DESERT

DESERT
Online at http://jdesert.ut.ac.ir

DESERT 15 (2010) 151-158

Predicting the distribution of plant species using logistic regression (Case study: Garizat rangelands of Yazd province)

M.A. Zare Chahouki[a*], A. Zare Chahouki[b]

[a] *Assistant Professor, Faculty of Natural Resources, University of Tehran, Iran*
[b] *MSc.Graduate, Faculty of Natural Resources, University of Tehran, Iran*

Received: 29 November 2009; Received in revised form: 17 October 2010; Accepted: 10 November 2010

Abstract

The aim of this research was to study the relationships between presence of plant species and environmental factors in Garizat rangelands of Yazd province and providing their predictive habitat models. After delimitation of the study area, sampling was performed using randomized-systematic method. Accordingly, vegetation data including presence and cover percentage were determined in each quadrate. The topographic conditions were recorded in plot locations. Soil samples were taken at depths of 0-30 and 30-80 cm in each plot. The measured soil variables included texture, lime, saturation moisture, gypsum, acidity (pH), ECe and soluble inos (Na^+, K^+, Mg^{2+}, Cl^-, Co_3^{2-}, HCo_3^- and So_4^{2-}). Logistic regression technique was used to analyze the collected data. The results showed that the vegetation distribution is mainly related to soil characteristics such as texture, gravel, EC, gypsum, lime and OM. The presence of *Artemisia sieberi- Zygophyllum eurypterum* has relation with gravel, lime, available water and pH. *Ephedra strobilaceae-Zygophyllum eurypterum* has positive relation with gypsum. *Rheum ribes-Artemisia sieberi* has relation with clay and OM. *Cornulaca monacantha* has also relation with elevation above sea, gravel and gypsum. The presence of *Seidlitzia rosmarinus* has relation with lime. Electrical conductivity is the most factors effect on presence of *Tamarix ramosissima*.

Keywords: Logistic regression; Environmental factors; *Artemisia sieberi*; *Tamarix ramosissima*; *Ephedra strobilacea; Zygopyllum eurypterum*; *Rheum ribes*

1. Introduction

Predictive modeling of plant species' distributions based on their relationship with environmental variables is important for a range of management activities. Examples include management of threatened species and communities, risk assessment of non-native species in new environments, and the estimation of the magnitude of biological responses to environmental changes (Ferrier, 2002; Barry and Elith, 2006).

The analysis of species–environment relationship has always been a central issue in ecology. The importance of climate to explain plant distribution was reported earlier on (Humboldt and Bonpland, 1807; de Candolle, 1855). Climate in combination with other environmental factors has been much used to explain the main vegetation patterns around the world (e.g. Salisbury, 1926; Cain, 1944; Good, 1953; McArthur, 1972; Box, 1981; Stott, 1981; Walter, 1985; Woodward, 1987; Ellenberg, 1988). The quantification of such species–environment relationships represents the core of predictive geographical modeling in ecology. These models are generally based on various hypotheses as to how environmental factors control the distribution of species and communities (Austin, 2002).

Numerous methods have been developed for building predictive species habitat models. Guisan and Zimmermann (2000) presented a comprehensive review and classified the methods into two categories: (1) regression-

* Corresponding author. Tel.: +98 261 2223044, Fax: +98 261 2223044.
E-mail address: mazare@ut.ac.ir

based methods; and (2) environmental envelope methods. Regression methods relate species response to single or multiple environmental predictors. These methods include frequently used approaches such as logistic regression (LR; Hosmer and Lemeshow, 1989), generalized additive modeling (GAM; Hastie and Tibshirani, 1990), and classification and regression tree (CART; Breiman et al., 1984).

Logistic regression is a frequently used regression method for modelling species distributions (Guisan and Zimmermann, 2000; Rushton et al., 2004). This is a particular case of Generalised Linear Models (GLM, McCullagh and Nelder, 1983). GLM has been recognized in ecology for some time as having great advantages for dealing with data with different error structures particularly presence/absence data that is the common type of data available for spatial modelling of species distributions (Nicholls, 1989, 1991; Rushton et al., 2004). In the other hands, logistic regression is one of the methods that can predict the probability of occurrence of each plant species related to site condition factors.

Ecologists believe that the relationships between plant species and environmental factors is non-linear (McCune, 2004). Function of logistic regression is a sigmoid curve. This method has been used by Wu and Huffer (1998), Bio et al., 2002; Austin et al., 1990; Carter et al., 2006; Lassueur et al., 2006 for predictive species modeling. In this paper, we examined the relationship between occurrence of plant species with environmental factors in Garizat rangelands of Yazd province. Then, provided prediction maps for species using LR models.

2. Material and methods

2.1. Study area

This research was conducted in Garizat rangelands. The study area was 94130 ha. Garizat rangelands are located in the southern of Garizat region of the Yazd province in center of Iran (31º 04´ 53´´N, 53º 40´ 04´´E to 31º 21´ 26´´N, 54º 14´ 58´´E) (Fig. 1). The maximum elevation of 2100 m and the minimum elevation is 1400 m. Average annual precipitation ranges from 200 mm to 45 mm (Zare Chahouki, 2006).

2.2. Data collection

Sampling was done in homogeneous units resulted from overlaying of hypsometric, aspect, slope and geologic maps. Within each unit 3-5 parallel transects with 300-500 m length, each containing 30-50 quadrates (according to vegetation variations) were established and the sampling procedure was randomized–systematic. The quadrate size was determined for each vegetation type using the minimal area method; hence suitable quadrate size for different species ranged from 1*2m to 10*10m (2-100 m^2). Floristic list, density and canopy cover percentage were determined in each quadrate. Soil samples were taken from 0-30 and 30-80 cm in starting and ending points of each transect.

These samples were air-dried, thoroughly mixed, and passed through a 2mm sieve to get rid of gravel and boulders. The weight of gravel in each sample plot was determined and expressed as a percentage of the total weight of the soil sample. The portion finer than 2mm was kept for physical and chemical analysis according to Jackson (1967) and Allen and Stainer (1974). Soil texture was determined by the hydrometer analysis (Bouyoucos, 1962), saturation moisture (weighting method), organic carbon (determined using Walkely and Black rapid titration, Black, 1979), pH in saturation extract (determined by pH meter), electrical conductivity (ECe) (determined by conductivity meter), lime (determined using 1N HCl, Jackson, 1967), soluble calcium and magnesium (determined by titration with solution EDTA method), soluble chlorine (determined by titration with $AgNO_3$), soluble carbonate and bicarbonate (determined by titration with H_2SO_4 using methylorange and phenolphthalein, respectively) and soluble sodium and potassium (determined by flame photometry method).

2.3. Data Analysis

Logistic regression (LR) is a kind of generalized linear model (GLM) suitable for analysis when response data are binary. It uses a logit link to describe the relationship between the response and the linear sum of the predictor variables (Miller and Franklin, 2002). This is accomplished by applying the following regression equation, in which presence/absence of an object is transformed into a continuous probability *y* ranging from 0 to 1. Values close to 1 represent high probability of presence, whereas values close to 0 represent high probability of absence. In order to discrete *y* into presence and absence, a posterior threshold is assigned.

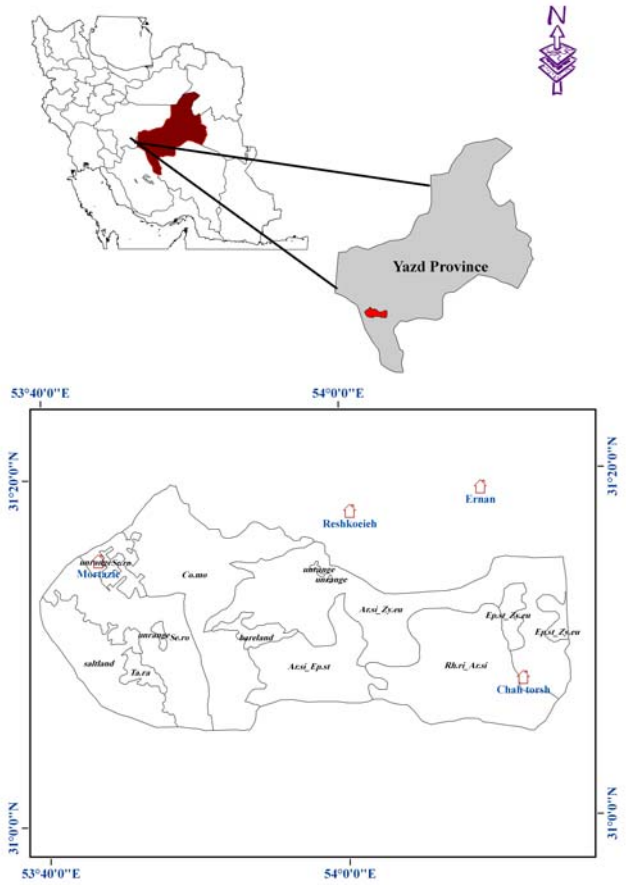

Fig. 1. Location of the study area

Occurrence probability of each plant species is calculated with respect to the combined effect of site conditions with the following equation:

$$Y = \frac{\exp(LP)}{(1+\exp(LP))} = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n)}{1+\exp(b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n)} \quad (1)$$

or

$$Y = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+...+b_nx_n)}} \quad (2)$$

Where y is the probability; x_n is explanatory variable; b_0 is the constant; and exp is an exponential function.

To assess possible collinearity problems, model coefficients and their errors are checked for irregularities and approximate variance inflation factors (VIF) are calculated for the final regression models (De Veaux and Ungar, 1994; Dallal 2001). VIF=1/(1−R^2), with a coefficient of determination (R^2) obtained from the regression of each explanatory variable (the way it appears in the model; i.e. in its linear or quadratic form) against all other explanatory variables present in the model (also using linear or quadratic terms as modeled). A VIF greater than 10 is often considered to point at possible collinearity problems (Dallal, 2001). Notice that the computed variance inflation factors are approximate. We used exclusively linear models (with first and second-order model terms) for VIF calculation, as R^2 is not readily available for GLM.

Models are calculated with individual selected variables and their combination (SPSS, 13.0). The best model is selected based on two criteria; *i.e.* approximate variance explained (Nagelkerka R square) and goodness of fit (Hosmer and Lemeshow test statistic; for details refer to SPSS 13.0).

2.4. Mapping prediction models

To plants predictive mapping, it is necessary to prepare the maps of all affective factors of models. Topographic data (elevation, slope, and aspect) were derived from DEM with accuracy 10m.

To mapping soil characteristics, geoestatistical method including variogram analysis and Kriging interpolation were used by GS^+ Ver. 5.1.1 .

Based on obtained predictive models for each species (through LR method) related predictive maps were prepared in GIS (Fig. 3).

2.5. Model evaluation

The best measure of agreement between observed (actual vegetation types) and predicted presence/absence is Kappa (Cohen 1960; Monserud and Leemans, 1992; Bell and Fiedling, 1997; Zimmermann and Guisan, 2000; Moisen and Frescino, 2002; Robertson et al. 2003; Liu et al. 2005). Its calculation is based on marginal probability of a contingency table. Kappa is used as the main measure, in the study, to evaluate the models. Monserud and Leemans (1992) suggested the following ranges of agreement for the κ statistic: no agreement, <0.05; very poor, 0.05–0.20; poor, 0.20–0.40; fair, 0.40–0.55; good, 0.55–0.70; very good, 0.70–0.85; excellent, 0.85–0.99; and perfect, 0.99–1.00. Negative values indicate extremely poor agreement (Monserud and Leemans, 1992). We used these ranges to describe the levels of agreement reported here using two tests.

3. Results

3.1. LR models

In the study area, 6 communities; *Rheum ribes-Artemisia sieberi*, *A. sieberi-Zygophyllum. Eurypterum*, *Ephedra strobilaceae-Z. eurypterum*, *Cornulaca monocnatha*, *Seidlitzia rosmarinus* and *Tamarix ramosissima* from north to south of study area, respectively, were distinguished.

Predictive models of abovementioned communities are represented in equation 1-6. Regarding equation 1, occurrence of *R. ribes-A. sieberi* community has inverse relation with clay and positive relation with OM. *A. sieberi-Zygophyllum. E.* community is significantly affected by the presence of gravel, lime, available water and pH (Eq. 2). Equation 3 shows that occurrence of *E. strobilaceae-Z. eurypterum* community is dependent to second layer gypsum.

Factors affecting the distribution of *C. monocnatha* is elevation, gravel and gypsum of first layer.

Occurrence of *S. rosmarinus* community has positive relation with lime of second layer.

T. ramosissima community is significantly affected by the presence of EC of first layer.

$$P(Rh.ri-Ar.si)=\frac{1}{1+e^{-(-4.041\mathrm{clay}_1+12.577\mathrm{OM}_1+27.653)}} \quad \text{(Eq.1)}$$

$$P(Ar.si-Zy.eu)=\frac{1}{1+e^{-(-0.781\mathrm{gravel}_1+0.592\mathrm{gravel}_2-0.734\mathrm{lim}_2+0.483\mathrm{AW}_1+5.452\mathrm{pH}_2-40.798)}} \quad \text{(Eq.2)}$$

$$P(Ep.st-Zy.eu)=\frac{1}{1+e^{-(1.880\mathrm{gyps}_2-47.277)}} \quad \text{(Eq.3)}$$

$$P(Co.mo)=\frac{1}{1+e^{-(-0.115\mathrm{abs}+6.548\mathrm{gravel}_1-6.229\mathrm{gyps}_1+94.378)}} \quad \text{(Eq.4)}$$

$$P(Se.ro)=\frac{1}{1+e^{-(2.362\mathrm{lime}_1-70.517)}} \quad \text{(Eq.5)}$$

$$P(Ta.ra)=\frac{1}{1+e^{-(1.218\mathrm{EC}_1-24.408)}} \quad \text{(Eq.6)}$$

3.2. Mapping prediction models

For soil characteristics mapping, at the first, spatial structure of data was evaluated in GS^+ for windows and component of variogram was determined (Table 1 and Fig. 2). In next stage, point map of soil characteristics were prepared in Arc GIS 9.2. Finally, using component variogram and kriging interpolation, soil map characteristics in 1:50000 scale were provided.

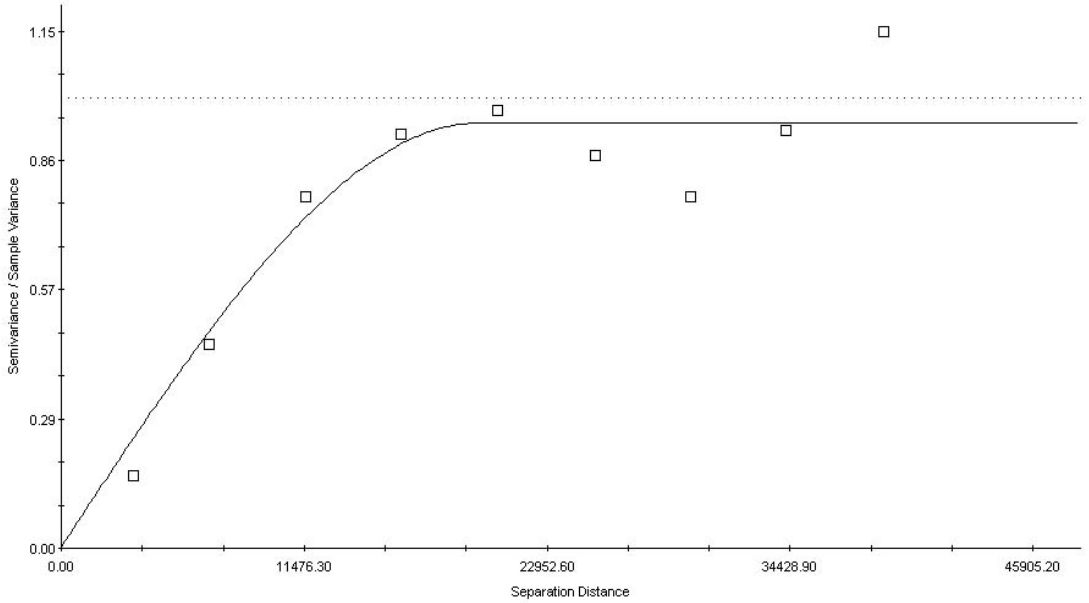

Fig. 2. Variogram model for gravel content of the first depth of soils in Garizat rangelands

Table 1. Components of variogram analysis for selected soil characteristics. For vegetation types and variables abbreviations, see Appendix A.

Characteristic	Model	Nugget effect (%)	Sill (%)	Effective range (m)	Lag distance (m)	Correlation Coefficient
gr1	Spherical	0.00287	0.94532	19740	4590.52	0.89
gr2	Spherical	0.0444	0.95820	15640	4590.52	0.81
clay1	Spherical	0.00276	1.03634	23680	4590.52	0.95
clay2	Spherical	0.10491	0.97463	16880	4590.52	0.84
A.W.1	Exponential	0.32999	10.42999	18990	4590.52	0.78
A.W.2	Spherical	0.17000	8.72000	15720	4590.52	0.79
O.M1	Spherical	0.00850	3.79235	91710	4590.52	0.96
O.M2	Spherical	0.00215	3.919989	91710	4590.52	0.97
lime1	Spherical	0.26279	2.16991	101100	4590.52	0.73
lime2	Spherical	0.34209	1.94708	101100	4590.52	0.69
gy1	Spherical	0.003754	2.12352	852900	4590.52	0.90
gy2	Spherical	0.24988	2.23559	81740	4590.52	0.78
pH1	Spherical	0.13731	2.60071	101100	4590.52	0.97
pH2	Spherical	0.66694	1.49709	92480	4590.52	0.93
EC1	Spherical	0.32951	1.64325	93530	4590.52	0.82
EC2	Spherical	0.36977	1.71377	97030	4590.52	0.88

3.3. Model evaluation

The accuracy of the predicted maps were tested with actual vegetation maps. In this study, the adequacy of vegetation type mapping was evaluated using kappa statistics. The values of Kappa coefficient based on LR of predicted and actual maps of vegetation cover indicates the accordance of predicted map for *C. monocantha* habitat is excellent and for *E. strobilcaea-Z. eurypterum*, *S. rosmarinus* and *T. ramosissima* habitat is good. While for *R. ribes-A. sieberi* and *A. sieberi-Z. eurypterum* habitat is Fair (Table 2).

Table 2. Kappa coefficient and accordance classes for predicted vegetation types. For vegetation types abbreviations

Vegetation type	Kappa Coefficient	Accordance class
R. ribes-A. sieberi	0.51	Fair
A. sieberi-Z. eurypterum	0.42	Fair
E. strobilacea-Z. eurypterum	0.58	good
C. monocnatha	0.90	Excellent
S. rosmarinus	0.60	good
T. ramosissima	0.56	good

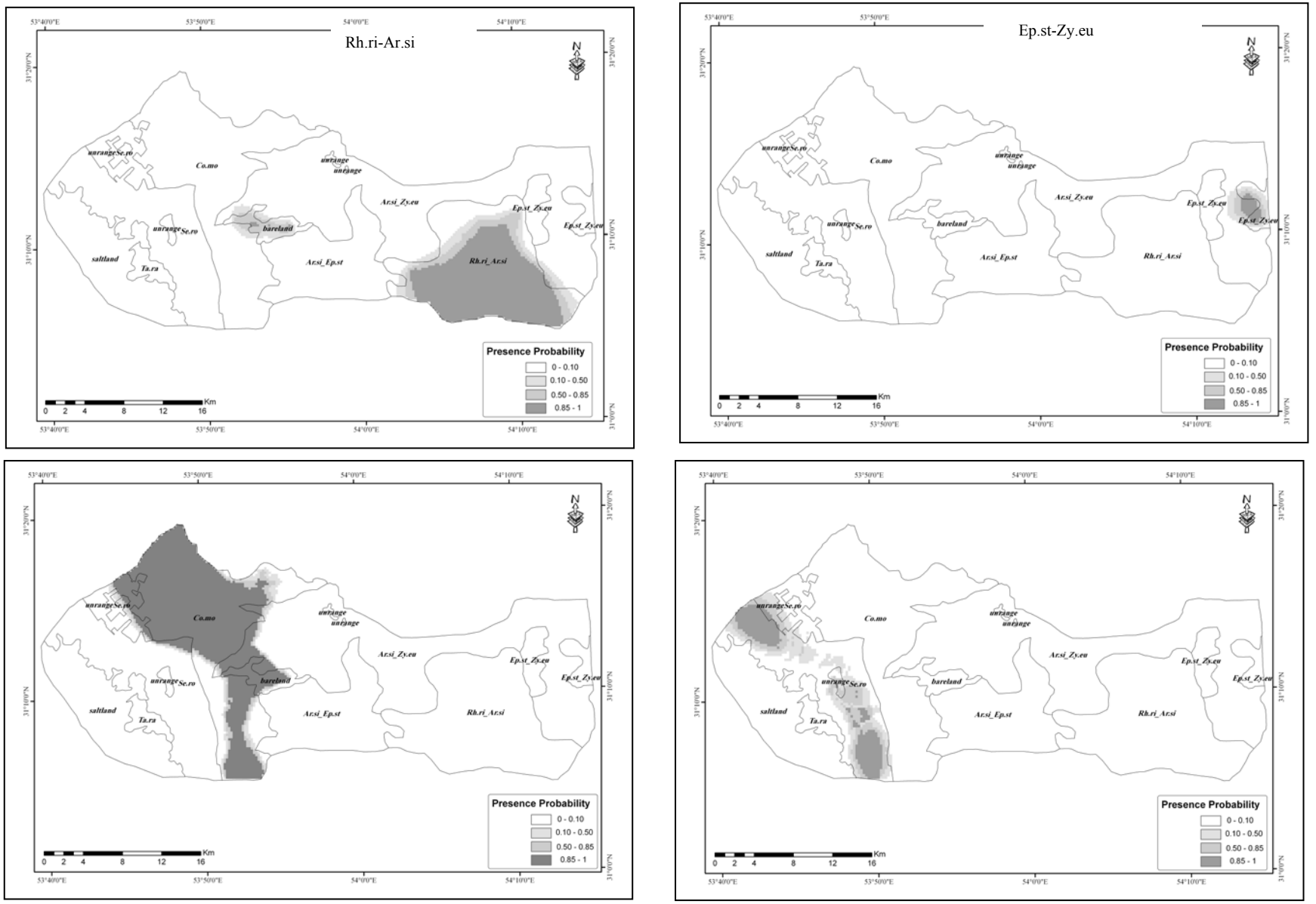

Fig. 3. Predicted map of vegetation types provided by logistic regression

4. Discussion and conclusion

The results showed that the presence of vegetation types was related to edaphic and topographic factors. In general, the most important ecological factors affecting on vegetation of Garizat rangelands are elevation, gravel, texture, lime, EC, pH, gypsum, available water and organic matter. Soil physical properties are affecting on water availability (Enright et al, 2005). Similar to our findings, Tavili et al (2009) and Enright et al (2005) showed that those environmental factors that affect on water availability were the most effective on distribution of vegetation in arid rangelands.

Predictive maps of *E. strobilacea-Z. eurypterum, C. monachantha, S. rosmarinus* and *T. ramosissima*, which have narrow amplitude, high accordance with actual vegetation map prepared for the study area. Among species of study area, predictive model of *R. ribes-A. sieberi* and *A. sieberi-Z. eurypterum*, due to its ability to grow in different habitat conditions, is not possible.

In conclusion, our work has shown that evaluating a habitat suitability model based only on presences is possible. The result of this research confirmed that both GIS expert system and logistic regression model are capable of predicting the spatial distribution of plant species (Yang, 2004).

Logistic regression is a suitable method in prediction of different plant species occurrence. In this study, predictive models of vegetation types were provided based on absence-presence of species, using logistic regression. Comparing results of prediction with real vegetation map of the study area shows that in logistic regression method absence-presence of species, as a dependent variable, is an effective factor for predictive species modeling. In this method, it is not necessary to use quantification attributes like density, frequency, biomass and canopy cover. These properties are severely affected by sampling method, shape, size and number of quadrate, as well as precipitation, while absence-presence is not dependent to above-mentioned factors.

Based on the prediction models, it is possible to estimate the probability of presence/absence of plant species in response to environmental factors (He et al, 2007). In case of calibration of the resulted models for each species, it is possible to apply such models in introduction of plant species for rangeland rehabilitation activities considering the environmental factors affecting establishment of vegetation cover.

References

Allen, M.M. & Stainer, S.T., 1974. Chemical Analysis of Ecological Materials. Blackwell Scientific Publications, Oxford, London, 565pp.

Austin M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecol. Model., 157:101–118.

Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realised qualitative niche: environmental niches of five Eucalyptus species. Ecol. Monogr. 60, 161–177.

Bio A.M.F., P.D. Becker, E.D. Bie, W. Huybrechts & M. Wassen, 2002. Prediction of plant species distribution in lowland river valleys in Belgium: modeling species response of site conditions. Journal of Biodiversity and Conservation, 11: 2189-2216.

Black, C.A., 1979. Methods of soil analysis. American Society of Agronomy 2, 771-1572.

Bouyoucos, G.J., 1962. Hydrometer method improved for making particle size analysis of soil. Journal of Agronomy 54, 464–465.

Box, E.O., 1981. Macroclimate and Plant Forms: An Introduction to Predictive Modeling in Phytogeography, Junk, The Hague, 258 pp.

Boyce, M.S. and McDonald, L.L., 1999. Relating populations to habitats using resource selection functions. Trends Ecol. Evol. 14, 268–272.

Breiman, L., J.H. Friedman, R.A. Olshen, & C.J. Stone, 1984. In: Classification and Regression Trees. Wadsworth, California.

Cain, S.A., 1944. Foundations of Plant Geography. Harpers and Brothers, New York, London.

Carter G. M., E.D. Stolen & D.R. Breininger, 2006. A rapid approach to modeling species–habitat relationships. Journal of Biological Conservation, 127: 237 -244.

Cohen, J., 1960. A coefficient of agreement of nominal scales. Educ. Psychol. Measure. 20, 37–46.

Dallal, G.E., 2001. Collinearity. http://www.tufts.edu/≈gdallal/collin.htm.

De Candolle, A.I., 1855. Ge´ographique botanique raisonne´e. Masson, Paris.

De Veaux, R.D. & Ungar, L.H., 1994. Multicollinearity: A tale of two nonparametric regressions. In: P. Cheeseman, P. and Oldford, R.W. (eds) Selecting models from data: AI and Statistics IV, pp. 293-302, Springer-Verlag. New York, NY, US.

Ellenberg, H., 1988. Vegetation ecology of Central Europe, fourth ed. Cambridge University Press, Cambridge.

Enright, N.J., B.P. Miller & R. Akhter, 2005. Desert vegetation and vegetation-environment relationships in Kirthar National Park, Sindh, Pakistan. Journal of Arid Environments, 61, 397–418.

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, 24: 38-49.

Good, R., 1953. The Geography of the Flowering Plants, second ed. Longman, London.

Guisan, A. & N.E. Zimmermann, 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135: 147-186.

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

Hastie, T., R. Tibshirani, & J. Friedman, 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

He, M.Z., J.G. Zheng, X.R. Li & Y.L. Qian, 2007. Environmental factors affecting vegetation composition in the Alxa Plateau, China. Journal of Arid Environments, 69, 473-489.

Hosmer D.W. & S. Lemeshow, 1980. Goodness-of-fit tests for the multiple logistic regression models. Communications in statistics-Theory and Methods A9:1043-1069.

Jackson, M.L., 1967. Soil Chemical Analysis-Advanced Course. Washington Department of Soil Sciences, 498pp.

Lassueur T., S. Joost & C.F. Randin, 2006. Very high resolution digital elevation models: Do they improve models of plant species distribution? Journal of Ecological Modeling (Article in press).

McArthur, R.H., 1972. Geographical Ecology: Patterns in the Distribution of Species. Harper and Row, New York.

McCullagh, P. & J.A. Nelder, 1983. Generalized Linear Models. Chapman and Hall, London, p. 261.

McCune, B., 2004. Nonparametric multiplicative for habitat modeling. Orogon state university, USA, 43p.

Miller, J. & J. Franklin, 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling 157 (2-3): 227-247.

Moisen, G.G. & T.S. Frescino, 2002. Comparing five modeling techniques for predicting forest characteristics. Ecol. Modell, 157: 209-225.

Monserud, R.A. & R. Leemans, 1992. Comparing global vegetation maps with the Kappa statistic. Ecol. Model, 62: 275–293.

Nicholls, A.O., 1989. How to make biological surveys go further with generalised linear models. Biological Conservation, 50(1-4): 51-75.

Robertson, M. P., C.I. Peter, M.H. Villet & B.S. Ripley, 2003. Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modeling techniques, Ecological Modelling, 164: 153–167.

Rushton, S.P., S.J. Ormerod & G. Kerby, 2004. New paradigms for modelling species distributions? J. Appl. Ecol. 41, 193–200.

Salisbury, E.J., 1926. The geographical distribution of breeding plants in relation to climatic factors. Geographical Journal 57, 312–335.

Stott, P., 1981. Historical Plant Geography. An Introduction. George Allen and Uwin, London.

Tavili A., M. Rostampour, M.A. Zare Chahouki and J. Farzadmehr, 2009. CCA application for vegetation-environmental relationships evaluation in arid environments (Southern Khorasan rangelands). Journal of Desert, 14(1): 101-111.

von Humboldt, A., Bonpland, A., 1807. Essai sur la ge´ ographie des plantes, Paris.

Walter H., 1985. Vegetation of the Earth and Ecological Systems of Geobiosphere, third ed. Springer, Heidelberg.

Woodward F.I., 1987. Climate and Plant Distribution. Cambridge University Press, Cambridge, 174 pp.

Wu H. & F.W. Huffer, 1998. Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. Journal of Biometrics, 54: 509-525.

Yang X., 2004. Modelling the Spatial Distribution of *Tricholoma matsutake*. MSc. Thesis in Geo-information Science and Earth Observation, International Institute for Geo Information Scince and Earth Bservation Enschede, The Netherlands, 55p.

Zare Chahouki M.A., 2006. Modelling the spatial distribution of plant species in arid and semi-arid rangelands. PhD Thesis in Range management, Faculty of Natural Resources, University of Tehran, 180 p.