GIS and logistic regression as tools for environmental management: a coastal cliff vegetation model in Northern Spain

R. Álvarez-Arbesú[1] & A.M. Felicísimo[2]
[1]*Instituto de Recursos Naturales y Ordenación del Territorio, Universidad de Oviedo, Spain*
[2]*Área de Ingeniería Cartográfica, Geodesia y Fotogrametría, Universidad de Extremadura, Spain*

Abstract

One of the objectives of environmental planning is to present criteria for the conservation of biodiversity, the reduction of fragmentation of the vegetation, and the preservation of endangered communities. Correct planning requires knowledge of which are the most appropriate areas for each type of community or biotope, expressed cartographically by means of suitability models. These may be constructed by means of methods of spatial analysis which relate the current presence or absence of biological communities to a set of relevant environmental variables. Geographical information systems are an ideal tool for this task, since they integrate spatial database handling with the possibility of applying techniques of statistical analysis for model building.
From a map of the environmental variables, we elaborated a suitability model for marine cliff vegetation present in the study area (a part of the Autonomous Community of Asturias, N Spain) using logistic regression techniques integrated into a geographic information system.
The analysis of suitability models is useful in environmental management since it allows one to set up criteria for action in conservation and restoration by recommending priority zones and measures for the preservation of habitats declared of interest in the European Union. The present work describes in detail the methods used in constructing the models, and the results obtained. We discuss the models' capacity to serve as aids to decision making in environmental management, as well as the problems presented by the method that was used and the lines of future investigation aimed at reducing those problems.

1 Introduction

1.1 The littoral and its status

Europe's littoral and sea-cliffs are ecotopes with special characteristics: a strong marine influence, intense human use, great environmental value, and occupying small areas with marked environmental gradients.
The littoral of the Iberian Peninsula has been heavily modified, particularly in the South where mass tourism has been the main economic resource for decades. Along the Cantabrian coast (N Spain), the level of degradation is less, and it is still possible to act to recover the environmental quality provided by its natural vegetation.
The importance of the littoral is recognized explicitly in European regulations, from those of the widest level to those of local application. For instance, the introduction of Decree 107/93 of the Principality of Asturias (N Spain) proposes as a general strategy "to consider the entire littoral [...] as a territorial unit of high value, which, though partially degraded, it is necessary to preserve by a judicious choice of the opportune means of protection ...". Amongst the measures to be taken, it refers to conservation, restoration, reduction of fragmentation of communities, and conservation of diversity.
The main problem in defining the environmental action which will allow these goals to be achieved is the overall sparsity of information, and the poor quality or reliability of what information there is available. The usual case is to delimit by hand the zones of action where repopulation or plantation should be carried out, based on subjective criteria. The question is straightforward: Where and how can I act so as to have a reasonable guarantee of success? Ignoring this question has led to failure being the general norm of many reforestation programs, either because the species introduced are ill-adapted to the environmental conditions, or because the zones chosen for action were unsuitable.

1.2 The actual and the potential vegetation

The term actual or real vegetation is applied to plant communities that exist at a specific location at the time of its description. The real vegetation is a patchwork of different classes or categories of communities. Classifying these communities in accordance with some key allows one to construct a vegetation map, which is to be interpreted as a cartographic model of the real vegetation that is normally far more complex. While they are a simplification of reality, vegetation maps are indispensable documents for correct environmental management of a territory.
Vegetation maps have a limited temporal validity. Even when there is no human influence, vegetation dynamics are complex and intense, especially in littoral zones which are subject to sharp environmental gradients. In this sense, the ecological term of succession is applicable, defining the natural sequence in which an organism or group of organisms replaces another in some habitat over the passage of time. Then, the potential vegetation is defined as the stable community which would exist in an area as a consequence of progressive geobotanical succession if man ceased to affect and alter plant ecosystems.

Cliff vegetation is a case in which the potential vegetation never reaches the climax phase. Cliffs are ecotopes with steep slopes, little soil, and subject to a strong marine influence. This situation impedes succession culminating in a wooded phase (generally interpreted as the climax phase), and stability consists of a permanent non-climactic community.
For this reason, when restoration actions are undertaken, one has to ensure that the zones to be planted are suited to the type of vegetation, i.e., that they form part of its area of potential development. The task of delimiting the potential area has, until relatively recently, been carried out subjectively, and at scales that are of little use in planning (e.g. [1] , Rivas-Martínez, 1987). Given the interest for planners of this type of information, priority should be given to the development of methods which are both objective and applicable to small working scales.

2 Objectives

The specific objective of the present work is to construct a model of the potential distribution of subclimactic plant formations in a cliff zone of Northern Spain. This potential vegetation model will constitute a tool for delimiting the most suitable zones for environmental restoration.
In constructing the model, the distribution of the vegetation is assumed to be conditioned by a set of limiting factors that may be known or modeled, at least partially. The basic information is the present distribution of vegetation, and a set of topographical, climatic, and lithological data which together are organized into a geographic information system. The information in the GIS allows one to analyse the relationships between the presence or absence of plant communities and the values of the limiting factors.
The result will be a suitability or probability model for the presence of this type of vegetation in the study area. The suitability is expressed on a scale of 0 (incompatible) to 1 (ideal). The suitability value at a site will depend on the values of the factors that favour or limit the implantation of each type of vegetation.
From the suitability model, one will define an area of potential distribution, where the environmental factors have values that are well-suited to the implantation and growth of the plant community. Normally, the area of potential distribution is greater than the area of actual distribution, since the vegetation has been eliminated in many places.

3 Study area

The study area is in the zone of Cape Vidio (N Spain), within a space that is protected under the regulations of the Principality of Asturias, and known as Western Coast Protected Landscape. The topography is characterized by sharp changes in orientation and exposure to wave action. The length of coastline studied is about 15 km.
There were various reasons for choosing this zone to study. Firstly, there is good information available on some of the factors involved in the model: vegetation,

bathymetry, geomorphology, and topography. Secondly, there is great spatial heterogeneity, favouring the existence of diverse environments, and allowing a wide variety of environmental conditions to be included in the study. And thirdly, it is an area of environmental interest since it is within a protected area and includes plant communities classified as of Communitary interest by EU regulations.

4 Material and methods

The distribution model was constructed from different maps and digital models of the terrain that allow the actual vegetation (the dependent variable) to be related to the other environmental factors (independent variables). The cartographic material used was the following:

4.1 The dependent variable: the map of actual vegetation

The map of actual vegetation was constructed in the field on 1:5000 scale topographical maps with the aid of aerial photographs at approximately 1:18 000 scale taken in September 1983. During the first phase of the work, the plant communities were identified in the field, delimiting them on the aerial photographs. Then, in the laboratory, the information was tesselated onto the topographical maps.
The vegetation was classified according to height, structural characteristics, and successional significance (Table 1).

Table 1: Types of vegetation present in the study zone, successional significance, and percentage cover.

class	Description and successional significance	% cover
1	Tree and shrub formations of more than 2 m in height, representing subclimactic vegetation, conserved only in a part of their potential area	15
2	Halophytic low-stature shrub (26%) or herbaceous (3%) formations; vegetation adapted to salinity and sea-spray, characteristic of cliffs	29
3	Non-halophytic low-stature shrub or herbaceous formations (21%); vegetation present in other ecotopes not subject to marine influence	21
4	Zones with low vegetation cover (35%); degraded zones potentially available for environmental restoration action	35

4.2 The independent variables: topography, bathymetry, and geomorphology

The independent variables topography and bathymetry were modeled from the information in the topographical and nautical maps. These variables were included in the analysis for two main reasons: a) they are potentially influential variables in the vegetation's distribution, since they can act as a limiting factor on the different plant communities; and b) they are modelable variables, i.e., their distribution of values over the terrain can be estimated from the available information using a variety of methods.

The variables used were the following:

- Altitude: a 25 m cell size digital elevation model (DEM) was constructed from the topographical map and the nautical chart; the range of altitudes was from 0 to 75 m.
- The slope was estimated from the DEM; the resulting range was 0-70°.
- The orientation was estimated from the DEM.
- Soil depth was estimated from a 1:10 000 scale lithological map [4] (Menéndez Fernandez, 1995).
- Wave intensity (see below).

We calculated an intensity for the wave fronts to be used as an estimator of exposure to marine influence, and to differentiate coastal zones with different degrees of protection against wave action. To this end, we constructed manually a wave refraction model for the most frequent wave values in this zone of the Bay of Biscay (Table II).

There are basically three factors involved in refraction: the rate of re-orientation of the wavefronts induced by the bathymetry; the orientation of the wavefronts in the open sea, which depends on the wind direction; and the wave period, which depends on the distance from the cliff to the point at which a storm gave rise to the waves.

Table 2. Most frequent values of the variables involved in the calculation of the wave refraction model.

	Class	Relative frequency
Sector	NW	0.31 (of all sectors)
Significant height	1-2 m	0.41 (of the NW sector)
Wave period	10-12 s	0.27 (of the NW sector)

The refraction model was constructed by simulating the path of a wave-train for each direction and wave period, and then analysing the changes in orientation that it undergoes in passing from deep to shallow waters. The method used [5] (CERC, 1984) draws lines that are perpendicular to the advance of the wavefront, and whose separation is such that the energy dissipated between each of these normals is constant and takes the value unity. When they reach the cliff, the separation between these normals is inversely proportional to the amount of energy reaching each section of the coastline. In the refraction model, the values of the potential refraction range from 0.1 (the cells with the least exposure to waves) to 1.0 (the most exposed cells) .

All the independent variables were standardized so that their ranges of variation were similar.

4.3 Methods: logistic regression

Multiple logistic regression (MLR) has been used as a predictive method to generate likelihood models in a wide variety of fields, such as epidemiology [6] (Thomson et al., 1999), geological prospecting [7] (Agterberg, 1992), sylviculture [8] (Wilson et al., 1996), or wildlife conservation [9] (Mladenoff et al., 1999).

An MLR is well-suited to the present purpose since the dependent variable is dichotomous (presence/absence of vegetation), and the technique allows the independent variables to have non-Gaussian distributions. Also, the result of the regression ranges from 0 to 1 so that it is appropriate for the generation of a likelihood model [10] (Jongman et al., 1995).

The introduction of the spatial component into the MLR in order to generate cartographic models is fairly recent, and is usually included as a development tool in geographic information systems. Guisan et al. [11] (1998) use an MLR to generate a model of the distribution of a plant species, Carex curvula, in the Swiss Alps. Another similar study was applied to aquatic vegetation by van de Rijt et al. [12] (1996) using the GIS Grass (US Army Construction Engineering Research Laboratory).

The logistic model for the dependence of the probability of presence, $P(i)$, on the value of n explanatory variables is:

$$P(i) = \frac{1}{1 + e^{-[b(0)+b(1)\cdot x(1)+\ldots+b(n)\cdot x(n)]}} \tag{1}$$

where $P(i)$ is the probability of presence of the plant community, $x(1) \ldots x(n)$ represent the values of the environmental variables, and $b(1)..b(n)$ the corresponding coefficients. The results for each point of the terrain vary between the extremes 0 (incompatible, null likelihood of presence) and 1 (ideal).

Using the MLR to allocate a suitability value to each point of the terrain, one obtains a spatial distribution model.

The value may vary between 0 and 1. Values close to 0 indicate that the environmental variables make the site unsuitable for the development of the plant formation being analysed. Values close to 1 indicate sites where the combination of environmental values is ideal for the presence of that formation.

In the present work, the values of the coefficients of the logistic regression were calculated from a random sample of the zones, observing whether or not the plant community was present. The procedure was the following:

1. We formed two random samples of 990 and 998 points over the terrain. The first set was used for the calculation of the logistic model, and the second to test the fit of the model in order to guarantee independence of the test.
2. The logistic regression was performed using a commercial statistics program with a forward stepwise maximum likelihood method.
3. The goodness of fit was tested by comparing the results of the model with the testbed sample.

5 Results

5.1 Logistic regression tests

The best fitting logistic model finds all of the independent variables to be significant. The coefficients and statistics for the independent variables are listed in Table III.

Table 3. Variables, regression coefficients (b) and statistics for the logistic regression.

Variable	coefficient *b*	Standard error	Wald statistic	Significance
Orientation	0.5274	0.0590	79.82	$P<0.0001$
Slope	-0.6494	0.1088	35.62	$P<0.0001$
Refraction	-1.9500	0.2747	50.38	$P<0.0001$
Geomorphology	1.6073	0.1592	101.93	$P<0.0001$
Altitude	1.0908	0.1330	67.23	$P<0.0001$
Constant	-7.4554	0.7504	98.71	$P<0.0001$

5.2 ROC curve and classification error rates

The predictive capacity of the model can be evaluated as a function of the percentages of correct classifications, both for presences and for absences. These figures are deduced from the classification table (Table 4).

Table 4. Classification table to evaluate the model's predictive capacity via the proportions of false positives and false negatives (1: presence, 0: absence).

	predicted			
observed	0	1		
0	n_{00}	n_{01}	$n_{00}/(n_{00}+n_{01})$	sensitivity
1	n_{10}	n_{11}	$n_{11}/(n_{10}+n_{11})$	specificity

The sensitivity and specificity of the model depend on the threshold or cut-off which is set so as to classify each point according to its likelihood value.
Typically, as the threshold or cut-off point is changed, sensitivity and specificity vary inversely to each other. The Receiver Operating Characteristic (ROC) analysis summarise the performance of the logistic model. ROC curve is a plot of the relationship between sensitivity and specificity across all cut-off points of the model. The area under the curve is a estimator of the model accuracy. An area of 1 is perfectly accurate, whereas one of 0.5 is performing no better than chance.
The area corresponding to the present logistic model is 0.966, which implies a good predictive capacity (Fig. 1).

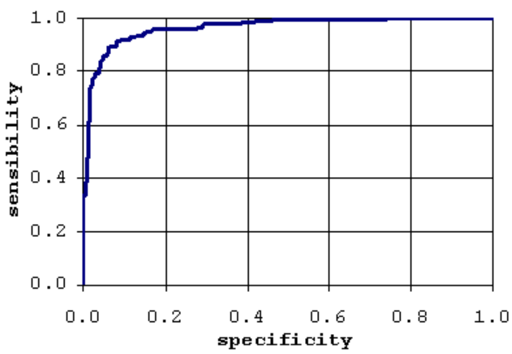

Figure 1. Receiver Operating Characteristic (ROC) curve. The area under the curve (an estimator of the model's accuracy) is 0.966.

The area of potential distribution is the sum (n_{01}+n_{11}). The value n_{11} represents the potential area which is actually occupied by tall woody formations. The false positives n_{01} represent the area of potential distribution which is not currently occupied (available potential area).

The values of each term depend on the classification threshold or cut-off point. A low threshold reduces the false negatives n_{10} and increases the area of potential distribution. This threshold can be arbitrarily lowered, reducing the error, but the potential area will then grow until it loses any real meaning.

The point of compromise can be obtained according to the expression put forward by Pereira & Itami [13] (1991) to evaluate the gain of the model: the improvement of the model over a random model is estimated by $(1-n_{10})-n_{01}$. In our case, the greatest improvement is 83% obtained for a threshold of 0.38.

5.3 Likelihood (or probability) surfaces

The final step in the model development process is the generation and cartographic representation of the probability values. The coefficients derived for the model were applied to the spatial information system database to generate a set of suitabilities (probabilities) ranging from 0 to 1 over the study area. The result is shown in Fig. 2.

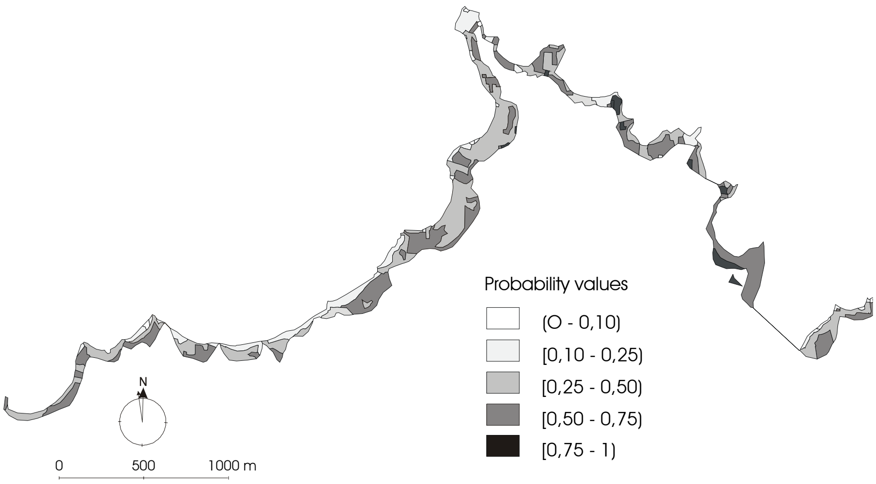

Figure 2. Suitability model for tall tree/shrub formations in the study zone.

6 Applications in environmental management

The final suitability map allows one to define the priority zones in the environmental management of a territory, and gives an orientation as to appropriate actions to take. As an example, two types of action can be defined in the present zone:

- Conservation of part of the present vegetation, applicable to the zones occupied by native pre-forest communities (fundamentally, shrub formations of bay laurel, Laurus nobilis).
- Reduction of the fragmentation of these communities via the recuperation of zones occupied by plantations of exotics (especially eucalyptus and pine), or by substitution formations - scrub or shrubs (mainly Cytisus sp.) where the degradation is moderate, and gorse or heath (Ulex sp. and Erica sp.) where degradation and loss of soil is greater.

The first type of action (conservation) would be applied to zones where the current vegetation is interpreted as pre-forest. The suitability model allows the zones for the second type of action to be defined, since one would choose those occupied by substitution formations and which also present high values of suitability that would guarantee success.

7 Discussion

The process of constructing the model of potential distribution is carried out under the hypothesis that the actual distribution of tree and shrub communities is a sufficient sample of the potential distribution.

The communities modeled in the present work are widely distributed over the cliff faces of the Cantabrian littoral. They are, however, fragmented and declining in area with each year that passes. The suitability model that was constructed is a fairly accurate reflection of their current spatial distribution, and the prediction of the area of potential distribution is biologically reasonable in extension and characteristics. Another argument in support of the apparent trustworthiness of the model is the analysis of the actual vegetation in the potential area. A large part of this area consists of communities dominated by ferns, an element which is regarded as an indicator of forest in zones which have lost the direct halophytic influence of the sea. In any case, they may be considered to be substitution formations in the sense of the ecological succession of the communities that were modeled.
The transition from model to action is a problem which requires complementary information, in particular about the ownership of the land and its present economic management. The opportunity must not be missed, however, of recovering, at least in the environmental sense of the expression, zones which mostly lie within the public domain.

8 References

[1] Rivas-Martínez, S., *Memoria del mapa de series de vegetación de España*, Publicaciones Ministerio Agricultura, Pesca y Alimentación: Madrid, 1987.
[2] CERC. *Shore Protection Manual*. U.S. Army Coastal Engineering Research Center: Vicksburg, Mississippi. 1984.
[3] Agterberg, F.P. Estimating the probability of occurrence of mineral deposits from multiple map patterns". *Use of Microcomputers in Geology,* eds. Merriam, D.F. & Kürzl, H., Plenum Press: New York, pp. 73-92. 1992.
[4] Wilson, W.L., Day, K.R., & Hart, E.A. Predicting the extent of damage to conifer seedlings by the pine weevil (Hylobius abietis L.): a preliminary risk model by multiple logistic regression. *New Forests*, 12, pp. 203-222, 1996.
[5] Mladenoff, David J., Sickley, Theodore A. & Wydeven, Adrain P. Predicting gray wolf landscape recolonization: logistic regression modelos vs. new field data, *Ecological Applications*, 9(1), pp. 37-44, 1999.
[6] Jongman, R.H.G., Ter Braak, C.J.F. & van Tongeren, O.F.R. (eds.) *Data analysis in community and landscape ecology*. Cambridge University Press: Cambridge, 1995.
[7] Guisan, A., Theurillat, J.-P. & Kienast, F. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*, 9, pp. 65-74, 1998.
[8] van de Rijt, C.W.C.J., Hazelhoff, L. & Blom, C.W.P.M. Vegetation zonation in a former tidal area: a vegetation-type response model based on DCA and logistic regression using GIS. *Journal of Vegetation Sci.*, 7, 505-518, 1996.
[9] Pereira, J.M.C. & Itami, R.M. 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, 1991.