View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Crossref

ENVIRONMETRICS
*Environmetrics*, **10**, 67–77 (1999)

# VARIABLE SELECTION IN LARGE ENVIRONMENTAL DATA SETS USING PRINCIPAL COMPONENTS ANALYSIS

JACQUELYNNE R. KING[1]* AND DONALD A. JACKSON[2]

[1]*Department of Fisheries and Oceans, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, British Columbia, Canada V9R 5K6*
[2]*Department of Zoology, University of Toronto, 25 Harbord Street, Toronto, Ontario, Canada M5S 1A1*

## SUMMARY

In many large environmental datasets redundant variables can be discarded without the loss of extra variation. Principal components analysis can be used to select those variables that contain the most information. Using an environmental dataset consisting of 36 meteorological variables spanning 37 years, four methods of variable selection are examined along with different criteria levels for deciding on the number of variables to retain. Procrustes analysis, a measure of similarity and bivariate plots are used to assess the success of the alternative variable selection methods and criteria levels in extracting representative variables. The Broken-stick model is a consistent approach to choosing significant principal components and is chosen here as the more suitable criterion in combination with a selection method that requires one principal component analysis and retains variables by starting with selection from the first component.
Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS variable selection; principal components analysis; Procrustes analysis; environmental data; limnology; discarding variables

## 1. INTRODUCTION

The use of multivariate statistics in ecology is increasing. Often in multivariate analyses of ecological data, particularly with environmental parameters, the number of variables available for measurement is large ($\geqslant 20$). For reasonable stability and reliability of multivariate analyses of ecological and biological data, a data matrix with a 3:1 ratio of observations to variables should be used (Gibson *et al*. 1984; Grossman *et al*. 1991; Williams and Titus 1988). Solutions based on proportionately more variables will be less stable and the resulting eigenvector coefficients will be less reliable. As a consequence, the interpretability of the analysis will be compromised. In multivariate analyses, having more than 10 variables means that choosing a subset of variables will not often change results substantially because discarded variables are often redundant (Jolliffe 1972). Often in very large datasets there are variables present that do provide additional information. For example, in principal components analysis, if two variables, $x_1$ and $x_2$, are correlated such that $x_1 = x_2 + \varepsilon$ (where $\varepsilon$ is a random disturbance), then either $x_1$ or $x_2$ can be selected with little information lost or with little change to the first few principal components (Jolliffe 1972). When variables begin to outnumber the observations, decisions

* Correspondence to: J. R. King, Department of Fisheries and Oceans, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, British Columbia, Canada V9R 5K6.

CCC 1180–4009/99/010067–11$17·50
Copyright © 1999 John Wiley & Sons, Ltd.
*Received 15 July 1997*
*Accepted 20 June 1998*

should be made about which variables to keep for analysis. Such decisions should be made routinely in ecology to strengthen the reliability of the results (i.e. to reduce error). To some extent, the identification of redundant variables can save time and money if fewer variables need to be measured.

As part of a limnological study on the relationship between long-term climate and lake thermal stratification (King *et al*. 1997), a large climate dataset was compiled for Manitoulin Island, Ontario (45°33′N; 82°00′W). The final analysis in that study used canonical correlation analysis to examine the overlapping variation in climate and stratification variables. The climate dataset offered a potential of 36 variables measured over 37 years, and for reliability in the final canonical correlation analysis, climate variables were discarded to adhere to the 3:1 (obs:var) goal identified above. Here, we examine the use of different selection methods and criteria levels for discarding variables in the climate dataset, with the intention of highlighting the use of variable selection techniques and providing recommendations on which techniques are most effective.

One method for choosing which variables to retain is to use a statistical tool which both identifies those variables that express a large amount of variation and identifies the redundant variables. Principal components analysis (PCA) summarizes the major variation or information that is contained in many dimensions into a reduced number of uncorrelated dimensions. PCA is an appropriate tool for variable selection and the use of PCA to discard redundant variables has been outlined in Jolliffe (1972), Krzanowski (1987) and McCabe (1984). Criteria for choosing $p$ principal components in order to obtain a reduced subset containing $p$ variables have been identified (Jolliffe 1972; Krzanowski 1987), but not directly compared to determine the most consistent and accurate criterion. This study examines several of these criteria for variable retention. In order to identify the most representative, and also the success of each selection method and criteria, Procrustes analysis (PA) and a measure of similarity ($Q$) are used as measures of fit between the reduced subsets and the original climate dataset.

## 2. METHODS

### 2.1. The environmental dataset

King *et al*. (1997) compiled a large climate data set to examine the relationship between climate variability and lake thermal stratification patterns for South Bay (Lake Huron), Manitoulin Island, Ontario. Monthly mean air temperature (°C), wind speed (km $h^{-1}$), monthly prevailing wind direction (eight compass point direction), and percent of total monthly hours that wind blew along the bay's fetch were obtained from Environment Canada for the South Baymouth meteorological station for 1955–92. Monthly mean incoming solar radiation (MJ $m^{-2}$) data for Manitoulin Island were obtained from the Industrial Climate Research Group, Environment Canada. Data from April to November were used because these months encompass the lake stratification season. Iceoff Julian dates were obtained from the Ontario Ministry of Natural Resources and were included in the climate dataset because iceoff dates act as a summary variable of the weather conditions just prior to iceoff (e.g. warm conditions are correlated with early iceouts). Iceoff date also signifies the date when the physical barrier of ice is removed and the water surface is exposed to meteorological conditions. In total, the climate dataset was comprised of 36 variables spanning 37 years. All variables were tested for normality and transformed as required.

Copyright © 1999 John Wiley & Sons, Ltd. 

## 2.2. The methods of variable selection

Principal components analyses of the correlation matrices were performed using the SAS PRINCOMP procedure (SAS 1989). Four methods of variable selection (B1Backward, B1Forward, B2 and B4) based on PCA were used and are extensively described in Jolliffe (1972). All four selection methods and the various criteria are summarized in Table I.

### *2.2.1. Methods B1Backward and B1Forward*

In method B1Backward, a PCA is performed on the original matrix of $K$ variables and $n$ observations. The eigenvalues ($\lambda$) are used to select the number of component axes to evaluate based on some criterion $\lambda_o$. If $p_1$ components have eigenvalues less than $\lambda_o$, then the eigenvector coefficients (i.e. loadings) on the remaining $K–p_1$ components are evaluated starting with the last component (i.e. the component with the lowest $\lambda$ value). The variable associated with the highest eigenvector coefficient (i.e. the highest loading) is then discarded from each of the $K–p_1$ components. Another PCA is performed on the remaining $K–p_1$ variables and the selection

Table I. Summary of selection methods and criteria for the number of variables to discard or to retain (Modified from Jolliffe 1972)

| Selection method | Method of selecting $p$ variables from $K$ original variables | Criteria for deciding on the value of $p$ |
|---|---|---|
| B1Backward | Similar to B2 but reject fewer variables initially, then do another PCA, reject a few more and repeat until no more variables are to be discarded | 1. The number of principal components with eigenvalues $\geq \lambda_o$ (here $\lambda_o = 0{\cdot}70$, $0{\cdot}65$, $0{\cdot}60$)<br>2. The number of principal components required to account for some proportion ($\alpha_o$) of the total variance (here $\alpha_o = 0{\cdot}90$, $0{\cdot}80$) |
| B1Forward | Similar to B4 but retained fewer variables initially, then do another PCA, retain a few more and repeat until no more variables are to be retained | Criteria 1 and 2 |
| B2 | Associate ($K–p$) variables with each of the last ($K–p$) components and discard these variables | Criteria 1 and 2<br>3. Arbitrarily select $p$ such that the ratio of number of observations to $p$ is 3:1 (here $p = 12$)<br>4. Use the broken-stick model of Frontier (1976) to assign $p$ as the number of principal components with eigenvalues exceeding the expected value generated by a random (Broken-stick) distribution. See Legendre and Legendre (1983) for a table of eigenvalues based on the Broken-stick distribution |
| B4 | Associate $p$ variables with each of the first $p$ components and retain these variables | Criteria 1–4 |

Copyright © 1999 John Wiley & Sons, Ltd. 

process is repeated with the same criteria, such that $K–p_1–p_2$ variables remain. Principal component analyses are repeated until all components have eigenvalues higher than $\lambda_o$.

Jolliffe (1972) did not evaluate method B1Backward, because he considered it too time consuming, but did suggest $\lambda_o = 0{\cdot}70$ as a suitable criteria and argued that a cut-off of $\lambda_o = 1{\cdot}0$ retains too few variables. We used cut-off values of $\lambda_o = 0{\cdot}70$, $0{\cdot}65$ and $0{\cdot}60$. As an alternative to the $\lambda_o$ criterion, the cumulative proportion of variance ($\alpha$) explained by $p$ principal components was also used to select the number of component axes for interpretation. Cut-offs of $\alpha_o = 0{\cdot}80$ and $0{\cdot}90$ (i.e. 80 and 90 per cent of total variance) were used to select the $p$ components. Because the last component's cumulative per cent variation will always be $1{\cdot}0$ (i.e. 100 per cent), $K–(p_1 + 1)$ components could be kept. Though not outlined in Jolliffe (1972), the B1Backward method was also performed 'Forward', so that instead of *discarding* variables starting with the final component, variables are *retained* starting with the first.

### 2.2.2. Method B2

Method B2 is the same as B1Backward but requires only one PCA to be performed on the original $K$ by $n$ matrix. Based on some cut-off criteria, if $p$ variables are to be retained then $K–p$ variables are rejected backwards from the last component. In addition to the eigenvalue ($\lambda_o$) and proportion of variance ($\alpha_o$) cut-off criteria outlined above, we also used the Broken-stick model (Frontier 1976). Jackson (1993) identified the Broken-stick model to be a consistent approach for determining a suitable number of components for interpretation. The Broken-stick model assumes that total variance is proportioned among the components and that the expected eigenvalue distribution should follow a Broken-stick distribution. Observed eigenvalues are considered interpretable when they exceed the expected eigenvalues generated by the Broken-stick model. For a table of eigenvalues generated by the Broken-stick model, consult Legendre and Legendre (1983) or calculate as:

$$b_k = \sum_{i=k}^{p} \frac{1}{i} \tag{1}$$

where $p$ is the number of variables and $b_k$ is the size of eigenvalue for the $k$th component under the Broken-stick model.

Krzanowski (1987) has suggested that it is acceptable to arbitrarily choose $p$ variables from $K$ variables, provided that the amount of variance retained by the resulting $p$ components is acceptable. A good rule of thumb for the ratio of observations to variable is 3 : 1 so here for 36 observations, 12 variables (36 : 12 = 3 : 1) were retained by rejecting $K$-12 variables.

### 2.2.3. Method B4

Method B4 retains variables by starting with the first component and keeping the variable with the highest loading. All $K–p$ remaining variables are rejected. The seven criteria levels used to select $p$ in methods B2 ($\lambda_o = 0{\cdot}70, 0{\cdot}65, 06{\cdot}0$; $\alpha_o = 0{\cdot}80, 0{\cdot}90$; Broken-stick model; 3 : 1 ratio) were also used for B4.

Copyright © 1999 John Wiley & Sons, Ltd. 

### 2.3. Evaluating the reduced variable subsets

#### *2.3.1. Procrustes analysis*

Procrustes analysis uses a rotational-fit algorithm that minimizes the sum-of-the-squared residuals ($m^2$ statistic) between two matrices and allows for the direct comparison between a pair of data matrices. Here we used the Ordinary Least-squares (i.e. rigid Procrustean) rotation in GRF-ND (Slice 1994), to compare the concordance between the original climate dataset and the resultant subsetted (reduced variable) matrices. In order to compare similar dimensions, the PCA scores from the first $p$ principal components of the original dataset were compared in turn to PCA scores from the 24 reduced-variable subsets, where $p$ denotes the number of retained variables. The resultant $m^2$ statistic ($0 < m^2 < 1{\cdot}0$) is a measure of goodness-of-fit of the two matrices where lower $m^2$ values indicate better fit.

#### *2.3.2. Weighted measure of similarity*

A correlation matrix for the first $p$ principal components of the original dataset and the total $p$ principal components for each of the 24 reduced variable subsets (resulting from the 24 selection analyses) was obtained through the CORR procedure in SAS (SAS 1989). A weighted measure of similarity was calculated as

$$Q = \frac{\sum_{i=1}^{p} v_i \cdot r_i}{\sum v_p} \tag{2}$$

where $v_i$ = proportion of total variance explained by the $i$th component, $r_i$ = the correlation coefficient between the $i$th components and $p$ = the number of variables retained.

In order to compare methods and criteria, we classified the $m^2$ statistic and $Q$ into poor, average and good-fit classes identifying the degree of concordance of the subset to the original data matrix ($m^2$ statistic: 0·0–0·2, good-fit, 0·2–0·8 average, 0·8–1·0 poor; $Q$: 0·0–0·2 poor-fit, 0·2–0·8 average, 0·8–1·0 good). Though this is admittedly subjective, we felt that a consistent criteria of evaluation facilitated comparisons.

#### *2.3.3. Bivariate plots*

As an final evaluation, we examined the bivariate plot of PCA scores along the first two principal components for the original 36 variable matrix and with the plot for a subset matrix evaluated by the $m^2$ statistic and the $Q$ value as 'good' and with the plot for a subset matrix evaluated as 'poor'. This provides a visual examination of the patterns of variance retained in the subsets compared to the original matrix's pattern of variance.

## 3. RESULTS

The results of applying the four selection methods and the various criteria levels are summarized in Table II, which gives the number of analyses required, the number of variables selected and the

Copyright © 1999 John Wiley & Sons, Ltd. 

Table II. The variables retained by each criteria level (i.e. a critical value for eigenvalues ($\lambda_o$) or proportion of variance ($\alpha_o$); the Broken-stick model (BS); selecting $n = 12$ variables for a 3:1 ratio of obs:var) for all four selection methods (B1Backward; B1Forward; B2; B4). The number of PCAs required to select the variables is also indicated; methods B2 and B4 required only one PCA

| | B1Backward | | | | | B1Forward | | | | | B2 | | | | | | | B4 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\lambda_o$ | | | $\alpha_o$ | | $\lambda_o$ | | | $\alpha_o$ | | $\lambda_o$ | | | $\alpha_o$ | | 12 | BS | $\lambda_o$ | | | $\alpha_o$ | | 12 | BS |
| | 0·70 | 0·65 | 0·60 | 0·90 | 0·80 | 0·70 | 0·65 | 0·60 | 0·90 | 0·80 | 0·70 | 0·65 | 0·60 | 0·90 | 0·80 | | | 0·70 | 0·65 | 0·60 | 0·90 | 0·80 | | |
| No. of PCAs performed | 5 | 5 | 4 | 7 | 5 | 6 | 5 | 5 | 5 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| No. of variables retained | 6 | 7 | 9 | 10 | 7 | 4 | 5 | 6 | 9 | 6 | 15 | 15 | 16 | 17 | 12 | 12 | 7 | 15 | 15 | 16 | 17 | 12 | 12 | 7 |
| Variables retained | | | | | | | | | | | | | | | | | | | | | | | | |
| Iceoff date | | | | | | | | • | • | | | | | | | | | • | • | • | • | • | • | • |
| April air temperature | • | • | • | • | • | | | | | | • | • | • | • | • | • | • | | | | | | | |
| May air temperature | | | | | | • | • | • | • | | | | | | | | | • | • | • | • | • | • | • |
| June air temperature | | | | | | | | | | | | | | | | | | | | | | | | |
| July air temperature | | | | | | | | | | • | | | | | | | | • | • | • | • | • | • | • |
| Aug. air temperature | | | | | | | | | | | | | • | • | | | | | | • | • | | | |
| Sept. air temperature | | | | | | | | | | | | | | | | | | | | | | | | |
| Oct. air temperature | | | | | | | | • | | | | | | | | | | • | • | • | • | | | |
| April solar radiation | • | • | • | • | • | | | | • | | • | • | • | • | • | • | • | | | | | | | |
| May solar radiation | • | • | • | • | • | | | | • | • | • | • | • | • | • | • | | • | • | • | • | | | |
| June solar radiation | | | | | | | | | | | | | | | | | | | | | | | | |
| July solar radiation | | | | | | | | | | | | | | | | | | | | | | | | |
| Aug. solar radiation | | | • | • | | | | | | | • | • | • | • | | | | | | | | | | |
| Sept. solar radiation | | | | | | | | | | | • | • | • | • | | | | | | | | | | |
| Oct. solar radiation | | | | | | | • | • | • | • | | | | | | | | • | • | • | • | • | • | • |
| April wind speed | | | | | | | | | | | • | • | • | • | • | • | • | | | | | | | |
| May wind speed | | | | | | | | | | | • | • | • | • | • | • | • | • | • | • | • | • | • | • |
| June wind speed | | | | | | | | | | | | | | | | | | | | | | | | |
| July wind speed | | | | | | | | | | | | | | | | | | | | | | | | |
| Aug. wind speed | | | | | | | | | | | | | | • | | | | | | | • | | | |
| Sept. wind speed | | | | | | | | | | • | | | | | | | | • | • | • | • | • | • | |
| Oct. wind speed | | | | | | | | | | | • | • | • | • | • | • | • | | | | | | | |
| April wind direction | | | | | | | | | | | | | | | | | | | | | | | | |
| May wind direction | • | • | • | • | • | | | | • | | • | • | • | • | • | • | | • | • | • | • | • | • | • |
| June wind direction | | | | | | | | | | | | | | | | | | | | | | | | |
| July wind direction | | | | | | | | • | • | | | | | | | | | • | • | • | • | • | • | |
| Aug. wind direction | | | | • | | • | • | | | | • | • | • | • | • | • | | • | • | • | • | • | • | |
| Sept. wind direction | • | • | | • | • | | | | | | • | • | • | • | • | • | • | | | | | | | |
| Oct. wind direction | • | • | • | • | • | | | | | • | • | • | • | • | • | • | | • | • | • | • | • | • | |
| April wind along fetch | | | • | | | | | | | | | | | | | | | | | | | | | |
| May wind along fetch | | | | | | | | | • | | • | • | • | | | | | • | • | • | • | | | |
| June wind along fetch | • | • | • | • | • | | | | | | • | • | • | • | • | • | • | | | | | | | |
| July wind along fetch | | | | | | • | • | | | | | | | | | | | • | • | • | • | • | • | • |
| Aug. wind along fetch | | | | | | | | | | | | | | | | | | | | | | | | |
| Sept. wind along fetch | | | • | • | | | | • | • | • | • | • | • | • | • | • | | • | • | • | • | • | • | |
| Oct. wind along fetch | | | | | | | | | | | | | | | | | | | | | | | | |

Copyright © 1999 John Wiley & Sons, Ltd.

actual variables selected. A total of 24 different combinations of the four selection methods and cut-off criteria were analyzed.

There is considerable difference in the size of the reduced subsets chosen by the various selection methods and the various levels of criteria. The selection methods B1Backward and B1Forward retained fewer variables than the B2 and B4 selection methods (Table II). Both the B1Backward and the B1Forward methods tend to select too few variables such that the per cent of variation retained is fairly low (Table II). Overall, the selection methods (with the exception of B1Forward) were consistent in retaining particular variables across cut-off criteria (Table II). The $\lambda_o$ and $\alpha_o$ criteria retained a higher number of variables than the Broken-stick model.

Overall, the B4 selection method offered subsets with the both the high measures of fit (i.e. low $m^2$ statistic) and measures of similarity (i.e. high $Q$ value) (Table III). The B4 selection method with the Broken-stick criterion yielded the best results. This selection method had a relatively low $m^2$ statistic and accounted for at least 60 per cent of the variance (Table III). The two lowest $m^2$

Table III. Percent of variation explained by $p$ principal components (%),weighted total squared distance ($m^2$) and measure of similarity ($Q$) for each reduced subset with $p$ selected variables. Also included is a categorization of fit based on the $m^2$ statistic (poor: 1·0–0·8; average: 0·8–0·2; good: 0·2–0) and of similarity based on the $Q$ value (poor: 0–0·2; average: 0·2–0·8; good: 0·8–1·0)

| Method | Criteria | $p$ | % | $m^2$ | $Q$ |
|---|---|---|---|---|---|
| B1Backward | $\lambda_o = 0{\cdot}70$ | 6 | 58·53 | 0·853381 | 0·239 |
| | $\lambda_o = 0{\cdot}65$ | 7 | 64·44 | 0·845931 | 0·255 |
| | $\lambda_o = 0{\cdot}60$ | 9 | 70·94 | 0·849199 | 0·302 |
| | $\alpha_o = 0{\cdot}90$ | 10 | 74·42 | 0·853646 | 0·213 |
| | $\alpha_o = 0{\cdot}80$ | 7 | 63·44 | 0·845931 | 0·215 |
| | | | | Poor | Poor |
| B1Forward | $\lambda_o = 0{\cdot}70$ | 4 | 47·04 | 0·729253 | 0·131 |
| | $\lambda_o = 0{\cdot}65$ | 5 | 52·89 | 0·748841 | 0·103 |
| | $\lambda_o = 0{\cdot}60$ | 6 | 58·53 | 0·842396 | 0·121 |
| | $\alpha_o = 0{\cdot}90$ | 9 | 70·94 | 0·842647 | 0·206 |
| | $\alpha_o = 0{\cdot}80$ | 6 | 58·56 | 0·842647 | 0·113 |
| | | | | Poor–average | Poor |
| B2 | $\lambda_o = 0{\cdot}70$ | 15 | 87·93 | 0·837267 | 0·427 |
| | $\lambda_o = 0{\cdot}65$ | 15 | 87·93 | 0·837267 | 0·427 |
| | $\lambda_o = 0{\cdot}60$ | 16 | 89·61 | 0·839893 | 0·355 |
| | $\alpha_o = 0{\cdot}90$ | 17 | 91·09 | 0·840425 | 0·673 |
| | $\alpha_o = 0{\cdot}80$ | 12 | 80·82 | 0·910071 | 0·497 |
| | Choose 12 | 12 | 80·82 | 0·910071 | 0·497 |
| | Broken-stick | 7 | 63·44 | 0·903964 | 0·425 |
| | | | | Poor | Average |
| B4 | $\lambda_o = 0{\cdot}70$ | 15 | 87·93 | 0·886095 | 0·537 |
| | $\lambda_o = 0{\cdot}65$ | 15 | 97·93 | 0·886095 | 0·537 |
| | $\lambda_o = 0{\cdot}60$ | 16 | 89·61 | 0·887852 | 0·459 |
| | $\alpha_o = 0{\cdot}90$ | 17 | 91·09 | 0·888010 | 0·315 |
| | $\alpha_o = 0{\cdot}80$ | 12 | 80·82 | 0·762973 | 0·456 |
| | Choose 12 | 12 | 80·82 | 0·762973 | 0·456 |
| | Broken-stick | 7 | 63·44 | 0·750874 | 0·430 |
| | | | | Poor–average | Average |

Copyright © 1999 John Wiley & Sons, Ltd.

values (i.e. evaluated as having the best fit) correspond to the B1 Forward method with the $\lambda_o = 0{\cdot}70$ ($m^2 = 0{\cdot}72953$) and $\lambda_o = 0{\cdot}65$ ($m^2 = 0{\cdot}748841$) criteria. However, these subsets account for only 47·04 per cent and 52·89 per cent of the total variation in the original dataset which is fairly low. The B1Backward method tends to produce reduced subsets with relative poor fit (i.e. $m^2$ statistic $> 0{\cdot}8$) and poor similarity (i.e. $Q < 0{\cdot}20$) to the original dataset irrespective of criteria and the B1Forward selection method does not provide improved results (Table III). All criteria levels in method B2 produced subsets with average similarity measures, but with poor fits assessed by the $m^2$ statistic.

Bivariate plots of PCA scores along the first two principal components for the original 36 variable matrix (Figure 1), the seven variable 'B4-Broken-stick' selected matrix (Figure 2) and the six variable 'B1Forward-80% variance' selected matrix (Figure 3) illustrate the patterns of variance retained in a 'good' and a 'poor' subset compared to the original matrix's pattern of variance. In Figure 1, two groups are identified with six solid circles for 6 years that lay at the extreme left and six squares for 6 years that lay at the extreme right of the bivariate plot from the original variable matrix. The 'B4-Broken-stick' selected subset was identified by Procrustes analysis and the measure of similarity as a better representative subset than the 'B1Forward-80 per cent variance' subset. The 'B4-Broken-stick' bivariate plot reasonably retains the position of the extreme groups with the solid circle years towards the extreme left and the square years (with the exception of 1965 and 1972) at the extreme right and the two groups are still reasonably separated. In contrast, the 'B1Forward-80 per cent variance' bivariate plot (Figure 3, note scores along the Principal component 1 axis reversed for comparison purposes) illustrates the loss of the original pattern with less of a separation between solid circle years and square years. Though the bivariate plots represent only two dimensions of multidimensional matrices, the comparison of subset plots illustrate the relative success in retaining the original pattern of variance.

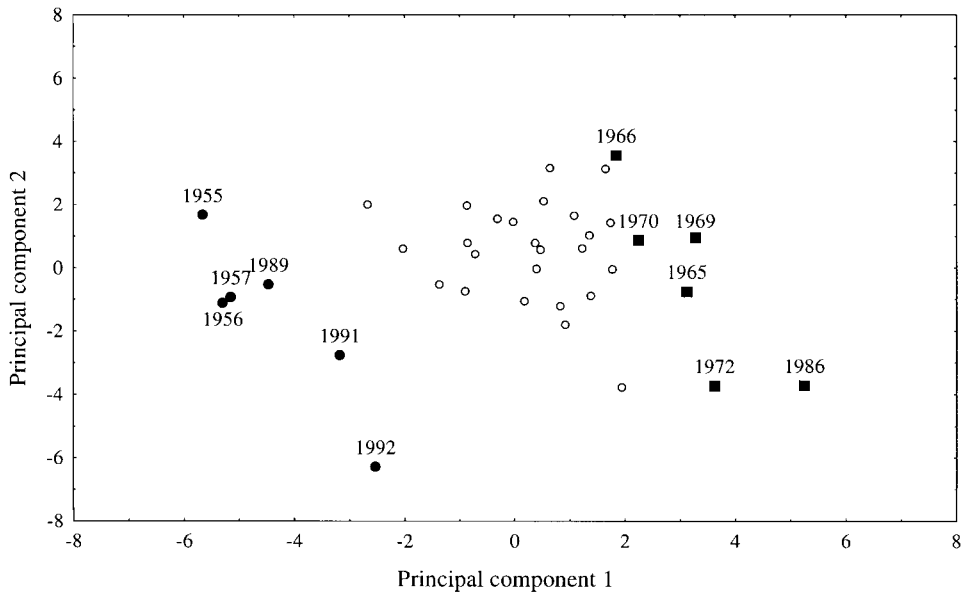


Figure 1. Bivariate plot of scores for the first two principal components from a PCA on the original 36 variable matrix. Selected years at the two extremes of the plot are denoted with circles and squares

Copyright © 1999 John Wiley & Sons, Ltd. 

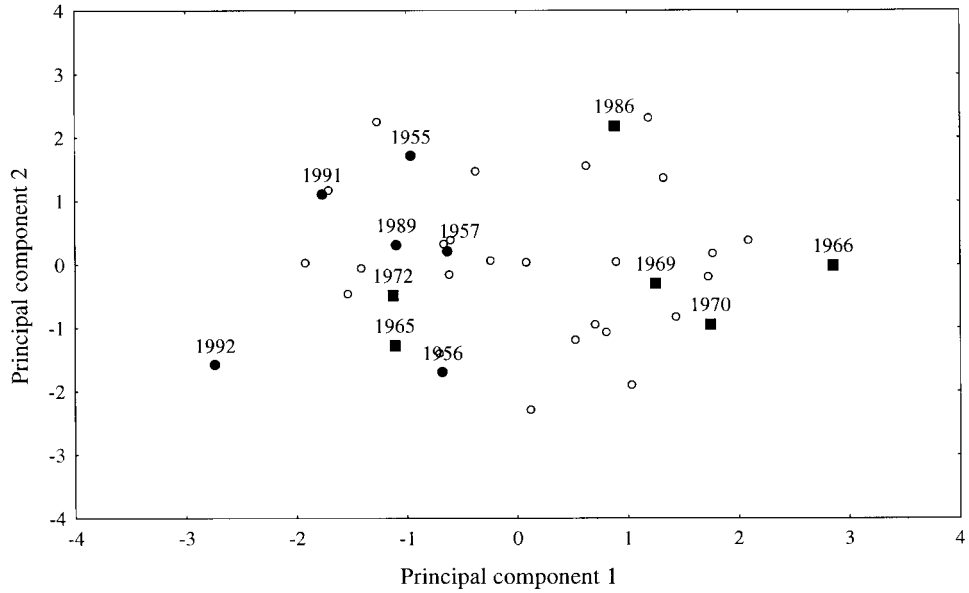


Figure 2. Bivariate plot of scores for the first two principal components from a PCA on the seven variable subset selected by the 'B4-Broken-stick' method. The circled years that were at the extreme left on the original matrix plot of PCA scores are still towards the right extreme. Years with squares (with the exception of 1965) that were at the extreme right on the original matrix plot of scores are retained to the right in this subset matrix plot

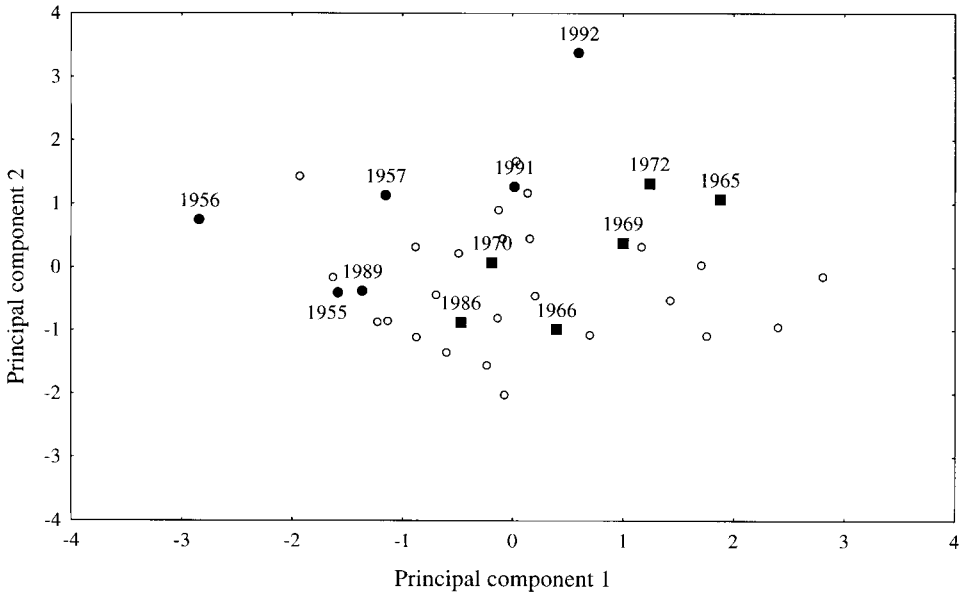


Figure 3. Bivariate plot of scores for the first two principal components from a PCA on the six variable subset selected by the 'B1Forward-80 per cent variance' method. For convenience of comparison with the other two bivariate plots, the scores along Principal component 1 axis have been reversed (i.e. multiplied by −1). The pattern of the circled and squared years in the original matrix plot of scores is lost in this subset matrix plot of scores

Copyright © 1999 John Wiley & Sons, Ltd. 

## 4. DISCUSSION

Though the speed of personal computers makes the time to both B1Backward and B1Forward selection methods manageable, performing repetitive principal component analyses is tedious. Since the resulting reduced subsets generally had poor fits and similarity to the original dataset, B1Backward and B1Forward is not a recommended method for variable selection. Jolliffe (1972) identified B2 and B4 as consistently good methods of variable selection and the results from our analyses support Jolliffe's findings. B4 tended to produce reduced subsets with better fit to the original dataset than B2, while B2 produced subsets with slightly better measures of similarity.

Given that using a cut-off criteria of $\lambda_o \leqslant 0{\cdot}70$ retains too many variables for our goal of a 3 : 1 observations to variables ratio and that a cut-off of $\lambda_o > 0{\cdot}70$ leads to the retention of too many components (Jackson 1993) we do not suggest using the eigenvalue ($\lambda_o$) criteria. Our results here also suggest that cumulative proportion of variance ($\alpha_o$) criteria is also inappropriate. In assessing stopping rules for significant principal components, Jackson (1993) warns that regardless of which cumulative proportion of variance level is used, it does not appear to be a promising approach, since many of the components that are retained will summarize noise. The Broken-stick model has already been identified as a consistent approach for selecting significant components (Jackson 1993) and if extended to variable selection, this criterion worked well with the B4 method. The 'B4-Broken-stick' subset contained few variables, but a suitable amount of per cent of variance, and had good measures of fit and reasonable similarity.

From an ecologist's view, variable selection is useful for reducing the number of variables required for statistical analyses since it can improve the reliability and stability of final results (Gibson *et al*. 1984; Grossman *et al*. 1991; Williams and Titus 1988). PCA can be used to identify those variables that contain the most information. If resources become limited, the selected variables may provide a suggestion for future data collection in ecological studies. We suggest that ecologists use the B4 selection method and Broken-stick cut-off criteria to select subsets of variables from large environmental datasets to use in subsequent statistical analyses.

## REFERENCES

Frontier, S. (1976). Étude de la decroissance des valeurs propres dans une analyze en composantes principales: comparison avec le modèle de baton brisé. *J. Exp. Mar. Bio. Ecol*. **25**, 67–75.

Gibson, A. R., Baker, A. J. and Moeed, A. (1984). Morphometric variation in introduced populations of the common myna (*Acridotheres tristis*): an application of the jackknife to principal component analysis. *Syst. Zool*. **33**, 408–421.

Grossman, G. D., Nickerson, D. M. and Freeman, M. C. (1991). Principal component analyses of assemblage structure data: utility of tests based on eigenvalues. *Ecology* **72**(1), 341–347.

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**, 2201–2214.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis I: artificial data. *Appl. Statist*. **21**, 160–173.

King, J. R., Shuter, B. J. and Zimmerman, A. P. (1997). The response of the thermal stratification of South Bay (Lake Huron) to climatic variability. *Can. J. Fish. Aquat. Sci*. **54**, 1873–1882.

Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist*. **36**, 22–33.

Legendre, L. and Legendre, P. (1983). *Numerical Ecology*. Elsevier: Amsterdam.

McCabe, G. P. (1984). Principal variables. *Technometrics* **26**, 137–144.

Copyright © 1999 John Wiley & Sons, Ltd. 

SAS (1989). *SAS/STAT Guide*, version 6·08. SAS Institute, Cary, NC.
Slice, D. E. (1994). *Generalized Rotational Fitting of n-Dimensional Landmark Data*. Revision 11-01-94. State University of New York at Stony-Brook, Stony Brook, NY.
Williams, B. K. and Titus, K. (1988). Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology* **69**(4), 1275–1285.

Copyright © 1999 John Wiley & Sons, Ltd. 