# *Factor Analysis in Hydrology—An Agnostic View*

JAMES R. WALLIS

*IBM Watson Research Center*
*Yorktown Heights, New York 10598*

*Abstract.* It is suggested that factor analysis, if used in the classical manner, will never be of great value for hydrologic analysis. However, factor analysis used as a numerical procedure for screening variables and building effective regression equations is a useful and powerful tool for hydrologic analysis that can be expected to yield equations that outperform others when used as predictors for control samples. (Key words: Reduced rank regression; antifactor analysis)

## INTRODUCTION

Advocates of factor analysis assume that they have a random sample from a homogeneous population in which they have measured many variables. They believe that these variables can be transformed into fewer more meaningful factors that will then be invariant as to choice of attributes and entities. Factor analysis is often considered a part of statistical methodology. But in view of the fervor shown by its advocates as well as by its detractors, it is suggested that factor analysis might better be classified as a religion.

In this paper, we assume that the reader is familiar with the terminology and objectives of factor analysis, and in addition is aware of the schism that exists between believers such as *Cattell* [1965] and nonbelievers such as *Matalas and Reiher* [1967]. For our present purpose it is necessary to make a clear distinction between classical factor analysis and an identical numerical screening and modeling procedure (called antifactor analysis in this paper) that is based upon different criteria and expectations.

Because of the special nature of hydrologic data, there appears to be little justification for hydrologists to contemplate using classical factor analysis. Hydrologic data are different from psychological data in two important respects. First, in hydrology we rarely have large random samples taken from a homogeneous population; and second, measurement errors on hydrologic variables tend to be much smaller than those in typical psychometric studies. To make classical factor analysis work with hydrologic data, it is necessary either to define the factors in nonmetric terms or to define the factors in terms of the variables and accept the idea that factorial invariance cannot be obtained. Both alternatives make the classical factor analyst's concepts of factorial invariance and common factors operationally meaningless.

In contrast, antifactor analysis as defined and used in this paper appears to be a most useful tool for certain types of hydrologic research. The method has resulted from work done with reduced rank regression prediction equations and to my knowledge has not been separated from this parent.

## MODEL BUILDING USING REDUCED RANK REGRESSION

Since reduced rank regression was first demonstrated in the early 1960's [*Kendall,* 1961], empirical evidence has accumulated to show that reduced rank regression equations have a number of desirable properties:

1. The equations tend to make sense, that is, variables that are expected to have positive effects on the dependent variable have positive signs, and conversely [*Kendall,* 1961; *Wallis,* 1965*a*].

2. The remaining regression coefficients are comparatively stable as excess variables are removed from the analysis [*Wallis,* 1965*a*].

3. Excellent results are obtained, even with small sample sizes [*Burket,* 1964; *Wallis,* 1967].

4. The correlation of observed $Y$'s with the

$\hat{Y}$'s for a control sample not used in establishing the regression tends to be higher than correlations obtained by full rank conventional regression using the same variables [*Burket*, 1964; *Wallis*, 1966]. Of course, for the initial sample the correlation of observed $Y$'s with the $\hat{Y}$'s will always be higher for the full rank regression than it is for the comparable principal component reduced rank regression.

The only disadvantage to the original procedure [*Kendall*, 1961] was that the actual reduced rank that should be used in generating the regression equation was not analytically prescribed. Kendall had recommended that all eigenvalues that were 'small' be discarded. But, unfortunately, 'small' was not defined. And in any case it is feasible for even the smallest positive eigenvalue to be associated with an eigenvector that is the best predictor for a specified criterion ($Y$ variable). An analytic definition of 'small' has been presented (Beale *et al.*, unpublished manuscript), but comparisons between it and the alternative method suggested below have not yet been made.

To overcome the above disadvantage to reduced rank regression, another procedure can be used. The procedure is numerically equivalent to factor analysis but greatly different in spirit and criteria.

The object of the following suggested method of analysis is to obtain a subset of the predictor variables that have approximately the same apparent rank as the whole set of predictor variables. In addition, a further elimination of predictor variables and hence rank is sometimes possible, because some of the identified predictors may not be needed for predicting the specific criterion variable. And this method can, at will, be further reduced to give the misnamed 'orthogonal regression procedure' [*Mustonen*, 1967].

## REDUCED RANK REGRESSION WITH VARIMAX FACTOR WEIGHT MATRIX

Given:

$R_{xx}$, correlation matrix of predictor variables $(n, n)$;
$R_{xy}$, correlation matrix of the predictor variables with the criterion variables $(n, k)$;
$A$, the unrotated reduced rank factor weight matrix $(n, m)$ with $m \leq n$. Note that if $m = n$ the subsequent regression coefficients become identical to those of conventional regression;
$T$, orthonormal transformation matrix. Note that under this transformation the regression coefficients do not change from those obtained using the unrotated $A$ matrix;
$B$, rotated factor weight matrix $(n, m)$;
$\beta$, matrix of square roots of rotated factor contributions for each of the $m$ factors on each of the $Y_k$;
$C$, matrix of standardized regression coefficients for each of the $n$ predictor variables and each $Y_k$;
$X$, standardized matrix of predictor variables;
$e$, matrix of errors $(n, n)$.

Then:

$$R_{xx} = AA' + e$$
$$B = AT$$
$$\beta = (B'B)^{-1}B'R_{xy}$$
$$C = B(B'B)^{-1}\beta$$
$$\hat{Y} = XC$$

*The matrix A.* Commonly used methods of obtaining the matrix $A$ include (1) principal components analysis [*Kendall*, 1961, pp. 71–74], (2) principal factor analysis [*Harman*, 1960], (3) alpha factor analysis [*Kaiser and Caffrey*, 1965], and (4) canonical factor analysis [*Rao*, 1955]. Principal factor, alpha, and canonical factor analyses use communality factoring, a concept that has only recently been introduced into hydrology [*Matalas and Reiher*, 1967]. Limited tests have been made with communality factoring and with the antifactor analysis approach [*Wallis*, 1966]. The work has been criticized for insufficient replication and because one could expect the choice of an inappropriate model to lead to the observed superiority of reduced rank estimators. It has been further suggested that if the data fit a linear model it would be impossible (a priori?) for reduced rank equations to predict better than the full rank counterpart, and that hence there is little need for producing more robust regression methods in the face of nonlinear data. For much hydrologic research the validity of this latter suggestion appears questionable. From additional unpublished tests it appears that communality factoring introduces unnecessary complications and uncertainties into an otherwise straightforward procedure and should generally not be used in antifactor analysis. A possible rare exception to this last prohibition would be those problems for which the original data and final equation have a very low coefficient of determination ($R^2$). Recalculation of the regression coefficients by alpha or canonical factor analysis using the previously obtained effective rank and choice of variables may lead to better (less biased) estimators and hence give an equation that could be expected to predict better with control observations.

*The matrix T.* There are a large number of available transformation matrices, and the criteria for selecting which to use are somewhat arbitrary. Of the presently available methods, the criterion of Varimax is clearly superior because it, as *Kaiser* [1964, p. 42] has stated, 'never results in catastrophe.' In this context, a catastrophe would be a method of rotation that does not fulfill the following four criteria.

1. The rotation should be specific, so that all researchers who start with the same data get to an identical end point.

2. The resulting columns of the rotated factor weight matrix ($B$) should be easily interpretable in terms of the predictor variables; that is, if possible, only large and small values should appear.

3. The $B_{ij}$'s for $m$ of the factors should remain relatively unchanged even if $m + 1$ or $m + 2$ factors are rotated, and some columns of the enlarged matrix should have only small values. For an example of factor stability see the Appendix. To recognize similar columns, the user should have at least two defining variables per factor, as most computer programs do not specify the order of computer printout of the factors.

4. Unique variables should define separate factors and not be ignored as they are by the cluster analysis methods of factor analysis. It is important to recognize that the above criteria are less stringent than a requirement for factorial invariance. We have not stated that our factor-defining variables should remain constant over successive samples or for other choices of variables. Present evidence suggests that for watershed studies we often have small biased samples, and that the concept of factorial invariance is operationally meaningless [*Wallis and Anderson*, 1965; *Eiselstein*, 1967; *Anderson*, 1965; and *Wong et al.*, 1963].

It is at this point in the analysis that the reported 'increased understanding' sometimes occurs. To illustrate: of the seven persons who have worked with me using their data and antifactor analysis, six have reported the serendipitous discovery of anomalies in variables or data. This sample is too small to allow me to make firm conclusions concerning the method's ability to provide such secondary benefits, especially as serendipity is hard to measure and even harder to correlate with any specific method of analysis.

*Recommended method of use.* 'Rules for model building using principal components analysis and Varimax rotation of the factor weight matrix are still in an embryonic state of development. The tentative procedure suggested below will doubtless be modified' [*Wallis*, 1965*a*, p. 453]. The suggested steps of the procedure are:

1. Know as much as possible about the system being investigated.

2. Use only variables that can reasonably be expected to relate to a single underlying process.

3. Transform variables to approach multivariate normal distribution.

4. Make a principal components analysis of the predictor variable correlation matrix, using a high proportion of the explained variance as the cutoff for the initial estimate of $m$. The initial cutoff should be high enough to ensure that at least one of the rotated factors obtained at step 5 of the procedure has only small loadings. Most experimental hydrologic data have sufficient multicollinearity for the 0.995 explained variance cutoff to be effective. If multicollinearity among the predictor variables is absent, then $m$ should be set equal to $n$, and the Mustonen 'orthogonal regression procedure' used.

5. Make a Varimax rotation of the principal component factor weight matrix.

6. Retain no more than two defining variables per factor. Defining variables are those with high factor loading. An important exception to this rule occurs when bias in the sample results in functionally unrelated or anomalous variables appearing as definers of a single factor, in which case they should all remain in the analysis. Additional samples and analyses will be needed to isolate the individual variable effects of such composite factors.

7. Make a principal component analysis with Varimax rotation on the remaining variables. For this analysis, set the number of factors to retain equal to the number of single-factor defining variables plus one-half the number of paired defining variables.

8. If variables have factor loadings of greater than an arbitrary value of 0.40 on 2 or more

factors, they are composite variables, and one should attempt to redefine them to eliminate this confounding or look for additional observations where such confounding does not exist. If composite variables remain in the analysis, then the factor contributions to the explained variance of the $Y$'s are not clearly identifiable with the defining variables, and step (10) of the procedure becomes more difficult. Note that this step is a clear denial that we are either looking for or expecting to find 'factorial invariance.' For most data the elimination of composite variables does not appreciably influence the goodness of fit to the original sample, but it does help the goodness of fit when the equation is used for prediction with control observations. If communality factoring is used, then composite variables lead to Heywood cases (communalities $\geq 1.0$), and the effect of these upon subsequent regression coefficients is untested.

9. Investigate the defining variable or variables of each factor to see whether further transformation would increase their value as a predictor. *Box* [1955] has suggested some useful approaches for further model building.

10. If the model is still too complex, eliminate variables and factors whose contributions to the explained variance of the desired criterion variable are small. Steps (6) through (10) are recursive.

11. Check for autocorrelation using the Durbin and Watson statistic, and if autocorrelation is detected consult *Johnston*'s [1960] book for evasive measures.

12. Test the final equation with a new sample to determine how well it predicts. Significance or goodness of fit tests to the original sample are not justifiable for deductive model building procedures.

## DISCUSSION

The procedure given above appears complex, but it can be executed with great speed and low cost using existing computers and programs [*Wallis,* 1965*b*]. In addition, if later two sets of observations yield differing Varimax matrices, we should regard it as a warning that regressions developed from either set may be in trouble as predictors for the other set. In other words, although knowledge of the dissimilarities in the underlying structure of the variables does not change the prediction that is obtained by a regression, it might deter one from using a poor predictor, or even goad one into developing a more appropriate equation.

Discriminant function and other methods of numerical taxonomy have been used to separate groups of entities. In this regard it is suggested that the antifactor analysis procedure may have three advantages over its more conventional brethren. First, it can show an important difference between groups, even if the variable means and standard deviations of both groups are identical. Second, repetitious factor defining variables cannot inflate the difference between groups. Third, the factor contributions to the explained variance in the criterion variable can show whether observed differences between groups are relevant to a specific criterion variable.

Alternative screening procedures to antifactor analysis are numerous, and a concise summary of most alternatives is available [*Winokur,* 1967]. In hydrology only one alternative, stepwise regression using arbitrary '$F$' levels, and residual mean square error criterion has seen much use. The advocates of this latter method believe that the method 'is preferable if prediction of the dependent variable with minimum error is the desired result' [*Julian et al.,* 1967].

As a test of the relative predicting power of stepwise regression versus antifactor analysis, 2000 observations were generated by successive use of equation 1 (Appendix). The first 500 observations were set aside as an independent control. Random plus or minus errors were assigned to the six $X$ variables of the remaining 1500 observations at various levels of average error (20% for 501–1000, 40% for 1001–1500, and 60% for 1501–2000). All random numbers were from the IBM 360-50 APL rectangular random number generator scaled to have a mean value of 1.0 and range of 2.0 to 0.0. Data for each error level were subdivided into ten samples of 50 observations each, and these thirty sets of data were analyzed by stepwise regression (BMD 03R with 5% '$F$' to include and 10% '$F$' to exclude), and by antifactor analysis. Predicted $\hat{Y}$'s were correlated with observed $Y$'s for the 500 control samples by each of the sixty equations. Stepwise regression was found to be a better predictor two-thirds of the time, although its variance was also larger, so that the mean superiority in $R^2$ units was only 0.002. Similar results were obtained for antifactor analysis versus all variable full rank regression. It should be stressed that this test used data that fitted a linear model and had errors that were symmetric about the mean; these are optimal conditions for full rank and stepwise regression, and yet their vaunted superiority as predictors was not evident.

It was suggested that hydrologic data often violate the assumptions of stepwise regression, and that there are at least three disadvantages to its use. First, stepwise procedures capitalize on the specific errors in the initial sample, and if the model is a poor representation of the true functional relationship, or if the errors are nonrandom, then the estimators are biased, and the equations tend to be sensitive to differences in sample size. Second, stepwise procedures favor composite variables, and these may introduce an additional source of prediction error when used with control observations. Third, stepwise procedures do not lead towards further model building, thinking, or understanding of the phenomena being studied To paraphrase an authority: having made one stepwise regression, there appears to be little else to do but to make another. It is suggested that antifactor analysis screening be substituted for stepwise procedures whenever true functional relationships are unknown or when errors may be nonrandom.

In summation, a procedure numerically equivalent to classical factor analysis is recommended for use in hydrologic analysis. The philosophy, expectations, and criteria of use are so different that the procedure might better be called antifactor analysis. The procedure can be used for attacking complex prediction problems. And it is one that integrates easily with all of the other statistical, numerical, and analytical techniques available to the hydrologist. Further, it is suggested that hydrologic data are so rarely a random sample of a homogeneous population that classical factor analysis of hydrologic data will most likely be unproductive, although if used intelligently it might lead to decision rules that reduce inventory and survey costs for specific areas and problems [*Dawdy and Feth*, 1967; *T.V.A.*, 1965].

TABLE 1. 10 Principal Component Factor Weight Matrices Based upon an Identical Model and with Random Numbers Selected from a Rectangular Distribution. Each Run Based upon 100 Observations (For Results of Varimax Rotation of These Matrices See Tables 2 and 3)

Run 1 Components

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $(X_1)$ | +0 33 | +0 94 | +0 02 | +0 02 | −0.11 | −0 02 |
| $(X_1)^2$ | +0 35 | +0 93 | +0 01 | −0 02 | +0 11 | +0.02 |
| $(X_2)$ | −0.77 | +0 20 | +0 59 | −0.11 | −0.03 | +0 04 |
| $(X_2)^2$ | −0.78 | +0 19 | +0 58 | +0.11 | +0.03 | −0 04 |
| $(X_3)$ | −0 77 | +0 22 | −0.59 | −0.06 | +0 01 | −0 09 |
| $(X_3)^2$ | −0.78 | +0 21 | −0.58 | +0.06 | −0 01 | +0.09 |

Run 2 Components

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $(X_1)$ | +0 56 | +0.63 | −0 52 | +0 03 | −0 08 | +0 08 |
| $(X_1)^2$ | +0 58 | +0 64 | −0 49 | −0 03 | +0.08 | −0.08 |
| $(X_2)$ | −0 52 | +0 75 | +0.39 | +0 09 | −0.04 | −0.06 |
| $(X_2)^2$ | −0.52 | +0 75 | +0 40 | −0.10 | +0 04 | +0 06 |
| $(X_3)$ | +0 69 | +0 02 | +0.72 | −0 07 | −0 08 | −0 05 |
| $(X_3)^2$ | +0.68 | +0.05 | +0.72 | +0 08 | +0 08 | +0.05 |

Run 3 Components

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $(X_1)$ | −0.35 | −0 93 | +0.07 | −0 02 | −0.12 | −0 03 |
| $(X_1)^2$ | −0 34 | −0.93 | +0 09 | +0 02 | +0 12 | +0 03 |
| $(X_2)$ | −0 76 | +0 15 | −0 62 | +0.01 | +0 03 | −0.10 |
| $(X_2)^2$ | −0 77 | +0 15 | −0.61 | −0 00 | −0.03 | +0.10 |
| $(X_3)$ | +0 77 | −0.28 | −0.59 | +0.13 | +0 02 | −0 00 |
| $(X_3)^2$ | +0.73 | −0.27 | −0.61 | −0.13 | +0.02 | +0.00 |

Run 4 Components

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $(X_1)$ | −0 75 | +0 11 | +0 64 | −0 00 | +0.08 | −0 09 |
| $(X_1)^2$ | −0.71 | +0 09 | +0 69 | +0 00 | −0 08 | +0.09 |
| $(X_2)$ | +0 70 | −0 26 | +0 66 | +0.12 | −0 05 | −0.04 |
| $(X_2)^2$ | +0 71 | −0.24 | +0.65 | −0.12 | +0.05 | +0.03 |
| $(X_3)$ | −0.27 | −0.95 | −0.10 | −0.07 | −0.08 | −0 06 |
| $(X_3)^2$ | −0.26 | −0.95 | −0.10 | +0.07 | +0.08 | +0 06 |

Run 5 Components

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $(X_1)$ | −0 54 | −0 62 | +0.55 | −0.11 | −0 06 | −0 04 |
| $(X_1)^2$ | −0 57 | −0 63 | +0.51 | +0.11 | +0.06 | +0 04 |
| $(X_2)$ | −0 49 | +0.77 | +0.39 | +0 08 | −0 08 | −0.05 |
| $(X_2)^2$ | −0 49 | +0.76 | +0.39 | −0.08 | +0 08 | +0 06 |
| $(X_3)$ | −0.73 | −0 03 | −0.67 | −0.01 | +0.07 | −0 09 |
| $(X_3)^2$ | −0.69 | −0 05 | −0.72 | −0.00 | −0.08 | +0.09 |

Run 6 Components

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $(X_1)$ | +0.64 | −0.25 | +0 72 | +0 06 | −0.11 | −0.01 |
| $(X_1)^2$ | +0.64 | −0.24 | +0.72 | −0.06 | +0.11 | +0.01 |

TABLE 1. (Cont'd)

| Variable | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $(X_2)$ | +0.54 | +0 82 | −0.17 | +0 01 | +0 02 | −0.12 |
| $(X_2)^2$ | +0.53 | +0 83 | −0.12 | −0.01 | −0 02 | +0.12 |
| $(X_3)$ | +0.61 | −0.46 | −0.63 | −0.11 | −0.06 | −0.01 |
| $(X_3)^2$ | +0.60 | −0.48 | −0.63 | +0.11 | +0.06 | +0.02 |
| **Run 7 Components** | | | | | | |
| $(X_1)$ | +0 18 | +0.90 | +0.37 | −0.14 | −0.02 | −0 01 |
| $(X_1)^2$ | +0.19 | +0.90 | +0.35 | +0.14 | +0 02 | +0 01 |
| $(X_2)$ | +0.70 | −0.39 | +0.58 | −0.01 | +0 12 | −0 05 |
| $(X_2)^2$ | +0.70 | −0.41 | +0.53 | +0 01 | −0.11 | +0 06 |
| $(X_3)$ | −0 78 | −0.15 | +0.60 | −0.02 | +0 07 | +0.10 |
| $(X_3)^2$ | −0.76 | −0.15 | +0.62 | +0.02 | −0.07 | −0.10 |
| **Run 8 Components** | | | | | | |
| $(X_1)$ | −0.49 | −0.73 | −0.46 | +0.08 | +0.04 | −0.09 |
| $(X_1)^2$ | −0.45 | −0.73 | −0.49 | −0.08 | −0.03 | +0.09 |
| $(X_2)$ | −0.53 | +0.68 | −0.50 | +0.10 | +0.02 | +0.08 |
| $(X_2)^2$ | −0.50 | +0.70 | −0.50 | −0.10 | −0.02 | −0.08 |
| $(X_3)$ | −0.78 | +0 00 | +0.61 | −0.05 | +0.11 | +0 01 |
| $(X_3)^2$ | −0.78 | −0.02 | +0.61 | +0.04 | −0.11 | −0.01 |
| **Run 9 Components** | | | | | | |
| $(X_1)$ | −0 71 | +0.29 | −0.62 | −0.05 | −0 11 | −0.05 |
| $(X_1)^2$ | −0.73 | +0 28 | −0.61 | +0 05 | +0.10 | +0.06 |
| $(X_2)$ | −0.29 | −0.94 | −0.12 | +0.12 | −0.05 | +0.01 |
| $(X_2)^2$ | −0 28 | −0 94 | −0.15 | −0.12 | +0.05 | −0.01 |
| $(X_3)$ | +0.75 | −0.09 | −0.64 | −0.04 | −0.05 | +0.11 |
| $(X_3)^2$ | +0.75 | −0.07 | −0.65 | +0.04 | +0.05 | −0.11 |
| **Run 10 Components** | | | | | | |
| $(X_1)$ | +0 68 | +0.40 | −0.61 | −0 08 | +0.01 | −0 09 |
| $(X_1)^2$ | +0 66 | +0.41 | −0 62 | +0.08 | −0 01 | +0.09 |
| $(X_2)$ | +0.44 | −0 87 | −0.17 | +0.10 | +0.07 | −0 05 |
| $(X_2)^2$ | +0.44 | −0.88 | −0 14 | −0.09 | −0.07 | +0.05 |
| $(X_3)$ | −0.69 | −0.17 | −0.69 | −0.06 | +0.10 | +0.04 |
| $(X_3)^2$ | −0.71 | −0.16 | −0.67 | +0.05 | −0.11 | −0.04 |

## APPENDIX. EXAMPLE OF FACTOR STABILITY

We require (1) that the factors should tend to have only high and low loadings with a minimal number of intermediate values (0.35 to 0.75); and (2) that $n$ of the hyperplanes should remain pointing in the same direction as the number of rotated factors ($k$) is increased from the effective rank ($m$) towards the order ($n$); and (3) that ($k - m$) excess hyperplanes should generate factors with only low loadings.

To estimate how well principal components and Varimax rotation of the factor weight matrix meet the above requirements, a simple test was devised (Tables 1, 2, 3).

Ten samples each of 100 observations were generated using equation 1

$$y_j = \sum_{i=1}^{3} (X_{ij} + X_{ij}^2) \qquad j = 1, 2, 3, \cdots 1000 \qquad (1)$$

and a computer random number generator with rectangular distribution for the $X$'s. Three groups of two highly correlated $X$ variables resulted, with the groups being only slightly correlated. The 10 principal component factor weight matrices for these $X$ variables are given in Table 1, where it can be seen that we have many intermediate loadings, and that the underlying simple relationship between the variables is not evident. It should be evident from Table 1 that the numerical value of an eigenvector is primarily a function of small random fluctuations in extreme values rather than of the total underlying variable relationships, and that interpretations and explanations of eigenvectors in terms of variables are nonsensical and should never be attempted.

TABLE 2. Mean and Standard Deviation of 10 Varimax Factor Weight Matrices, Each of Which Was Based on an Identical Model, and with Random Numbers Selected from a Rectangular Distribution (100 Observations per Sample, Order = 6, Rank = 6)

Factor

| Variable Name | 1 Mean | 1 S. D. | 2 $\bar{X}$ | 2 $\sigma X$ | 3 $\bar{X}$ | 3 $\sigma X$ | 4 $\bar{X}$ | 4 $\sigma X$ | 5 $\bar{X}$ | 5 $\sigma X$ | 6 $\bar{X}$ | 6 $\sigma X$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $(X_1)$ | +0.991 | ±0 001 | −0 006 | ±0.033 | −0.009 | ±0.040 | +0.124 | ±0.007 | −0.001 | ±0.005 | −0.000 | ±0 003 |
| $(X_1)^2$ | +0.991 | ±0.002 | −0.003 | ±0.022 | −0.010 | ±0.045 | −0.124 | ±0.007 | +0.002 | ±0.005 | +0.000 | ±0.003 |
| $(X_2)$ | −0.002 | ±0 023 | +0.989 | ±0.003 | +0 000 | ±0.069 | −0.002 | ±0.005 | +0.124 | ±0.008 | −0.000 | ±0 005 |
| $(X_2)^2$ | −0.007 | ±0.030 | +0.989 | ±0.004 | −0.007 | ±0.069 | +0.002 | ±0.005 | −0.125 | ±0.008 | +0.001 | ±0 005 |
| $(X_3)$ | −0.009 | ±0.041 | −0 002 | ±0.073 | +0.989 | ±0.003 | +0.000 | ±0.003 | −0.000 | ±0.006 | +0.123 | ±0.006 |
| $(X_3)^2$ | −0.011 | ±0.041 | −0.005 | ±0.067 | +0.989 | ±0.004 | +0.000 | ±0.003 | −0.001 | ±0.006 | −0.112 | ±0.031 |

The principal component factor weight matrices of Table 1 were rotated by Varimax with the rank set to 6 (Table 2), and then with the effective rank of 3 (Table 3). For this test it can be seen that principal components with Varimax meet the factorial stability requirements required by antifactor analysis. Similar results have been obtained with other models.

TABLE 3. Mean and Standard Deviations of 10 Varimax Factor Weight Matrices, Each of Which Was Based on an Identical Model, and with Random Numbers Selected from a Rectangular Distribution (100 Observations per Sample, Order 6, Rank 3)

| | Factor | | | | | |
|---|---|---|---|---|---|---|
| Variable | 1 | | 2 | | 3 | |
| Name | Mean | S.D. | $X$ | $\sigma X$ | $X$ | $\sigma X$ |
| $(X_1)$ | +0.991 | ±0.001 | −0.006 | ±0 033 | −0.009 | ±0 040 |
| $(X_1)^2$ | +0.991 | ±0.002 | −0 003 | +0.022 | −0.010 | ±0.045 |
| $(X_2)$ | −0.002 | ±0.023 | +0.989 | ±0 003 | 0.000 | ±0.069 |
| $(X_2)^2$ | −0.007 | ±0.030 | +0.989 | ±0.004 | −0 007 | ±0.069 |
| $(X_3)$ | −0.009 | ±0.042 | −0.002 | ±0 073 | 0 989 | +0.003 |
| $(X_3)^2$ | −0.011 | ±0.041 | −0.005 | ±0.067 | 0 989 | ±0 004 |

*Acknowledgments.* The work reported here was supported by the United States Forest Service, The Charles E. Bullard Fellowship Fund, Harvard University, and IBM. Concepts expressed grew from conversations with William Meredith, University of California, Berkeley; Henry Anderson, U. S. Forest Service, Berkeley; Harold Thomas, Jr., Harvard University; and Herbert S. Winokur, Jr., Harvard University. A useful review of an earlier draft was provided by N. C. Matalas of the U. S. Geological Survey.

REFERENCES

Anderson, H. W., Watershed modeling approach to evaluation of the hydrological potential of unit areas, *Proc. Int. Symp. Forest Hydrol.*, Pennsylvania State Univ., 1965.

Box, G. E. P., The exploration and exploitation of response surfaces, *Biometrics, 10,* 16–40, 1955.

Burket, G. R., A study of reduced rank models for multiple prediction, *Psychometric Mono., 12,* 66 pp., 1964.

Cattell, R. B., Factor analysis: An introduction to essentials. 1. The purpose and underlying models, *Biometrics, 21,* 190–215, 1965.

Dawdy, D. R., and J. H. Feth, Application of factor analysis in the study of chemistry of groundwater quality, Mojave River Valley, California, *Water Resources Res., 3*(2), 505–510, 1967.

Eiselstein, Leo, A principal component analysis of surface runoff data from a New Zealand alpine watershed, Submitted for publication in *Proc. I.A.S.H. Symp., Fort Collins, Colo.*, 1967.

Harman, H. H., *Modern Factor Analysis,* University of Chicago Press, Chicago, 1960.

Johnston, J., *Econometric Methods,* 300 pp., McGraw-Hill Book Company, New York, 1960.

Julian, R. W., et al., Prediction of water yield in high mountain watersheds based on physiography, *Colo. State Univ. Hydrol. Paper 22,* 20 pp., 1967.

Kaiser, H. F., Psychometric approaches to factor analysis, *Proc. Invitational Conference on Testing Problems,* 37–45, Educational Testing Service, Princeton, New Jersey, 1964.

Kaiser, H. F., and J. Caffrey, Alpha factor analysis, *Psychometrika, 30,* 1–14, 1965.

Kendall, M. G., *A Course in Multivariate Analysis,* Charles Griffin Co. Ltd., London, 1961.

Matalas, N. C., and Barbara J. Reiher, Some comments on the use of factor analysis, *Water Resources Res., 3*(1), 213–224, 1967.

Mustonen, S. E., Effects of climatologic and basin characteristics on annual runoff, *Water Resources Res., 3*(1), 123–130, 1967.

Rao, C. R., Estimation and tests of significance in factor analysis, *Psychometrika, 20,* 93–111, 1955.

T. V. A., Division of water control planning, design of a hydrologic condition survey using factor analysis, *Tenn. Valley Auth., Res. Paper No. 5,* 1965.

Wallis, J. R., Multivariate statistical methods in hydrology—A comparison using data of known functional relationship, *Water Resources Res., 1*(4), 447–461, 1965*a*.

Wallis, J. R., WALLY-1. . . . A large principal components regression program with Varimax rotation of the factor weight matrix, *U. S. Forest Serv. Res. Note PSW-92,* 6 pp., 1965*b*.

Wallis, J. R., Effect of random errors on the accuracy of predictions made from reduced rank regression equations (unpublished report on file at Pacific SW Forest and Range Exp. Station, U. S. Forest Service, Berkeley, California), 1966.

Wallis, J. R., When is it safe to extend a prediction equation?—An answer based upon factor and discriminant function analysis, *Water Resources Res., 3*(2), 375–384, 1967.

Wallis, J. R., and H. W. Anderson, Multivariate analysis in sediment network design, *I.A.S.H. Publ. No. 67,* 357–378, 1965.

Winokur, H. S., Jr., Statistical methods for specifying a regression equation, *Social Systems Res. Inst.*, University of Wisconsin, 8 pp., 1967.

Wong, S. T., J. R. Shaeffer, and H. Gotass, *Multivariate Statistical Analysis of Metropolitan Water Supplies,* Northeastern Illinois Area Planning Commission, Chicago, Illinois, 1963.

(Manuscript received November 15, 1967.)