Factor Analysis for Soil Test Data: A Methodological Approach in Environment Friendly Soil Fertility Management

Hari Dahal, Ph D[2]

Abstract

Soil test data were used in factor analysis employing the Principal Component Analysis technique for the reduction and summarization of soil variables. Principal component analysis was found to be highly suggestive in analyzing soil test data on which a rational fertilizer nutrients recommendation can be made for a sustainable soil fertility management reign.

Introduction

"No single resource is more important to achieving a sustainable agriculture than the soil which contains essential nutrients, stores the water for plant growth and provides the medium in which plants grow" (TAC/CGIAR, 1989).

The above statement shows in itself that soil resource management is the key to environmental friendly and sustainable agricultural system. Soil is natural resource that provides essential nutrients to crop growth, needs proper care, conservation and management in order to maintain a high degree of soil fertility systems. One of the ways to assess the soil fertility status is to get soil sample tested for different soil nutrient variables.

It is a common practice in soil testing that the test data are interpreted against a soil analysis rating chart (Annex 1). The chart gives ratings of values such as *very low, low, medium, high* and *very high* based on which recommendation of nutrients are made. Soil test data however as such have little to say about the rate at which fertilizer nutrients are recommended without calibrating into different soil fertility indices (Dahal, 1996). Besides, the rating itself can not give over all summarization of soil variables in terms of their individual contribution to soil fertility status in the enumeration units.

When a large number of variables are correlated redundancy of information due to the set of dependent variables may occur and in some cases even give rise to multicollinearity thus making the results difficult for interpretation. There is however technique in multivariate analysis in which variables are not classified as *dependent* or *independent* but the whole set of interdependent relationships are investigated. One of the most important techniques of interdependence relationship is *factor analysis*.

Objective

The purpose of this paper is to show how a number of soil variables could be reduced to a smaller number of dimensions by employing factor analysis so as to make an easier interpretation of the problems.

Factor Analysis

Factor analysis is a generic name denoting a family of statistical techniques primarily concerned with the reduction and summarization of observed variables in terms of common underlying dimensions or factors. The main objective of factor analysis is to obtain a way of condensing the information contained in a number of original variables into a smaller set of

[2] Joint Secretary, Gender Equity and Environment Division, Ministry of Agriculture and Cooperatives, Singh Durbar, Kathmandu. Email: drdahal_h@yahoo.com

variates (factor) with a minimum loss of information (Hair et al, 2003). Factor analysis that includes both principal component and common factor analyses is an extremely powerful analytical technique and can often indicate which variables in a set of data are important and which are having little significance.

Principal Component Analysis (PCA)

As is said earlier factor analysis includes both common factor analysis and principal component analysis and is functionally very similar in that they are used for the same purpose of data reduction but are quite different in their underlying assumptions. Principal component analysis assumes that the *total variance* of the observed variables should be used in the analysis while in common factor analysis (CFA) only the *common variance* is considered (Hair et al, 2003; George and Mallery, 2006). Unlike CFA, principal component analysis makes no assumption of a model; it is a mathematical linear transformation of the original variables, the objective of which is to account for the maximum share of the variability present in the original set of variables with a minimum number of composite variables known as *principal components*. Principal component analysis technique is far more common than common factor analysis particularly after the advent of high-speed computers. It is however to note that *factors* and *components* have been used interchangeably in factor analysis.

What technique to select is based mainly on the purpose of the analytical work? If it is just to reduce a large set of observed variables to a smaller set of uncorrelated variables then the use of PCA is appropriate. When the intention is to identify latent variables as to model some meaningful underlying constructs CFA is the suitable technique. For the purpose of this study principal component analysis as processed and analyzed in SPSS package will be illustrated here.

PCA is one of the multivariate methods of data analysis that transforms a number of correlated variables into smaller set of uncorrelated variables called principal components while maintaining most of the information in the original variables. Although a complex mathematical procedure, principal component can be expressed as:

$$PC_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1k}x_k$$
$$PC_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2k}x_k$$
$$\ldots\ldots\ldots \quad \ldots\ldots\ldots \quad \ldots\ldots\ldots$$
$$PC_k = a_{k1}x_1 + a_{k2}x_2 + ... + a_{kk}x_k$$

Subject to $a^2{}_{11} + a^2{}_{12} + ... + a^2{}_{kk} = \max imum$.

Where, PCs are principal components, which are uncorrelated (orthogonal) with each other and also known as eigenvectors. The coefficients a_{11}, a_{12}, a_{1k} etc are assigned to the original *Xs* variables. The principal components are formed in decreasing order of importance which means the first principal component (PC_1) accounts for the maximum variance, PC_2 has the next maximum variance and so on.

While performing the PCA if the variables are not all in the same units they should be standardized so that;

$z_j = \frac{(x_j - \bar{x}_j)^2}{s_j}$, Where z_j is the z-scores for the j^{th} variable, x_j is the observed j^{th} variable and $\bar{x}_j$ is the mean of the j^{th} observed variables which is calculated as,

$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$, Where j = 1, 2,.....k and *n* is sample and s_j is the standard deviation of the j^{th} variate. The Pearson product moment formula is used for obtaining the correlation coefficients. If the correlation between 1 and 2 variables are worked out then;

$$r_{12} = \frac{\sum_{i=1}^{N} (z_i 1 \times z_i 2)}{N}$$

Where, z_i1 and z_i2 are the z scores for sample i on variables 1 and 2.
N = sample size

Methodology

The soil test data for the principal component analysis were included from Kathmandu, Lalitpur and a Terai district of Banke for the fiscal year 2004-05. The samples were analyzed for organic matter (%), total nitrogen (%), phosphorus (kg/ha), potassium (kg/ha) and pH (rating scale) in the soil laboratory, Soil Management Directorate of the Department of Agriculture, Nepal.

Results and Discussions

Analysis for Kathmandu soil variables from a total sample size of 50 was performed first to see the descriptive statistics and the whole factor analysis using PCA method in SPSS.10. The output is given in Table 1

Table 1 Factor Analysis Using Kathmandu Soil Test Data

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PH	50	3.80	7.30	5.3220	.8869
OM	50	1.10	8.07	3.4481	1.8612
N	50	.05	.40	.1724	9.351E-02
P205	50	4.39	263.63	89.5166	69.8746
K20	50	90.57	507.40	188.6806	97.3882
Valid N (listwise)	50				

Correlation Matrix [a]

		PH	OM	N	P205	K20
Correlation	PH	1.000	-.614	-.616	.181	-.292
	OM	-.614	1.000	1.000	-.144	.403
	N	-.616	1.000	1.000	-.140	.399
	P205	.181	-.144	-.140	1.000	-.336
	K20	-.292	.403	.399	-.336	1.000
Sig. (1-tailed)	PH		.000	.000	.104	.020
	OM	.000		.000	.159	.002
	N	.000	.000		.167	.002
	P205	.104	.159	.167		.009
	K20	.020	.002	.002	.009	

a. Determinant = 3.862E-04

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.666
Bartlett's Test of Sphericity	Approx. Chi-Square	365.449
	df	10
	Sig.	.000

Communalities

	Initial	Extraction
PH	1.000	.607
OM	1.000	.937
N	1.000	.938
P205	1.000	.810
K20	1.000	.600

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			action Sums of Squared Loadi			tation Sums of Squared Loadir		
Compone	Total	of Varianc	umulative %	Total	of Varianc	umulative %	Total	of Varianc	umulative %
1	2.811	56.216	56.216	2.811	56.216	56.216	2.556	51.130	51.130
2	1.082	21.642	77.857	1.082	21.642	77.857	1.336	26.728	77.857
3	.634	12.689	90.547						
4	.472	9.445	99.992						
5	38E-04	3.476E-03	100.000						

Extraction Method: Principal Component Analysis.

Component Matrix [a]

	Component	
	1	2
PH	-.766	-.141
OM	.936	.246
N	.935	.252
P205	-.334	.836
K20	.600	-.490

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix [a]

	Component	
	1	2
PH	-.762	.164
OM	.959	-.132
N	.960	-.126
P205	1.164E-02	.900
K20	.366	-.683

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

The *descriptive statistics* revealed the average values of soil attributes as acidic pH, and a medium level of both organic matter and nitrogen. The phosphorus content of the soils was high with medium level of potash in the soils.

As to *factor analysis*, there is substantial number of variables correlated among each other. The Bartlett's Test of Sphericity that is a test statistic and used to test the null hypothesis that variables are uncorrelated to each other. This null hypothesis is rejected with the approximate chi-square value of 365.449 at 10 degrees of freedom. Kaiser-Meyer-Olkin, which is the measure of sampling adequacy (MSA), is found to be 0.666 suggesting the factor analysis an appropriate method to analyze the correlation matrix. KMO measure of sampling adequacy varies between 0 and 1 and the values closer to 1 are better whereas below 0.50 is unacceptable. The KMO value can be increased in many ways such as-increasing the sample size or increasing the number of variables. If the average correlations among the variables are high or the numbers of factor is diminished the value of KMO becomes large (Hair et al, 2003).

Interpretation of Results

Communalities as shown in Table 1 measure the amount of variance a variable shares with all the other variables in the analysis. This is also the proportion of each variable's variance explained by the principal components. It is also noted that the communality (h^2) can be defined as the sum of squared factor (component) loadings. Large communality means a large amount of the variance in a variable is extracted by the factor solution. In other words, variables with high values are well represented in the common factor space while low value variables are not well represented (Malhotra N K, 2004). Communality can be calculated by squaring each component loading, adding and then multiplying by 100.

$$h^2 = (pc_1)^2 + (pc_2)^2 + ... = (pc_k)^2$$

Where, h^2 = communality
pc = principal component

Initial Eigenvalues

The eigenvalue is the variance explained by a component or factor and is denoted by lambda (λ).

$$\lambda_k = \sum_{i=1}^{n} A^2{}_{ik}$$

Where, A_{ik} is the factor loading for variable *i* on component *k* and *n* is the number of variables. A low eigenvalue contribute little to the explanation of variances in the set of variables being analyzed. The initial eigenvalues for the first and second components are 2.811 and 1.082 respectively (Table1). The percentage variance thus explained by each of the two components is about 56 and 22 with a cumulative percentage of 78. In rotated sums of squared loading the eigenvalues have been more evenly distributed, 2.556 for the first component and 1.336 for the second leaving the total amount of variance the same. With the varimax rotation the interpretation of the component matrix has been simplified. The sum of the eigenvalues, as expected is equal to the number of variables being analyzed.

Components Loadings

The component matrix table shows the components loadings that are the correlations between the variables and the components. This is the central output of factor or principal component analysis, which is also the basis for imputing a label to the different factors of components (Nargunkar, 2005). It is the rule of thumb that larger the size of the component loading for a variable, the more important the variable is in interpreting the component. The first component is generally more highly correlated with the variables than the second components and so on. In interpreting, loadings above 0.6 are considered *high* where as those below 0.4 are *low*.

In the above example, the first table *component matrix* gives the unrotated solution and the second the rotated solution, which is the main basis for component interpretation. Looking at the *rotated component matrix* the first component has high loadings for three variables - pH, organic matter, and nitrogen.

In second component, P_2O_5 is strongly associated and K_2O is high but negatively correlated. Although there are many criteria to select components from the output analysis only two components were selected based on the Kaiser criterion. The Kaiser rule is to select those components or factors, which have eigenvalues *greater* than one (Gaur and Gaur, 2006).

NAMING THE COMPONENTS

In component 1, nitrogen (N) and organic matter (OM) have very high loadings and both of which have positive signs. Thus nitrogen and organic matter vary together while pH having negative sign move in opposite direction. It means increased in soil pH is associated with the decreased in the availability of nitrogen and organic matter in soil systems.

It is however to be noted that the optimum range of soil pH where most of the plant nutrients including nitrogen are available at the vicinity of 6.5 and 7.5 which is near to neutrality zone. Since soil pH is a reaction, a condition of nutrients availability the principal component 1 is represented by nitrogen and organic matter, which have very high loadings and are positively associated. The first component therefore can be named as "*growth component*" a factor responsible for vegetative growth and succulence in crops. The second component is primarily represented by P_2O_5 and in lesser extent by K_2O that is negatively associated to the component. Both of these variables vary in opposite direction but have similar functions in crop production. The main responsibility of these variables includes vigor and resistance in the crops and can be named as *"product quality component"*. It is therefore the components associated to soil fertility maintenance and crop growths in Kathmandu soils are identified as *growth* and *product quality* components out of five original variables involved in the principal component analysis.

LALITPUR DATA (N = 46)

The average soil reaction, organic matter, and nitrogen contents were similar to Kathmandu soils except that P_2O_5 and K_2O contents were very high in the soils. Since the KMO measure of sampling adequacy was acceptable level (0.588) and the Bartlett's test of sphericity was significant, the data were put into factor analysis. Two components were identified based on Kaiser criterion (Annex Table 2). The rotated component matrix shows all nutrient components except pH were associated with the first component explaining the variance of about 49 percent where as the second component shares about 21 percent. The first component may be named as *"primary nutritional factor"* and the second component as *"soil reaction factor"* together explaining the soil fertility variation by 70 percent.

KATHMANDU VALLEY DATA (N = 105)

Since the total amount of variance accounted for by the components in the above examples did not cross 80 percent it is thought the sample size in both cases to be inadequate. The data of both districts therefore were merged and adding nine more cases the sample size of 105 was made to represent the coterminous valley.

Table 2. Factor Analysis Using Valley Data

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PH	105	3.00	8.20	5.3210	.9447
OM	105	1.10	8.07	3.5932	1.6854
N	105	.05	.40	.1798	8.456E-02
P205	105	1.32	1843.79	167.3671	259.5458
K20	105	89.14	3857.63	470.9298	605.3024
Valid N (listwise)	105				

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.607
Bartlett's Test of Sphericity	Approx. Chi-Square	839.801
	df	10
	Sig.	.000

Communalities

	Initial	Extraction
PH	1.000	.660
OM	1.000	.967
N	1.000	.967
P205	1.000	.780
K20	1.000	.815

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.665	53.301	53.301	2.665	53.301	53.301	2.213	44.256	44.256
2	1.525	30.500	83.801	1.525	30.500	83.801	1.977	39.545	83.801
3	.579	11.582	95.383						
4	.230	4.608	99.991						
5	4.704E-04	9.409E-03	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix [a]

	Component	
	1	2
PH	-.269	.767
OM	.978	.108
N	.978	.108
P205	.286	.836
K20	.383	.817

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

The descriptive statistics as shown in Table 2 revealed the soils as acidic, with a medium levels of organic matter and nitrogen content, very high level of P_2O_5 and high level of K_2O per unit of area. Of the 10 correlations 8 were significant and KMO and Bartlett's test were in acceptable levels. With the application of latent root criterion two components were retained explaining a total of about 84 % variance. The VARIMAX rotated components matrix revealed very high loadings of organic matter and nitrogen to the first component responsible for 44 percent of the total variance. The second component accounts for about 40 percent of the variance has high loadings of P_2O_5, K_2O and pH. As of Kathmandu soils the first component may be called as *"growth component"* and second *"product quality component"* with a positive pH value.

BANKE DATA (N = 70)

As of now the analysis was limited to Kathmandu valley soils, it was interesting to see if the analysis was markedly different for a low land Terai district of Banke. The measure of sampling adequacy (MSA) was miserably acceptable (0.524) with a significant Bartlett's test of sphericity. The descriptive statistics revealed acidic soils with low levels of both organic matter and nitrogen. The P_2O_5 content was very high while the K_2O was medium.

Table 3 Showing the Factor Analysis Output for Banke Soil Data

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PH	70	4.80	7.50	5.5357	.4553
OM	70	.59	4.08	2.2103	.6715
N	70	.03	.20	.1099	3.343E-02
P205	70	3.14	370.27	116.0833	78.1732
K20	70	104.79	456.51	274.1207	81.4031
Valid N (listwise)	70				

Factor Analysis

Correlation Matrix [a]

		PH	OM	N	P205	K20
Correlation	PH	1.000	.060	.051	-.416	-.173
	OM	.060	1.000	.997	.104	.181
	N	.051	.997	1.000	.115	.179
	P205	-.416	.104	.115	1.000	.396
	K20	-.173	.181	.179	.396	1.000
Sig. (1-tailed)	PH		.311	.338	.000	.076
	OM	.311		.000	.195	.067
	N	.338	.000		.172	.069
	P205	.000	.195	.172		.000
	K20	.076	.067	.069	.000	

a. Determinant = 3.833E-03

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.524
Bartlett's Test of Sphericity	Approx. Chi-Square	370.006
	df	10
	Sig.	.000

Communalities

	Initial	Extraction
PH	1.000	.562
OM	1.000	.984
N	1.000	.982
P205	1.000	.709
K20	1.000	.476

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.123	42.463	42.463	2.123	42.463	42.463	2.047	40.934	40.934
2	1.590	31.808	74.270	1.590	31.808	74.270	1.667	33.336	74.270
3	.780	15.592	89.863						
4	.504	10.079	99.942						
5	2.889E-03	5.779E-02	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix [a]

	Component	
	1	2
PH	.167	-.731
OM	.991	4.468E-02
N	.990	5.357E-02
P205	7.693E-02	.838
K20	.226	.652

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Two components were extracted explaining about 74 percent variation (Table 3). Organic matter and nitrogen have very high loadings with the first component accounting for 41 percent of variance while 33 percent was accounted for second component in which P_2O_5, K_2O and pH have high loadings. As usual the first component may be termed as *"growth component"* because of its contribution in vegetative growth, chlorophyll and protein formation in the crops. The second component again may be called as *"quality product component"*. The negative sign for pH in the second component indicates the reverse relationship between pH and other two variables. It is important to note that as the pH decreases the deficiency of P_2O_5 and K_2O becomes pronounced.

VALIDATING COMPONENT ANALYSIS

Although PCA was highly meaningful in analyzing important soil variables in a set of test data the degree of generalizability is a critical issue. One of the important questions is to obtain suitable data in terms of number of variables and sample size. It has been observed that the sample size fewer than 50 is not suitable for analysis. In general a *ten-to-one ratio* between the number of observations and variable is said to be more appropriate. The analyses in the preceding paragraphs have also been constrained in the sense that the sample size for each of the districts was not sufficiently large. The number of variables were also limited to major nutrients only whereas adding cation exchange capacity (CEC), textural classes and micronutrients to the analysis could have given better results to interpret the components for soil fertility management. The generalization of these findings through the principal

component analysis therefore should be done cautiously. Furthermore, the samples were tested as and when received and were not randomly selected to represent the population.

Conclusions

As expected principle component analysis has been shown to be a useful technique in reduction and summarization of soil variables. Among all five variables nitrogen (N) was the most important soil nutrient followed by organic matter (OM). Since nitrogen and organic matter are strongly associated number one component was highly loaded by these two variables in almost all cases. If organic matter is not considered as a nutrient, the second most important nutrient was Phosphorus. In most cases P_2O_5 and K_2O with pH have high loadings with the second component. Although K_2O is a third important nutrient, its loading in general was next to pH. Its means in order to maintain soil fertility, balancing pH was more important than adding more of K_2O fertilizer as its average levels was already medium to very high range in the soils.

Acknowledgement

The contribution in supplying soil test data for the purpose of this paper by the Chief and staffs of the Directorate of Soil Management, Department of Agriculture is gratefully acknowledged.

REFERENCES

1. Dahal, H: ECOLOGICAL APPROACH TO SUSTAINABLE AGRICULTURE THROUGH INTEGRATED NUTRIENT RESOURCE MANAGEMENT: A MICRO-LEVEL STUDY IN THE EASTERN TERAI FARMING SYSTEM, NEPAL. A Ph D Dissertation Submitted to the School of Environment, Resources and Development, Asian Institute of Technology (AIT), Bangkok, Thailand, April, 1996.
2. George D and P Mallery, 2006: *SPSS for Windows- Step by Step*. Pearson Education, Darling Kindersley (India).
3. Gaur A S and S S Gaur, 2006: *Statistical Methods for Practice and Research*. Response Books - A Division of SAGE Publication, New Delhi, India.
4. Hair J F, R E Anderson, R L Tatham and W C Black, 2003: *Multivariate Data Analysis*. Pearson Education, Singapore and India.
5. Malhotra N K, 2004: *Marketing Research - an applied orientation*. Pearson Education, Singapore and India.
6. Nargundkar R, 2005: *Marketing Research-text and cases*. Tata McGraw Hill Publishing Company Ltd, New Delhi, India.
7. TAC/CGIAR, 1989: *Sustainable Agricultural Production: Implication for International Agricultural Research*. Food and Agriculture Organization (FAO), Rome.

ANNEX

Annex Table 1 Soil Analysis Rating Chart

Rating	Nitrogen %	Organic Matter %	Phosphorus kg/ha	Potassium kg/ha
Very Low	<0.05	<1.0	<10	<55
Low	0.05-0.10	1.0-2.5	10-30	55-110
Medium	0.10-0.20	2.5-5.0	30-55	110-280
High	0.20-0.40	5.0-10.0	55-110	280-500
Very High	>0.4	>10.0	>110	>500

Source: Ministry of Agriculture and Cooperatives (Fertilizer Unit)/HMG/Nepal

Table 2 Factor Analysis Using Lalitpur Data

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PH	46	3.00	7.00	5.1022	.7446
OM	46	1.33	6.39	3.4123	1.2567
N	46	.07	.32	.1707	6.288E-02
P205	46	1.32	556.48	166.4380	140.1981
K20	46	89.14	1987.76	589.7039	442.5319
Valid N (listwise)	46				

Factor Analysis

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.489	49.782	49.782	2.489	49.782	49.782	2.473	49.461	49.461
2	1.025	20.505	70.287	1.025	20.505	70.287	1.041	20.826	70.287
3	.971	19.416	89.703						
4	.514	10.282	99.985						
5	7.473E-04	1.495E-02	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix [a]

	Component	
	1	2
PH	5.306E-02	.974
OM	.915	7.301E-02
N	.915	7.575E-02
P205	.629	-.228
K20	.631	.171

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.