When Is It Safe To Extend a Prediction Equation?—An Answer Based upon Factor and Discriminant Function Analysis

JAMES R. WALLIS

Pacific Southwest Forest and Range Experiment Station
Forest Service, U.S. Department of Agriculture
Berkeley, California

Abstract: Prediction equations for hydrologic events developed from one population of observations (watersheds) are often solved for another population that is removed either in time or in space. Predictions of this kind are never certainties, although some predictions are obviously more uncertain than others. This paper proposes an empirical uncertainty classification that may be found useful for separating probably successful from probably unsuccessful extensions of prediction equations. The classification system is illustrated by a prediction equation for suspended sediment discharge developed from some watersheds in California, and by a discriminant function for marine versus nonmarine sediments based upon microelements. (Key words: Statistics; principal components analysis; linear discriminant functions).

INTRODUCTION

Hydrologists often wish to predict the behavior of one or more variables (criteria) for one or more of a new population of watersheds. They do so by solving a mathematical equation for the appropriate predictor variables, but when the predictions are checked they are often found to be associated with large errors. Such errors can be attributed to two main causes: first, errors in the specification of the original model; and second, unwarranted assumptions concerning the degree of similarity between the original and the new group of watersheds. This paper will concern itself almost entirely with this second cause of errors.

Figure 1 shows two groups of watersheds for which we wish to know whether a prediction equation (model) developed from the watersheds of group 1 is applicable to all or any of the watersheds of group 2. For the current discussion we can subdivide this one question into two easier subquestions: (1) Is there a significant difference between the two groups of watersheds? (2) Is a significant difference in the groups relevant to the specific criterion being predicted?

To illustrate relevance versus significance, consider two groups of watersheds: one selected entirely from the Mojave Desert, and the second entirely from the Sonoran Desert. The Mojave can be distinguished from the Sonoran watersheds by the vegetation (for example, presence or absence of Joshua trees and Saguaro cacti). A distinction as to vegetation type would normally be both significant and relevant to differences in rainfall-runoff. But for the Joshua tree-Saguaro cactus example we could have high significance without necessarily having relevance to the rainfall-runoff relationship.

THE LINEAR DISCRIMINANT FUNCTION

Procedures for distinguishing between groups of objects and for testing the significance of the difference between groups of objects were developed in the 1920's and 1930's by Fisher, Hotelling, and Mahalanobis. Their work has been reviewed in an excellent monograph [*Hodges*, 1950], and a brief summary can be found in Appendix 1.

HOW RELEVANT IS A SIGNIFICANT DIFFERENCE BETWEEN TWO GROUPS OF WATERSHEDS?

Discriminant analysis normally ceases once the classification has been made and the significance of the difference between groups obtained. But in studies where we wish to take a prediction equation developed from one set of watersheds and apply it to another set removed in either time or space, we also have to consider whether or not a significant difference between

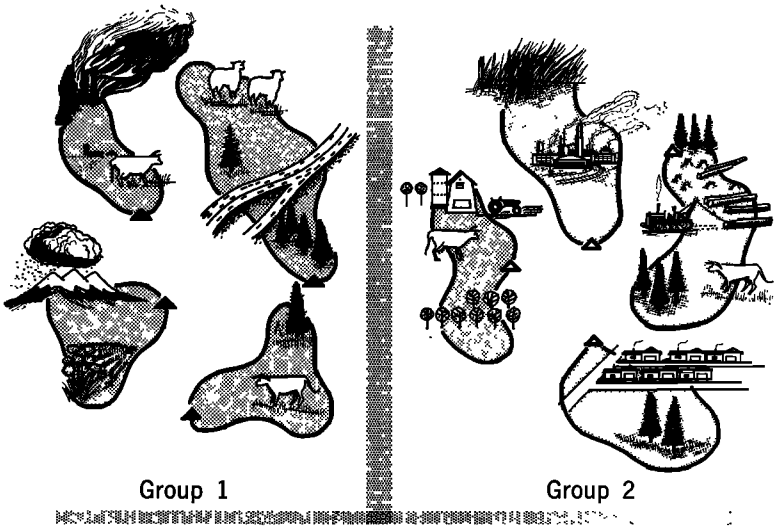

Fig. 1. Is an equation developed from the watersheds of group 1 useful for predicting with group 2?

groups is relevant to the desired prediction. In the rest of this paper the concept of relevance will be further illustrated, and an empirical procedure for assigning proposed extensions of prediction equations into high or low uncertainty categories will be given.

AN EXAMPLE OF SIGNIFICANCE VERSUS RELEVANCE

An analysis of annual suspended sediment loads for 46 California streams and 41 variables has been made by *Wallis and Anderson* [1965]. Their definitions for 16 of the variables used in the study are given below.

A = Contributing sediment area in square miles, equal to watershed area if no lakes or reservoirs are large enough to be efficient sediment traps.

LC = Mean flow path length in miles. Calculated by the grid sampling technique.

SC = Standard deviation of flow path length in miles. Grid sampling was used with a population of from 30 to 40 streams per watershed.

MAP = Mean annual precipitation for the basin. Obtained from a 100-point sampling of the unpublished California Department of Water Resources Rainfall Map.

MAQ = Mean annual runoff in cfs per square mile for the area (*A*).

I = Probable maximum 24-hour precipitation.

RRA = Relative rain area modified for California latitudes.

RSM = Relative snowmelt area modified for California latitudes.

PUSED = Percentage of area covered by unconsolidated sediments. In practice assumed to be all sediments equal to or younger than the Cenozoic.

PSED = Percentage of area covered by consolidated sediments. In practice assumed to be all sediments older than Cenozoic.

GRASS = Acres classified as grassland per square mile of watershed area (*A*).

IGS = Composite interaction variable made up of percentage slope times percentage grassland area.

LOG 1 = Acres classified as logged with roads and landings predominantly in draws per square mile of watershed area (*A*).

SED = Tons per square mile per year of

suspended sediment load, based upon long-term flow and specific flow-concentration relationship.

UROADS = Acres classified as unimproved roads per square mile of watershed land (*A*).

FIRE 10 = Acres of wildfire per square mile of watershed area (*A*) in the 10-year period to the year of the suspended sediment analysis.

They divided the 46 watersheds into two groups (primarily on the basis of presence or absence of the variable *LOG* 1) and developed a principal components regression equation for predicting suspended sediment (*SED*) from the data of one of the groups. The equation should be applicable to the second group, provided that there are no large differences in sediment-producing variables between the initial and second group.

A principal components regression, with varimax rotation, was made for all the variables of their study. In the analysis in the present study an additional dummy criterion variable (*Z*) was used to discriminate between the two groups of watersheds. Factors with small contribution to the explained variance of both (*SED*) and (*Z*) were eliminated from further consideration, as were most composite variables. Table 1 shows the varimax factor weight matrix that resulted from the above procedure. Each of six factors (area, sedimentary rock type, grassland, and two precipitation characteristics) had large contributions to the explained variance of the discriminant variable (*Z*). For the (*Z*) variable separate '*F*' tests showed significance of the 0.05 level for each of the 13 variables used to define the six factors of Table 1.

Two of the factors—area and amount of sedimentary rock—represented by five variables (*A*, *LC*, *SC*, *PUSED*, and *PSED*) had factor contributions of 0.00 to the explained variance of watershed sediment loads per square mile. In other words, significant differences in area and in area of sedimentary rock type exist between the initial and control group of watersheds. But these differences are not demonstratively relevant to differences in the production of sediment per square mile of watershed area (*SED*).

These two groups of watersheds have large differences in four unrelated factors (amount of runoff, type of precipitation, area in grassland, and amount of sediment producing logging method) that are relevant to sediment production. It is therefore unreasonable to expect that a sediment prediction equation developed from one group would be applicable to the other. For predictions of another variable unrelated to annual sediment load (for example, the mean August concentration of dissolved solids), however, it would be necessary to reconsider the question of the relevance of the significant differences between the two groups.

TABLE 1. Varimax Factor Weight Matrix and Factor Contributions to Variation in Annual Suspended Sediment Loads (*SED*), and to Discriminant Function (*Z*) for Two Groups of 23 California Watersheds.*

Variable	Varimax Factor Loadings (Only Those Greater than 0.3 Shown)					
Name	1	2	3	4	5	6
A	0.984					
LC	0.979					
SC	0.969					
MAP		0.938				
MAQ		0.923				
I			0.837			
RRA			0.953			
RSM			0.980			
PUSED			0.432	−0 810		
PSED		0.322	0.510	0.727		
GRASS		−0.333			0.888	
IGS					0.963	
LOG 1						0.900
	Factor Contributions to the Explained Variance					
SED	0.00	0.10	0.29	0.00	0.05	0.07
Z	0.12	0.16	0.04	0.12	0.12	0.09

* Data from *Wallis and Anderson* [1965].

UNCERTAINTY CATEGORIES

A four-step procedure for assigning uncertainty categories to predictions made from one group of watersheds to another group removed in either time or space is given below. It should be understood that this classification scheme is entirely empirical.

1. Measure variables in both groups of watersheds that are thought to be influencing the differences in the item to be predicted.
2. Make a discriminant analysis of these variables, using principal components regression and varimax rotation.

3. Simplify and reanalyze the rotated factor weight matrix by eliminating
   a. Composite type variables.
   b. Unnecessary factor defining variables (those with high factor loadings, but with small factor contribution to the discriminant, say 0.01 or less).
4. Classify the proposed extension into its uncertainty category.
   a. Low uncertainty extensions: Factor contributions to the discriminant coming from no more than one factor that is related to the criterion to be predicted. Under this circumstance the group 1 equation is likely to give successful predictions for group 2 data.
   b. High uncertainty extensions: Factor contributions to the discriminant coming from two or more factors that also contribute to differences in the variable that is to be predicted. Under this circumstance it is unlikely that a prediction equation developed using group 1 watersheds will give satisfactory predictions for group 2.

EXAMPLES OF THE USE OF UNCERTAINTY CATEGORIZATION

The uncertainty categories suggested above do not represent absolute quantities. For successful results they must be interpreted by each individual researcher with intelligent caution.

For instance, consider the (*B*) and (*V*) discriminant function for marine versus environment that was developed from a population of modern sediments [*Potter et al.*, 1963]. We should like to know whether or not the discriminant can be safely used to characterize marine versus nonmarine environments for ancient sediments. A discriminant analysis of modern versus ancient sediments using equation 4 in appendix 1 and Potter's data was made (Table 2).

Results for the pooled data given in Table 2 indicate that (*B*) and (*V*) can be used to make an effective separation of marine from nonmarine environments of deposition. The high *R* squared for the variable (*ZMAR*) is equivalent to a 91% correct classification. Furthermore, that (*B*) and (*V*) are essentially unrelated to differences in age of sediments (*ZAGE*) can be seen by their low factor contributions to the discriminant (0.04 and 0.01). The total *R* squared for the discriminant of 0.05 is equivalent to a classification system that is only slightly better than the 50% that could be expected to arise from a random guess. In other words, a discriminant based upon the (*B*) and (*V*) levels of modern sediments can safely be applied to predicting environments of deposition for ancient sediment, an example of a low uncertainty category extension of a prediction equation.

TABLE 2. Microelements (B) and (V) as Discriminators of Modern versus Ancient (*ZAGE*), and Marine versus Nonmarine (*ZMAR*) Sediments*

Variable	Varimax Rotated Factor Weight Matrix	
	(1)	(2)
B	0.95	0.31
V	0.31	0.95

	Factor Contribution to the LDF		Total
ZAGE	0.04	0.01	0.05
ZMAR	0.37	0.14	0.51

* Pooled data taken from *Potter et al.* [1963].

A second example of the use of uncertainty categories is summarized in Table 1, where important factor contributions to the explained variance of both (*SED*) and (*Z*) come from four unrelated factors (columns 2, 3, 5, and 6 of Table 1). A prediction equation for (*SED*) developed from group 1 watersheds would be a high uncertainty equation for predicting (*SED*) for group 2 watersheds. If checked, the observed and predicted values of (*SED*) for group 2 watersheds do not agree with each other, except for those few watersheds of group 2 that are classified as being like group 1 watersheds by the (*Z*) discriminant.

Uncertainty categories as defined and used above are not foolproof, although I believe that they will give an adequate answer for most real-world, experimental data. The computer runs necessary to produce Table 1 took less than 2 minutes of IBM 7094 computer time. Even at $500 per hour this amounts to less than $17; a wrongly applied prediction equation could lead to far more expensive results.

I suggest that tests using relevant, unrelated variable *LDF*'s might well be routinely made and reported each time a prediction equation is extended in either time or place.

APPENDIX 1. THE LINEAR DISCRIMINANT FUNCTION

Fisher formulated a method for finding the linear combination of variables for which the separation between two groups of objects is a maximum. He called this quantity the linear discriminant function (*LDF*), and it is usually symbolized as Z in equation 1

$$Z = \lambda_1 X_1 + \lambda_2 X_2 \cdots \lambda_p X_p \qquad (1)$$

in which X's are the variables and λ's are the necessary proportionality coefficients. The normal equations for the *LDF* (equation 2) are similar to the normal equations for the standard regression model (equation 3). The only difference is the substitution of λ's for b's on the left-hand side, and of d's for the sum of the XY cross product terms. The d's of equation 2 are the differences between the variable means for the two groups, and the S_{ij} are the customary sums of squares and cross products

$$\begin{aligned} S_{11}\lambda_1 + S_{12}\lambda_2 + \cdots S_{1p}\lambda_p &= d_1 \\ S_{21}\lambda_1 + S_{22}\lambda_2 + \cdots S_{2p}\lambda_p &= d_2 \\ S_{p1}\lambda_1 + S_{p2}\lambda_2 + \cdots S_{pp}\lambda_p &= d_p \end{aligned} \qquad (2)$$

$$\begin{aligned} S_{11}b_1 + S_{12}b_2 + \cdots S_{1p}b_p &= Sx_1y \\ S_{21}b_1 + S_{22}b_2 + \cdots S_{2p}b_p &= Sx_2y \\ S_{p1}b_1 + S_{p2}b_2 + \cdots S_{pp}b_p &= Sx_py \end{aligned} \qquad (3)$$

Computer programs to solve for the *LDF* are numerous (for example, *BIMD* 05), but a similar equation is obtained by using a dummy y variable and standard regression. And if the dummy variable has dichotomous values based upon the class frequencies (f_1 and f_2) given by equation 4

$$f_2/(f_1 + f_2) \qquad -f_1/(f_1 + f_2) \qquad (4)$$

then the resulting coefficients will be proportional to the λ's of equation 1 [*Anderson*, 1950, p. 140]. The original Mahalonobis D^2 statistic (equation 5) made the assumption of independence between the X variates

$$D_p^2 = 1/p \sum_{i=1}^{p} (\bar{X}_{1i} - \bar{X}_{2i})^2/\sigma_i \qquad (5)$$

When the correlation matrix (R) of the population is known, then the form of the Mahalonobis D^2 given in equation 6 can be used to test for the significance of the total difference in the p mean values of the X variables for all s-t comparisons, and where Rst is the cofactor of the stth element of R, and the sample estimates of R, σ's, and $\bar{X}$'s are substituted for the corresponding unknown population parameters

$$D_p^2 = \frac{1}{|R|} \sum_{s=1}^{p} \sum_{t=1}^{p} R_{st} \frac{\bar{X}_{1s} - \bar{X}_{2s}}{\sigma_t} \frac{\bar{X}_{1t} - \bar{X}_{2t}}{\sigma_t} \qquad (6)$$

Equation 7 can be used with p and $n + n - p - 1$ degrees of freedom and an 'F' distribution to test for the significance of the difference between the two groups

$$\frac{N_1 N_2 (N_1 + N_2 - p - 1)\ D^2}{p(N_1 + N_2)(N_1 + N_2 - 2)} \qquad (7)$$

To classify a new set of observations, equation 1 is solved for each new observation and for the mean values of each of the initial groups. The new Z values are compared with the mean Z values and the item assigned to the group to which its Z value is closest [*Potter et al.*, 1963, p. 691].

APPENDIX 2. METHODS OF PICKING VARIABLES FOR DISCRIMINANTS, AND THE EFFECT OF SAMPLE SIZE UPON THEM

The stepwise procedures. The stepwise procedures can be subclassified into, first, stepwise upwards, in which variables are added in order of their descending contribution to the discriminant; and second, stepwise downwards, in which all the variables are initially included, and then those with the least significant 'F' values are successively dropped from the analyses. The stepwise upwards procedure can also be subclassified into those methods that do not allow a previously included variable to be dropped [*Schultz and Goggans*, 1961], and those that do [*Weiner and Dunn*, 1965]. Stepwise procedures terminate at some preset arbitrary 'F' level, and the number of variables included in the final *LDF* is controlled by this 'F' level.

The different stepwise procedures do not necessarily pick the same variables for inclusion in the final *LDF*, although if composite variables exist among the choices, then the stepwise procedures tend to select these in preference to the more independent variables. But when composite variables are not present, the stepwise procedures result in the same choice of variables as that produced by the unrelated measurement methods.

THE UNRELATED MEASUREMENT PROCEDURES

If unrelated variables exist in an array they can be isolated by any of the numerous methods of factor analysis [*Horst*, 1965], but the discussion that follows is restricted to methods based upon principal components analysis and varimax rotation of the factor weight matrix. Justification for these restrictions are, first, that alternative methods depend upon the concept of communality, which has not yet appeared necessary for the analysis of hydrologic data; and, second, as *Kaiser* [1964, p. 43] has stated, 'old reliable work-horse varimax will never result in catastrophe, while other criteria in this class (methods of rotation), perhaps better in specific problems, can be disastrous in others.'

A computer program to make a principal components regression upon a varimax rotated factor weight matrix is available [*Wallis*, 1965*b*]. By using a dummy discriminant variable as a criterion (equation 4), it is possible to eliminate predictor variables until there remains a set of comparatively unrelated variables [*Wallis*, 1965*a*, p. 454]. These variables can then be used to form the *LDF*. Note that the final decision on whether to include dimension-defining variables with only small factor contribution to the discriminant function can be made, as with the stepwise procedures, by means of an '*F*' test. A test based upon a percentage decrease in the number of misclassifications, however, would appear to be more appropriate.

Two modifications of the above procedure that give discriminant functions, which may be more suitable for small samples, are outlined below.

First, the procedure given above is repeated, but if more than one factor-defining variable is included per factor, then the discriminant function used is one based upon reduced rank (that is, number of principal components retained less than the number of predictor variables used). This procedure produces discriminant functions with regression coefficients that are comparatively stable over different samples of objects and variables.

A second modification of the unrelated variable method is to make principal components and varimax rotation analyses, and then to generate factor scores by equation 8 [*Horst*, 1965, p. 479]. Factor scores are then used as input variables for a *LDF*. For a new sample the resulting *LDF* is used with factor scores obtained by equation 8, using the original sample means and standard deviations

$$G = A'A$$
$$B = AG^{-1}$$
$$Y = SB$$

in which A is the varimax factor weight matrix, S is the normalized data matrix, and Y is the least-square factor score matrix (8)

Analyses by this method are comparatively simple if a large, flexible factor analysis program is available (e.g., the University of California program known as *BCTRY*), but otherwise probably should not be attempted.

TWO EXAMPLES OF PICKING VARIABLES FOR AN LDF

Potter et al. [1963] found that the sedimentary rocks can be separated into marine or nonmarine origin of environment on the basis of the amount of microelements present in the rocks. They used a stepwise procedure to determine that only two elements—boron (B) and vanadium (V)—were necessary to make a reliable separation of recent marine from recent nonmarine sediments. A similar result was obtained by making principal components regression analyses with varimax rotation of the factor weight matrices (Tables 3, 4). In Table 3 the composite type variables were (*GA*) and (*NI*), but these variables were not helpful to a discriminant function of sediment type. In fact, one can infer from Table 3 that factor 5 (*NI*) and factor 6 (*PB*) contribute nothing to the discriminant, and that the small contributions coming from dimensions 2, 3, and 4 (*CR*, *CU*, and *GA*) were probably caused by the slight loadings for (*V*) that existed on each of these

TABLE 3. Microelements as a Discriminator of Recent Marine from Nonmarine Sediments*

Element	Varimax Rotated Factor Weight Matrix						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
B	0.972	0.105	0.099	0.129	0.052	−0.082	0.092
CR	0.081	0.877	0.250	0.245	0.038	0.266	0.174
CU	0.116	0.258	0.905	0.167	0.072	0.216	0.145
GA	0.255	0.468	0.262	0.730	0.100	0.204	0.252
NI	0.287	0.610	0.348	0.318	0.467	0.130	0.297
PB	−0.103	0.206	0.191	0.117	0.035	0.934	0.154
V	0.207	0.390	0.277	0.322	0.112	0.404	0.670
	Factor Contributions to the LDF							Total
	0.29	0.04	0.03	0.04	0.01	0.00	0.15	0.56

* Data from *Potter et al.* [1963].

factors. These suppositions appear to be confirmed by the factor contributions [defined in *Wallis,* 1965*a,* p. 461] shown in Table 4, whose sum 0.54 (total explained variance) is only slightly below the 0.56 given by the 7-variable discriminant of Table 3. For these data it is apparent that both stepwise and unrelated variable methods choose the same variables for inclusion in the final *LDF*.

In the next example of picking variables for an *LDF*, we shall see data from 75 hypothetical hollow cylinders [*Wallis,* 1965*a*]. The 15 variables defined in that study were:

Variable	*Symbol*	*Function*
–	*K*	Constant (π)
1	*H*	Height
2	*HH*	(Height)^2
3	*2KROH*	Outside curved surface
4	*2KRIH*	Inside curved surface
5	*D*	Density
6	*DD*	(Density)^2
7	*DDIAGO*	Density times diagonal of outside cylinder
8	*DDIAGI*	Density times diagonal of inside cylinder
9	*RO*	Radius of outside cylinder
10	*KRORO*	End area of outside cylinder
11	*DIAGO*	Diagonal of outside cylinder
12	*RI*	Radius of inside cylinder
13	*KRIRI*	End area of inside cylinder
14	*DIAGI*	Diagonal of inside cylinder
15	*W*	Weight

For testing discriminant functions obtained by different methods, the first 60 of these hollow cylinders were sorted into two weight groups: a light group with weights below 133.24, and a heavy group with weights above 145.44.

Linear discriminant functions were made for these two weight groups using all 14 predictor variables, the stepwise procedure, and the three unrelated measurements methods discussed above. The *LDF* produced by the stepwise procedure with 0.01 *'F'* level to enter and a 0.05 *'F'* level to eliminate variables resulted in equation 9. This equation consists of three *HPV* variables

TABLE 4. Microelements B and V as Discriminators of Recent Marine from Nonmarine Sediments*

Element	Varimax Rotated Factor Weight Matrix	
	(1)	(2)
B	0.984	0.176
V	0.176	0.984
	Factor Contributions to the LDF	Total
	0.30 0.24	0.54

* Data from *Potter et al.* [1963].

$$Z = -0.0004324(2KROH) + 0.0003086(2KRIH) - 0.004554(DDIAGO) \quad (9)$$

Equation 10 gives the corresponding *LDF* for the unrelated measurements method with a single variable retained for each of the four dimensions (height, outside radius, inside radius, and density)

$$Z = -0.01255\ (H) - 0.08096\ (D) - 0.0004915\ (KRORO) + 0.01011\ (RI) \quad (10)$$

The relative efficiencies of equations 9 and 10 can be compared in two different manners. First, solve both equations for all the original observations, and the equation that results in the fewest misclassified cylinders is declared to be 'Best.' Second, solve both equations for a new population of cylinders, and the equation that misclassifies the fewest new cylinders is then declared 'Best.' These two methods of picking the 'Best' *LDF* do not necessarily give the same answer. The 'Best' equation therefore depends upon the definition of 'Best' that is used.

For this study 300 new random hollow cylinders were generated, using the technique previously described [*Wallis*, 1965*a*]. Of these new cylinders 145 had weights below 133.24, whereas 148 had weights above 145.44. The remaining seven had weights between 133.24 and 145.44 and were excluded from subsequent analyses. (Listings for the program used to generate and the resulting 300 new hollow cylinders are contained in a file report. Requests for copies should be sent to the Director, Pacific Southwest Forest and Range Experiment Station, P. O. Box 245, Berkeley, California 94701.)

Comparisons of the effectiveness of equations 9, 10, and A 14-variable *LDF* (equation not given) were made for the original 60 and control set of 293 cylinders. The results, along with those for an 8-variable, 4-component reduced rank discriminant and the varimax factor score *LDF*, are given in Table 5. Loss of power resulting from including too many variables in the *LDF* is shown for the 14-variable *LDF* (all variable selection of Table 5). It is apparent that the stepwise method gives the lowest percentage misclassification for the original sample, but not for the control group.

EFFECT OF SAMPLE SIZE UPON DIFFERENT DISCRIMINANT FUNCTIONS

The ability to classify correctly a new sample is a property of the sample size used in formulating the discriminant function, as well as of the method used in making the discriminant function. Samples of three different sizes were made from the first 60 of the initial hollow cylinders. The first sample consisted of the cylinders numbered 1–40, the second of those numbered 41–60, and the third was a random sample of 10 (numbers 16, 8, 39, 25, 3, 4, 27, 26, 35, and 7). Stepwise and unrelated variable *LDF*'s were made for each sample. The ability of each *LDF* to classify correctly the 293 control cylinders was then determined. To test the effectiveness of *LDF*'s made from large initial samples, the process was then reversed by using 293 cylinders to generate the *LDF*'s, whereas the initial 60 were used as an independent test. Similar analyses were made by the reduced rank and factor score method (Figure 1).

TABLE 5. Percentage of Hollow Cylinders Misclassified by Five Different Discriminant Functions, on an Original Sample of 60, and on a Control Set of 293 Others

	Sample Percentage Misclassified by Each Discriminant Function				
	Linear Discriminant Functions				
	All Variable Selection	Stepwise Selection	Unrelated Measurement	Reduced Rank Method	Factor Score Method
Original 60	8.4	10.0	11.7	13.3	13.3
Control 291	50.5	24.7	15.5	20.5	35.0

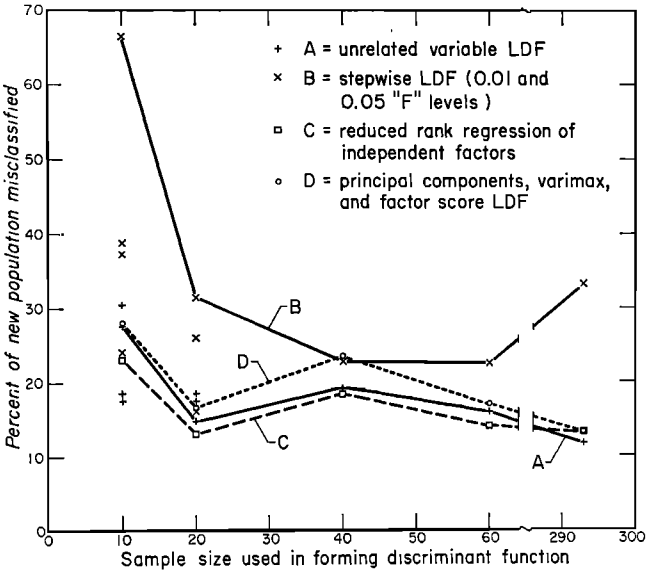

Fig. 2. Effect of original sample size and method of forming discriminant function upon the per cent of a control population misclassified.

The stepwise procedure with *'F'* levels set at 0.01 and 0.05 resulted in different choices of variables for each sample. The equations went from a single variable for the sample of 10 to an 8-variable equation for the population of 293. Furthermore, their ability to classify a new population correctly was also highly variable (line B, Figure 1). Additional stepwise *LDF*'s were obtained for three new populations of 10 cylinders and two new populations of 20 cylinders. The ability of each of these five new *LDF*'s to predict correctly the previously used population of 293 cylinders was determined, and the per cent misclassified was plotted on Figure 2.

When used for prediction, the unrelated variable *LDF*'s do not show extreme sensitivity to differences in the initial populations (Figure 1) or to differences in initial sample size (line A, Figure 1).

In summation, stepwise *LDF*'s tend to be poor predictors for new populations. The stepwise procedure is not recommended either for very small or for very large populations, or if composite variables exist among the choices. Unrelated variable *LDF*'s when applied to new populations tend to give fewer misclassifications and have smaller variability than stepwise *LDF*'s. There is some evidence to suggest that among the predictors tested in this paper unrelated variable reduced rank discriminators may be the most efficient.

REFERENCES

Anderson, T. W., *An Introduction to Multivariate Statistical Analysis,* 373 pp., John Wiley & Sons, New York, 1958.

Hodges, Joseph L., *Survey of Discriminatory Analysis,* 115 pp., USAF School of Aviation Medicine, Randolph Field, Texas, 1950.

Horst, Paul, *Factor Analysis of Data Matrices,* 730 pp., Holt, Rinehart, Winston, New York, 1965.

Kaiser, Henry F., Psychometric approaches to factor analysis, *Proceedings Invitational Conference on Testing Problems,* pp. 37–45, Educational Testing Service, Princeton, N. J., 1964.

Potter, P. E., N. F. Shimp, and J. Witters, Trace elements in marine and fresh-water argillaceous sediments, *Geochim. Cosmochim. Acta, 27,* 669–694, 1963.

Schultz, E. F., and J. F. Goggans, A systematic procedure for determining potent independent variables in multiple regression and discriminant analysis, *Auburn Univ. Agric. Expt. Sta. Bull. 336,* 75 pp., 1961.

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 ro-

tation of the factor weight matrix, *U. S. Forest Serv. Res. Note PSW-92,* 6 pp., 1965*b*.

Wallis, J. R., and H. W. Anderson, An application of multivariate analysis to sediment network design, *Int. Assoc. Sci. Hydrol., Publ. 67,* 357–378, 1965.

Weiner, J. M., and O. J. Dunn, Elimination of variates in linear discriminant problems, *Biometrics,* p. 258, March 1965.

(Manuscript received August 8, 1966; revised December 19, 1966.)