Comparing Discriminant Analysis, Ecological Niche Factor Analysis and Logistic Regression Methods for Geographic Distribution Modelling of *Eurotia ceratoides* (L.) C. A. Mey

Lyla Khalasi Ahvazi\(^A\), Mohammad Ali Zare Chahouki\(^B\), Faeze Ghorbannezhad\(^C\)

\(^A\)Ph.D Student, Department of Rangeland Management, Rangeland and Watershed Management Faculty, Gorgan University of Agricultural Sciences & Natural Resources, Gorgan, Iran. (Corresponding Author). Email: khalasi@alumni.ut.ac.ir

\(^B\)Department of Rehabilitation of Arid and Mountainous Regions, University of Tehran, Iran.

\(^C\)Educated in master degree, Department of GIS and Remote Sensing, Geography faculty, Tarbiat Modares university, Iran.

Received on: 11/04/2013
Accepted on: 17/06/2013

**Abstract.** *Eurotia ceratoides* (L.) C. A. Mey is an important plant species in semi-arid lands in Iran. New approaches are required to determine the distribution of this plant species. For this reason, geographical distributions of *Eurotia ceratoides* were assessed using three different models including: Multiple Discriminant Analysis (MDA), Ecological Niche Factor Analysis (ENFA) and Logistic Regression (LR). The study area was located in northeast rangelands of Semnan, Iran. Sampling was performed in each vegetation type using randomized-systematic method. Vegetation data in addition to environmental factors' data such as topography and soil were prepared. The MDA and LR methods were performed with SPSS software as predictive modelling methods based on presence and absence data. The ENFA model was performed by the means of necessary statistical analysis in Biomapper (Version 4.0) software only by presence data. The plant predictive mapping needs the maps of all effective factors based on model parameters. Mapping of soil characteristics was done by geo-statistical method. The accuracy of the predicted map was tested with the actual vegetation map. Predictive maps of *E. ceratoides* (based on the LR and MDA methods) with Kappa coefficients as 0.56 and 0.64 had a good accordance with actual vegetation map prepared for the study area. Kappa coefficient of potential habitat map (based on ENFA method) of *E. ceratoides* was 0.85; hence, it had a very good accordance. The results obtained by all methods showed that this species is distributed in the rangeland with pH as 7.8-8, EC as 0.17-0.26 dc/m and silty-sandy texture in 1600-2200 m elevation. Organic matter in the depth of 20-80 cm and pH in the depth of 0-20 cm did not significantly influence the differences. Minimum sampling is needed using these methods which provide worth while data about the presence of the plant species in the other places.

**Key words:** Actual vegetation map, Geo-statistical method, ENFA, Kappa coefficient, LR, MDA.
Introduction
Anticipation of plant species in arid environment can be assessed based on environmental variables. Prediction models foresee the plant distribution and their habitats; consequently, they can be applied for conservative and management objectives. Distribution of *Eurotia ceratoides* was evaluated in this study. The aim of the ecological researches has been to get vision into functionality and complexity of the ecosystem by observing individual factors which affect it. *Eurotia ceratoides* is one of the most important range plants that is often seen as the associated species and rarely seen as the dominant species in rangelands. It is very critical for soil conservation and grazing pressure. *Eurotia ceratoides* plant due to bushy form, resistance to drought, proper protein percent and easy propagation is one of the native plants of desert rangelands that may have a special importance in arid and semi-arid rangelands regarding these characteristics that caused wide uses of this plant in rangeland amendment programs (Bespalova, 1964).

Uniyal *et al.* (2005) suggested that in arid regions where it takes several years for a plant to grow, degradation of *Eurotia* will cause desertification. Three usual methods have been used to model this species ecosystem. The results of each one have been evaluated.

The importance of species distribution modelling has increased in recent years with specific examples ranging from the studies on climate changes (Huntley *et al.*, 2007) to assess distribution of rare species (Guisan *et al.*, 2006) and niche theory research (Pearman *et al.*, 2008). Most efforts of species distribution modeling are still based on the generalized linear models or generalized additive models, but there has been much progress in terms of modeling process (Ke´ry *et al*., 2010).

MDA assigned a linear combination between variables with normal errors. It is one of the cluster analysis methods which classify the cases into dependent categories and this method shows how data are classified. Researchers such as Joy and Death (2003) and Maron and Lill (2004) have used this method. ENFA method is a multivariate approach to study distribution modelling of species with only presence data. This method considers Eco-Geographical Variables (EGV) as well as presence data for plant species in various locations; then, it can predict the desirable ecosystem for the specified species. In fact, ENFA method works like the Principal Component Analysis (Hirzel *et al*., 2002). The first factor in ENFA method is the ‘marginality' meaning. It reflects the ecological distance between the mean of each factor in the desired species distribution and the mean for the entire region in the same element-based ENFA method while it takes into account the eco-geographical predictor as well as the presence of data for plant species in various locations and then, it can predict the desirable ecosystem for the specified species (Hirzel *et al*., 2001; Arnese, 2007).

GLM is a generalization of multiple regression analysis with a binomial distribution and logistic link that may fit for polynomials with higher degree than a linear one. The presence/absence of the species is explained by a sum of weighted eco-geographical factors. Jongman *et al.* (1995) and Nicholls (1989) suggested that the weights are tuned in order to generate the best fit between the model and the calibration data set.

In recent years, ecological studies used Regression methods (Sawchik *et al*., 2003; Gutierrez *et al*., 2005).
Multiple Discriminant Analysis (MDA) method is similar to regression methods. But it has few differences. Dependant variable in regression method is always quantitative with normal distribution while in MDA, it is qualitative with limited classes. MDA is also similar to multiple analysis of variance (MANOVA) sharing many of the same assumptions and tests (Srinivas, 2009). Vegetation distribution models tend to describe vegetation patterns based on environmental variables (Manel et al., 1999). Environmental factors interact with systems in such a complex way that the whole system achieves a broader functionality that cannot be deduced by considering individual environmental factors (Tan et al., 2006). In fact, the use of distribution modelling of species has some purposes such as predictive occurrence of multiple stable states of ecosystem processes and habitat selection or distribution of species (Tan and Beklioglu, 2005; Baran et al., 1996; Lek et al., 1996; Özesmi and Özesmi, 1999). The other authors conducting their studies on vegetation distribution as a discussion in the ecological studies such as generalized additive and linear models are Seaone et al., (2003); Seaone et al., (2004); Dunk et al., (2004); Meggs et al. (2004) and Tan and Beklioglu, (2005). Austin (2007) suggested that to perform the comparisons of methods, they rarely use the same type of data (counts or presence/absence) while applying the regression method in the same way (multiple linear versus curvilinear terms) or a common set of predictors. *Eurotia ceratoides* helps in soil stabilization and also produces considerable forage for wild and domestic animals and new approaches are required to determine what environmental factor is needed for each plant species and how the plant species can be obtained and used to make decisions about land use, habitat, grazing, etc. Naturally, it has become clear that recognizing the ecosystem is very necessary to manage the rangelands (Christensen et al., 1996; Yaffee, 1999) so that plant distribution models must be very useful for the distribution and abundance of plants. The aim of this study was to predict the distribution of *E. ceratoides* with three different models (MDA, ENFA and LR) by conducting some ecological studies on the northeast rangelands of Semnan. Studies in this field can be necessary for rangelands restoration goals if the proposed models have a good accuracy and reliability and are tested in various regions.

### Materials and Methods

#### Study area

This research was performed in the northeast of Semnan rangelands with an area of 74000 hectares (see Fig. 1) which is located in the center of Iran (35° 53’ N, 54° 24’ E to 35°50’ N, 53°43’ E). The maximum elevation of the study area is 2260 m a.s.l. and the minimum elevation is 1129 m a.s.l. Mean annual precipitation of the study area ranges from 275 mm in the mountains to 128 mm in the saline lowlands. Minimum temperature occurs in December (around -6°C) while the highest temperature reaches +45°C in June.

#### Data collection

The survey of vegetation quantities was initiated in 2009 for a one-year period. Sampling was performed in each vegetation type using randomized-systematic method based on field surveys and then, dominant vegetation types were determined. Fifteen quadrates were located within 50 m distance in length. Three 750 m transverse transects with 45
quadrates with a distance of 50 m from each other were located in each vegetation type (according to vegetation variations). Quadrate size was determined for each vegetation type using minimal area method (Cain, 1938). Floristic list, density and canopy cover percentage were determined in each quadrat. Assessing soil properties was done based on collecting six soil samples (0-80) in each separate region as well as 0-20 and 20-80 cm samples from starting and ending points of each transect. Available moisture (weighting method), Electrical Conductivity (EC), lime (Jackson, 1967), pH in the saturation extract, soil organic matter (Black, 1979) and soil texture (determined by Bouyoucos hydrometer) were measured in laboratory as soil elements (Black, 1979). The elevation, slope (using GPS) and slope direction were determined at the location of each quadrat.

Methods of data analysis
At first, it was necessary to prepare the maps of all effective factors used in this research (Fig. 2). Topographic data (elevation, slope and aspect) were derived from Digital Elevation Models (DEM) with a resolution of 10 m. Geo-statistical methods were used to map soil characteristics. Block Kriging method had been applied by GS+ (Version 5.1.0) and GIS (Version 9.3) software to predict soil factor. Two plant distribution models (MDA and LR) with a binomial probability distribution and ENFA method only with presence data were fitted for *Eurotia ceratoides* based on 22 topography and soil predictor variables.
Multiple Discriminant Analysis (MDA)

MDA method is usually applied to ornithological data (Buckton and Ormerod, 1997; Buckton et al., 1998; Manel et al., 1999). The explanatory variables collection was selected here to maximize within-group variance in order to classify the groups (Venables and Ripley, 1997). In order to classify the cases into groups, this method uses Discriminant prediction equation. In addition, it employs sequential Discriminant analysis for investigating the differences between or among groups, determining the most parsimonious way to distinguish groups and ascertaining the percent of variance in the dependent variable explained by the independents (Manel et al., 1999) as well as assessing the relative importance of the independent variables for classifying the dependent variable and discarding variables which are relatively related to group distinctions.

MDA was performed with SPSS 15 (Venables and Ripley, 1997). Discriminant function is shown below (Equation 1).

\[ F = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + c \]  

(1)

Where \( F \) is the latent variable formed by the discriminant function, \( b \) is discriminant coefficients, \( x \) is the discriminating variables and \( c \) is a constant.

The discriminant function coefficients are partial coefficients reflecting the unique contribution of each variable to the classification of the criterion variable. The standardized discriminant coefficients like beta weights in regression are used to assess the relative classification importance of the independent variables.

Ecological Niche Factor Analysis (ENFA)

Niche-based species distribution models are the ones that relate the observations of species gathered over a certain period of time with various attributes of the environment such as topography and soil factor (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005). ENFA model only requires “presence” data but not “absence” ones during the calculation process. ENFA was entirely performed with the Biomapper (Version 4.0) software (Hirzel et al., 2001). So, all data layer formats were changed to raster layers in IDRISI Kilimanjaro (Version 14.02) software for entering the Biomapper (Version 4.0) software. Then, the predictors were first normalized by the Box–Cox algorithm (Sokal and Rohlf, 1981). Ecological niche factors were then computed on these normalized predictors. By importing the information layers into the appropriate model and doing necessary statistical analysis in Biomapper (Version 4.0) software, the potential habitat map was created.

Logistic regression

Presence and absence data were related to 22 environmental and habitat factors using a generalized linear model and multiple logistic regression with a logic link and binomial error distribution (McCullagh and Nelder, 1989; Jongman et al., 1995). The logic transformation of the probability of presence/ absence (\( p \)) was modeled as a linear function of 22 possible explanatory variables (Equation 2).

\[ Y = \frac{\exp(LP)}{1 + \exp(LP)} = \frac{\exp(b_0 + b_1 x_1 + \ldots + b_n x_n)}{1 + \exp(b_0 + b_1 x_1 + \ldots + b_n x_n)} \]  

(2)

\( b_0 \) and \( b_i \) are the regression constants, \( b_0 \) is the constant and \( \exp \) is an exponential function. \( b_1, b_2, \ldots, b_n \) are the logic coefficients of \( x_1, x_2, \ldots, x_n \) variables, respectively. Presence/absence of an object is transformed into a continuous probability ranging from 0 to 1.

The step function used in the statistical package SPSS (version of 15.0) provides a
Comparing Discriminant procedure for this purpose using two criteria: (1) approximate variance explained ($R^2$) and (2) goodness of fit (Hosmer and Lemeshow test statistics). Significant variables at each step had to significantly reduce the scaled deviance. Model was fitted using a maximum likelihood method (McCullagh and Nelder, 1989). To select the variables in the final model, the backward elimination was used (Green et al., 1994; Austin and Meyers, 1996). At the end, comparisons of predicted (probability scale) and observed (presence–absence) values were based on Kappa coefficient maximized over the full range of possible probability thresholds (hereafter max Kappa; Gusian et al., 1999). The accuracy of the predicted maps and adequacy of vegetation mapping types were evaluated using the Kappa statistic. Actual vegetation map was prepared by Department of Natural Source, Semnan province in 2008.

Results

Multiple Discriminant Analysis (MDA)

For MDA, classification of each case was derived from euclidean distances to the centroids of the ‘positive’ and ‘negative’ groups. In all approaches, scores for correct assignment were expressed as percentages of the total number of cases (Fielding and Bell, 1997). Equation 3 showed that discriminant function by this analysis and occurrence of *E. ceratoides* are dependent on the percentage of gravel, lime, organic matter in the soil depth of 0-20 cm (D 1) and available moisture percentage in the depth of 20-80 cm (D 2). Eigenvalue amount of this function is 1.215 that explains 100% of variance and also, canonical correlation amount is 0.741. These values could be used to give us measures of sensitivity (=percentage of true correctly identified presences) and specificity (=percentage of true correctly identified absences). (Fig. 3), shows the predicted map of *E. ceratoides* using the MDA model.

$$F = 1.15Gravel_1 + 2.45Lime_1 + 23.45OM_1 + 2.42Moisture_1 - 46.25$$  \(3\)

![Fig. 3. Predicted map of *E. ceratoides* using MDA model](http://www.opoosoft.com)
ENFA based modelling
Marginality coefficients showed that the most important variables are essentially linked to high gravel, elevation and texture. The next factors account for some more specialization regarding lime and gravel frequency in 20-80 deep and the altitude (Specialization 1), EC, gypsum and silt are main factors in the second one (Specialization 2) so that EC and gypsum have negative impacts as well as lime in the depth of 0-20 (Specialization 3) showing sensitivity to the shifts away from their optimal values in these variables (Table 1).

<table>
<thead>
<tr>
<th>Eco-Geographical Variables</th>
<th>Depth (cm)</th>
<th>Marginality</th>
<th>Specialization 1</th>
<th>Specialization 2</th>
<th>Specialization 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravel (%)</td>
<td>0-20</td>
<td>0.362</td>
<td>0.244</td>
<td>0.101</td>
<td>-0.263</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>0.357</td>
<td>0.016</td>
<td>0.0907</td>
<td>-0.258</td>
</tr>
<tr>
<td>Clay (%)</td>
<td>0-20</td>
<td>-0.009</td>
<td>0.100</td>
<td>0.114</td>
<td>0.179</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>-0.320</td>
<td>-0.316</td>
<td>0.259</td>
<td>0.150</td>
</tr>
<tr>
<td>Silt (%)</td>
<td>0-20</td>
<td>0.345</td>
<td>-0.191</td>
<td>0.027</td>
<td>0.247</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>0.032</td>
<td>-0.002</td>
<td>0.062</td>
<td>0.269</td>
</tr>
<tr>
<td>Sand (%)</td>
<td>0-20</td>
<td>0.223</td>
<td>0.083</td>
<td>-0.158</td>
<td>-0.243</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>0.026</td>
<td>0.181</td>
<td>-0.186</td>
<td>-0.235</td>
</tr>
<tr>
<td>Lime (%)</td>
<td>0-20</td>
<td>0.042</td>
<td>0.164</td>
<td>-0.342</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>-0.188</td>
<td>0.010</td>
<td>-0.343</td>
<td>0.160</td>
</tr>
<tr>
<td>Organic Matter (%)</td>
<td>0-20</td>
<td>-0.056</td>
<td>0.139</td>
<td>0.374</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>0.210</td>
<td>0.296</td>
<td>0.358</td>
<td>0.176</td>
</tr>
<tr>
<td>Available Moisture (%)</td>
<td>0-20</td>
<td>0.203</td>
<td>-0.241</td>
<td>0.114</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>0.172</td>
<td>0.239</td>
<td>0.120</td>
<td>0.287</td>
</tr>
<tr>
<td>Gypsum (%)</td>
<td>0-20</td>
<td>-0.071</td>
<td>0.092</td>
<td>-0.068</td>
<td>0.366</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>-0.071</td>
<td>0.092</td>
<td>-0.068</td>
<td>0.264</td>
</tr>
<tr>
<td>Electrical Conductivity (ds/s)</td>
<td>0-20</td>
<td>-0.062</td>
<td>0.096</td>
<td>-0.069</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>-0.077</td>
<td>0.101</td>
<td>-0.072</td>
<td>0.235</td>
</tr>
<tr>
<td>pH (%)</td>
<td>0-20</td>
<td>0.064</td>
<td>-0.679</td>
<td>-0.113</td>
<td>-0.126</td>
</tr>
<tr>
<td></td>
<td>20-80</td>
<td>0.032</td>
<td>0.274</td>
<td>-0.134</td>
<td>-0.210</td>
</tr>
<tr>
<td>Elevation (m)</td>
<td>-</td>
<td>0.338</td>
<td>0.025</td>
<td>-0.259</td>
<td>-0.114</td>
</tr>
<tr>
<td>Slope (%)</td>
<td>-</td>
<td>-0.578</td>
<td>0.183</td>
<td>0.255</td>
<td>-0.144</td>
</tr>
</tbody>
</table>

Suitability map was built from these four factors for the northeast Semnan (Fig. 4). The results showed that 15000 hectares of study site may be potential habitat of *Eurotia ceratoides* which constitutes 20 percent of the study site. To evaluate the verity of this model, Boyce index was used and model rectitude in this test was determined as 93.2 percent. The mean and the standard deviation of the accuracy assessment were calculated for modal validation.

**Logistic Regression-based modeling**
The predicted occurrence probability of *E. ceratoides* was showed in equation 4. Regarding equation 4, the occurrence of *E. ceratoides* is dependent on the
percentage of gravel in the soil depth of 0-20 cm (D1) and slope percentage. This function explains 100% of variance and also canonical correlation amount that is 0.93. Goodness of fit (Hosmer and Lemeshow test statistics) was high and function was significant.

Predictive vegetation map based on the predictive model obtained using LR method was generated in GIS environment. (Fig. 5), shows the predicted map of *E. ceratoides* using the logistic regression model.

\[
P(E. ceratoides) = \frac{\text{Exp}(0.118 \times \text{slope} + 0.631 \times \text{gravel} - 2.864)}{1 + \text{Exp}(0.118 \times \text{slope} + 0.631 \times \text{gravel} - 2.864)}
\]
Discussion and Conclusion

The best measure of agreement between the observed and predicted presence-absence is Kappa ($\kappa$) statistic (Monserud and Leemans, 1992; Guisan and Zimmermann, 2000; Robertson et al., 2003; Liu et al., 2005). Kappa coefficients were prepared by MDA, ENFA and LR methods successively given as 0.64, 0.85 and 0.56 indicating that MDA and LR methods have a good accordance and ENFA has a very good accordance with actual vegetation map prepared for the study area (Monserud and Leemans, 1992).

ENFA Analysis achieved the best prediction success (kappa as 0.85) although this model was difficult to use for plant distribution modelling purposes due to high complexity in comparison to other models. This method was more practical and economical than LR and MDA models.

If the rate of occurrence is reduced, a positive prediction error will raise (Fielding and Bell, 1997) while some techniques such as logistic regression are more sensible to these effects than others. The results of this study showed that the important factors affecting the distribution of *E. ceratoides* by ENFA method are lime, organic matter, gypsum, sand and pH. These are slope and gravel given by LR method and organic matter, lime, gravel and available moisture achieved by MDA method. But ENFA method separates each of these effective factors into specialization, marginality and global tolerance of *E. ceratoides*. Marginality can be defined as ecological distance between the means of distribution of *E. ceratoides* in each environmental factor and same factor in the whole study area (Songlin et al., 2007). This index shows that *E. ceratoides* prefers silty-sandy texture and a higher amount of gravel and elevation than the mean value of these variables in the study area. On the other hand, specialization shows the species specialty in the range of its used resources. This index is the inverse of the tolerance level of species and its low amount indicates that the specified species has a high endurance for environmental factors such as organic matter, gypsum, sand, lime and altitude.

Comparing modeling methods needs the attention to correctly applying logistic regression and discriminant analysis considered in our study and we ensured that explanatory variables were linearized and normalized by transformation and incorporation into principal components' analysis prior to further analysis. Our data were also collected from sites randomly. Many ecologists focus on evaluating the species distribution models solely that may compel us to reaffirm the value of testing models with partitioned data (Kohavi, 1995).

ENFA method is one of the new modelling techniques that use presence data. It is widely used because it saves time and cost. Besides its capability in computing the number of desirable habitats, it presents important ecological factors such as Specialization, Marginality and global tolerance that are great ecological concepts.

Tan et al. (2006) have compared such methods and their results showed that GLM and connectionist neural network models appear to be most suitable and robust provided that a predictive variable reflecting time dependent dynamics will be included in the model either implicitly or explicitly. Ko et al. (2009) showed that nonlinear models (GARP, ANN and LR) provided better predictions than linear (MDA) ones. In this paper, we find that each model can be used in certain situation described above. But ENFA method shows a greater precision.
The summarized results of three methods showed that *E. ceratoides* is distributed in rangelands with pH as 7.8-8, EC as 0.17-0.26 dc/m and silty-sandy texture in 1600-2200 m elevation. Organic matter and pH in 20-80 and 0-20 cm deep did not significantly influence the differences.

The application of vegetation distribution models will remain important because vegetation types are frequently used in nature conservation, management and legislation (Peters et al., 2009). Specifically, the conservation of *E. ceratoides* becomes important not only for the stability of the ecosystem but also for the sustainability of rangeland. Therefore, studies on the distribution of *E. ceratoides* are of immediate need for the conservation and proper management of this vegetation habitat. Studies in this field can introduce suitable plant species for rangeland restoration if predictive models with acceptable accuracy are prepared and tested in different areas.

**References**


تغییرات توزیع پراکنش گونه Eurotia ceratoides (L.) C.A. Mey.

نتایج استاتیستیکی تحلیل عامل تصادفی-سیستماتیک انجام شد. اطلاعات پوشش گیاهی و همچنین اطلاعات عوامل محیطی یکی از مهم‌ترین عوامل مشخص مناطق شکست و نیمه خشک در ایران است. روشهای جدیدی برای تعیین توزیع پراکنش این گونه‌های گیاهی مورد نیاز است. در این مطالعه (MDA) این روشهای جدید برای تعیین توزیع پراکنش این گونه‌های گیاهی مورد نیاز است. یکی از مهم‌ترین عوامل محیطی یکی از مهم‌ترین عوامل مشخص مناطق شکست و نیمه خشک در ایران است. روشهای جدیدی برای تعیین توزیع پراکنش این گونه‌های گیاهی مورد نیاز است. این روشهای جدید برای تعیین توزیع پراکنش این گونه‌های گیاهی مورد نیاز است.