Multivariate Environmental Statistics
G.P. Patil and C.R. Rao, editors

Chapter 17

CORRESPONDENCE ANALYSIS APPLIED TO ENVIRONMENTAL DATA SETS

DONALD E. MYERS*

University of Arizona, Tucson, AZ

Environmental data sets are nearly always multivariate and distributional information is frequently lacking. There are at least two objectives in analyzing such a data set; one, to reduce the number of variates by eliminating the redundancies in the data set and two, identify and characterize spatial clustering or patterns. The analysis is frequently complicated by missing values, censored values or by truncation. Correspondence Analysis is a multivariate technique that is frequently useful for both of the above objectives. In its original form it was limited to categorical data but an alternative derivation allows the use of continuous data. Although similar to Principal Components Analysis it has some advantages for environmental data. The algebraic formulation is shown, diagnostics described and applications to environmental data sets reviewed. Practical aspects such as the availability of software are discussed.

Key Words: Multivariate analysis, principal factor, variation, error profile, singular value decomposition

1. Introduction

The objectives in analyzing environmental data sets are many and varied as are the methods and techniques that are used. There are features of such data sets that are distinctive. They are nearly always multivariate and while nearly all variates or analytes are of interest there are often reasons for searching for a smaller number of variates that will still adequately represent the information in the data set. Correspondence Analysis is a particular form of multivariate analysis and is particularly useful for such exploratory analysis. The most common developments of Correspondence

*Donald E. Myers is Professor of Mathematics at the University of Arizona, he is also a member of the Applied Mathematics Program. His principal research area is multivariate geostatistics.

Analysis assume that the data is categorical whereas most environmental data sets represent continuous valued variates. A purely algebraic derivation is reviewed and examples are given of the applicability of Correspondence Analysis to multivariate environmental data sets.

2. Environmental data sets

In order to describe both the objectives and the problems pertaining to the analysis of multivariate environmental data sets certain notation will be introduced. In general such data sets are presented as an array as follows

$$\begin{matrix} x_{11}, & \cdots & , x_{1p} \\ \vdots & & \vdots \\ x_{n1}, & \cdots & , x_{np} \end{matrix}$$

Ordinarily the columns correspond to analytes or variates, and the rows (usually) correspond to sample locations or time points. The x_{ij}'s then are assumed to be real-valued. Because many organic as well as inorganic compounds are toxic at very low concentrations and because the analytical techniques can not ascertain such low concentrations, some of the x_{ij}'s may be what are called non-detects or non-quantified. That is, there is a concentration level such that the analytical technique is not adequate to determine whether the analyte is present or absent but only that the concentration is below the detection limit. These are sometimes replaced with a value equal to the detection limit and sometimes are coded as a non-detect. Because most multivariate methods and Correspondence Analysis in particular requires that each x_{ij} have a value, different schemes are utilized to cope with the non-detect values and one must be concerned with the robustness of the analysis with respect to these schemes. In some instances it may be adequate to replace all non-detects with zeros, in others to use half the non-detect value. If a risk analysis is the objective then it may be preferable to determine the most conservative replacement. These consequences are illustrated empirically in the application of Correspondence Analysis to the Lake Chautauqua data, Avila and Myers (1991). Some analytical techniques not only have a detection limit but also a quantification limit, that is, below this limit the technique will not reliably determine the concentration. Reported concentrations below the quantification limit but above the detection limit are thus not reliable but may significantly affect the data analysis. If the probability distribution for a given analyte were known or assumed then intermediate values could be simulated

Because many of the analytical techniques used for environmental samples are quite expensive, a particular sample may not be completely analyzed. That is, some of the x_{ij}'s may be missing. Many multivariate

techniques can not incorporate missing values and it may be necessary to delete one or more entire rows from the data array. While concentrations are necessarily non-negative valued if a log transformation is utilized the new data array may contain negative values. The initial normalization step in Correspondence Analysis precludes the use of negative values. One way of coping with negative values for a particular variate is to add a constant to all values of that variate. This technique is illustrated in the application of Correspondence Analysis to the Eastern Lake Survey data, Rhodes and Myers (1991).

Analytical results are nearly always reported as averages, that is, a concentration per unit volume or area. The reported concentration averages may or may not be on the same scale as the physical sample since it is common practice in the laboratory to either use aliquots or to be composited. Unless concentrations in situ are assumed to be constant over the area or volume represented by the sample, the use of composites or subsamples for analysis can change the variability. Another way of interpreting this point is that it corresponds to a change of units. Most multivariate techniques do not incorporate this change in the sample support or a change of units, empirical results are given for Correspondence Analysis in Rhodes and Myers (1991) as well as consideration of whether there may "natural" units for the variates in a particular application.

If the correlation matrix is used in Principal Components Analysis then the normalization step will have converted the data to dimensional free values although the normalization process is associated with the R mode, i.e., analysis of the columns. Correspondence Analysis incorporates a different form of normalization but does not distinguish between the analysis of the columns and the rows (variates and samples). Environmental data sets often include a number of different types of variates, for example chemical and physical. Correspondence Analysis provides a natural way to treat some of the variates as supplementary and hence allows separating the variates into at least two classes. This is illustrated in the application to the Lake Chautauqua data.

3. Objectives

In some instances particularly in the case of environmental data, data is collected or observations are made on multiple variates to compensate for the lack of other information or knowledge. For example in many environmental applications there are no applicable state equations. This may be because of insufficient information with respect to the environment into which the pollutants have been dispersed or concerning the mode or process by which the pollutants have been dispersed. Multivariate data sets will in

general be more complex and one objective may simply be data reduction. This may be coupled with the objective of identifying non-observeable variates. The data reduction or the non-observeable variates may aid in the identification or characterization of spatial patterns. Multivariate methods may be used in some instances to identify sources or to apportion sources.

Multivariate methods such as Principal Components Analysis (PCA), Factor Analysis (FA) and Correspondence Analysis (CA) implicitly incorporate or quantify some form of row column association. This association can often be used to identify outliers or anomalous values, more generally they can be used to detect or characterize anomalous regions. Often multivariate data sets will include data on different kinds of variables, for example physical vs chemical. The multivariate analysis may be used to characterize relationships between active and supplemental variables. Finally some forms of multivariate analysis may be more efficient in obtaining particular results than another. The results may be produced in simpler or more useful forms.

Most of these objectives will be illustrated in the application of Correspondence Analysis to three environmental data sets.

4. The method

Consider an $n \times p$ array, X, of nonnegative data values such that each row and column has at least one non-zero entry. In Principal Components Analysis the data set is visualized as n points in R^p or p points in R^n. The usual standardization is easily interpreted and the eigenvectors have a geometric interpretation. The distance between points is given by the Euclidean norm. While the geometric interpretation for rows and columns is still valid for Correspondence Analysis a different metric is used. Here X is replaced by F where

$$x_{ij} \to f_{ij} = x_{ij}/L$$

$$L = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}.$$

The analysis is principally concerned with the row and column profiles which are given by

$$f_{i1}/f_{i+}, \ldots, f_{ip}/f_{i+}$$

and

$$f_{1j}/f_{+j}, \ldots, f_{nj}/f_{+j}$$

where

$$f_{i+} = \sum_{j=1}^{p} f_{ij}; \quad j = 1, \ldots, p \tag{5}$$

and

$$f_{+j} = \sum_{i=1}^{n} f_{ij}; \quad i = 1, \ldots, n. \tag{6}$$

In the case of frequency data the profiles have a natural interpretation in terms of probabilities and there is a connection between the metric and the chi-square test statistic for independence on a contingency table. For environmental data an purely algebraic formulation is preferable. The objective is to find a form of a singular value decomposition of the normalized data matrix $F = [f_{ij}]$ of the following form:

$$f_{ij} = f_{i+} f_{+j} (1 + \sum_{k=1}^{p-1} (\lambda_k)^{.5} \phi_k \psi_k). \tag{7}$$

When the summation is from $k = 1$ to K, $K \leq p - 1$, then an approximation of F is obtained, F_K, in a certain least squares sense. Such a decomposition occurs in other contexts, in particular consider the case of a function defined on R^2, Lancaster (1958) has shown that such functions are representable in terms of the eigenfunctions of certain operators. The uniqueness of the eigenvalues and the eigenvectors is assured by a theorem in F. Avila and D. Myers (1991) which in turn is based on the Eckart-Young decomposition theorem. In comparison with the representation given by Principal Components Analysis, this representation is symmetric with respect to rows and columns. Moreover both the R and Q modes are obtained at the same time. In this algebraic form it is not necessary to limit the analysis to categorical data, however in that case there are close connections with the chi-square test statistics used in testing for independence.

THEOREM 1. *Let F be an $n \times p$ (assume $n \geq p$) matrix with non-negative entries f_{ij} such that $\sum_{i=1}^{n} \sum_{j=1}^{p} f_{ij} = 1$. Let D_p and D_n be diagonal matrices with diagonal entries f_{+j} and f_{i+} respectively. Let 1_n be a vector in R^n with all the coordinates equal to one. Then, there exist $(p-1)$ triplets $(\lambda_1, \phi_1, \psi_1), \ldots, (\lambda_{p-1}, \phi_{p-1}, \psi_{p-1})$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p-1} \geq 0$, $\phi_1, \ldots, \phi_{p-1}$ are vectors in R_n and $\psi_1, \ldots, \psi_{p-1}$ are vectors in R_p, such that:*

(i) *For every $k, l = 1, \ldots, p-1$*

$$\psi_k^{\mathrm{T}} D_n \psi_l = \delta_{kl} \tag{8}$$

and

$$\phi_k^T D_p \phi_l = \delta_{kl}. \tag{5}$$

(ii) *For every* $k = 1, \ldots, p-1$

$$F D_p^{-1} F^T \psi_k = \lambda_k D_n \psi_k \tag{6}$$

and

$$F^T D_n^{-1} F \phi_k = \lambda_k D_p \phi_k \tag{7}$$

$$\mathrm{Tr}[(F - F_K) D_p^{-1} (F - F_K)^T D_n^{-1}] = \sum_{k=K+1}^{p-1} \lambda_k. \tag{8}$$

The ϕ's and the ψ's are called factors and sometimes standardized scores multiplying the factors by the square roots of the $\lambda_k's$ we get the coordinates which can be plotted in the usual cartesian system.

The factors are unit vectors (in the norms induced by the matrices $\ldots$ and D_p and can be obtained from an eigenvalue-eigenvector problem. It is seen that only one set of factors need be obtained, since the other can be computed from transition formulas. There are only $p-1$ nontrivial factors since one factor, corresponding to an eigenvalue equal to zero or one (depending on the matrix used to extract them), is discarded because it represents the induced "correlation" due to the closure of the data. The optimality of the nontrivial factors is expressed in iv) of the theorem. The case $K = p-1$ gives the reconstruction formula (1). When $K \leq p-1$ factors are kept, we can estimate the error of the approximation when the model F_K is used, by looking at the matrix norm of the difference $F - F_K$

$$\mathrm{Tr}[(F - F_K) D_p^{-1} (F - F_K)^T D_n^{-1}] = \sum_{k=K+1}^{p-1} \lambda_k. \tag{9}$$

5. Diagnostics

Having generated the eigenvalues and factors an approximation is obtained for each choice of K. While the matrix norm leads to the "percent variation" explained in terms of the eigenvalues of the factors retained, it is useful to have additional diagnostics to aid in choosing K and to aid in interpreting the factors and the representation. The percent variation explained is a global measure but it is also useful to quantify how well the variables and samples are reconstructed by a given number of factors and to be able to easily identify the principal components of a factor in terms of the variables and in terms of the samples.

i) The cumulative percentage of variation is given by

$$(\sum_{k=1}^{K} \lambda_k)/(\sum_{k=1}^{p-1} \lambda_k) \tag{1}$$

which is a global measure of fit when K factors are retained; each giving the contribution of a particular factor. This is related to the Frobenius norm of $F - F_K$.

ii) For every $k = 1, \ldots, p-1$

$$RC^k(j) = (\lambda_k \phi_{jk}^2)/\left(\sum_{l=1}^{p-1} \lambda_l \phi_{jl}^2\right); \quad j = 1, \ldots, p \tag{2}$$

and

$$RC^k(i) = (\lambda_k \psi_{ik}^2)/(\sum_{l=1}^{p-1} \lambda_l \psi_{il}^2); \quad j = 1, \ldots, p. \tag{3}$$

These are called the relative contributions, or squared correlations, of factor k with column j or row i. They provide a measure of the row or column variation explained by a particular factor. They also characterize the contribution of a factor to the representation of a variable or factor.

iii) For every $k = 1, \ldots, p-1$

$$AC^k(j) = f_{+j} \phi_{jk}^2; \quad j = 1, \ldots, p \tag{4}$$

and

$$AC^k(i) = f_{i+} \psi_{ik}^2; \quad i = 1, \ldots, n. \tag{5}$$

These are called the absolute contributions of column j or row i to factor k. They help in understanding the composition of a particular factor, and are quoted as percentages.

6. Supplementary variates

These can be either rows or columns. A given supplementary row $(f_{s1}, f_{s2}, \ldots, f_{sp})$ can be projected onto the kth principal axis, with its projection (coordinate) being equal to

$$\psi_{sk}^* = \sum_{j=1}^{p} (f_{sj} \phi_{kj})/(f_{s+}). \tag{1}$$

Analogously, for a supplementary column $(f_{1s}, f_{2s}, \ldots, f_{ns})^{\mathrm{T}}$ its projection onto the kth principal axis is

$$\phi^*_{sk} = \sum_{j=1}^{p} (f_{is}\psi_{ik})/(f_{+s}).$$

7. Example 1

Chautauqua is a narrow 24 kilometer long lake in northwestern New York. It was sprayed with sodium arsenite as a herbicide from 1955 to 1960. Ninety-eight sediment grab samples collected in 1972 along transits made at half-mile intervals. These were analyzed for europium, sodium, manganese, potassium, bromine, arsenic, gallium, lanthanum, hafnium, cesium, terbium, scandium, iron, tantalum and antimony. In addition percentages of sand, silt ,clay and organic matter, depth of water were recorded for each sample. This data set was initially analyzed and reported by Hopke et al. (1976), the primary method used was Principal Components Analysis and Factor Analysis followed by Varimax rotation. As reported in Avila and Myers (1991) these results are obtained somewhat more easily by the use of Correspondence Analysis, the chemical variables and the physical variables can be separated by making the latter supplementary. In particular no rotation is necessary.

Five sample locations were identified as anomalous. One sample was high in sodium but with nominal values for all other variables. Three samples had very high manganese values, two of these came from the deepest part of the lake where there are iron-manganese nodules. Two of the anomalous samples came from the sandy northern part of the lake. A new data set was generated by making these five samples supplementary. In the original data the second factor was primarily manganese whereas in the reduced data set manganese is identified with the third factor. In both cases these factors explain more than 99% of the variation. This data set provides a particularly good example of the use of supplemental variables. When the physical variables were made supplementary they were still well represented by the factors generated by the chemical variables and vice versa.

8. Example 2

The Eastern Lake Survey-Phase I was conducted in order to evaluate possible effects on surface waters in the Eastern United States due to acid deposition. After the exclusion of lakes with a surface area of less than 4 ha and lakes that were close to urban areas or other possible close sources of air pollution, a probability sample of lakes was selected and sampled

taken from each of these lakes. The acidity of the lakes and the consequent impact on marine life was of principal interest. There are two versions of the data set, one of which has fewer variables and does not include a number of peripheral variables. This smaller version of the data set was analyzed by Rhodes and Myers (1991). Nineteen of 26 variables were used in the analysis. The data set for the Northeastern region is further partitioned into five subregions. Correspondence Analysis was used for three objectives. The primary objective was to reduce the number of analytes in order to make subsequent analyses easier and the results easier to interpret. A second objective was to identify outlying or atypical samples (lakes) and analytes. Finally Correspondence Analysis was used to assist in portraying the data geometrically. Seven factors were required to explain at least 96% of the variation.

Ordinarily acidity is characterized by the pH but in lakewater chemistry a second variable ANC (Acid Neutralizing Capacity) is often more useful. The Correspondence Analysis clearly brought out the value of ANC vs pH. ANC is a computed value and does depend on pH but in general pH was found not to be useful. Outliers or atypical samples and analytes were found by selectively choosing subsets of samples or analytes to be treated as supplementary variables. Because some variables had been log transformed negative values occurred in the data set. The robustness of the technique with respect to the addition of a positive constant to all values of a particular analyte is demonstrated. The changes resulting from this shift in the values of the one variable were largely predictable by considering the relative change in the coefficient of variation. In general the larger the coefficient of variation the greater the contribution of a variable (or sample) to the factors which explain the largest part of the variation.

The importance of the coefficient of variation was seen in another way. In general the units used in reporting the various concentrations are different and hence it was important to ask whether the choice of the units produced a representation that was an artifact of the units. While the units used are those that are considered to be chemically appropriate the change in units results in multiplying the concentrations by a factor. When the factor is larger than one the coefficient of variation will increase and hence the variable is more important in determining the principal factors. Conversely if the factor is less than one the coefficient of variation will decrease. Although some changes in the compositions of the factors occurred the groupings of the variables and samples were essentially unchanged.

Two additional characteristics of the data set were analyzed. A variable or sample was considered "superfluous" if when deleted from the data set the composition of the factors remained essentially the same. A variable or sample was considered unresolved, i.e., not well represented by a small

number of factors if the sum of the relative contributions remained less than 50% for that number of factors retained. In all three alkalinity classes (of lakes) three variables were superfluous; pH, DIC and DOC (Dissolved Inorganic and Organic Carbon). One variable was unresolved in all three classes, namely NH_4. This is a particularly useful property of the relative contribution statistic, the identification of variables or samples that are strongly identified only with those factors that correspond to small eigenvalues.

9. Example 3

In atmospheric chemistry one of the major problems is to differentiate between sources and source regions for trace substances. Whereas the R-mode results from Correspondence Analysis provide information about the analytes, the Q-mode results provide information about the samples, e.g., the sample locations in terms of their chemical signature. Dutot et al. (1988) considered air pollution data collected in the Lorraine region of France near the Moselle river. The analyses included concentrations for fifteen analytes and also information on size fractions. Correspondence Analysis was used to determine the composition of the emission sources and the mean chemical profile of each source.

10. Practical aspects

As indicated above CA is either similar to or coincident with a number of multivariate methods, these include Dual Scaling, Reciprocal Averaging, Homogenity Analysis, Canonical Scoring, Canonical Correlation Analysis, Log-Linear Modeling and of course, Principal Components and Factor Analysis. Most of these techniques emphasize exploratory analyses as opposed to model development. While not particularly computer intensive in the computations, these methods are most useful if combined with powerful graphics packages and provide for interactive application.

CA has attracted a lot more interest in the last ten years because of the appearance of two important books; LeBart et al. (1984) and Greenacre (1984) both of which reflect the work of Benzecri. The first of these two is an English version of one which appeared in French nearly ten years earlier.

As is true of nearly all statistical software, increases in computing power has made algorithms such as CA more accessible. David et al. (1977) published a program for CA but it was clearly intended for a mainframe or at least a minicomputer. In 1988 BMDP added CA as an option in the mainframe version and SAS followed with such an option as well. More recently Hoffman (1991) provided a review of four CA programs designed

for use on a PC with DOS, these programs carry a price of $120 to $495. With the exception of David et al. (1979) all these programs emphasize the analysis of categorical data. The analyses reported in Avila and Myers (1991), Rhodes and Myers (1991) were all obtained using a PC version. A version of this program that uses the GEO-EAS file format and screen management style is available from the author. The inclusion of various utilities for displaying the results is very important and greatly aids in the interpretation of the results.

Notice. Although the research described in this article has been funded wholly or in part by the U.S. Environmental Protection Agency through a Cooperative Research Agreement with the University of Arizona, it has not been subjected to Agency review and therefore does not reflect the views of the Agency and no official endorsement should be inferred.

References

Avila, F. and Myers, D.E., 1991, Correspondence Analysis applied to environmental data sets: a study of Chautauqua Lake sediments. *Chemometrics and Intelligent Laboratory Systems* **11**, 229-249.

Avila, F., Myers, D.E. and Palmer, C., 1991, Correspondence Analysis and adsorbate selection for chemical sensor arrays. *J. Chemometrics* **5**, 455-465.

David, M., Campiglio, C. and Darling, R., 1974, Progress in R- and Q- mode analysis: Correspondence Analysis and its application to the study of geological processes. *Can. J. Earth Sci.* **11**, 131-146, 603 and 1497-1499.

David, M., Dagbert, M. and Beauchemin, Y., 1977, Statistical analysis in geology: correspondence analysis method. *Quarterly of the Colorado School of Mines* **72**(1), 1-57.

Dutot, A.L., Bergametti, G. and Buat-Menard, P., 1988, Application of Correspondence Analysis to apportion sources of ambient particles. *Atmospheric Environment* **22**, 1737-1743.

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

Hoffman, D., 1991, Statistical Computing Software Reviews (Four Correspondence Analysis Packages). *The American Statistician* **45**, 305-311.

Hopke, Ph. K., Ruppert, D.F., Clute, P.R, Metzger, W.J. and Crowley, D.J., 1976, Geochemical profile of Chautauqua lake sediments. *J. of Radioanalytical Chemistry* **29**, 39-59.

Hopke, Ph. K., 1976, The application of multivariate analysis for interpretation of the chemical and physical analysis of lake sediments. *J. Environ. Sci. Health* **A11**(6), 367-383.

Lancaster, H.O., 1958, The structure of bivariate distributions. *Ann. Math. Statist.* **29**, 719-736.

Lebart, L., Morineau, A., and Warwick, K.M., 1984, *Multivariate Descriptive Statistical Analysis.* New York: John Wiley and Sons.

Rhodes, H.R. and Myers, D.E., 1991, Correspondence Analysis used in the evaluation of lakewater chemistry in the Adirondacks. *J. Chemometrics* **5**, 273-290

Zhou, D., Chang, T. and Davis, J. C., 1983, Dual extraction of R- mode and Q-mode factor solutions. *Math. Geology* **15**(5), 581-605.