Factor analysis and logistic regression for forest categorical and quantitative data

Kyriaki Kitikidou (Corresponding author)

Dimokritos University of Thrace, Greece, Department of Forestry and Management of Environment and Natural Resources

Northern Canada University

Pandazidou 193, 68200, Orestiada, Greece

Tel: +30-6932042106 E-mail: kkitikid@fmenr.duth.gr

Nikolaos Arambatzis

Prefecture of Chalkidiki, Department of Agricultural Development, Greece

Abstract

Analyzing units described by a mixture of sets of quantitative and categorical variables is a relevant challenge. In this paper, data were extracted from the management study of the University forest of Taxiarchis (Northern Greece). The variables used in the analyses were: site quality, exposure, rock, soil, altitude, incline, tree age, tree crown ratio, tree height, and section size. Principal components analysis (PCA) was used to include these two types of variables in order to study the correlations of a large number of forest variables by grouping the variables in factors. A few components accounted for most of the variation, and these components were used to replace the original variables. Thus, this method reduced the number of variables in the data file. Further on, hierarchical cluster analysis (HCA) was used for detecting groupings in data. The objects in these groups were cases (forest sections). Finally, a multinomial logistic regression model was fitted to the treated data, and it was proved to be quite useful for site quality prediction in the above-mentioned forest.

Keywords: Forest variables, hierarchical cluster analysis, multinomial logit, principal components analysis.

1. Introduction

In different studies, statistical units are simultaneously described by heterogeneous variables, quantitative and categorical. A particular case, used as an application of the method that we propose, arises in experimental designs where plots established in forests are described with both quantitative and categorical data.

The problem of mixed data has already been studied. Gower (1971) first proposed a solution for balancing quantitative and categorical variables, whatever their type. Specific distances are used for categorical and quantitative variables; the range of variation of every distance is standardized to 1 before aggregation in a global distance. Moreover, by using a set of a priori weights, users can favour those variables that they consider to be important. Podani (1999) extends Gower's general coefficient to ordinal variables. Grabmeier & Rudolph (2002) follow Gower's proposal to standardize the range of variation of any distance, but considering a larger range of distances; sets of variables are considered but only to group the variables depending on the associated distance. Balancing is performed at the variable level, in order to give the variables equal importance (or the importance decided by the user).

The introduction of a mixture of quantitative and categorical poses the problem of unit weighting. In our study, we applied Principal Components Analysis (PCA) (Jackson, 2003; Morrison, 2004) and Hierarchical Cluster Analysis (HCA) (Sharma, 1995; Vizirgiannis, Haldiki, & Gynopulos, 2003), in order to apply a multinomial regression model (Agresti, 2002; Santer, & Duffy, 2004) for forests' site quality estimation. The aim of this work was to:

1) study the correlations of a large number of forest variables by grouping the variables in factors
2) detecting groupings in forest sections
3) estimate a model to analyze "site quality" variable, predicting its value with several variables measured in Taxiarchis University forest (Northern Greece).

2. Materials and Methods

This study was carried out in Taxiarchis University forest, Northern Greece (40˚27' N, 23˚32' E). The study forest has an area of 3895 ha (University Forests Administration, 1994). For the data collection point sampling, with systematic point selection, was used (Matis 2004). In a forest map of Taxiarchis 50 parallel lines 500 km apart were drawn. A point was selected every 200 m at length of the parallels. The sample that resulted had a size of 350 points, corresponding to 30 different forest sections, according to the management plan (University Forests Administration 1994).

Data on these 350 points is extracted on site quality (Site Quality=1 for best, 5 for worst), exposure (Exposure=1 for North, 2 for South, 3 for East, 4 for West), rock (rock=1 for mica schist, 2 for mica schist and gneiss, 3 for mica schist and limestone, 4 for talc schist), soil (1 for sandy, 2 for clay, 3 for loamy), altitude (m), incline (%), tree age (years), tree crown ratio (%), tree height (m), and section size (ha).

First, PCA was applied in our data. PCA is a method of decomposing a correlation or covariance matrix, and it is often used in exploratory data analysis to:

- Study the correlations of a large number of variables by grouping the variables in factors so that variables within each factor are more highly correlated with variables in that factor than with variables in other factors.
- Interpret each factor according to the meaning of the variables.
- Summarize many variables by a few factors. The scores from the factors can be used as input data for *t* tests, regression, ANOVA, discriminant analysis, and so on. In our study, PCA factors are used as independent variables in a multinomial regression model.

Cluster analysis is a multivariate procedure for detecting groupings in data. The objects in these groups may be cases or variables. In HCA, initially, each object (case or variable) is considered as a separate cluster. Then two closest objects are joined as a cluster and this process is continued (in a stepwise manner) for joining an object with another object, an object with a cluster, or a cluster with another cluster until all objects are combined into one single cluster. This hierarchical clustering is then displayed pictorially as a tree referred to as the Hierarchical tree. The term 'closest' is identified by a specified rule in each of the Linkage methods. Hence in different linkage methods, the corresponding distance matrix (or dissimilarity measure) after each merger is computed by a different formula. In our study, HCA is used in order to reduce data by grouping similar forest sections.

Finally, a multinomial logistic regression model is applied, having site quality as dependent variable and PCA factors as independents. PCA factors weighted our data variables and HCA weighted our data cases, respectively.

3. Results - Discussion

Component loadings (rotated) are correlations of the variables with the principal components (factors) and they are given in Table 1.

The Variance explained for each component is the eigenvalue for the factor. The first factor accounted for 58.4% of the variance; the second, 22.1%. The Total Variance is the sum of the diagonal elements of the correlation matrix. By summing the Percent of Total Variance Explained for the two factors (58.4+22.1= 80.5), we can say that more than 80% of the variance of all eleven variables was explained by the first two factors.

To interpret each factor (Figure 1), we look for variables with high loadings. The five variables that load highly on factor 1 (exposure, rock, soil, altitude, incline) can be said to measure "site"; while the three that load highly on factor 2 (age, tree crown ratio, tree height), "trees". This data set included a variable (section size) that did not loaded highly on any specific factor.

In the factor scree plot (Figure 1), the eigenvalues are plotted against their order (or associated component). We used this display to identify large values that separated well from smaller eigenvalues. This can help to identify a useful number of factors to retain. Scree is the rubble at the bottom of a cliff; the large retained roots are the cliff, and the deleted ones are the rubble. The points in the factor loadings plot are variables, and the coordinates are the rotated loadings. We looked for clusters of loadings at the extremes of the factors. The five variables at the right of the plot load highly on factor 1 and all reflect “site”. The three variables at the top of the plot load highly on factor 2 and reflect “trees”.

Using hierarchical cluster analysis, the forest sections were joined first at a distance of 0.087. The last entry represents the joining of the largest two clusters to form one cluster of all 30 sections. The clusters are best illustrated in a tree diagram (Figure 2). The scale for the joining distances is printed at the bottom. Sections formatted three groups at a distance of 0.860. Finally, at a distance of 0.908, all sections were added to form one large cluster.

Parameter estimates from the multinomial logistic regression are given in Table 2. All parameter estimates are significantly different from zero (p-values <0.05). In a multinomial context, we will want to know how the probabilities of each of the outcomes will change in response to a change in the covariate values. This information is provided in the derivatives (Table 3), which tells us, for example, that when “site” factor increases by one unit, the probability of Site Quality category 3 goes up by 0.042, and Site Quality categories 1 and 2 go down by 0.017 and 0.025, respectively.

4. Conclucion

The need to simultaneously consider mixed data, composed of sets of quantitative and categorical variables, arises in a wide range of applications. This can be done by combining Principal Components Analysis, and Hierarchical Cluster Analysis, so that quantitative and/or categorical variables may be taken into account. This methodology balances the influence of variables, allowing the units to be clustered based on all the variables, and can be considered as a solution proposed to the unit weighting problem that arises when type of data is quantitative and categorical.

This technique can be very useful in forest data analysis: it is a natural requirement for a researcher to be able to combine several measurements in different scales, in order to estimate an important variable for management purposes, such as site quality.

References

Agresti. A. (2002). *Categorical data analysis. 2nd ed.* New York: John Wiley & Sons.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857–871.

Grabmeier, J., & Rudolph, A. (2002). Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, 6, 303–360.

Jackson, J. (2003). *A user’s guide to principal components*. New York: Wiley Interscience.

Matis, K. (2004). *Sampling of natural resources. 3rd ed*. Pigasos publications: Thessaloniki, Greece.

Morrison, D. (2004). *Multivariate statistical methods. 5th ed*. CA: Duxbury Press.

Podani, J. (1999). Extending Gower’s general coefficient of similarity to ordinal characters. *Taxon*, 48(2), 331–340.

Santer, T., & Duffy, D. (2004). *The statistical analysis of discrete data*. New York: Springer-Verlag.

Sharma, S. (1995). *Applied multivariate techniques*. New York: John Wiley & Sons.

University Forests Administration. (1994). Management study of Taxiarchis – Chalkidiki 2002-2011. Thessaloniki, Greece.

Vizirgiannis, M., Haldiki, M., & Gunopulos, D. (2003). *Uncertainity handling and quality assessment in data mining*. London: Springer-Verlag.

Table 1. Component loadings from Principal Components Analysis.

	1	2
exposure	0.9191	0.1638
rock	0.8998	0.2599
soil	0.8992	0.2295
altitude	0.8871	0.2507
incline	0.8404	0.1806
age	0.1068	0.8403
tree crown ratio	0.2509	0.7496
tree height	0.2508	0.8866
section size	0.1807	0.1704

Table 2. Parameter estimates of the multinomial logistic model.

Parameter	Estimate B	Standard error	Exp(B)	p-value	95% Confidence Interval	
					Lower	Upper
1 CONSTANT	2.5435	0.9834	2.5864	0.0097	0.6161	4.4709
2 SITE	-0.1917	0.0768	-2.4956	0.0126	-0.3423	-0.0411
3 TREES	-0.0003	0.0001	-2.1884	0.0286	-0.0005	0.0000

Table 3. Individual variable derivatives averaged over all observations, from the multinomial logistic model.

PARAMETER	1	2	3
1 CONSTANT	0.2033	0.3441	-0.5474
2 SITE	-0.0174	-0.0251	0.0425
3 TREES	0.0000	0.0000	0.0001

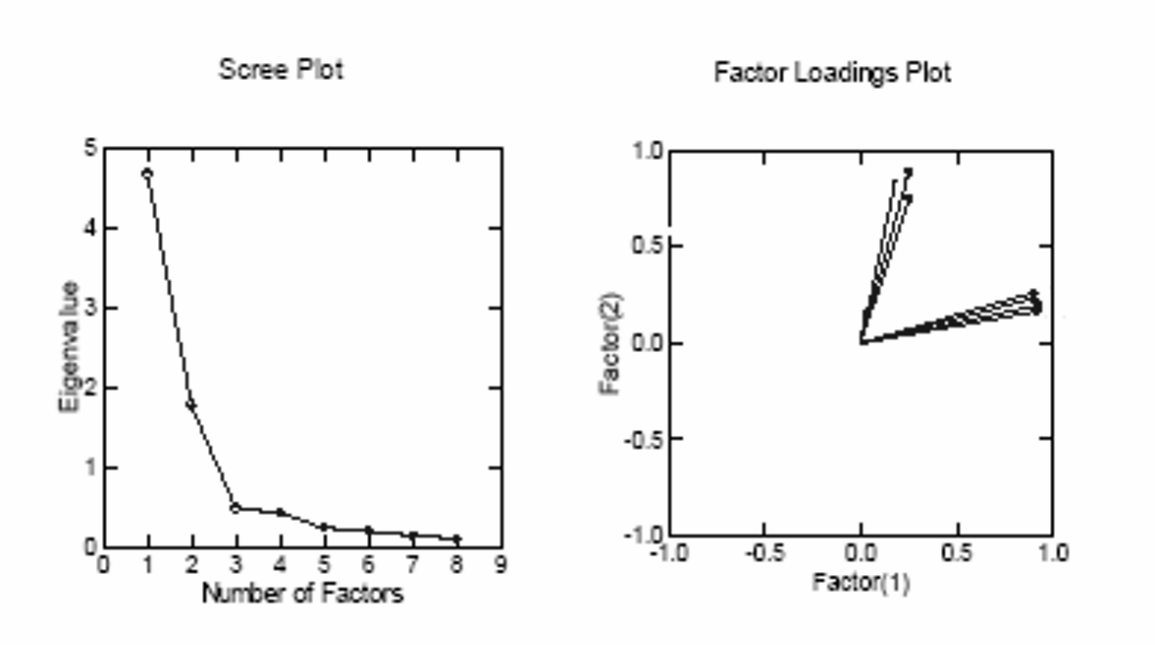

Figure 1. Principal Components Analysis plots.

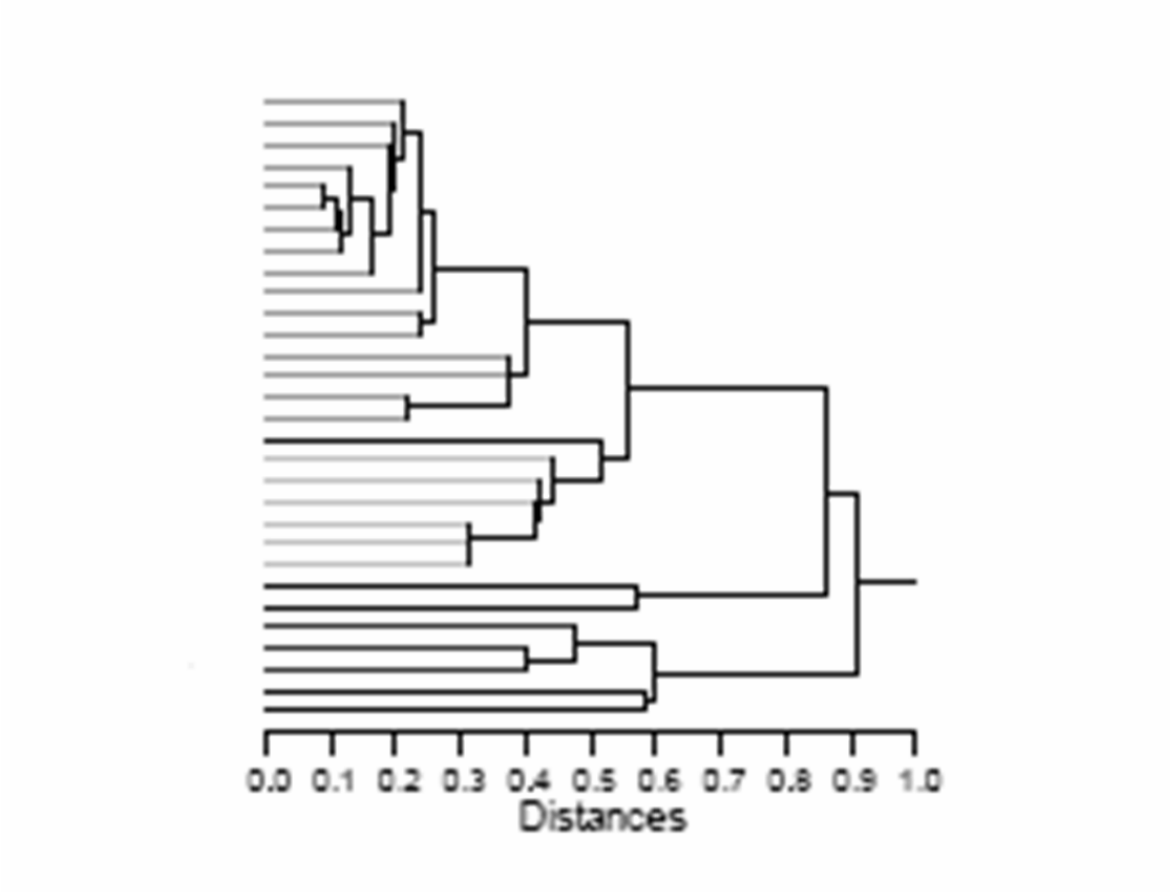

Figure 2. Dendrogram from Hierarchical Cluster Analysis.