APPLYING PRINCIPAL COMPONENT ANALYSIS TO SOIL-LANDSCAPE RESEARCH-QUANTIFYING THE SUBJECTIVE

R. David Hammer

John W. Philpot

Jon M. Maatta

Applying Principal Component Analysis to Soil-Landscape Research--Quantifying the Subjective

R.David Hammer
Assistant Professor of Soil Science, University of Missouri, Columbia
John W. Philpot
Professor of Statistics, University of Tennessee, Knoxville
Jon M. Maatta
Assistant Professor of Statistics, University of Missouri, Columbia

ABSTRACT

Principal component analysis is a multivariate statistical procedure that can be used to identify factors (correlated subsets of variables) in large data sets. This statistical method appears useful for scientists investigating soil processes, but it has received little attention. Reported applications of principal component analysis share a common fault--subjective, user-specified analytical options apparently are not recognized, for they are not discussed. Reported data sets are often small, have low observations-per-variable ratios, and lack tests of robustness. A large soil data set is used to demonstrate systematic procedures for an optimum rotated principal component solution. This solution retained 21 variables aligned among four "clean" and "logical" factors, and extracted 79% of the variance. Robustness was confirmed by comparison with common factor analysis solutions. When carefully applied, the presented guidelines should enhance scientists' abilities to identify and transfer knowledge about multivariate data sets, and should allow different scientists to independently arrive at similar factor solutions.

Key words: *factor analysis, communality, variance, robust.*

1. Introduction

Factor Analysis

Common factor analysis (FA) is a statistical procedure designed to identify factors as correlated subsets of variables in large data sets. Principal component analysis (PCA) provides a set of components that, when rotated, can be used for the same purpose. It should be made clear that all references to PCA in this paper refer not to the initial principal components that

Contribution of Missouri Agric. Exp. Stn. Journal Series No. 11,238.

can be extracted, but to the rotated components, or factors, that define the correlated subsets of variables; hence the use of the word "factors" throughout the paper. The essential difference between FA and PCA lies in the definition of the variance that is targeted for extraction, but both can be considered forms of factor analysis. In fact, major software designers (SAS, SPSS, STATGRAPHICS,and SYSTAT) have made principal components the default option in their factor analysis procedures.

Harman (1976) defined FA as an exploratory statistical device which will provide the best results when the "practitioner" understands " . . . at least in principle, what is going on in the analysis of a body of data. . ." In other words, some knowledge of the system being examined will be important in conducting and interpreting FA.

Burden of Proof

A survey of the literature in which PCA/FA are used to investigate soil-landscape and soil-site relationships revealed a distressing lack of consideration for important PCA/FA analytical procedures (Arp, 1984; Campbell et al., 1970; Fourt et al, 1971; La Bastide and Van Goor, 1970; Litaor et al., 1989; Nortcliff, 1978; Ovalles and Collins, 1988; Page, 1976; Richardson and Bigler, 1984; Rowe and Sheard, 1981; Sarkar et al., 1966; Severson, 1981; Sondheim et al., 1981; Sondheim and Standish, 1983; Williams and Rayner, 1977). Since clear guidelines for determining the "optimum" final PCA/FA solution currently do not exist, we believe that the burden of proving that reported results are optimum lies with the investigators. A need exists for more complete and precise reporting of statistical methodology, particularly with "nontraditional" procedures. Lack of awareness of analytical methods used to produce reported results reduces confidence in the results and the methodology. More importantly, the ability to transfer knowledge from system to system is hindered. Among the deficiencies that can be found in the literature are:

1. Observations-per-variable ratio is not addressed (in some cases, investigators used fewer observations than variables);
2. Data sets often are extremely small, a problem accentuated when;
3. Robustness of the data is seldom addressed or tested;
4. Research objectives sometimes are not suited for PCA/FA;
5. Important methodologies are not discussed, including the procedures for determining the optimum number of extracted factors and criteria to determine which variables contribute to the variance extracted, and;
6. The criteria for an "optimum" solution are not discussed.

Objectives

The objective of this manuscript is to present and discuss a methodology for a more systematic PCA/FA analysis. Specific emphasis will be given to determining which variables to retain in the data set and in extracting the optimum number of factors. A large data set of soil chemical and physical properties will be used as the template. Four complete rotated factor loading matrices will be presented so that the reader may begin to develop an appreciation of the the subtle changes which occur as the data are compressed into an optimum solution.

2. Methodology

The Data Set

Data were obtained as part of a research project designed to statistically evaluate a forest land classification for the Mid-Cumberland Plateau in Tennessee (Smalley, 1982; 1984). All soil profiles had A (surface) horizons, Bt (argillic horizons with high clay content), and transition horizons that were labelled AB horizons. A total of 138 grid points was sampled.

Soil physical properties measured in the field included horizon matrix color (from the Munsell charts), coded (Buntley and Westin, 1965) for statistical analyses, and horizon thicknesses.

Basic Nomenclature and Principles

Principal component analysis and common factor analysis ordinarily begin with the correlation matrix and are mathematically similar in many respects. The methods share the common goals of; 1) attempting to summarize intercorrelations among the variables, and 2) reducing a large number of variables to a smaller number of factors. The major difference in the two methods is the variance that is targeted for explanation. PCA attempts to account for all of the variance in the data set, as represented by the 1.0 values in the diagonal of the correlation matrix. Common factor analysis attempts to account for only that variance that is common to the variables sampled. For example, the shared variance for a given variable is often represented by the squared multiple correlation, R^2j, between the jth variable and the rest of the variables. Clearly R^2j is less than or equal to 1.0. The focus on only that portion of the variance that is shared by the variables led to the phrase "common factor analysis". Detailed discussion is beyond the scope of this paper.

The remainder of this paper will center, for the sake of convenience, upon PCA and orthogonal rotations of the principal components into factors. The presented methodology applies equally well to FA. Factor analysis could be substituted for PCA throughout the manuscript without changing the relevance of interpretations.

The general PCA model, after orthogonal rotation, can be represented by:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots a_{jm}F_m; \qquad (1)$$

where any variable (z_j) is represented by the sum of its loadings ($a_{ji} \ldots a_{jm}$) onto the factors ($F_1 \ldots F_m$). The a_{jk} values form an n-by-m factor matrix, each row of which shows how that variable loads on the m factors. Each column of the matrix shows the relative strength of the variables in that factor. The factor is *characterized* by those variables which load primarily upon it. "Primary loaded" variables should account for most of the variance extracted by the factor. The total variance (s^2_k) the factor extracts from the data is the sum of the squared factor loadings in the kth column, or:

$$s^2_k = a^2_{1k} + \ldots a^2_{nk}; \qquad (2)$$

where "1" through "n" represent individual variables. A single factor is a vector in space which contains positive and negative loadings. These "directions," as well as the variables themselves, will determine the user's interpretation of the factor.

The estimation of the correlation matrix during PCA is characterized by the use of communalities to estimate the "1's" on the diagonal. The communality (h^2_j) of a variable, j, is the sum of the squared factor loadings of the variable (row j of the matrix):

$$h^2_j = a^2_{j1} + a^2_{j2} + \ldots a^2_{jm}, \qquad (3)$$

and represents the amount of variance of that variable that is captured by the factors. Low communalities indicate that the variable is somewhat unique, and doesn't cluster well with the other variables in the data set. Such variables are inappropriate for factor analysis.

Statistical Methods

The procedures and definitions detailed below have been used for a number of years in a graduate level Factor Analysis course taught at the University of Tennessee by Dr. Philpot. The procedures have been tested on data from the behavioral and biological sciences. The PCA analysis described below was performed on a MacIntosh SE with an accelerator board and SYSTAT version 3.2 software (Wilkinson, 1986). The FA results with which the PCA analysis is compared were obtained using SAS (SAS Institute Inc, 1982) mainframe software, using the same iterative techniques.

Assumptions made prior to these analyses and adhered to during the search for the "best" PCA/FA solution were: 1) the optimum number of factors and variables is not known, and will be determined during the analysis; 2) judgement and statistical procedures will be used together to determine the final solution; and 3) the final solution, if valid, will be "logical" (it will make sense) and "clean" (will consist of variables strongly aligned upon factors and will have few secondary loadings).

Interpreting the Factor Matrix

Each stage leading to a final PCA solution requires determining the optimum number of factors to extract and deciding which variables to retain in the data set. The rotated PCA solution should consist of factors which each contain several correlated variables which are "primarily" loaded upon that factor. Factors should have few secondary loadings. Each variable will load onto all factors in the solution, but should load primarily onto one factor. The objective here is to achieve a simple, easily interpreted factor matrix via orthogonal rotation.

The factor matrix can be developed almost like a print is developed in the darkroom. The procedure for finding the structure inherent in the factor matrix is:

1. Calculate the mean and standard deviation of the communalities of all variables involved. For any variables having communalities more than two standard deviations below the mean, the corresponding row of the factor matrix should be deleted.
2. Examine the remaining rows of the factor matrix. Highlight the(absolute) maximum loading in each row. These are primary loadings.
3. Find the (absolute) minimum of the primary loadings. This value determines the lower boundary of the salient loadings.
4. Highlight all other (absolute) loadings in the factor matrix that equal or exceed the boundary value found in step 3. These are secondary loadings. The salient loadings consist of the primary loadings for each variable plus any secondary loadings.
5. Re-analyze using only those variables with robust communalities and only those factors that are well defined by three or more salient loadings.

Table 1 is an example of a rotated loading score matrix. The primary loading of a variable is the highest absolute value of all its loading scores. In this case, the primary loading of organic matter is 0.712 on factor 2. Calcium (Ca) loads primarily onto factor 1, and magnesium (Mg) loads primarily onto factor 2. The minimum primary loading is 0.448, which becomes the boundary value. A secondary loading occurs when the absolute value of a non-primary loading of a variable equals or exceeds the boundary value. In table 1, calcium has a secondary loading on factor 2 because the loading score of 0.452 exceeds

0.448. Large numbers of secondary loadings in an analysis create ambiguity and decrease the clarity of interpretations.

Factor Extraction

The initial number of factors to extract is often based on one of two criteria--a scree plot or the number of eigenvalues greater than 1. Alternatively, an extraction with n/3 factors, where n is the number of variables, should provide a good starting point. After the initial extraction, the number of factors will be depend on the pattern of salient loadings. A factor should contain at least three variables with salient loadings to be considered for retention.

Variable Retention

Two criteria were used to determine which variables to retain--variable "behavior" and the distance of the communality of each variable from the mean communality for that extraction. Variable "behavior" refers to the persistence of alignment of individual variables with correlated variables upon a factor. Variables should not "jump around" from one stage of the analysis to the next.

Robustness and Observations-per-Variable Ratio

Observation-to-variable ratio is an important topic not well addressed in the literature. Tabachnick and Fidell (1983) recommended a minimum of 100 samples and suggested that 1000 observations would be optimum. Kendall (1975) suggested that the number of observations should be 10 times the number of variables. The rule of thumb Philpot developed through teaching and research experience is that n*(n-1)/2 is the preferred number of observations for n>7 (where n is the number of variables), and 3n is the minimum acceptable ratio. This study utilized 138 observations on three soil horizons. A total of 38 soil variables resulted (Table 2).

Tabachnick and Fidel (1983) suggested that robustness be verified by splitting the data set and performing, then comparing, separate analyses. Ideally, similar results should be achieved. The time and expense required to build a robust soils data set from field and laboratory analyses generally precludes this technique. Robustness of the data analyzed in this project was tested by subjecting the data to FA solutions using Minimum Residual (MINRES), Maximum Likelihood, and Image analyses, and comparing them with PCA. This approach is patterned after the recommendations of Harris (1967).

3. Results and Discussion

Changes in variable alignment, factor loadings and variance extracted were observed as the data were systematically compressed into the optimum solution. Careful observation of the rotated factor loading matrices provides insight into relationships among the variables. More importantly, the changes observed in successive stages underscore the importance of bringing careful, objective logic to PCA/FA analyses. Four of the rotated factor loading matrices are compared and discussed to illustrate this point.

The analyses required 13 stages to reach a final solution. Table 3 summarizes the progression of steps to the final (optimum) PCA solution, which contained 17 fewer variables and four fewer factors than the second stage. The fluctuation in the minimum communality is noteworthy, and ranged from 0.163 (third stage) to 0.723 (ninth stage). As variables with low communalities were progressively dropped, the minimum communality rose steadily. The optimum solution, which had a minimum communality of 0.701 and no secondary loadings, indicates that all of the retained variables extracted high variance and were strongly aligned with a particular factor.

Table 4 is the rotated factor loading matrix of the 38 original variables on six factors, and is from the third stage (Table 3). Calcium in the A horizon has three secondary loadings (on factors 1,4, and 5) and is aligned with K and Mg on factor 6. The three secondary loadings indicate that excessive "noise" exists. The number of secondary loadings in the matrix (20) indicates great ambiguity in this solution.

The communalities of pH (H_2O) in the Bt and extractable acidity in the AB (0.163 and 0.245), respectively, were more than two standard deviations from the mean, so these variables were dropped prior to the fourth stage. The fourth stage revealed that two more variables (pH (KCl) in the Bt horizon and thickness of the AB horizon) had communalities sufficiently low to warrant their removal. The fifth stage (Table 5) contained 34 variables and 6 factors.

Comparison of tables 4 and 5 reveals interesting changes in the composition of the rotated factor loading matrix after dropping four variables. Factor 5 remained virtually unchanged, accounting for 7.8% of the variance and loaded primarily by the same variables. Factors 1 and 3 seemed to change places between the third and fifth stages. The variables aligned upon factor 1 in Table 4 aligned upon factor 3 in Table 5, and the amount of variance extracted increased to 18.6% from 17.7%. Four fewer variables are contributing to the variance, but the stronger alignment of the remaining variables increased the variance extracted by factor 3. The variable "A horizon clay" loaded primarily onto factor 6 in the third stage, but by the fifth stage it had aligned onto factor 1 with several other textural variables. Variables associated with factor 2 did not change, but the signs of the loading scores were reversed. The total amount of variance extracted by the factors increased from 70.2% to 75.8% during the two stages. The changes resulting from removal of four variables illustrate the importance of some degree of uniformity in deciding how to determine and report the optimum PCA/FA solutions of large multivariate data sets.

A subjective decision was made at this juncture in the analysis. Alignment of the textural variables was not so "clean" as had been expected. A colleague (G.J. Buntley, personal communication) suggested that fine silts would behave like clays and that fine sands and very fine sands would behave like coarse silts. The 10 textural variables were combined into four classes. Within horizons, fine sands, very fine sands, and coarse silts were combined into a single variable called "silt" and clay and fine silt were combined into a variable called "clay."

Results of combining the textural variables were very satisfactory (Table 6). The textural variables together aligned primarily onto a single factor which accounted for 14.4% of the extracted variance. The alignment of textural variables left the A horizon cations (except Na) aligned upon a single factor with organic matter. The new factor 1, loaded primarily by cations from the AB and Bt horizons, extracted 21.6% of the variance. This was the first stage in which the ultimate variable alignment began to become apparent. Three variables (Na in the AB and Bt horizons, and A horizon color) possessed communalities more than two standard deviations below the mean, and were dropped from the subsequent stage.

Factor 4 of the seventh stage had primary loadings for only the two A horizon pH variables, and no secondary loadings. The correlation matrix revealed a correlation coefficient between these two variables of 0.911. The high correlation indicated that the two variables could be considered as one in this horizon. Since there were fewer than three salient loadings, the decision was made to reduce the number of factors by one.

The eighth stage revealed that organic matter and Na in the A horizon had low communalities, so they were dropped. The ninth stage contained one factor (number 5) loaded primarily by only two variables from the A horizon--extractable acidity and thickness. The factor accounted for only 7% of the variance, and the correlation coefficient between the two variables was only -0.464, so a subjective decision was made to compress the data further by removing one more factor. The two variables were dropped subsequently because of low communalities. They had loaded onto a shared factor not because they were highly correlated, but because they were not strongly correlated with variables aligning upon the more robust factors. The preceding discussion illustrates the importance of consulting the correlation matrix when considering relationships among variables in the rotated factor loading scores matrix.

The twelfth stage resulted in no secondary loadings and no communalities more than two standard deviations from the mean. Four robust factors each contained several correlated primary loaded variables (Table 7). This was deemed the "optimum" solution after subsequent compression to 3 factors resulted in an increase in secondary loadings (from 0 to 4) and a reduction in variance extracted (from 79.3% to 70.9% (Table 3)). The final solution was "clean" and "logical." The four factors could be named according to the variables loaded upon them, as indicated in table 8.

The variables dropped from the data set during compression to the final solution accounted for a total of only 22.4% of the variance extracted from the first stage of 12 factors and 38 variables.

One could logically argue that the eleventh stage would have been an acceptable final solution. It contained no secondary loadings and differed from the 12th stage primarily in that A horizon thickness was dropped. Comparison of this rotated PCA solution with several FA solutions and a rotated PCA solution using SAS (SAS, 1982) software indicates that A horizon thickness was retained by Maximum Likelihood analysis and Minimum Residuals, but was dropped in Image analysis and in the other PCA analysis (Table 9). Results of the four procedures were very similar. The data used in this analysis appear to be robust.

One might question how the data would have behaved had the textural variables been combined prior to the first stage rather than part way through the analyses. To test this concern, the analysis was repeated from the beginning, using the same procedures and guidelines upon the data, but with the textural variables combined. The same results were obtained. The same variables were retained, and the same alignment was achieved

4. Conclusions

The various FA procedures and rotated PCA produced "clean," "logical" final solutions. The methods used to obtain the optimum solution produced similar results with all procedures, indicating that different investigators could probably arrive independently at the same conclusions following the suggested guidelines.

Knowledge of the subject matter and reference to the correlation matrix were essential in arriving at the final solution. Deming's (1960) observation ". . .mathematics, judgement, and substantive knowledge work together to the best advantage." is particularly apropos for multivariate analyses. One should not expect the statistical procedures to sort all the noise from the data. In the hands of the careful, competent investigator, the subjective nature of these multivariate statistical procedures is not a detriment. On the contrary, the astute investigator obtains valuable insight into the structure and relationships of the variables as the optimum solution is developed.

The analyses reduced a large, complex data set to four logical factors, and indicated that approximately one fourth of the variables made minimal contributions to the structure of the data set. Subsequent analysis of soil-landscapes in this region could focus upon the correlated variables. This would allow investigators to build larger data sets without additional expenditures for laboratory analyses.

5. Acknowledgements

Field and laboratory research were conducted while the senior author was a research associate at the University of Tennessee, Knoxville. The authors thank R.J. Miles for assistance with field sampling and G.W. Smalley, W.L. Parks, and G.J. Buntley for encouragement and support during field and laboratory efforts. The research was a cooperative agreement between the Institute of Agriculture and the U.S.D.A Forest Service. Partial funding was provided through grant number FS-19-81-34 from the U.S. Forest Service, Southern Forest Exp. Stn., New Orleans, LA.

6. References

Arp, P.A. 1984. Forest floor variability and factor analysis: A case study. Can. J. Soil Sci. 64: 457-461.

Buntley, G.J. and F.C. Westin. 1965. A comparative study of developmental color in a Chestnut-Chernozem-Brunizem soil climosequence. Soil Sci. Soc. Am. Proc. 29: 579-582.

Campbell, N.A., M.J. Mulcahy, and W.M. McArthur. 1970. Numerical classification of soil profiles on the basis of field morphological properties. Aust. J. Soil Res. 8: 43-58.

Deming, W.E. 1960. Sample design in business research. Wiley Classics Library Ed. pub. 1990. Wiley Interscience , NY.

Fourt, D.F., D.G. M. Donald, J.N.R. Jeffers, and W.O. Binns. 1971. Corsican pine (*Pinus nigra* var. *maritima* (Ait.) Melville) in Southern Britain. Forestry 44: 189-207.

Gorsuch, R.L. 1974 Factor Analysis W.B. Saunders Co., Philadelphia.

Harman, H.H. 1976. Modern factor analysis. 3rd Ed., Univ. of Chicago Press, Chicago.

Harris, C.W. 1967. On factors and factor scores. Psychometricka 32: 363-379.

Johnson,R.A , and D.W. Wichern. 1988 Applied Multivariate Statistical Analysis, 2nd Ed. Prentice Hall, Englewood Cliffs, N.J.

Kendall, M. 1975. Multivariate analysis. Hafner Press, NY.

Kim, J O., and C.W. Mueller. 1978 Factor Analysis: Statistical Methods and Practical Issues. Sage Publications, Beverly Hills

La Bastide, J.G.A. and C.P. Van Goor. 1970. Growth-site relationships in plantations of *Pinus elliottii* and *Araucaria angustifolia* in Brazil. Plant and Soil 32: 349-366.

Litaor, M.I., Y. Dan, and H. Koyumdjisky. 1989. Factor analysis of a lithosequence in the northestern Samaria Steppe (Israel). Geoderma 44: 1-15.

Nortcliff, S. 1978. Soil variability and reconnaissance soil mapping: A statistical study in Norfolk. J. Soil Sci. 29: 403-418.

Ovalles, F.A. and M.E. Collins. 1988. Variability of northwest Florida soils by principal component analysis. Soil Sci. Soc. Am. J. 52: 1430-1435.

Page, G. 1976. Quantitative evaluation of site potential for spruce and fir in Newfoundland. Forest Sci. 22: 131-143.

Richardson, J.L. and R.J. Bigler. 1984. Principal component analyses of priaire pothole soils in North Dakota. Soil Sci. Soc. Am. J. 48: 1350-1355.

Rowe, J.S. and J.W. Sheard. 1981. Ecological land classification: A survey approach. Environ. Manag. 5: 451-464.

Sarkar, P.K., O.W. Bidwell, and L.F. Marcus. 1966. Selection of characteristics for numerical classification of soils. Soil Sci. Soc. Am. Proc. 30. 269-272.

SAS Institute Inc. 1988. SAS/STAT Users' Guide: Release 6.03 ed. SAS Institute Inc., Cary, NC.

Severson, R.C. 1981. Evaluating chemical character of soil material for suitability in rehabilitating mined land in the San Juan Basin, New Mexico. Soil Sci. Soc. Am. J. 45: 396-404.

Smalley, G.W. 1982. Classification and evaluation of forest sites on the mid-Cumberland Plateau. U.S.D.A. Forest Serv. Gen. Tech. Rep. SO-38.

Smalley, G.W. 1984. Landforms: A practical basis for classifying forest sites in the interior uplands. pp. 92-112. In Proceedings, 12th Annual Hardwood Sympopsium, Hardwood Research Council, Chashiers, North Carolina.

Sondheim, M.W., G.A. Singleton and L.M. Lavkulich. 1981. Numerical analysis of a chronosequence, including the development of a chronofunction. Soil Sci. Am. J. 45: 558-563.

Sondheim, M.W. and J.T. Standish. 1983. Numerical analysis of a chronosequence including an essessment of variability. Can. J. Soil Sci. 63: 501-517.

SPSS Inc. 1990 SPSS Reference Guide. SPSS Inc.,Chicago, Ill.

STST Inc. 1988 STATGRAPHICS User's Guide: Statistical Graphics System STSC Inc., Rockville, Md.

Tabachnick, B.G. and L.S. Fidell. 1983. Using multivariate statistics. Harper and Row, New York.

Wilkinson, L, 1986. SYSTAT: The system for statistics. SYSTAT, Inc., Evanston, Il.

Williams, C. and J.H. Rayner. 1977. Variability in three areas of the Denchworth soil map unit. III. Soil grouping based on chemical composition. J. Soil Sci. 28: 180-195.

Table 1. Examples of primary and secondary loadings on a rotated factor loading score matrix.

VARIABLE	FACTOR 1	FACTOR 2
Organic matter	-0.038	-0.712*
Ca	0.561*	-0.452∞
Mg	0.115	-0.448*

* Primary loading
∞ Secondary loading

Table 2. Variables used in the Principal Component Analysis. The "x" indicates that the variable was included in the data set.

	SOIL HORIZON		
VARIABLE	A HORIZON	AB HORIZON	Bt HORIZON
Organic matter	X	-	-
pH (H_2O)	X	X	X
pH (KCl)	X	X	X
K	X	X	X
Ca	X	X	X
Mg	X	X	X
Na	X	X	X
Thickness	X	X	X
Color	X	X	X
Extractable acidity	X	X	X
Fine Sand	X	-	X
Very Fine Sand	X	-	X
Coarse Silt	X	-	X
Fine Silt	X	-	X
Clay	X	-	X

Table 3. Characteristics of stages leading to final factor solution.

Stage	Factors	Variables	Communalities Mean	Communalities Max	Communalities Min	Secondary Loadings	Eigen. > 1	% Var. Extracted	Variables Dropped
1	12	38	—	—	—	—	8	79.98	————
2	8	38	0.764	0.944	0.420	7	8	76.54	————
3 (Table 4)	6	38	0.702	0.941	0.163	20	8	70.21	pH (H2O)Bt Extractable acid. (AB)
4	6	36	0.734	0.940	0.424	0	7	73.37	pH(KCl)Bt Thickness (AB)
5 (Table 5)	6	34	0.758	0.946	0.480	4	6	75.78	Combine textural variables
6 (Table 6)	6	28	0.781	0.920	0.557	0	6	78.13	Na (AB),Na(Bt) Color (A)
7	6	25	0.821	0.919	0.706	0	6	82.12	†
8	5	25	0.779	0.913	0.593	3	6	77.88	Organic matter, Na (A)
9	5	23	0.805	0.915	0.723	1	4	80.39	††
10	4	23	0.762	0.904	0.403	3	4	76.19	Extractable acid . (A)
11	4	22	0.781	0.906	0.575	0	4	77.53	Thickness (A)
12 § (Table 7)	4	21	0.793	0.909	0.701	0	4	79.26	————
13	3	21	0.702	0.844	0.377	4	4	70.86	

† Dropped one factor. Factor 4 was loaded only by the two pH variables in the A horizon.
†† Dropped one factor. Factor 5 was loaded only by the A horizon thickness and A horizon extractable acidity, and accounted for only 7% of the variance.
§ The optimum (final) solution.

Table 4. Rotated factor loading matrix of the 38 variables upon 6 factors. Primary loadings are in bold type, and secondary loadings are indicated with an asterisk (*). This was the third stage of the solution.

VARIABLE	FACTOR 1	2	3	4	5	6
Organic matter	0.017	-0.016	0.160	-0.040	0.011	**0.843**
pH (H2O) A	-0.194	0.009	0.132	**-0.806**	0.003	-0.180
pH (KCl) A	-0.289	0.026	0.195	**-0.787**	-0.158	-0.178
K A	0.333	0.045	0.139	0.354*	0.408*	**0.683**
Ca A	0.393*	0.008	0.105	0.467*	0.348*	**0.610**
Mg A	0.245	0.072	0.248	0.462*	0.213	**0.709**
Na A	0.127	**0.593**	-0.050	-0.273	0.277	0.348*
Ca AB	**0.918**	0.164	0.020	0.195	0.002	0.020
Mg AB	**0.811**	-0.142	0.234	0.123	0.187	0.205
Na AB	-0.112	-0.006	-0.349*	**-0.674**	0.092	-0.052
K AB	**0.860**	0.023	0.116	0.072	0.091	0.056
pH (H2O) AB	**-0.748**	0.354*	0.159	-0.075	-0.240	0.073
pH (KCl) AB	**-0.580**	-0.253	0.475*	-0.143	-0.110	0.306
Ca Bt	**0.885**	0.163	0.073	0.198	0.073	0.152
Mg Bt	**0.812**	-0.091	0.152	0.214	0.095	0.290
Na Bt	0.204	-0.203	-0.060	0.019	**0.633**	0.207
K Bt	**0.683**	0.081	0.305	0.210	0.278	0.314
pH (H2O) Bt	-0.032	-0.060	0.027	0.040	**0.347**	0.188
pH (KCl) Bt	−0.207	0.194	**0.568**	-0.125	0.061	0.008
Thickness A	0.428*	0.053	0.038	0.362*	**0.613**	0.219
Thickness AB	-0.074	-0.060	0.425*	0.064	**0.452**	-0.062
Thickness Bt	-0.144	**0.730**	0.118	0.149	-0.011	0.315
Color A	0.033	-0.066	0.048	0.358*	**-0.661**	-0.058
Color AB	0.183	**-0.575**	0.060	-0.433*	0.275	-0.253
Color Bt	0.124	**-0.705**	0.171	-0.259	0.390*	0.156
Coarse Silt A	0.063	-0.089	0.402*	0.116	0.094	**0.680**
Fine Silt A	0.129	-0.120	**0.807**	-0.240	0.059	0.026
Clay A	0.275	0.280	0.321	0.348*	-0.053	**0.504**
Coarse Silt Bt	0.215	-0.117	**0.728**	0.119	-0.083	0.263
Fine Silt Bt	0.248	0.037	**0.730**	0.064	0.205	0.292
Clay Bt	0.261	**-0.510**	0.246	-0.075	0.072	0.335
Extractable Acidity A	-0.338	0.175	-0.114	-0.097	**-0.655**	0.220
Extractable Acidity AB	**0.418**	0.026	0.113	-0.192	-0.131	0.055
Extractable Acidity Bt	0.324	**0.726**	-0.314	0.169	-0.124	-0.059
Very fine sand A	-0.079	**0.863**	-0.084	0.182	0.104	0.055
Fine sand A	-0.296	-0.086	**-0.712**	-0.192	-0.183	-0.517*
Very fine sand Bt	-0.151	**-0.789**	-0.211	0.190	-0.025	0.126
Fine sand Bt	-0.294	0.010	**-0.753**	-0.172	-0.171	-0.373*
Variance Extracted (%)	17.7	11.6	12.1	9.4	7.8	11.5

Table 5. Rotated factor loading matrix after 5 stages. Primary loadings are in bold type. Secondary loadings are indicated with an asterisk (*).

VARIABLE	FACTOR 1	2	3	4	5	6
Organic matter	0.250	0.028	0.005	-0.029	-0.048	**0.861**
pH (H2O) A	0.073	-0.017	-0.210	**-0.842**	0.044	-0.145
pH (KCl) A	0.114	-0.035	-0.309	**-0.825**	-0.108	-0.157
K A	0.272	-0.032	0.330	0.381	0.346	**0.670**
Ca A	0.231	0.005	0.396	0.487*	0.282	**0.599**
Mg A	0.378	-0.069	0.223	0.490*	0.161	**0.658**
Na A	-0.016	**-0.584**	0.103	0.159	0.329	0.366
Ca AB	0.107	-0.161	**0.925**	0.159	-0.004	0.009
Mg AB	0.354	0.136	**0.771**	0.130	0.199	0.144
Na AB	-0.387	0.020	-0.090	**-0.647**	0.125	0.033
K AB	0.179	-0.012	**0.859**	0.040	0.069	0.075
pH (H2O) AB	0.059	-0.353	**-0.753**	-0.069	-0.242	0.072
pH (KCl) AB	0.462*	0.234	**-0.661**	-0.089	-0.088	0.207
Ca Bt	0.180	-0.161	**0.888**	0.168	0.054	0.134
Mg Bt	0.283	0.091	**0.801**	0.198	0.075	0.245
Na Bt	0.015	0.216	0.194	0.078	**0.647**	0.211
K Bt	0.399	-0.073	**0.675**	0.182	0.218	0.297
Thickness A	0.119	-0.038	0.435	0.389	**0.565**	0.227
Thickness Bt	0.161	**-0.740**	-0.168	0.200	0.010	0.238
Color A	-0.040	0.054	0.057	0.317	**-0.679**	-0.118
Color AB	0.046	**0.572**	0.128	-0.387	0.354	-0.260
Color Bt	0.213	**0.710**	0.091	-0.224	0.400	0.139
Coarse silt A	0.523*	0.078	0.015	0.139	0.061	**0.594**
Fine silt A	**0.807**	0.100	0.054	-0.293	0.059	-0.056
Clay A	**0.440**	-0.294	0.231	0.379	-0.054	0.382
Coarse silt Bt	**0.784**	0.097	0.157	0.069	-0.117	0.158
Fine silt Bt	**0.792**	-0.052	0.185	0.031	0.163	0.207
Clay Bt	0.321	**0.505**	0.209	-0.027	0.084	0.265
Extractable acidity A	-0.150	-0.170	-0.292	-0.150	**-0.721**	0.255
Extractable acidity Bt	-0.291	**-0.725**	0.342	0.185	-0.088	-0.046
Very fine sand A	-0.037	**0.861**	-0.078	0.207	0.094	0.024
Fine sand A	**-0.813**	0.100	-0.229	-0.187	-0.147	-0.407
Very fine sand Bt	-0.185	**0.801**	-0.113	0.197	-0.062	0.134
Fine sand Bt	**-0.852**	0.013	-0.212	-0.712	-0.159	-0.234
Variance extracted (%)	15.4	12.8	18.6	10.5	7.8	10.6

Table 6. Rotated factor loading matrix after 6 stages. Textural variables havebeen combined. Primary loadings are in bold type.

VARIABLE	FACTOR 1	2	3	4	5	6
Organic matter	-0.005	0.029	0.232	0.020	0.002	**0.900**
pH (H2O) A	-0.187	0.113	0.039	**-0.868**	0.020	-0.117
pH (KCl) A	-0.287	0.110	0.071	**-0.847**	-0.138	-0.124
K A	0.313	-0.098	0.315	0.389	0.393	**0.631**
Ca A	0.378	-0.065	0.269	0.503	0.325	**0.561**
Mg A	0.204	-0.110	0.418	0.495	0.196	**0.618**
Na A	0.090	**-0.607**	0.170	-0.283	0.375	0.231
Ca AB	**0.921**	-0.161	0.154	0.146	0.006	-0.017
Mg AB	**0.753**	0.173	0.380	0.160	0.203	0.101
Na AB	-0.084	0.003	-0.356	**-0.641**	0.165	0.003
K AB	**0.846**	0.027	0.211	0.070	0.065	0.082
pH (H2O) AB	**-0.755**	-0.330	0.133	-0.121	-0.260	0.068
pH (KCl) AB	**-0.681**	0.300	0.418	-0.058	-0.102	0.214
Ca Bt	**0.882**	-0.156	0.241	0.152	0.065	0.104
Mg Bt	**0.791**	0.098	0.288	0.223	0.091	0.240
Na Bt	0.183	0.176	0.021	0.109	**0.672**	0.167
K Bt	**0.664**	-0.034	0.429	0.186	0.213	0.301
Thickness A	0.425	-0.075	0.172	0.373	**0.577**	0.174
Thickness Bt	-0.180	**-0.759**	0.352	0.092	0.029	0.074
Color A	0.056	-0.018	-0.039	0.321	**-0.668**	-0.133
Color AB	0.114	**0.673**	-0.016	-0.294	0.327	-0.243
Color Bt	0.073	**0.743**	0.115	-0.123	0.417	0.157
Extractable acidity A	-0.278	-0.218	-0.178	-0.169	**-0.681**	0.327
Extractable acidity Bt	0.339	**-0.789**	-0.107	0.085	-0.059	-0.146
Silt A	-0.267	0.312	**-0.756**	-0.018	-0.047	-0.123
Clay A	0.179	0.000	**0.828**	0.052	-0.047	0.242
Silt Bt	-0.170	0.183	**-0.823**	-0.087	-0.241	-0.031
Clay Bt	0.207	0.283	**0.773**	0.055	0.130	0.262
Variance extracted (%)	21.6	11.51	14.4	11.4	9.7	9.5

Table 7. Rotated factor loading matrix of final PCA result. Primary loadings are in bold type. This was the twelfth stage.

VARIABLE	FACTOR 1	2	3	4
pH (H2O) A	-0.164	**-0.837**	0.083	0.176
pH (KCl) A	-0.272	**-0.862**	0.078	0.123
K A	0.225	**0.688**	0.529	0.042
Ca A	0.296	**0.753**	0.446	0.036
Mg A	0.102	**0.727**	0.577	-0.047
Ca AB	**0.906**	0.168	0.201	-0.136
Mg AB	**0.710**	0.251	0.438	0.241
K AB	**0.811**	0.147	0.288	0.060
pH (H2O) AB	**-0.776**	-0.160	0.076	-0.386
pH (KCl) AB	**-0.749**	-0.030	0.377	0.264
Ca Bt	**0.847**	0.228	0.330	-0.117
Mg Bt	**0.718**	0.356	0.389	0.140
K Bt	**0.602**	0.333	0.561	0.036
Thickness Bt	-0.175	0.092	0.393	**-0.719**
Color AB	0.167	-0.293	-0.077	**0.763**
Color Bt	0.047	0.046	0.167	**0.871**
Extractable acidity Bt	0.398	0.019	-0.084	**-0.778**
Clay A	0.092	0.091	**0.849**	0.005
Silt A	-0.227	-0.037	**-0.789**	0.275
Clay Bt	0.124	0.144	**0.807**	0.323
Silt Bt	-0.167	-0.064	**-0.815**	0.111
Variance extracted (%)	25.307	17.009	22.473	14.472

Table 8. Names of, and amounts of variance extracted by, the factors of the final Principal Component Analysis solution.

FACTOR NUMBER	FACTOR NAME	VARIANCE EXTRACTED (%)
1	Subsurface chemistry	25.3
4	Drainage and thickness	14.5
3	Soil texture	22.5
2	A horizon	17.0

Table 9. Comparison of variable retention, variable loading, and variance extracted among three factor analysis methods (SAS) and principal component (SYSTAT). The number represents the factor upon which the variable loaded. The dashed line indicates that the variable was not retained by the analytical procedure.

	STATISTICAL PROCEDURE			
Variable	Maximum Likelihood	Image	Minimum Residuals	Principal Component
A HORIZON				
Organic matter	2	-	2	-
pH (H_2O)	2	2	2	2
pH (KCl)	2	2	2	2
Ca	2	2	2	2
Mg	2	2	2	2
K	2	2	2	2
Na	-	-	-	-
Thickness	2	-	2	-
Color	-	-	-	-
Extractable acidity	-	-	-	-
Silt	3	2	3	3
Clay	2	2	3	3
AB HORIZON				
pH (H_2O)	1	1	1	1
pH (KCl)	1	1	1	1
Ca	1	1	1	1
Mg	1	1	1	1
K	1	1	1	1
Na	-	-	-	-
Thickness	-	-	-	-
Color	4	4	4	4
Extractable acidity	-	-	-	-
Bt HORIZON				
pH (H_2O)	3	-	3	-
pH (KCl)	3	3	3	-
Ca	1	1	1	1
Mg	1	1	1	1
K	1	1	1	1
Na	-	-	-	-
Thickness	4	4	4	4
Color	4	4	4	4
Extractable acidity	4	4	4	4
Silt	3	3	3	3
Clay	3	3	3	3
Total variables retained	25	22	25	21
Percent variance extracted	71.3	79.2	71.7	79.3