AN APPLICATION OF FACTOR ANALYSIS FOR INTERPRETATION OF SOIL ANALYSIS DATA

T.P. ABRAHAM AND A. HOOBAKHT
Soil Institute, Iran

(Received in August. 1972; Accepted in January, 1973)

With the advent of high speed digital computers multivariate techniques are being increasingly used in the analysis of biological data. Factor analysis and component are among the various multivariate methods used widely especially in the field of psychology. However, these techniques are of wide applicability in the field of biology and soil science as a large number of variables are involved in these fields of study. The attempt to study one variable at a time has no doubt given useful results and indiscriminate use of multivariate techniques can do no good. However, due to the intercorrelations among many of the variables measured, it is not possible to obtain complete information from single variate approach only. Pearce and Holland (1960) have discussed some applications of multivariate methods in Botany taking as example the use of component and factor analysis on measurements taken on tree crops· Abraham and Khosla (1965) have used component analysis in the study of pests and diseases. Holland (1969) discussed the application of component analysis to interpretation of soil data taking nutrient composition of soils as an example. With the establishment of soil data banks and computer facilities it is anticipated that advantage will be taken of multivariate techniques to simplify the data by bringing out underlying patterns from the apparently complex inter-correlations among the variables. The object of the present paper is to illustrate the use of factor analysis in the interpretation of soil data collected by the Soil Institute of Iran.

METHOD

The techniques of component analysis and factor analysis have been discussed at great length in standard statistical books such as (Kedall and Stuart 1968) ; Rao (1965), Anderson (1959) and may not be repeated here in detail. In addition, Cattel 1965(a) and 1965(b) have given expository articles on the essentials of factorial analysis and its role in research.

The aim of the factor analysis is to explain observed relationships among numerous variables in terms of simpler relations. The simplification can take the form of producing a set of classificatory categories or creating a smaller set of hypothetical variable. Measurements are taken on p variables over N entities (objects) and the inter correlations among these variables are calculated as they vary over the N entities. There will be $\frac{p(p-1)}{2}$ pairings of the variables producing a square symmetric correlation matrix R. In analysing the structure of this matrix two approaches can be taken, one the principal component analysis and the other the factor analysis. In component analysis the self correlations (diagonal elements of R) are taken as unities and the matrix is used to find out a linear transformation of the original variables to new variables with the property that the new variables are orthogonal to each other, the first component accounts for the largest variation, the second variable giving the second largest variation and so on. The components correspond to the Eigen values and accompanying Eigen factors of R. If the first few components say q out of $p(q<p)$ can account for most of the variation, we have in effect reduced the dimensionality of the p dimensional hyper space to a q-space thus simplifying the data. However, in practice the components isolated do not necessarily correspond to any meaningful physical phenomenon. Further, by equating the self correlations obtained from sample to unity we are ignoring the experimental and sampling errors involved.

Factor analysis on the other hand seeks to explain the matrix of correlations R by a small number of hypothetical variables or factors. The possibility of extraneous variation is also taken into account and the diagonal elements of R will no longer be unities, but are replaced by smaller number known as communalities which represent the proportion of variation of each variate which is due to factors operating on some or all of the other variates, the rest of it is being assumed due to chance. The problems associated with the decision on the number of factors to be extracted and the successive interaction technique for determining the values of the communalities which in turn depends on the number of factors, etc., are discussed in the references cited above. Suppose we have a $p \times N$ matrix of N observations on a $(p \times 1)$ vector X. We suppose that the observed x's are in fact linear combinations of some underlying ρ which are known as factors there being $k<p$ of them.

We have $$x_j=\sum_{m=1}^{k} l_{jm}\,\rho_m+e_j$$

$$V(e_j)=\sigma^2,$$

$$V(\rho_m)=1,\ m=1,\ldots,k$$

We assumed that ρ's and e's are normal with zero means and the x's are also normal. We have then

$$\text{Cov}\,(x_j, x_m) = E\left[\sum_{t=1}^{k} I_{jt}\ \rho_t + e_j\right]\left[\sum_{t=1}^{k} I_{mt}\ \rho_t + e_m\right]$$

$$= \sum_{t=1}^{k} I_{jt}.I_{mt} \qquad j \neq k$$

$$V\,(x_j) = \sum_{t=1}^{k} I^2_{jt} + \sigma^2_j$$

Thus $R = LL' + \triangle$,

where L is the $p \times k$ matrix of coefficients I_{jm} and $\triangle$ is $p \times p$ diagonal matrix of σ^2_j. The problem is to estimate the I's and σ^2's. Due to the indeterminate nature of the system further constraints are imposed such as normalizing the I's. The procedure is explained in detail in Kendal and Stuart (1968).

In addition to finding out the number of factors required to account adequately for an observed set of variables we may be interested in defining what these variables are. When a predetermined number of factors are extracted from a correlation matrix, we have a matrix of factor loadings with k columns for the k factors and p rows for the p original variables. These factor loadings for variables $x_1, \ldots\ldots x_p$ on factors 1, 2, ..., k are the correlations of the newly discovered factors with the original variables. A particular element in the factor loading matrix indicates the extent to which the factor is represented in a given matrix and is, in an unrotated factor matrix, largely a function of the particular method used in extracting the latent roots and vectors of the correlation matrix and may have no empirical meaning. The concept of simple structure was developed among other techniques for an orthogonal rotation of the original factor loading matrix into such a position that the factors extracted are readily identifiable in terms of the original variables. Simple structure is realized by several computational techniques of which the best known is the varimax rotation which aims to maximize the fourth power of the factor loadings and prevents a variable being simultaneously highly loaded on two factors.

MATERIAL

The data considered here were obtained from analysis of soil samples taken from 46 sites in Shiraz area of Fars region in Iran. This is an important agricultural area where wheat is the main crop. The average rainfall is a little over 40 cm. Soil are alluvial with undeveloped profiles and mainly of coarse to medium texture. The soil data were collected from scattered farmers' fields and are thus representative of the tract. The measurements taken were

x_1 = Saturation percentage

x_2 = Electrical conductivity (mmhos/cm)

x_3 = *pH* (determined in water saturated soil paste by using glass electrodes)

x_4 = Available phosphorus (ppm Olsen's method)

x_5 = Available potash (with a flame photometer in an ammonium acetate soil extract)

x_6 = Organic carbon % (determined by quick titration method of Walkley and Blackley)

x_7 = Calcium carbonate

x_8 = CEC

x_9 = Available nitrogen

Analysis And Results

The means and standard deviations are given in Table 1.

The correlation matrix is given in Table 2.

TABLE 1

Mean and Standard Deviations of Soil Variables

Character	Mean	S.D.
Saturation %	44·9	3·93
Electrical Cond.	1·42	0·78
pH	7·95	0·14
Av P	10·31	6·27
Av. K	454·87	137·46
Org. Carbon	0·79	0·17
$CaCO_3$	34·93	6·02
C.E.C.	14·51	2·26
Av. N	126·56	51·30

TABLE 2

The Correlation Matrix

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1·00000	—0·08874	0·14014	—0·02633	0·35928	0·50303	—0·49912	0·17170	0·26650
X_2		1·00000	0·36187	—0·14146	—0·25696	—0·18489	0·15103	—0·38593	—0·34615
X_3			1·00000	0·08372	0·28881	—0·03392	—0·35712	0·12564	—0·05326
X_4				1·00000	0·60188	0·29498	—0·13503	0·36427	0 16524
X_5					1·00000	0·43502	—0·42455	0 42901	0·32308
X_6						1·00000	—0·18418	0 09255	0·24214
X_7							1·00000	- 0·43385	0 07027
X_8								1·00000	0 00211
X_9									1·00000

The factors were extracted one by one. With four factors the Eigen values were

$$\lambda_1=2.96359$$
$$\lambda_2=1.60219$$
$$\lambda_3=1.27223$$
$$\lambda_4=1.09635$$

The rotated factor matrix with 4 factors is given in Table 3.

TABLE 3

Rotated Factor Matrix

Variable				
1	−0.19354	0.09147	—0.76084	—0.48702
2	—0.17827	0.84117	0 16004	0.26895
3	0.20958	0.76553	—0.00872	—0 33309
4	0.91260	—0.00146	—0.08755	—0,02435
5	0 70476	0.06050	—0.43456	—0,35837
6	0 20642	—0.03594	—0.76516	—0.07764
7	—0.04560	—0.14807	0.18631	0.87107
8	0,44514	—0.28345	0 13435	—0.70262
9	0.24389	—0.28376	- 0.63322	0.22678

The communalities are given in Table 4.

TABLE 4

Communalities of Different Variables

Variable	Communality
1	0.86190
2	0 83730
3	0.74099
4	0.84111
5	0.81764
6	0.63541
7	0 81748
8	0 79021
9	0.59214

We have taken only up to 4 factors. Since the 5th factor corresponds to an Eigen value of 0.79967 which, if we follow Guttman's lower bound principle that any $\lambda < 1$ should be ignored, need not be taken into account.

Ignoring the non-significant correlation the orthogonal factors extracted can be written as

$$\begin{aligned} f_1 &= 0.91260\, x_4 + .70471\, x_5 + .44514\, x_8 \\ f_2 &= 0.84117\, x_2 + .76553\, x_3 \\ f_3 &= -0.76084\, x_1 - .76516\, x_6 - .63322\, x_9 \\ f_4 &= -0.48702\, x_1 + .81707\, x_7 - 0.70262\, x_8 \end{aligned}$$

The communalities show that the variables 6 and 9 viz. organic carbon and available nitrogen show large residual variation apparently due to high error of determination of these characters.

In factor analysis one of the major difficulties is to give a physical meaning to the factors extracted. In the present case the first factor is basically related to the cation exchange capacity. As the soils under consideration are generally low in organic matter the cation exchange capacity is likely to reside largely in the mineral fraction such as potassium. The second factor is a combination of Electrical Conductivity and *pH* which represents the salinity-alkalinity complex. The third factor represents saturation percentage, organic carbon and available nitrogen percentage. In other words this is a factor basically related to textural and nitrogen status. The meaning of the fourth factor is not so clear although it also appears primarily related to the exchangeable cations other than calcium. If we ignore the relatively low weightings 0.44514 in f_1 and $-.48702$ in f_4 the explanation of the basic factors seems more clear.

These are

(1) f_1^1 signifying nutritional status of the soil excepting nitrogen.

(2) f_2^1 the salinity-alkalinity factor.

(3) f_3^1 saturation percentage, 0.6% and available nitrogen (basically textural and nitrogen status).

(4) The exchangeable base status as measured only by cation exchange capacity and calcium.

Thus factor analysis appears to have brought out some of the basic factors associated with soil status and could be considered an important tool in explanatory work. It may be added that further resolution of the rotated orthogonal matrix to oblique axis may sometimes bring out more clearly the underlying factors. However, in the present case even without such rotation meaningful factors have been extracted. Scores based on these factors could be used for comparing different soil series basically in respect of internal soil characters generally used for crop-response soil test relations.

Summary

The technique of factor analysis has been applied to extract basic factors underlying observed soil variables in Shiraz area of Fars region in Iran. It was found that four factors corresponding to (1) signifying nutritional status, (2) salinity-alkalinity complex, (3) Textural and organic carbon, and (4) Exchangeable base status could be brought out as the underlying factors, so that, scores based on these basic factors could be used for comparison of internal soil variables.

References

1. Abraham, T. P. and Khosla, R. K. [1965] : On the possible use of component analysis techniques in Pest and Disease, J. Ind. Soc. Agri. Stat., Vol. 17.

2. Cattel, R. B. [1965(a)] : Factor analysis. An introduction to essentials. Biometrics, Vol. 21. The purpose and underlying models No. 1.

3. Cattel, R. B. [1965(b)] : Factor analysis. The role of factor analysis in Research Biometrics, Vol. 21, No. 2.

4. Holland, D. A. [1969] : Component analysis—An approach to the Interpretation of Soil Data, J. Sci. Fd. Agric., Vol 2C.

5. Kendall, M. G. and Stuart, A. [1968] : The advanced theory of statistics, Vol. 3.

6. Pearce, S. C. and Derek, A. Holland [1960] : Some applications or multivariate methods in Botany-Applied Statistics, Vol. IX, No. 1.

7. Rao, C. R. [1965] : Linear statistical inference and its applications.