Proceedings of the

SOIL SCIENCE CONFERENCE OF MALAYSIA 1995

B.Y. Aminuddin, A.B. Ismail,
A.R. Ahmad and M.Z. Ghazali
(Editors)

Malaysian Society of Soil Science

Proceedings of the

SOIL SCIENCE CONFERENCE OF MALAYSIA 1995

B.Y. Aminuddin, A.B. Ismail, A.R. Ahmad and M.Z. Ghazali (Editors)

Published by:
MALAYSIAN SOCIETY OF SOIL SCIENCE
P.O. BOX 12644, 50784 KUALA LUMPUR, MALAYSIA
1996

ALL RIGHT RESERVED
Copyright 1996 by Malaysian Society of Soil Science. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form by any means, electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the copyright owner.

Published by:

Malaysian Society of Soil Science
P.O. Box 12644, 50784 Kuala Lumpur, Malaysia

Perpustakaan Negara Malaysia Cataloguing-in-Publication Data

Proceedings of the Soil Science Conference of Malaysia 1995 / B.Y. Aminuddin ... [et al.].
ISBN 967-9945-17-0
1. Soil science--Malaysia--Congresses. 2. Soil management--Malaysia--Congresses. 3. Soil fertility--Malaysia--Congresses. 4. Soil conservation --Malaysia--Congresses. 5. Plants--Nutrition--Congresses. I. Aminuddin B.Y. (Aminuddin bin Yusoff). II. Soil Science Conference of Malaysia (1995 : Langkawi).
631.409595

Suggested citation:

Proc. Soil Science Conference of Malaysia 1995, eds. B.Y. Aminuddin *et al.*. Kuala Lumpur. Malaysian Society of Soil Science.

Soil Science Conference of Malaysia 1995, Langkawi

Using Factor Analysis in Soil Science: A Tool to Summarize Information

T.B.S. Christopher and A.M. Mokhtaruddin
Soil Science Department, UPM, 43400 Serdang

Key words: Factor analysis, multivariate, statistics, summarize information, data reduction, tilth, structure, consistency, soil quality, index

Abstract

Some soil qualities are quantified or evaluated by using many indices or by measuring many soil properties because these soil qualities are influenced by many interacting soil factors. The information produced from such a variety of measurements can be summarized using the multivariate statistical tool called factor analysis. Factor analysis simplifies complex sets of data by representing them with fewer variables called common factors. This paper is to explain the fundamentals of factor analysis, its use and its interpretation. To illustrate factor analysis, secondary data was used to derive a single measure of soil tilth from the measurement of several soil properties. Data of five variables: humus, porosity (PORE), bulk density (BD), sand and clay content of twenty soil samples were re-analysed with the statistical program SPSS. Factor analysis summarized the information contained in those five variables into two common factors: the structure and consistency factors. The variables associated to the structure factor were humus, PORE and BD; and to the consistency factor were sand and clay. Thus, the measure of soil tilth is: Tilth index = structure index - |consistency index| because the better the soil structure the better the soil tilth, but extreme values of consistency are detrimental to soil tilth.

Abstrak

Sesetengah kualiti tanah dinilai dengan menggunakan pelbagai indeks atau dengan mengukurkan pelbagai sifat tanah kerana kualiti-kualiti tanah tersebut dipengaruhi oleh banyak faktor tanah yang saling berinteraksi. Maklumat yang terhasil akibat daripada pelbagai ukuran itu boleh diringkas dengan suatu teknik statistik multivariat dikenali sebagai analisis faktor. Analisis faktor memudahkan interpretasi set-set data yang kompleks dengan mewakili set-set data ini dengan kurang bilangan pembolehubah dipanggil faktor sepunya. Kertas ini bertujuan untuk menerangkan asas-asas analisis faktor, penggunaannya dan interpretasinya. Untuk menjelaskan analisis faktor, data sekunder dipakai untuk menghasilkan satu pengukuran kemudahan pembajakan tanah daripada ukuran-ukuran pelbagai sifat tanah. Data lima pembolehubah: humus, porositi (PORE), ketumpatan pukal (BD), pasir dan lempung daripada dua puluh sampel tanah dianalisa semula dengan program statistik SPSS. Analisis faktor meringkaskan maklumat yang terkandung dalam lima pembolehubah tersebut kepada dua faktor sepunya: faktor struktur dan konsistensi. Pembolehubah-pembolehubah yang berkait dengan faktor struktur adalah humus, PORE dan BD; dan dengan faktor konsistensi adalah pasir dan lempung. Maka, pengukuran kemudahan pembajakan tanah adalah: Indeks kemudahan pembajakan tanah = indeks struktur - |indeks konsistensi| kerana lebih baik struktur tanah, lebih mudah pembajakan tanah, tetapi nilai-nilai ekstrem konsistensi menyusahkan pembajakan tanah.

Introduction

There are certain soil qualities such as soil fertility, soil tilth and aggregate stability that cannot be directly observed or measured, or be adequately represented by just a single index or measurement method. These soil qualities are such because they are influenced by many factors that interact with one another with varying degrees or importance. Consequently, to quantify or to evaluate this soil quality, what is usually done is to use various indices, or to measure most of the influential soil properties—with the belief that using this many indices or measuring this many soil properties would characterize the soil quality more thoroughly or fully.

Reasonable this belief may be, however, using several indices are likely to produce contradictory ranking or evaluation of the soil quality among the soil samples. Measuring many soil properties, on the other hand, raises the question on how to combine the various soil properties into a single variable representative of the soil quality. Furthermore, the interpretation of data is further burdened because of these many indices or soil properties.

These problems, however, can be solved using factor analysis. Factor analysis, which is a multivariate (*i.e.*, multi*variables*) statistical technique, solves these problems by summarizing or simplifying complex sets of data by determining if there is a smaller number of variables called *factors* that can represent or explain this sets of data. Consequently, instead of dealing with many variables, one is now dealing with fewer and more manageable (yet representative) factors.

Despite its potential, factor analysis remains one of the least understood and used because of its advanced mathematics and algorithms. To the non-specialists, textbooks and journal papers about factor analysis assume too much of a reader's background; therefore they are difficult and become unreadable. Thus, our objective is to explain, as clearly and plainly as possible and with minimal use of mathematics, the fundamentals of factor analysis, how to use it, and how to interpret its results. Our purpose here is not to explain comprehensively about factor analysis but to introduce and yet to explain enough so to overcome the steep learning curve of factor analysis. To avoid discussion of algorithms, the use of computers is stressed.

Fundamentals of Factor Analysis: The Factor Model

The goal of factor analysis is to summarize the information resulted from using many variables into fewer variables called *factors*. Factor analysis is based on the principle that the behaviour (covariance) of the measured or observed variables is influenced by a smaller number of *common factors* (Kim & Mueller, 1978).

In Figure 1, for example, the behaviour (covariance) of the observed variables X_1 and X_2 are both influenced by the common factor F. Likewise, X_1 is also influenced by the factor U_1, and X_2 by the factor U_2. Because the influence of U_1 and U_2 are exclusive to their respective observed variable, they are called *unique factors*. And in algebraic form: $X_1 = b_1F + d_1U_1$ and $X_2 = b_2F + d_2U_2$ with b_1, b_2, d_1 and d_2 being the linear weights or degree of association. These weights are the correlations between factors and variables only when a single common factor is involved, or when the multiple

common factors are uncorrelated to each another. Furthermore, this model shows that there is no covariation between F and U_1, F and U_2, and between U_1 and U_2. Therefore, the covariation between X_1 and X_2 is *completely* determined by F; without this common factor, X_1 and X_2 will cease to correlate with each other.

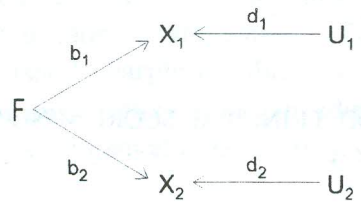

Figure 1. A two-variable, one common factor model

In summary, factor analysis (from Figure 1) has revealed that a single common factor is enough to explain the behaviour (covariance) of the two observed variables because these variables, although different, have a mutual focus or emphasis (represented by F). And because this common factor alone cannot account for all the variance in the observed variables (since the two variables are different), the unique factors must account for the rest. In short, the factor model cleaves each variable into two parts: a part due to the common factor (called *communality*) and a part due to the unique factor.

Factor analysis therefore determines: (a) the number of common factors responsible, or that can explain the behaviour (covariance) of the observed variables, and (b) the degree each observed variable is influenced by these common factors. Unique factors, being of little scientific interest, are usually disregarded.

Materials and Method

To best illustrate the use and interpretation factor analysis, secondary data was taken from the work of Suh *et al.* (1977). They used factor analysis to evaluate the soil tilth of various soils by deriving a numerical measure of soil tilth. This single tilth index was derived (or *summarized*, in factor analytic term) from the various soil properties important to tilth.

However, the work of Suh *et al.* (1977), like others, demands too much of a beginner. For clarity and brevity, an excerpt of data was selected and re-analyzed. Data of five soil properties (variables): total porosity (PORE), humus content, bulk density (BD), sand and clay content of twenty soil samples were factor analyzed with the statistical program SPSS. Note that other statistical programs such as SAS could also be used. Before factor analysis, all variables were standardized (having means of zero and variances of one) by subtracting the mean from the values and dividing the result by the square root of the variance, *i.e.*,

$$\frac{x_i - \bar{x}}{\sqrt{\text{variance } x}}$$

The full SPSS command syntax for factor analysis is:

```
FACTOR
  /VARIABLES <variable list> /MISSING LISTWISE /ANALYSIS <variable list>
  /PRINT INITIAL CORRELATION EXTRACTION ROTATION FSCORE
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /ROTATION VARIMAX
  /SAVE REG(ALL) .
```

The SAS command syntax is:

```
PROC FACTOR METHOD=PRINCIPAL SCORE MINEIGEN=1 ROTATE=VARIMAX;
  VAR <variable list>;
```

Results and Discussion

The various relationship groups or patterns in the correlation matrix (Table 1) were separated and analyzed by factor analysis to determine the common factors and their variances (Table 2). Several methods exist to extract or to determine the common factors and their variances. These methods, explained in Kline (1994) and Rummel (1970), are not covered here so to maintain the focus and simplicity of this paper. The method used here, as also by Suh *et al.* (1977), was the Principal Components extraction method. This method, however, is peculiar to the rest of the methods, but the appeal of this extraction method is it produces uncorrelated factors. Also, it can be shown mathematically that this method follows the factor model when dealing with standardized variables and when the not all the extracted factors are used (Afifi & Clark, 1984).

Table 1. Correlation coefficient matrix between all pairs of the five variables

	Humus	PORE	BD	Sand	Clay
Humus	1.00				
PORE	.72	1.00			
BD	-.80	-.93	1.00		
Sand	-.27	-.16	.25	1.00	
Clay	.12	.14	-.23	-.92	1.00

Table 2. Initial factors by the Principle Components extraction method *(eigenvalues greater than 1.0 are bolded)*

Factor	Eigenvalue, *e*	% variance (*e*/5 x 100%)	Cumulative % variance
F_1	2.88	57.7	57.7
F_2	1.68	33.7	91.4
F_3	0.33	6.5	97.9
F_4	0.07	1.4	99.3
F_5	0.04	0.7	100.0

In Table 2, the *eigenvalue* of each factor is how much variance it accounts for, and when divided by the number of variables, it indicates the proportion of variance the factor accounts for. To account for *all* the variance of the observed correlations in the correlation matrix, the number of factors will be equal to the number of variables. Therefore, in this example, five factors account for all the variance. However, not all of these factors were used; only the first two: F_1 and F_2. This is because of the large variance F_1 and F_2 accounted for (91.4%). The third factor F_3 and the subsequent factors should be disregarded because of the small variances they accounted for; also because their variances diminished sharply or abruptly as compared to the first two factors. In addition, a widely used criterion to consider the number of factors is to consider those factors having an eigenvalue equal to or greater than 1.0 (such as F_1 and F_2).

The values in Table 3 (except the eigenvalues and h^2) are the *factor loadings* or the correlation coefficient r between the factors and variables (r = 0.84 between humus and F_1). Also, 57.7% of the variance in F_1 is explained by the five variables. The communality (h^2) for each variable is the sum of squares of each factor. For example, 80% of the variance in the humus variable is explained by the two common factors F_1 and F_2 (*i.e.*, $0.84^2 + 0.31^2$). The higher the communality, the more these two factors can explain the variance of a variable.

Table 3. Factor matrix *(high correlations are bolded)*

	F_1	F_2	Commu-nality h^2
Humus	**0.84**	0.31	0.80
PORE	**0.86**	0.39	0.90
BD	**-0.92**	-0.31	0.95
Sand	-0.57	**0.80**	0.96
Clay	0.51	**-0.84**	0.96
Eigenvalue	2.88	1.68	Σh^2=4.57
% variance explained	57.7	33.7	91.4

Table 4. Factor matrix rotated with Varimax method *(high correlations are bolded)*

	F_1	F_2
Humus	**0.89**	0.11
PORE	**0.95**	0.05
BD	**-0.96**	-0.15
Sand	-0.14	**-0.97**
Clay	0.07	**0.98**

From Table 3, the variables humus, PORE and BD correlated closely with F_1 but not with F_2. On the contrary, the variables sand and clay correlated closely to F_2 but not with F_1.

To further ease the understanding of the relationship between the factors and variables, the factor matrix was *rotated* (Table 4). Rotation can be thought of as a rotation of a computer 3-D graph until its best or clearest perspective is viewed. Again, there are several methods to rotate, each attempting to produce the clearest view. But all of these methods, although different, essentially attempt to increase the "contrast", that is to maximize the correlations that are initially large and to minimize the correlations that are initially small. The rotation method used here, as also by Suh *et al.* (1977),

was the Varimax rotation which is suitable for uncorrelated factors (Kline, 1994; Afifi & Clark, 1984; Kim & Mueller, 1978; Rummel, 1970).
After rotation, it is now clearer that the variables humus, PORE and BD correlated closely to F_1 but lowly to F_2 (Table 4). Factor F_2, on the other hand, correlated closely to sand and clay but lowly to humus, PORE and BD.

At this point, however, it is insufficient just to conclude that the two factors F_1 and F_2 are enough to explain the five variables. These factors must also be identified. What F_1 and F_2 represent are revealed in their associations to the variables. As stated earlier, F_1 correlated closely to the first three variables but not to the last two (Table 3 and 4). Therefore, since low BD, high PORE and high humus content are usually associated with good soil structure, F_1 can be interpreted as the factor representing soil structure. F_2, however, correlated poorly to the first three variables but highly to the last two. F_2 can thus be interpreted as the factor representing soil consistency. Both soil structure and soil consistency, important factors of soil tilth, were incorporated into soil tilth by calculating the factor scores using the *regression procedure* (Rummel, 1970).

Factor score coefficients (Table 5) were calculated by factor analysis based on the interrelations among the variables and the factor loadings. For example, since the variables humus, PORE and BD correlated closely to each other, as well as to F_1, their scores on F_1 were higher than of the sand and clay variables which were close to zero (Table 5). Likewise, because the variables sand and clay correlated closely to each other and to F_2, their scores on F_2 were higher than of the other three variables. Also note that the scores of humus, PORE and BD on F_2 were close to zero. Therefore, a soil's factor score on structure index F_1 and on consistency index F_2 are:

$$\text{structure index} = 0.34\text{humus} + 0.37\text{PORE} - 0.37\text{BD} + 0.04\text{sand} - 0.07\text{clay}$$

and

$$\text{consistency index} = -0.03\text{humus} - 0.07\text{PORE} + 0.02\text{BD} - 0.51\text{sand} + 0.52\text{clay}$$

where standardized variables were used for both equations.

Table 5. Factor score coefficient matrix *(high scores are bolded)*

	F_1	F_2
Humus	0.34	-0.03
PORE	0.37	-0.07
BD	-0.37	0.02
Sand	0.04	-0.51
Clay	-0.07	0.52

Thus, from Table 6, soils with low BD, high PORE and high humus content had high values of structure index (soil A, B and F), but soils with high sand content had low values of consistency index (soil R and T).

The measure of soil tilth was calculated from F_1 and F_2 based on the rationale that the higher the structure index the better the soil tilth, but extreme values (large positive or negative values) of consistency such as in cohesive clayey soils and loose sandy soils are detrimental to soil tilth. Therefore the single numerical measure of soil tilth is:

Tilth index = structure index - | consistency index .

Table 6. Relation between the tilth index, factor scores and other soil properties

Soil sample	Measured variables					Factor scores F_1	F_2	Tilth index
	Humus	PORE	BD	Sand	Clay	structure index	consistency index	F_1-|F_2|
A	25.21	80.10	0.42	52.40	14.10	2.15	-0.72	1.44
B	13.65	76.30	0.61	51.10	28.00	1.09	-0.10	0.99
C	12.15	74.70	0.61	43.20	24.10	0.95	-0.08	0.88
D	12.21	75.40	0.61	53.10	26.70	0.99	-0.17	0.82
E	17.68	67.60	0.72	38.00	19.50	0.81	-0.11	0.70
F	15.95	73.20	0.60	25.20	33.20	1.03	0.58	0.45
G	2.57	78.80	0.56	38.10	33.80	0.65	0.37	0.28
H	11.27	76.80	0.60	29.00	42.20	0.90	0.83	0.07
I	10.89	67.00	0.78	34.40	31.00	0.30	0.41	-0.11
J	7.03	59.60	0.69	25.20	44.40	-0.16	1.13	-1.29
K	3.93	65.60	1.03	18.20	39.70	-0.52	1.08	-1.60
L	4.38	56.80	1.11	37.30	41.60	-0.95	0.87	-1.81
M	5.01	57.10	1.09	14.40	35.70	-0.88	1.07	-1.95
N	1.62	60.50	1.04	87.40	4.20	-0.58	-1.43	-2.01
O	1.48	61.00	1.03	94.80	3.10	-0.54	-1.61	-2.14
P	2.40	55.70	1.19	18.40	35.30	-1.20	1.01	-2.21
Q	2.64	54.20	1.21	86.00	8.50	-1.04	-1.19	-2.23
R	1.35	58.80	1.15	91.80	5.90	-0.82	-1.43	-2.24
S	2.34	53.40	1.22	20.00	37.30	-1.34	1.07	-2.41
T	1.05	58.30	1.16	92.30	1.50	-0.84	-1.59	-2.43

The larger the tilth index, the better or easier the soil tilth. Interestingly, each soil's factor scores (structure index, consistency index or tilth index) can be used for subsequent statistical analyses such as ANOVA, multiple mean comparisons and correlations.

Criteria to use factor analysis

The equation to compute the soil's factor score on structure index or consistency index appears very similar to that of multiple regression. However, the variables used in factor analysis must be dependent on each other. Strong relationships among the variables are encouraged in factor analysis but will cause *multicollinearity* in multiple regression; a situation that, among others, inflates the standard error.

Also, the number of samples used for factor analysis must be at least four to ten times larger than the number of variables used. To overcome cases of too many variables or too few samples, there are two possible methods that are briefly explained here. First, variables must be selected so that they contribute to the objective of the research. In the given example, the selected five variables are essential factors of soil tilth. Variables representing non-essential or too indirect factors should be disregarded. Second, if possible, the soil samples should not be bulked. Depending on the research purpose, some collected soil samples are usually mixed together or bulked. This bulk sample, representing one soil sample, is then analyzed with replications. To increase the number of soil samples, each of those collected soil samples should not be bulked but analyzed separately without replication.

Lastly, several statistical tests are available to determine if factor analysis should be used for a particular set of data, such as the Bartlett's sphericity test and the Kaiser-Meyer-Olkin sampling adequacy test (both tests are available in SPSS).

Conclusion

Factor analysis allows the quantification or evaluation of soil qualities that are otherwise inadequately or difficult to quantify because they are influenced by many interacting soil properties. In the given example of soil tilth, factor analysis revealed that the two factors, the structure and consistency factors, were enough to explain the five measured variables. The information contained in those five variables were summarized into two factors; thus those two factors adequately represent and may substitute the five variables. Also, with factor analysis, the variables associated to the structure factor were humus, PORE and BD; and with the consistency factor were sand and clay. From the rationale that the better the soil structure the better the soil tilth, but extreme values of consistency are detrimental to soil tilth, the numerical measure of soil tilth is:

Tilth index = structure index - | consistency index | .

Acknowledgement

We thank the National Council of Scientific Research and Development for the financial support, UPM for permission to present this paper, and the Malaysian Society of Soil Science for the grant to present this paper in the Soil Science Conference of Malaysia 1995 (Langkawi, April 17-19).

References

Afifi, A.A. & V. Clark 1984. *Computer-aided multivariate analysis*. Lifetime Learning Publications, California.

Kim, J.O. & C.W. Mueller 1978. *Introduction to factor analysis: What it is and how to do it*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-013. Sage, Beverly Hills and London.

Kline, P. 1994. *An easy guide to factor analysis*. Routledge, London.

Rummel, R.J. 1970. *Applied factor analysis*. Northwestern, Evanston.

Suh, Y., K. Kyuma & K. Kawaguchi 1977. A method of capability evaluation for upland soils. III. Numerical evaluation of soil tilth (2) Scaling of soil tilth and its evaluation with field characters. *Soil Sci. Plant Nutr.*, **23**, 263-273.