Coenoses 8(2):81-84, 1993

ON IDENTIFYING REPRESENTATIVE VARIABLES IN ECOSYSTEM STUDIES

Sun Ping

Department of Plant Protection, Beijing Agricultural University, Beijing 100094. P. R. China

Keywords: Classification, Ecosystem, Principal Components Analysis, Variable selection.

Abstract. In ecosystem analysis, we always need to classify an ecosystem into several sub-ecosystems using the variables that define the system. It is hoped that the variables describe the special features of the sub-systems. How to select variables that are of this kind? This question is discussed and a method is described. An example of the analysis of meteorological types from the Baoding discrict, Hebei Province, China, is also presented.

1. Introduction

We can select a smaller number of variables from a group of variables for statistical analysis according to different criteria (Orlóci 1978, p. 25; Dale 1986; Sun and Lin 1991). These criteria are relevant in a study of the wheat ecosystem in the Baodin district, Hebei Province, through which we wish to find the characteristic meteorological features of the district. The objective is to recognize meteorological types, and through this, to categorize the meteorological conditions in the district annually to help predict the dynamics of biological populations in wheat ecosystems. This could have practical value in management and improvement of wheat production.

The meteorological conditions in any one year are reflected by the states of at least 36 meteorological factors, including such typical kinds as the monthly average temperature, average relative humidity, and rainfall. If we try to use the 36 meteorological factors in principal components analysis to recognize meteorological types, the eigenvectors will still be vectors in a 36 dimensional space. These are too many to have practical usefulness. Therefore, the problem of variable selection prior to analysis arises.

In this paper we outline a new method and give an example to illustrate its application in the solution of a practical problem.

2. The Steps

The method involves a classification of the variables and a search for those regarded representatives. The method used here has similarity to an early method (Sun & Lin 1991), but differs from it in important respects. The aims are also different.

Suppose that $x_1, x_2, \ldots, x_m$ are the variables that we observed n times. $x_{i1}, x_{i2}, \ldots, x_{im}$ represents the ith observation. We wish to select a subset of the variables that satisfy strict criteria. The following are steps:

(1) Calculate a correlation matrix $\mathbf{R}=(r_{ij})$ with an element,

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sigma_i \sigma_j}$$

in which

$$\bar{x}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki}, \quad \sigma_i^2 = \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2$$

and $i=1$ to m, $j=1$ to m.

(2) Given the confidence limit α (significance level $1-\alpha$) and n, based on the F-distribution, we find the critical value $f=F(1, n-2, \alpha)$, and from $f=(n-2)r^2/(1-r^2)$ (Draper and Smith 1981) we obtain the critical $\hat{r}$ value. We inspect the ith row of R and find the value larger than $\hat{r}$. We group the corresponding variables together and name the group as C(i,k). In this group variable x_i is the representative element and there are k other elements $C(i,k)=\{x_i | x_{i1}, x_{i2}, \ldots, x_{ik}\}$. Every element in C(i,k) is linearly related to x_i with a significance level $1-\alpha$. Further details of this are described by Sun and Lin (1991). The end product is a set of m groups of variables.

(3) We order the m groups and denote the first group by C(1). C(1) should have the largest number of elements t+1. If there are several groups that have the same number of elements, such as C(i,t) and C(j,t), each having the largest number of elements, we can choose the order of the two groups after comparing the average value of correlation coefficients of the representative element and the other elements in the groups according to

$\bar{r}_i=(1/t)(r_{i,i1} +\ldots+r_{i,it})$

$\bar{r}_j=(1/t)(r_{j,j1} +\ldots+r_{j,jt})$

If $\bar{r}_i > \bar{r}_j$, then C(i,t) will be C(1) and C(j,t) will be C(2). Following these, we consider in the same manner the groups which have the second largest number of elements, and continue until the ordering of the m groups C(1), C(2),... C(m) is complete.

(4) The next step is sifting. We consider group C(m) firstly. If all the elements in C(m) appear in other groups, we reject C(m), otherwise we preserve C(m). Then we consider C(m-1),..., C(2), C(1). In general, if all the elements in C(i) appear in C(j) (j<i) or in the preserved groups, we reject group C(i). If not, we preserve group C(i).

The characteristics of the preserved groups are as follows:

1. Each of the m variables appears at least in one of the preserved groups.

2. The preserved groups are relatively distinct. In any one of the groups, there is at least one element which cannot be found in any other group.

The representative elements of the preserved groups are chosen as the representative variables of the m variables.

This type of variables selection is unique, and the variables with strong connection with other variables are protected.

3. Application

To establish the wheat production management system in Baoding district, we needed to classify the meteorological types in the district. We had ten years of data (1980-1989) and decided to use principal components analysis for this purpose. There is a total of 36 measured meteorological factors. We use x_1, ..., x_{12} to denote the monthly average temperature from July to June, x_{13},...x_{24} to denote the monthly average relative humidity from July to June, and x_{25},...,x_{36} to denote one-third of the monthly rainfall from July to June. i=1 to 10 denotes the year from 1980 to 1989 and x_{i1},... , x_{i36} corresponds to the ith year's observed values. The data of $x_{ij,}$ i=1 to 10, j=1 to 36 were listed in Sun and Lin (1991, Appendix 1).

We obtained six principal components in the analysis, which implies that there are six different meteorological types in the ten-year data. But since each of the six eigenvectors is a 36 dimensional vector, the differences of the meteorological type are blurred. It is prudent for us to select a smaller subset of variables before principal components analysis. To achieve this, we chose as a confidence limit α = 0.9 (significance level = 0.1) and analyze the data set of Sun and Lin (1991) accordingly. In the first three steps of variable selection, we obtain the following 36 groups:

C(1): $\{x_{25}| x_4, x_7, x_{13}, x_{14}, x_{27}\}$ r_{25} =0.708

C(2): $\{x_{30}| x_{13}, x_{18}, x_{21}, x_{32}, x_{33}\}$ r_{30} =0.699

C(3): $\{x_7 | x_4, x_{13}, x_{25}, x_{27}, x_{31}\}$ r_7 =0.66

C(4): $\{x_9 | x_8, x_{10}, x_{11}, x_{14}\}$ r_9 =0.703

C(5): $\{x_{13}| x_7, x_{14}, x_{15}, x_{20}\}$ r_{13} =0.667

C(6): $\{x_{14}| x_9, x_{13}, x_{25}, x_{26}\}$ r_{14} =0.666

C(7): $\{x_8 | x_9, x_{11}, x_{19}, x_{27}\}$ r_8 =0.660

C(8): $\{x_{27}| x_7, x_8, x_{15}, x_{25}\}$ r_{27} =0.625

C(9): $\{x_{21}| x_{12}, x_{30}, x_{32}\}$ r_{21} =0.713

C(10): $\{x_{32}| x_{20}, x_{21}, x_{30}\}$ r_{32} =0.690

C(11): $\{x_{31}| x_7, x_{19}, x_{36}\}$ r_{31} =0.655

C(12): $\{x_2 | x_{17}, x_{23}, x_{29}\}$ r_2 =0.654

C(13): $\{x_{35}| x_3, x_6, x_{23}\}$ r_{35} =0.632

C(14): $\{x_{17}| x_2, x_{12}, x_{29}\}$ r_{17} =0.619

C(15): $\{x_{16}| x_5, x_{28}\}$ r_{16} =0.809

C(16): $\{x_5 | x_{16}, x_{28}\}$ r_5 =0.794

C(17): $\{x_3 | x_6, x_{35}\}$ r_3 =0.718

C(18): $\{x_{28}| x_5, x_{16}\}$ r_{28} =0.703

C(19): $\{x_{11}| x_8, x_9 \}$ r_{11} =0.676

C(20): $\{x_{19}| x_8, x_{31}\}$ r_{19} =0.675

C(21): $\{x_{23}| x_2, x_{35}\}$ r_{23} =0.669

C(22): $\{x_{26}| x_{14}, x_{33}\}$ r_{26} =0.667

C(23): $\{x_4 | x_7, x_{25}\}$ r_4 =0.651

C(24): $\{x_6 | x_3, x_{35}\}$ r_6 =0.632

C(25): $\{x_{33}| x_{26}, x_{30}\}$ r_{33} =0.632

C(26): $\{x_{29}| x_2, x_{17}\}$ r_{29} =0.600

C(27): $\{x_{12}| x_{17}, x_{21}\}$ r_{12} =0.598

C(28): $\{x_{15}| x_{27}\}$ r_{15} = 0.819

C(29): $\{x_{18}| x_{30}\}$ r_{18} = 0.654

C(30): $\{x_{20}| x_{32}\}$ r_{20} = 0.645

C(31): $\{x_{10}| x_9 \}$ r_{10} = 0.641

C(32): $\{x_{36}| x_{31}\}$ r_{36} = 0.618

C(33): $\{x_{34}\}$

C(34): $\{x_{24}\}$

C(35): $\{x_{22}\}$

C(36): $\{x_1\}$

By completing the 4th step, there are 14 groups left, including C(1), C(2), C(4), C(5), C(6), C(9), C(11), C(12), C(13), C(14), C(33), C(34), C(35), C(36). The 14 representative elements are chosen as the representative variables in principal component analysis. Designating **X** the vector $(x_1, x_2, x_9, x_{13}, x_{14}, x_{16}, x_{21}, x_{22}, x_{24}, x_{25}, x_{30}. x_{31}, x_{34}, x_{35})'$, and using the SAS package (SAS Institute, 1985), we obtain the following eigenvectors:

U_1	U_2	U_3	U_4	U_5
-0.30	0.01	-0.50	0.03	0.15
-0.30	0.08	0.14	-0.02	-0.52
0.20	0.44	-0.31	-0.11	0.55
0.44	0.02	0.15	0.18	-0.03
0.40	0.26	-0.15	0.01	0.07
-0.11	0.39	0.21	-0.24	0.21
0.31	-0.37	0.00	-0.16	-0.25

0.14	-0.05	0.22	-0.37	0.55
-0.05	-0.28	0.19	0.37	0.46
0.27	0.32	0.16	0.37	-0.18
0.35	-0.39	-0.04	0.12	-0.05
-0.06	0.27	0.12	0.58	0.14
0.12	0.17	0.48	-0.31	-0.13
-0.30	-0.08	0.45	0.11	-0.01

The proportions of the corresponding characteristic roots are 0.32, 0.20, 0.15, 0.13 and 0.10. For convenience we introduce the following eigenvector matrix

$U = (U_1, U_2, \ldots\ldots U_k) = (u_{i,j})$

$i=1$ to m and $j=1$ to k. In the example, $m = 14$ and $k = 5$. We use V_i to denote the ith row vector of the eigenvector matrix.

As an outcome of the analysis, five meteorological types are recognized in Baoding. The meteorological type of eigenvector 1 represents lower average temperature in July and August of the previous year and lower relative humidity in October last year, less rainfall in May (since $u_{1,1}$, $u_{1,2}$, $u_{1,6}$, $u_{1,14}$ are smaller); higher relative humidity in last July and in March, and more rainfall in January (Since $u_{1,4}$, $u_{1,7}$, $u_{1,11}$ are larger). Other eigenvectors define other distinct meteorological groups.

For the purpose of comparison, we list the results which we obtained based on using the 36 variables. **X** is the vector $(x_1, \ldots, x_{12}, x_{13}, \ldots, x_{24}, x_{25}, \ldots, x_{36})'$. There are six principal components and the corresponding eigenvector matrix is:

0.08	-0.09	-0.31	-0.18	0.02	0.23
0.16	-0.19	0.04	0.16	0.21	0.11
0.26	-0.03	-0.03	-0.26	-0.03	-0.03
0.03	0.05	0.29	0.24	0.18	0.00
0.19	0.06	-0.11	0.26	-0.24	0.02
0.17	0.09	0.20	-0.16	-0.14	-0.16
0.15	0.26	0.21	-0.02	0.19	-0.06
0.04	0.31	-0.12	-0.21	0.14	-0.03
0.00	0.30	-0.24	0.06	0.11	-0.09
0.11	0.07	-0.31	0.02	0.28	0.07
-0.13	0.21	-0.15	-0.06	0.11	-o.25
-0.17	-0.28	-0.05	0.06	0.14	-0.15
-0.18	0.26	0.19	0.03	-0.03	-0.11
-0.15	0.29	-0.03	0.22	0.00	-0.05
0.05	0.15	-0.09	0.14	-0.14	0.31
0.21	0.02	-0.06	0.33	-0.14	-0.13
0.09	-0.24	-0.17	0.14	0.28	-0.06
-0.18	-0.09	0.17	-0.20	0.08	0.25
0.17	0.20	0.00	-0.26	0.00	0.07
-0.16	0.03	-0.15	-0.21	-0.08	0.08
-0.31	-0.05	0.12	-0.01	0.09	-0.19
-0.04	0.05	-0.04	-0.01	-0.39	-0.24
0.17	-0.17	0.19	0.14	0.01	-0.05
-0.02	-0.07	0.23	-0.04	-0.34	0.03
0.01	0.30	0.19	0.15	0.11	0.05
-0.21	0.08	-0.09	0.28	0.14	0.12
0.12	0.22	-0.06	0.05	0.09	0.25
0.09	-0.02	-0.17	0.27	-0.23	-0.04
0.14	-0.13	0.10	-0.01	0.40	-0.09
-0.32	0.04	0.15	-0.06	-0.04	0.00
0.19	0.20	0.15	-0.13	0.07	0.20
-0.30	-0.01	-0.10	-0.15	0.05	0.00
-0.23	-0.06	0.10	0.23	0.01	0.27
0.09	0.06	0.15	0.03	0.05	-0.42
0.22	-0.16	0.18	-0.15	-0.02	-0.05
0.02	0.07	0.28	0.08	0.07	0.27

The proportions of the corresponding characteristic roots are 0.22, 0.18, 0.15, 0.12, 0.10, 0.10. The six eigenvectors represent six different meteorological types, but there is linear correlation between the elements. For example, the row vectors

$V_{16}=(0.21\ 0.02\ -0.06\ 0.33\ -0.14\ -0.13)$

$V_{28}=(0.09\ -0.02\ -0.17\ 0.27\ -0.23\ -0.04)$

are very closely related. The regression equation of the two is

$y = -0.5 + 0.86\, x$,

and the correlation coefficient of the regression equation is 0.92. In this, the elements of V_{16} are used as the values of x and the elements of V_{28} are treated as the corresponding values of y.

Since the tendency of change of the elements in the two vectors is almost equal, it is obvious that the contribution that v_{16} made in classifying the meteorological types is similar to that what v_{28} has made. To classify the six meteorological features, one of the two elements can be deleted. In the process of variable selection, 22 variables are removed. So there must be some meteorological information lost.

If all the 36 variables are used, there are six principal components, i.e., six meteorological types. If the 14 representative variables are used, there are only five principal components, i.e., only five meteorological types. It is difficult for us to say exactly how much information is lost, but we think much of it was redundant. There is not much difference in the classification of meteorological features if six types or five types are used but to use 14 variables is much more convenient than to use 36 variables.

The positions of the 14 representative variables is clear enough for us to carry out further variable selection. For example, since variables x_{34}, x_{24}, x_{22} and x_1 only represent themselves, the meteorological changes in the whole year are less affected by these variables. If we just consider the first ten representative variables

(x_2, x_9, x_{13}, x_{14}, x_{16}, x_{21}, x_{25}, x_{30}, x_{31}, x_{35}), we obtain the eigenvector matrix:

$$\begin{array}{rrr}
-0.33 & 0.02 & 0.02 \\
0.24 & 0.42 & -0.26 \\
0.43 & 0.05 & 0.31 \\
0.42 & 0.28 & -0.17 \\
-0.16 & 0.40 & -0.30 \\
0.32 & -0.42 & -0.04 \\
0.26 & 0.37 & 0.39 \\
0.38 & -0.37 & 0.18 \\
-0.06 & 0.36 & 0.58 \\
-0.37 & -0.08 & 0.44
\end{array}$$

The proportions of the corresponding characteristic roots are 0.40, 0.26 and 0.15, respectively. The three eigenvectors represent three meteorological types. Interestingly, three meteorological types, described only by ten variables, proved sufficient to carry out ecosystem analysis effectively.

Acknowledgments. The author thanks Professor G. P. Patil very much for encouragement to prepare this paper, and for Professor L. Orlóci for notes and comments on the exposition and the contents of the paper.

References

Anderson, T. W. 1958. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Anderson, T. W. 1963. Asymptotic theory for principle component analysis. Ann. Math. Statist. 34 122-148.

Draper, N. R. and Smith, H., 1981. Applied Regression Analysis (2nd edition). Wiley, New York.

Dale, M. B. 1986. Comparison of some methods of selecting species in vegetation analysis. Coenoses 1: 35-51.

Orlóci, L. 1978. Multivariate Analysis in Vegetation Research. 2nd ed. Junk, The Hague.

SAS Institution. 1985. SAS User's Guide, Statistics, Version 6.02. SAS Institute, Cary, N.C.

Sun, P. and Changshan Lin. 1991. Classified and representative method - a useful way to select variables in ecology. Ecol. Modelling 55: 123-131.

Manuscript received: October 1992.