GEOLOGIJA 46/2, 385–390, Ljubljana 2003

Developing a suitability model for potential vegetation distribution based on GIS

Celestino ORDÓÑEZ GALÁN[1], Javier TABOADA CASTRO[1], Roberto MARTÍNEZ ALEGRÍA LÓPEZ[2]
[1]Dept. of Natural Resource and Environmental Engineering. University of Vigo. Spain
[2]Civil Protection. Auto. Govt. of Castilla & León. Spain
tino@lidia.uvigo.es

Key words: suitability model, potential, vegetation map, spatial analysis, WOFE, GIS

Abstract

The growing problem of deforestation around the world will have serious consequences for life on our planet if urgent steps are not taken to remedy this situation. In order to reforest affected areas, it is first necessary to decide the areas most suitable to each type of forest. This can be determined by using suitability models based on a series of morphological and environmental variables (altitude, gradient, insolation, etc.), all of which have a bearing on the presence or absence of a particular type of forest in a particular area.

Our study involves the creation of a potential vegetation model based on the environmental variables that have a bearing on the existence of a particular type of forest at any given point of terrain. These variables can be represented on maps that are included in a spatial database together with a distribution map of existing forests. A potential vegetation map for each study area is then drawn up using the integrated mathematical and statistical functions of the database. Two potential vegetation maps are created, one using discriminant analysis combined with correspondence analysis, and the other using logistic regression combined with weights of evidence. The results for each method are then compared to the real distribution of the forests in the area of study.

Introduction

When designing a reforestation project in order to reduce forest fragmentation and protect biodiversity in a particular area, the existing situation is a logical starting point. The surface area covered by each type of forest will be smaller than the original surface area since some of the forest will have been eliminated by artificial methods. Since reforestation projects are often based on working methods that are not entirely objective, they tend to give rise to different solutions depending on the technician who designed the plan. In a bid to eliminate this subjectivity, a number of authors have proposed the use of methodologies based on objective criteria for the development of suitability models for plant species (Van de Rijt et al., 1996; Felicísimo et al., 2002). These models take into account a series of physical and biological factors that have an impact on the distribution pattern of forests, among them lithology, altitude, gradient, temperature, rainfall, etc. All this information can be stored in a Geographic Information System (GIS) database, whose spatial analysis functions can be exploited to obtain distribution models based on objective methods (Felicísimo et al., 2002; Guisan et al., 1996).

In this study a potential vegetation model is created, based on the distribution of existing forests in the area of study and on the values of the morphological and environmental variables that have a bearing on this distribution. The proposed model was applied to the Liébana river basin in the northern Spanish region of Cantabria. This basin measures 629 km^2 and although its original forested area has diminished considerably (as in the entire Iberian Peninsula), it still contains autochthonous forests large enough to provide a representative sample.

Database construction

The first stage in the development of the model was the creation of maps for the GIS cartographic database.

The basic primary information was as follows:

- Vegetation map drawn up by the Department of Earth Sciences of the University of Cantabria and containing 180 classes of vegetation, of which 18 are forest. Only 6 of these forests were included in the study as they are the only ones still large enough to provide a suitable sample. Their growth patterns are also influenced by climatic rather than edaphic factors.
- Lithological map drawn up by the Spanish Technological and Geomining Institute, and containing 19 lithological classes for the study area.
- Topographic map drawn up by the Geography Division of the Spanish Army, and containing elevation markings every 20 metres.
- Rainfall map created using rainfall data for the last five years obtained from the Spanish National Meteorological Institute.

The following maps were then derived from the topographic map:

- Altitude maps calculated using a Delaunay triangulation algorithm converted to a raster structure with 50–metre cells.
- Gradient map obtained from an altitude map using a second-order finite difference scheme (Skidmore, 1989) with a resolution of 50 metres.
- Potential insolation map obtained from the DEM (Digital Elevation Model) and that takes into account the amount of sun received by each gridded cell in accordance with the path of the sun and topographic occlusion. Time resolution is 15 minutes and spatial resolution is 50 metres (Fernández-Cepedal & Felicísimo, 1987).
- Map of distances between the sea and the mid-point of each 50-metre cell. These distances were recorded to take into account the influence of the continental ocean gradient.

All these variables have a potential bearing on the distribution of forests and are included in the GIS cartographic database that is used in the simulation.

Development of a potential vegetation model using discriminant linear analysis

Discriminant analysis permits a series of observations to be classified into a number of previously defined classes. A linear combination is defined for each class that simultaneously maximises differences between means and minimises variances within classes (Fisher, 1936). This technique is widely used in fields such as taxonomy (Fisher, 1936), geology (Jordan et al., 1998), medicine (Mayer et al., 1998) and mining (Taboada et al., 2002), among others.

Each classification function provides a value for the cells in each class through the following formula:

$$S_i = c_i + \sum_{j=1}^{m} \lambda_{i,j} x_j$$

where S_i is the value of the discriminating function for class i; c_i is a constant, $\lambda_{i,j}$ is the coefficient for the n^{th} variable in class i; x_j is the value observed in the j^{th} variable in class i and m is the number of variables.

Once the discriminating functions have been established, each cell is assigned to the class associated with the greatest discriminating value.

This is the method used in some geographic information systems and satellite image processing programs to obtain supervised classifications using information from each of satellite's spectral bands.

We applied the method to a sample of 3320 cells taken randomly from the forested areas. The independent variables taken into consideration were altitude, gradient, potential insolation, distance from the sea and rainfall. Lithology was initially excluded, as

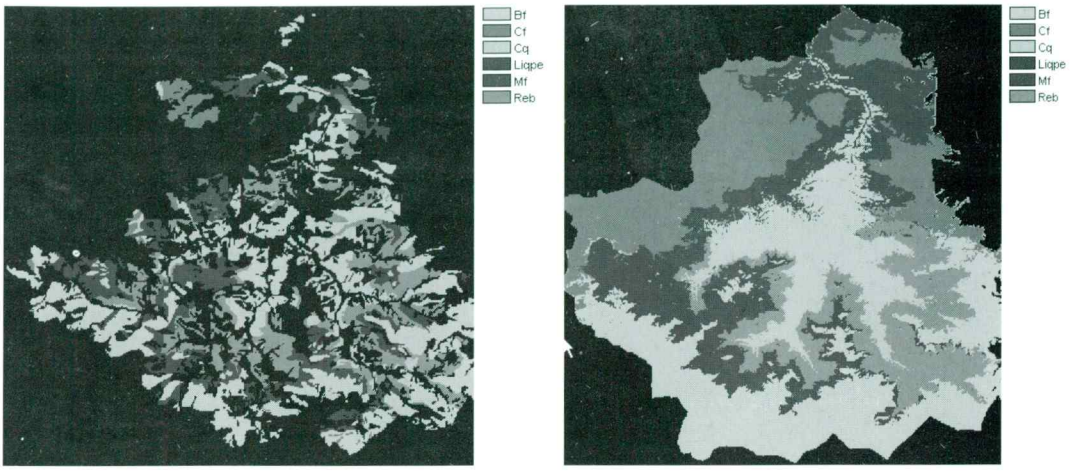

Figure 1. Distribution of existing forests (left) and potential distribution obtained using discriminant analysis, excluding lithology as an independent variable (right).

Table 1. Error matrices for the logistic regression (left) and discriminant analysis (right) models. The columns refer to actual forest and the rows to potential forest as determined by mathematical models.

	1	2	3	4	5	6	E.C.
1	20525	382	269	3998	32	2464	0.258
2	2881	2957	417	680	439	869	0.641
3	99	21	13666	100	377	2755	0.197
4	7132	359	752	9530	54	6476	0.608
5	342	934	2655	545	3986	1723	0.609
6	1956	21	2773	2570	1	8978	0.449
E.O.	0.377	0.367	0.334	0.453	0.185	0.614	

Kappa = 0.4688

	1	2	3	4	5	6	E.C.
1	23012	356	4	2807	0	866	0.149
2	2491	3354	16	211	255	477	0.506
3	5	28	13690	253	641	2623	0.206
4	5216	316	342	10390	30	3991	0.488
5	261	581	3217	260	3909	2230	0.626
6	1950	39	3263	3502	54	13098	0.402
E.O.	0.301	0.282	0.333	0.404	0.200	0.437	

Kappa = 0.5609

a non-quantitative variable. Figure 1 shows the potential vegetation map and the map showing the distribution of existing forests.

Comparison with the logistic regression model

The logistic regression method has recently been used to generate potential vegetation distribution models (Felicísimo et al., 2002). It permits an estimation of the probability of the presence of a particular type of forest on the basis of the values of n explicative variables. These variables can be measured on any scale (nominal, ordinal, interval or ratio), in accordance with the following expression:

$$P(Y=1|\mathbf{x}) = \frac{\exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}{1+\exp\left(\sum_{i=1}^{n} b_i x_i\right)}$$

where $P(Y=1|\mathbf{x})$ is the probability of forest existence at point x, b_0 is a constant and b_1 to b_n are coefficients for each explicative variable x_1 to x_n.

Results range between zero and one, indicating terrain-forest incompatibility and compatibility, respectively. As with the discriminant analysis method, the potential vegetation model is mapped by assigning to each cell the type of forest with the highest value on the corresponding suitability map.

Although it is not actually possible to test which of the models would result in better reforestation (it would mean waiting dozens of years for the forest to grow), we were able to test the validity of the model by analysing the extent to which the results matched the distribution of existing forests. Table 1 shows the error matrices and the Kappa index measurements (Rosenfield & Fitzpatrick-Lins, 1988) for the discriminant analysis and logistic regression models. It can be observed that the potential distribu-

Table 2. Error matrix and Kappa index for the logistic regression model and the weights of evidence method.

	1	2	3	4	5	6	E.C.
1	23117	394	480	5510	35	3132	0.2924
2	1870	2888	356	113	333	589	0.5303
3	21	7	12004	93	209	2291	0.1792
4	5185	343	709	8324	49	5116	0.5780
5	333	1024	3311	312	4225	1565	0.6077
6	2367	18	3668	3071	38	10572	0.4643
E.O.	0.2981	0.3821	0.4154	0.522	0.1358	0.5456	

Kappa = 0.4802

tion obtained using discriminant analysis provides a better match to the existing situation than the model obtained using the logistic regression method.

Inclusion of lithology as a variable in the model. Analysis of correspondences

The inclusion of the lithology variable in the models deserves special attention; this is because, as a non-quantitative variable, it receives a different treatment from the other variables.

For the logistic regression model it is necessary to define C-1 dummy variables for the C lithology classes present. In our case, this meant adding a further 18 variables to the 6 existing variables. This can be problematic in some cases - for example, when using the Idrisi Geographic Information System (version Idrisi32), which can only manipulate a maximum of 20 variables. For this reason we used the SPSS program (version 10.1) to calculate the logistic regression model. Combining the six logistic regression models into a single map and assigning to each cell the forest with the greatest value for this cell, we obtained results which were a poor match to the existing forest distribution, as can be seen in the error matrix in Table 4.

Felicísimo et al. (2002) use the weights of evidence method to multiply the probability value for each cell by a constant that depends on the different lithologies for this type of forest. The constant is greater than one when the weights are positive, less than one when they are negative and zero when the value is minus infinite (i.e. when a certain kind of lithology is not present in the forest type in question). Table 2 shows the matrix error for this method. As can be seen, the match is better than that of the logistic regression model without the lithology variable.

To introduce the lithology variable into the discriminant analysis, real values were

Table 3. Contingency table for actual forest type and lithology, with first-dimension lithology scores.

	FOREST							
LITHOLOGY	1	2	3	4	5	6	Total	Score
1	0	0	2	0	1	1	4	–0.7812
2	0	2	14	0	1	0	17	–0.7801
3	58	0	19	16	0	30	123	0.3276
4	55	3	84	53	3	81	279	0.1712
5	17	23	0	6	1	1	48	–1.3657
6	0	0	0	0	7	0	7	–3.0170
7	2	0	9	0	0	7	18	0.0078
8	1	0	0	0	0	0	1	0.3870
9	39	9	14	57	0	27	146	0.2005
10	1402	35	638	479	99	837	3490	0.1777
11	556	12	428	505	0	462	1963	0.2911
12	70	0	27	83	2	18	210	0.3644
13	130	0	0	37	0	30	206	0.4353
14	70	190	256	10	118	22	666	–1.4845
15	0	0	0	7	0	1	8	0.5877
16	21	34	19	0	76	20	170	–1.9161
17	12	5	0	0	3	0	20	–1.0161
18	0	0	0	0	11	2	13	–2.4842
19	0	0	0	0	4	0	4	–3.0170
Total	2442	313	1510	1253	326	1549	7393	

Table 4. Error matrix for model with lithology: discriminant analysis model with scored lithologies combined with correspondence analysis (left); logistic regression (right).

	1	2	3	4	5	6	E.C.
1	24119	362	5	2957	0	1098	0.155
2	1123	3252	16	19	238	133	0.319
3	9	57	13423	319	1036	3382	0.263
4	5339	409	305	10357	56	4143	0.497
5	70	546	3213	76	3069	700	0.600
6	2275	48	3570	3695	490	13809	0.422
E.O.	0.268	0.304	0.346	0.406	0.372	0.406	

Kappa = 0.5629

	1	2	3	4	5	6	E.C.
1	20112	337	5521	9383	150	10891	0.566
2	6124	2880	121	718	423	596	0.735
3	988	325	10272	232	1543	3209	0.380
4	4723	563	2846	6149	1553	6916	0.729
5	44	489	486	13	1177	306	0.532
6	944	80	1286	928	43	1347	0.709
E.0.	0.389	0.384	0.499	0.647	0.759	0.942	0.596

Kappa = 0.23

assigned to the different lithology classes in accordance with their capacity to discriminate the different types of forest through a correspondence analysis.

Correspondence analysis is a way of factoring categorical variables and representing them in a space that reflects their association in two or more dimensions (Greenacre, 1984). It tends to be used when a large number of rows and columns in the contingency table make it very difficult to understand associations between variables, but it can also be used to assign numerical values to categorical variables. The scores for the categories of one variable reflect their capacity to discriminate the other variable. Table 3 shows our contingency table, with an extra column added to record the scores for each lithology and forest type for the first dimension, which explains 68.8% of the variance (second dimension explains 16.3%). From these scores and the contingency table we can draw conclusions in regard to the relationship between both variables. Table 3 shows that lithologies 6 and 19, with very similar negative scores, are only present in forest 5.

The potential vegetation map is obtained by assigning first-dimension scores to each lithology variable and using this as a new quantitative variable in the discriminant analysis, together with the other variables (Figure 3). Comparing this error matrix model with the model without the lithology variable, we can observe that the errors are lower for four of the six forests. Of particular note is forest 2, with commission error falling from 50.56% to 31.98%.

Conclusions

The use of a deductive method supported by statistical methods to create potential vegetation models reduces subjectivity and thus represents an important advance in reforestation project design methods.

These models are created using multivariate analysis methods, such as discriminant

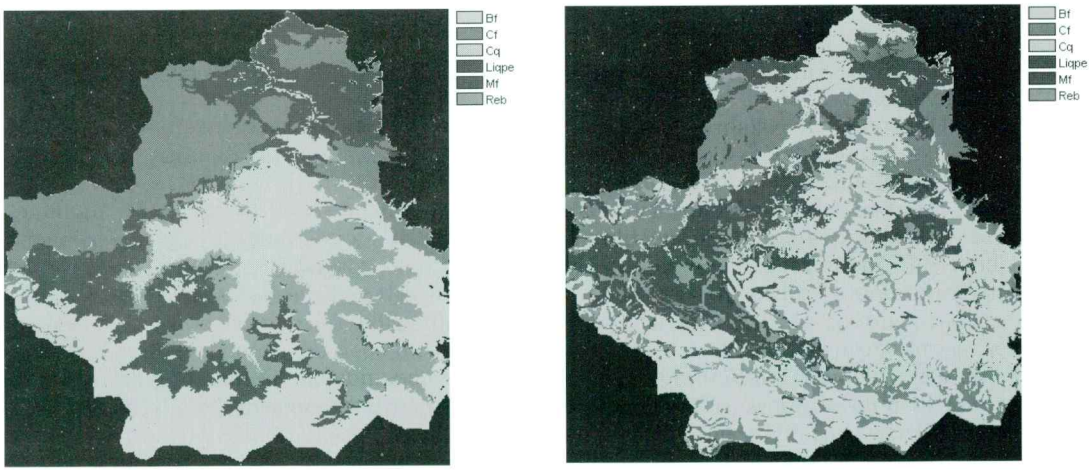

Figure 3. Potential maps including lithology as an independent variable: discriminant analysis model combined with correspondence analysis (left); logistic regression model (right).

analysis and logistic regression, which produce mathematical expressions associating independent variables with the dependent variable (in this case the forest).

When compared to the earlier logistic regression method proposed by some researchers, the discriminant analysis method combined with correspondence analysis produces a potential vegetation map that provides a better match to actual forest distribution. Matching results to existing distribution patterns is the only realistic way of corroborating results; the only other possibility would be to wait dozens of years for the forests to grow. The potential vegetation model only uses species which currently exist and in sufficiently large areas to provide a statistically representative sample, it not being possible, obviously, to take into account other poorly represented species or species which have already disappeared.

References

Felicísimo, A. M., Francés, E., Fernández, J.M., González-Díez A. & Varas J. 2002: "Modeling the Potential Distribution of Forests with a GIS". – Photogrammetric Engineering and Remote Sensing, *68*, 457–461.

Fisher, R.A., 1936: "The use of multiple measurements in taxonomic problems". – Annals of Eugenics, *7*, 179–188.

Fernández-Cepedal G. & Felicísimo, A.M. 1987: "Método de cálculo de la radiación solar incidente en áreas con apantallamiento topográfico". – Revista de Biología de la Universidad de Oviedo, *5*, 109–119.

Greenacre, M. J., 1984: Theory and Applications of Correspondence Analysis. London: Academic Press.

Guisan, A., J. P. Theurillat & F. Kiesnat, 1998: "Predicting the potential distribution of plant species in an alpine environment". – Journal of Vegetation Science, *9*, 65–74.

Taboada, J., Vaamonde, A., Saavedra, A., Ordóñez, C., 2002: "Geostatistical study of the feldspar content and quality of a granite deposit". – Engineering Geology, *65*, 285–292.

Jordan, M., Mateu, J. & Boix, A., 1998: "A classification of sediment types based on statistical multivariate techniques". – Wate, Air and Soil Pollution, *107*, 91–104.

Mayer, M., Wilkinson, I., Heikkinen, R., Orntoft, T. & Magid, E. 1998: "Improved laboratory test selection and enhanced perception of test results as tools for cost-effective medicine". – Clinical Chemistry and Laboratory Medicine, *36*: 683 – 690.

Morrison, D. F., 1976: Multivariate statistical methods, 2nd Ed.: McGraw-Hill, Inc., New York, 415 p.

Rosenfield, G.H. & K. Fitzpatrick-Lins, 1986: "A Coefficient of Agreement as a Measure of Thematic Classification Accuracy," – Photogrammetric Engineering and Remote Sensing, *52*, 223–227,

Skidmore, A. K., 1989. "A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model". – Int. J. Geographical Information Systems, *3*, 323–334.

Van de Rijt, C.W.C.J., L. Hazelhoff & C.W.P.M. Blom, 1996: "Vegetation zonation in a former tidal area: A vegetation-type response model based on DCA and logistic regression using GIS". – Journal of Vegetation Science, *7*, 505–518.