Selected multivariate statistical methods applied to runoff data from Montana watersheds
by Gary Lee Lewis

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE in Civil Engineering
Montana State University

Abstract:
A principal component analysis with varimax rotation of the principal factors was performed for watershed, storm, and runoff data from five central and eastern Montana watersheds. The analyses provided information about the relative importance of 29 independent variables to the peak discharge rates and runoff volumes produced by these variables.

Storm intensity, standard deviation of storm intensities , soil and air temperature, watershed azimuth, overland slope, watershed shape, reservoir area, and watershed area were among the most successively important variables. Correlations among some of the variables for the research watersheds were also indicated by the analyses. Principal component and rotated-factor regression equations for the runoff variables were developed, and are suggested as prediction equations for ungaged watersheds in central and eastern Montana.

184

SELECTED MULTIVARIATE STATISTICAL METHODS APPLIED TO RUNOFF DATA FROM MONTANA WATERSHEDS

by

Gary Lee Lewis

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

of

MASTER OF SCIENCE

in

Civil Engineering

Approved:

Head, Major Department

Chairman, Examining Committee

Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

August, 1968

ACKNOWLEDGMENTS

The author would like to express his sincere appreciation to all those who helped make this study possible. Special acknowledgment is extended to Professor T. T. Williams for his patient guidance and assistance during the problem selection, data collection, analyses, and manuscript preparation. Gratitude is also expressed to the undergraduate project assistants who helped with the data reduction, and also to Professor Williams for directing the assistants in the reduction of specific data for this study.

Financial support for the author was provided by a research assistantship granted by the Montana State Highway Department and the U.S. Bureau of Public Roads, and was administered by Professor Williams. The author is indebted to each for their generosity in providing the assistantship and computer time, without which the report could not have been possible.

Special gratitude is extended to the Committee members, Dr. Glen Martin and Dr. Tom Hansen for their careful examinations of the final draft and for their time and advice during the investigation.

The author's wife, Mrs. Gary Lewis, typed the many rough drafts and spent several sleepless nights typing

the final draft. The author's gratitude and appreciation for her patience is sincerely extended.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A principal component analysis with varimax rotation of the principal factors was performed for watershed, storm, and runoff data from five central and eastern Montana watersheds. The analyses provided information about the relative importance of 29 independent variables to the peak discharge rates and runoff volumes produced by these variables.

Storm intensity, standard deviation of storm intensities, soil and air temperature, watershed azimuth, overland slope, watershed shape, reservoir area, and watershed area were among the most successively important variables. Correlations among some of the variables for the research watersheds were also indicated by the analyses. Principal component and rotated-factor regression equations for the runoff variables were developed, and are suggested as prediction equations for ungaged watersheds in central and eastern Montana.

Chapter I

INTRODUCTION

In order to make reasonable predictions of peak discharge rates and total runoff volumes on small watersheds, an understanding of the factors causing the runoff is needed. If the important watershed and storm variables could be properly identified and measured, the relationships between these variables and the runoff could then be readily determined.

HISTORICAL BACKGROUND

Many attempts have been made to relate peak discharge rates and total runoff volumes from a watershed to their causative factors. The relationships which have been proposed usually take the form of graphs or equations relating the rates and volumes to factors that are believed to be important. The general trend in these methods is to choose the "important" factors, obtain measurements of each, and then relate the factors to the rates and volumes produced. However, the choice of factors is usually made from experience or judgment, and the proper or improper choice of factors leads to accurate or inaccurate results. A comparison of the methods indicates that there is considerable confusion as to which factors are to be used, and which

factors are more important than others. After analyzing several methods of discharge rate and runoff volume prediction, Sharp and Biswas (1965) wrote: "Exhaustive analysis of research data from small watersheds not only failed to reveal how various factors function in producing runoff, but failed even to reveal the parameters that should be used to estimate runoff." The wide range of factors used in prediction methods seems to substantiate this. Some factors appear in more of the methods than do other factors, but none of the methods agree on which set of factors are the most important.

PURPOSE AND METHODS

The study reported herein is an attempt to determine, for five central and eastern Montana watersheds, which of 29 factors are more important in producing peak discharge rates and total runoff volumes, and the regression equations relating peak discharge rates and runoff volumes to these factors. Multivariate statistical analyses are used to investigate the relative importance of each of the independent variables to the two dependent variables, peak discharge rate and total runoff volume. As explained in a later chapter, a principal component analysis of the correlation matrix of the independent variables is performed, and this is followed by a varimax rotation of the factor

weight matrix to determine which variables are most important. Regression equations for the dependent variables are found from the important variables, using the uncorrelated principal components and the rotated factor weight matrix.

Although only a few investigators have applied multivariate methods to hydrologic data, several have recommended their use in this field. The advantages which are presented in a later chapter indicate that they provide a significant improvement over some of the other methods used. The methods are employed here because they are well suited to the large volumes of data that have been generated on the watersheds being investigated. Because of the large amounts of data involved, multivariate analyses were not practical until the advent of the electronic, digital computer. Computer techniques were used extensively in the data reduction and analyses for this study.

SCOPE

Data for this study was taken from five small central and eastern Montana watersheds currently being studied for the Drainage Correlation Research Project, by the Department of Civil Engineering and Engineering Mechanics at Montana State University. The Drainage Correlation Research Project, which was initiated in 1963, is sponsored

by the Montana State Highway Commission and the Bureau of Public Roads, and is an investigation of the frequency of peak discharge rates for small watersheds in Montana. Data collection is expected to continue until September, 1969. Fifty runoff events on the five project watersheds, having peak discharge rates greater than 10 cfs, and occurring between April, 1964, and September, 1967, were studied for the investigation reported herein.

DEFINITIONS

Because certain terms and phrases are frequently used in this paper, several definitions are presented here. The definitions are those commonly used in the literature, and some are more thoroughly discussed in later chapters.

Dependent variable - the variable to be predicted from measurements of the independent variable(s) in a regression equation (e.g., peak discharge rate).

Independent variables - the variables on which measurements are obtained and substituted into the regression equation to calculate the prediction of the dependent variable (e.g., precipitation intensity, watershed area, etc.)

Regression equation - an equation for the dependent variable, derived from several measurements of this variable and the independent variable, or variables, in a manner which indicates the relationship of the independent variables to the dependent variable. (If more than one independent variable is involved, the equation is usually termed a "multiple regression equation")

Multiple linear regression equation - a multiple regression equation in which the dependent variable is

related to a sum of the independent variables, with each of the independent variables being multiplied by a different coefficient.

Coefficient of a variable - a constant to be multiplied by the measurement of a variable in a regression equation, component, or factor.

Multivariate studies - methods of studying more than two variables when the measurements of the variables are obtained simultaneously in time or space.

Linear correlation of two variables - a statistical measure of the closeness to a straight line of the graphical plot of measurements of both variables.

Components - Principal Components - Normalized Eigenvectors - derived, uncorrelated, independent variables written as the sums of the original independent variables, if each is multiplied by a coefficient.

Factor - a component whose coefficients on the independent variables have been multiplied by a constant. The squared coefficients within a factor total to the square of the constant.

Rotated factor - a factor whose large coefficients on the independent variables have been maximized.

Variate - a component, factor, or rotated factor.

Chapter II

LITERATURE REVIEW

In recent years a great many investigators have approached the problem of peak discharge frequencies from small watersheds, and have applied a great variety of techniques in analyzing the problem. Many of the investigators have used statistical methods, and several have applied multivariate analyses to hydrologic data. Reports of a large number of these investigations were reviewed in the course of the study reported herein. To discuss all the literature which was reviewed would result in an extremely voluminous and unwieldy report. It would seem more appropriate, therefore, to confine the review herein to the results of a few of the more relevant studies which bear directly on the present investigation.

INVESTIGATIONS WITH MULTIVARIATE ANALYSES

Prior to 1950, multivariate methods were well established, but were not practical in hydrology because of the time-consuming computations. However, as the high-speed computer became more accessible to investigators, the methods were recognized as a possible means of studying the peak discharge rate and runoff volume prediction problem. For example, Wong (1963) used multivariate methods to

analyze data from 90 basins in New England. The basins ranged in area from 10 to 2000 square miles, and measurements of eleven independent variables were taken on each of these basins to determine a regression equation for the mean annual flood with a recurrence interval of 2.33 years.

Wong found that a previous "ordinary" multiple regression on five of the eleven independent variables was not satisfactory because the variables, average land slope, mean altitude, tributary channel slope, stream density, and shape of basin were, according to Horton's Laws, "multicollinear," meaning that some were linearly related to others and should not be included. After obtaining measurements on six additional variables, drainage area, main channel slope, tributary channel slope, percentage of area in ponds and lakes, length of longest watercourse, and precipitation intensity; a principal component analysis and varimax rotation were performed on the data. This resulted in two new, unrelated variables or "components." Both of these components were linear functions of all the measured variables, but the new components were not linearly related to each other. Also, the first component was found to be significantly more important than the second, and each component was found to be more highly associated with certain variables. The first had large coefficients

on variables which expressed the area and length of the drainage basin, and the second was associated with the slope and topography of the drainage basin. Both components were about equally associated with mean annual flood, indicating that both were important to the dependent variable.

Because the two components indicated that size and length were not related to slope and topography, Wong decided that these two parameters would form a good set of independent variables for a multiple regression equation. He examined the correlations of mean annual flood with the components and the eleven independent variables, and found that the length of the main stream, L, was highly correlated with both the mean annual flood and the first component; and that the average land slope, S, was similarly related to the second component and mean annual flood. These two variables were consequently chosen for a multiple regression, giving:

$$\text{Log } Q_{2.33} = -1.02 + 1.29 \text{ Log } L + 0.97 \text{ Log } S$$

for the regression equation. This equation had a coefficient of determination of 0.80, which meant that 80 per cent of the variation in the mean annual flood could be explained by only two variables instead of eleven. The previous

regression on five independent variables had the same coefficient of determination, but involved linearly related variables. The study, therefore, reduced the number of "independent" variables needed to explain the same variation in mean annual flood for New England.

Eiselstein (1967) performed a similar analysis, although he was interested in runoff volume instead of mean annual flood rate. A 350-acre watershed was divided into 17 runoff plots on which data from 30 variables was obtained over a period of four years. The variables were grouped into five categories, storm variables, antecedent moisture variables, site variables, soil description variables, and the dependent variable, runoff in inches from each plot.

Because Eiselstein was aware of the inadequacy of ordinary multiple regression to provide a good prediction equation when the independent variables are not truly independent, he performed three separate analyses to show the different results that can be obtained. An ordinary linear regression analysis gave an equation in terms of all 29 "independent" variables, and 13 regression coefficients were found to be statistically significant, i.e., non-zero. This equation accounted for 77 per cent of the variation in the runoff volume, but Eiselstein was

not satisfied with the results because the non-significant variables had high coefficients of correlation with each other and with the significant variables. Also, the test for significance was not valid because correlated variables were used. This meant that the "significant" variables were a combined measure of several variables, and the "non-significant" variables could not honestly be discarded.

To attempt to separate the combined effects of several variables to the effect of each, a principal component analysis of the correlation coefficients of all combinations of the independent variables was performed. This resulted in 29 new, independent variables, or "components," which were each linear functions of all the 29 original "independent" variables. The first of these components had high coefficients in the rainfall variables, but the remaining components could not be readily associated with specific variables. This is the reason for the varimax rotation, which rotates the components to another set of reference axes so that only high or low coefficients exist on the original variables, and a better interpretation can be made.

Before rotating the "principal components," Eiselstein computed values for each component by using the original

variable data to solve the linear equations. This resulted in a numerical value of each component for each runoff event. Because the components were truly independent, a multiple regression on these values was performed. After the regression coefficient for each component was found, 18 of the 29 coefficients exhibited significance. Also, the regression equation explained 77 per cent of the variation, which was exactly the same amount explained by the ordinary regression equation. This analysis gave a good prediction equation because truly independent variables were used. However, because each significant component was a linear function of all 29 original variables, nothing could be stated about the importance of the original variables at this point in the analysis.

To further investigate the separate variables, Eiselstein's third analysis consisted of a multiple regression on rotated components instead of the original components. An initial, orthogonal variamax rotation of the original components yielded a set of components which all had low coefficients on 12 of the original variables. Because these variables were not significant to any of the rotated components, they were deemed unimportant to the runoff and discarded. Also, only the first seven components were deemed to be important because the rest each

contributed less than one per cent to the variation of the dependent variable. This resulted in seven linear equations for 17 of the original variables, giving essentially the same information as the first 29 components and variables.

A second varimax rotation of the seven components gave seven new components which could be interpreted in terms of the 17 remaining independent variables. One of the components accounted for 51 per cent of the variation in runoff, and had high coefficients on precipitation intensity and total precipitation. The second most important component accounted for 5 per cent of the variation in runoff and had high coefficients on slope, elevation, and "aspect." The third component, accounting for four per cent of the variation in runoff, had high coefficients on surface soil properties. The other components each explained only a small portion of the variation in runoff, and could not be associated with specific variables. The important interpretation from the components was that the rainfall variables were much more important to runoff than the aspect and soil properties, because they accounted for 51 per cent of the variation in runoff. Also, the variables within each component were linearly dependent and their combined effect was independent of the combined effect of the variables of other components. This meant that a

multiple regression on the components, or on certain variables from each component, would be a regression on truly independent variables. (Wong had made this same observation in 1963 and had written his final regression equation in terms of two nearly independent variables, one from each important rotated component. Other variables were significant to the components, but the two he chose were combined measures of those in each component.)

Eiselstein chose to find a multiple regression equation for all 17 of the variables in the seven components, rather than select one variable from each component to represent the total component. After substituting data from the original variables into the component equations, a multiple regression of the values computed gave an equation for runoff in terms of the components, and hence in terms of the 17 independent variables, because the components were linear functions of the variables. The coefficients of this final equation seemed to be realistic because no obvious fallacies in the signs of the coefficients could be detected. For example, the rainfall characteristics were directly and not inversely related to runoff as is the case in some ordinary multiple regression equations (Sharp, Gibbs, Owen, and Harris, 1960; Wallis, 1965). The final coefficient of determination was 0.67 indicating

a 10 per cent loss of information as a result of reducing 29 variables to 17, and 29 components to 7.

Snyder (1961) performed a principal component regression analysis relating total December runoff from a watershed to the rainfall in October, November, and December. A previous, ordinary multiple regression equation had negative coefficients on all the independent variables indicating that runoff increased as the monthly rainfall decreased. Because this was "intuitively" inaccurate, Snyder obtained three independent components from a principal component analysis, and derived a regression equation for these components, and hence for the variables, and found that all coefficients were positive. A reduction in the coefficients of determination from 0.83 to 0.75 for the principal components solution resulted, but Snyder felt that the loss in information about the variation in runoff was justified by the intuitively correct coefficients. This study was used to illustrate the possibilities for multivariate analyses in hydrologic studies, and no rotation of the principal components was made, because the component regression equation was deemed to be satisfactory.

The three above investigations used essentially the same techniques. In all of them, the ordinary multiple regression solution was unsatisfactory, and multivariate

techniques were used to overcome this difficulty. Each used a principal component analysis to obtain truly independent variables for the regression. Also, rotation of these components was used to provide information on the importance of certain variables in two of the papers.

RECOMMENDATIONS OF THE LITERATURE

The previous section discussed some of the applications of multivariate analyses. A short presentation of several recommendations found in the literature follows. The purpose here is to present the opinions of several authorities on the use of multivariate methods with hydrologic data.

Eiselstein's investigation was undertaken as a result of a paper by Wallis (1965), which suggested the advantages of multivariate methods over ordinary multiple regression for hydrologic studies. Wallis compared the presently used methods of obtaining an equation for the dependent variable in terms of several independent variables, and made several recommendations. First, he suggested that a linear-logarithmic transformation model, similar to Wong's, be used with hydrologic data. This should be followed by a principal component regression analysis with varimax rotation of the principal components for an initial analysis

of "multifactor hydrologic problems." These recommendations were made after a study of the adequacy of the methods in obtaining a known functional relationship, the weight of a solid cylinder in terms of its density and dimensions.

Wallis (1968) presented several other suggestions for the best utilization of multivariate statistical methods in hydrologic studies. First he suggests that no more than two variables important to any factor in the rotated factor table be retained for further study. After selecting the retained variables, Wallis suggests that a complete principal component analysis with varimax rotation of the principal factors be performed on the original measurements of only these variables. A regression on these factors is then suggested for the prediction equation.

After his analysis of 90 New England basins, Wong (1963) stated that, "multivatiate methods should be more widely encouraged in geomorphic and hydrologic research" (p. 198). Eiselstein recommended that, "a principal component analysis with varimax rotation of the factor weight matrix is a suitable statistical technique for the correlation of small watershed surface characteristics with surface runoff" (p. 484). Although Snyder did not use or recommend varimax rotation, he did state that principal component

regression analyses give "logical equations" for runoff when compared to ordinary multiple regression techniques.

METHODS APPLIED TO MONTANA WATERSHEDS

Multivariate methods have not been previously applied to small watersheds in central and eastern Montana. Boner (1963) presented a report of his study of the frequency and magnitude of floods in eastern Montana for the United States Geological Survey. This report presents an ordinary multiple regression equation for mean annual flood in terms of area, stream meander length, geographical region, and elevation of small watersheds in eastern Montana. The dependent variable was found to be directly proportional to the first three variables, and inversely related to elevation. Also, the flood having a recurrence interval, I, is obtained by multiplying the mean annual flood by a factor from a "composite flood frequency curve" for that interval. Variables other than those listed were not used because measurements were not available.

Boner and Omang (1967) presented a report on the magnitude and frequency of floods from watersheds smaller than 100 square miles in area in Montana. This report gives a method of obtaining the floods with recurrence intervals of 10 and 25 years for this region. The 10-year flood is

found from a regression equation involving the area, elevation, channel slope, and mean annual runoff of a watershed, and the 25-year flood is obtained from the 10-year flood upon multiplication by an empirically determined constant.

Both of the above studies employed ordinary multiple regression techniques with only four easily determined independent variables.

The studies of Wong, Eiselstein, and Snyder used more variables than four, but the developed equations could not reasonably be used in Montana. However, the methods should be applicable to any region. None of the literature reviewed indicates that attempts have been made to determine equations for both the peak discharge rate and the runoff volume in a single analysis.

Chapter III

THEORETICAL DEVELOPMENT

In this chapter, a survey of the available methods of analyzing runoff from small watersheds using measurements of several related variables is presented, followed by a discussion of the reasons the specific methods for this analysis were chosen. A complete theoretical development of the methods chosen concludes the chapter.

POSSIBLE METHODS OF ANALYSIS

Ordinary multiple linear regression is one of the few statistical methods of simultaneously analyzing several variables to estimate one or more of them. The analysis is simply an analytical method of plotting a line, plane, or hyperplane through a multitude of data points. In general, a linear equation with an unknown intercept and slope is assumed, and the intercept and slope are calculated from the data in a manner which minimizes the squared distances from the points to the line, plane, or hyperplane. Until recently, ordinary multiple linear regression of the logarithms of the variables has probably had the most use in predicting runoff from small watersheds (Sharp, et al, 1960).

The greatest objection to the use of multiple regression analyses with hydrologic data is that certain basic assumptions of multiple regression theory are violated. Multiple regression assumes that there are no correlations among the independent or "predictor" variables (Thurstone, 1947). Hydrologic variables, as shown by Wong, usually violate this assumption. If ordinary multiple regression is attempted, the coefficients of the multiple regression equation for the dependent variable have sometimes been found to be "absurd" and "grossly in error" (Thurstone, 1947, p. 61).

Many regression equations for runoff have been derived using many combinations of independent variables. Snyder (1962) studied some of these equations and concluded that multiple regression analyses do not yield "logical equations" when used in hydrology. In these cases, he was referring to the coefficients associated with each of the independent variables and not necessarily with the accuracy of the prediction equation. A presentation of the mechanics of multiple linear regression was given by DuBois (1957), and Baggaley (1964).

Another statistical approach to the problem of investigating several variables was developed by Harris, et al (1961). Their purpose was to present a method of

selecting the most important independent variables, and to use only these variables in the regression equation. A Taylor series expansion was used to determine successively the most important variables by initially removing the effects of the other variables. The method is an analytic approach to "graphical curvilinear multiple regression," and eliminates the usual "shotgun" search for important variables. The method provides a statistical means of determining the important variables, but because hydrologic variables are generally correlated, as discussed earlier, attempts at multiple regression are often not successful.

Wong (1963) discusses several other possible methods of analyzing runoff in terms of several independent variables. One of these, "stepwise multiple regression," is similar to Harris' analysis because the variable which contributes most to the variation in the dependent variable is determined. Its effects are then removed, and the second most important variable is found. This process is repeated until enough variables are found to account for all or most of the variation in the dependent variable, and a multiple regression is performed on these variables. Again, the relative importance of the independent variables is indicated, but the multiple regression

assumption may be violated. Ralston (1960) presents a development of stepwise multiple regression, and outlines the procedures for programming the method for digital computers.

Another method of handling this problem is by using multivariate statistical analyses. Several of these are outlined in the next section, along with the advantages and disadvantages of each. The methods chosen for use in this investigation are indicated, and the reasons for the selection are outlined.

MULTIVARIATE STATISTICAL METHODS

One commonly used multivariate statistical analysis is known as "component analysis." There are two varieties of component analysis, known as "principal component analysis" and "centroid analysis." Both are mathematical means of obtaining new variables or "components" from the intercorrelations of the chosen independent variables. The objective in the analysis is that the components obtained will, (1) be fewer than the number of independent variables under study, (2) represent or reproduce the original variables, (3) account for all the variation in the original variables and, (4) be uncorrelated even if the original variables were highly correlated with each other. Principal

component analysis extracts the new variables, or "variates," one by one. The first component is extracted in such a manner that it reproduces a maximum amount of the information in the original data. The second component accounts for a maximum amount of the information remaining after the extraction of the first component. This process is repeated until all the components have been extracted, and all the original information in the data is reproduced by the components.

Centroid analysis has been termed a "simplified approximation of the principal components solution" (Cooley and Lohnes, 1962, p. 153), and is used to avoid the involved calculations of a principal component analysis, namely, the solution of the "characteristic equation," defined later. Kendall (1957) gives an example of the "approximate" centroid solution compared to the principal component solution of the same problem, and shows the different solutions obtained. When computer availability obviates most concern for involved computations, the principal component analysis is the better method.

"Factor analysis" is a multivariate method similar to component analysis, but differing in the general method of approaching a problem. A factor is defined as a linear equation in terms of all or some of the original variables,

but is not the same as a component. Kendall (1957) states that component analyses attempt to proceed from the data to a model, while factor analyses begin with a model and investigate its agreement with the results. In factor analyses, an investigator examines the data and speculates on how many factors or groupings of variables might be present. Guilford (1952), in a discussion on when to factor analyze, illustrated this point by writing: "The initial planning should emphasize the formation of hypotheses as to what factors are likely to be found in the selected domain and to the probable properties of such factors" (p. 36). For example, if the relationship of peak discharge rate to several variables such as soil moisture, air temperature, wind speed, watershed area, storm duration, stream channel lengths, excess precipitation intensity, soil permeability, etc., is desired, then the variables might initially be viewed as being composed of two factors or variable "groupings," one of climate, and one of watershed characteristics. The results of the factor analysis are two or more factors, and an examination of the loadings of the factors will reveal if the two-factor assumption was correct. In hydrology, factors might easily be formulated, although none of the reviewed papers employed this method. Eiselstein (1967) grouped the original independent variables into four

categories, but he did not initially state that he expected a factor for each category. His first principal component could have been termed a "rainfall" factor, but the remaining components were not readily associated with the categories. His purpose, that of principal component analyses, was to derive new, truly independent variables for a multiple linear regression. Components are independent, and the combined effect of the variables within each component is independent of the effect of other components. This means that the important variables within a component are not necessarily from the same category, as is hopefully the case with factor analyses. No initial assumptions about the outcome are made with component analyses. The data is analyzed only for uncorrelated components, and the model differs from that of a factor analysis in the above manner. If factors result, they are coincidental, but are desirable because they give information about the importance of groups of variables.

Neither a factor analysis nor a component analysis provides information about the importance of each independent variable. The coefficients on the variables are correlations of the variables with the factors or components, but they generally are relatively large in magnitude for all the variables. If one or more variables in a component

analysis has a small correlation with all the components, then the variable is not important to any of the components, and hence to the total problem. Because the first component is found in a manner which yields high correlations with all the variables, no interpretations can be made, and the components are useful only when regression of uncorrelated variables is desired.

Because a principal component analysis does not usually provide information about the importance of each independent variable, another multivariate method, "rotation of the principal components," is in general use. The components can be rotated so that each component has high coefficients on certain variables and low coefficients on other variables, allowing interpretations for the important variables, and obtaining the "simple structure" of the components (Matalas, 1967). Graphical rotation presentations are given by Fruchter (1954) and Baggaley (1964). Both authors use two-dimensional plots which show the measurements of the variables as points, and rotated components are simply lines drawn through clusters or "streaks" of points so as to maximize zero-loadings on the components. However, the graphical solutions are approximate, and different investigators might obtain different results with the same data. Kaiser (1958, 1959), derived an

analytic rotation which guarantees the same results for different investigators. This method is known as "varimax rotation," and has been utilized in the literature (Rice, 1967; Wallis, 1965; Wong, 1967; Eiselstein, 1967).

Rotation of the variates from either a component or factor analysis has no effect on the amount of information retained by the variates. Only the interpretation of the loadings is affected (Cooley and Lohnes, 1962). Kaiser's "normal" varimax rotation not only minimizes the "in-between" loadings, but it also maintains an orthogonal or perpendicular reference frame (Wallis, 1965; Kaiser, 1958). This feature is desirable if a multiple regression on the rotated components is to be performed, because perpendicularity of the components means that they are uncorrelated, and the assumptions of multiple regression are not violated. "Oblique," or non-perpendicular analytical rotations such as the "Quartimin," "Oblimin," or "Covarimin" have been developed, but do not appear satisfactory (Cooley and Lohnes, 1962).

Multivariate methods are advantageous in several aspects. Ordinary multiple regression provides good prediction results, but gives no insight into the interrelationships of the variables. Multivariate methods obviate

the effects of highly correlated "independent" variables, while ordinary multiple regression analyses do not. Interpretations of the results of multivariate analyses allow the exclusion of unimportant variables and the recognition of the more important variables.

Multivariate methods also allow the reduction of the number of variates for multiple regression. Component analyses produce exactly as many new variates as there are independent variables. However, since the extraction is done on a "priority" basis, some of the latter variates may reproduce only a small portion of the information, and may be excluded. Wong (1967), Eiselstein (1967), and Rice (1967) were all able to considerably reduce the number of variates needed to reproduce almost all of the information present.

Besides allowing the interpretation of the importance of each variable to the original data, the orthogonality of the new variates is a principal achievement. Because the data can be expressed by uncorrelated variates, then multiple regression assumptions are not violated if the regression is performed on these variates. The new independent variates reproduce the original data and are truly uncorrelated. Because of this, the correlations among the independent variables are indicated with multivariate

methods by producing uncorrelated variates.

Still another advantage of multivariate methods with correlated variables is the improvement in the final regression equation coefficients. Snyder (1962) stated that multivariate methods yield "nice" coefficients which indicate the relative importance of each independent variable to the criterion. Wallis (1965) demonstrates that principal component regression coefficients tend to be "stable" when compared to ordinary regression coefficients. In a discussion of Rice's paper, Anderson (1967) stated that the "coefficients remained very distinctly realistic with regression on principal components" (p. 6). This, and the other advantages of multivariate methods seem to provide justification for their use in hydrologic studies.

Investigators of the many multivariate methods agree that the best statistical system of analyzing hydrologic data would start with a principal component extraction, followed by a varimax rotation for interpretation of the variables, and a multiple regression on either the principal components or the rotated factors for the prediction equation (Eiselstein, 1967; Wallis, 1965; Wong, 1967; Anderson, 1967). Baggaley (1964), however, states that good rotation results are not likely unless more than 20 variables are involved. Because 29 independent variables were available