 2 Two models for mapping the potential habitat of cycad-ferns in the study area with added easting and northing of predictor variables. (a) GARP-C4, (b) GARP-C5 ,(c) GARP-C6; (d) DT-C4, (e) DT-C5, (f) DT-C6 (C4: elevation, slope, TP, easting, northing; C5: elevation, slope, TP, easting; C6: elevation, slope, TP, northing; TP: terrain position)

5. CONCLUSIONS

The study developed the three models of DT, GARP and DA that related known tree sites to habitat characteristics and extrapolated the plant's potential sites in the study area. The accuracy of the DT model was higher than those of the GARP and DA. Accuracy assessment results indicated that the accuracies of DT with easting and northing added respectively were much greater than those of both GARP and DA with easting and northing added respectively, and GARP with easting and northing added respectively was also better than DA. More importantly, the accuracies of DT with easting and northing added respectively were greatly improved, those of DA with easting and northing added respectively were slightly improved, and the opposite was true with GARP. Easting was more effective than northing in improving model accuracy because of east-west distribution with cycad-ferns in the Kuan-Dau watershed of the Huisun area. However, easting and northing predictor variables were found to limit the ability of spatial extrapolation with models and make predictive results look more artificial. We shall attempt to incorporate predictor variables extracted from high spatial resolution, hyperspectral data into models to improve the ability of spatial extrapolation with models in a follow-up study.

The results show that the vegetation indices derived from SPOT-5 satellite images could not improve model accuracy for widely distributed tree species due to the limitations of spectral resolution and spatial resolution with SPOT-5 imagery. Airborne hyperspectral data and LIDAR data will be used in a follow-up study so that the model accuracy can be improved. Also we shall add more data samples taken from Tong-Mao Mountain about 10 km from the Huisun study area to test model accuracy and reliability.

REFERENCES

Asner, G. P., M. O. Jones, R. E. Martin, D. E. Knapp and R. F. Hughes, 2008. Remote sensing of native and invasive species in Hawaiian forests. Remote Sensing of Environment 112, pp.1912–1926.

Bourg, N. A., W. J. Mcshea and D. E. Gill, 2005. Putting a CART before search: successful habitat prediction for a rare forest herb. Ecology, 86 (10), pp.2793–2804.

De'ath, G. and K. E. Fabricius, 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. ERcology, 81 (11), pp.3178–3192.

Phillips, S. J., R. P. Anderson and R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions, Ecol. Model, 190, pp. 231–259

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

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

Guisan, A. and W. Thuiller, 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8, 993–1009.

Ke, Y., L. J. Qackenbush, J. Im., 2010. Synergistic use of QuickBird multispectral imagery and LIDAR data for object-based forest species classification. Remote Sensing of Environment, (114), pp.1141–1154.

Lelong C. C. D., J. M. Roger, S. Brégand, F. Dubertret, M. Lanore, N. A. Sitorus, D. A. Raharjo and J. P. Caliman, 2010. Evaluation of Oil-Palm Fungal Disease Infestation with Canopy Hyperspectral Reflectance Data. *Sensors 10*, pp.734–747.

Lillesand, T. M., R. W. Kiefer and J. W. Chipman, 2008. Remote Sensing and Image Interpreatation-6th ed, WILEY, pp.591.

Lisa, L., R. L. Lawrence, S. Podruzny and C. C. Schwartz, 2008. Mapping Regional Distribution of a Single Tree Species: Whitebark Pine in the Greater Yellowstone Ecosystem, Sensors 8, pp.4983–4994.

Lowell, K., 1991. Utilizing discriminant function analysis with a geographical information system to model ecological succession spatially. International Journal of Geographical Information System, 5 (2), pp.175–191.

Lu, J. C., Y. J. Liu and M. Y. Chen, 1986. Brainea insignis a new distribution and plant community composition in Taiwan. Quarterly of journal Chinese Forestry, 19 (1), pp. 121–126.

O'Brien, C. S., S. Rosenstock, J. J. Hervert, J. L. Bright and S. R. Boe, 2005. Landscape – level models of potential habitat for Sonoran pronghorn. Wildlife Society Bulletin, 33 (1), pp. 24–34.

Pearson, R. G., C. J. Raxworthy, M. Nakamura and A. T. Peterson, 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J. Biogeo 34, pp.102–117.

Pereira, J. M. C. and R. M. Itami, 1991. GIS-based habitat modeling using logistic multiple regression: a study of the Mt. Graham red squirrel. Photogrammetric Engineering & Remote Sensing, 57 (11), pp. 1475–1486.

Skidmore, A. K., 1990. Terrain position as mapped from a grided digital elevation model. Int. J. of Geographical Information Systems, 4 (1), pp. 33–49.

Sperduto, M. B. and R. G. Congalton, 1996. Predicting rare orchid (small Whorled Pogonia) habitat using GIS. Photogrammetric Engineering & Remote Sensing, 62 (11), pp.1269–1279.

Stockwell, D. R. B. and I. R. Noble, 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis, Math. Comput. Simul, 33, pp. 385–390.