# REGRESSION ON PRINCIPAL COMPONENTS AS A TOOL IN WATER SUPPLY FORECASTING 1/

By

Michael A. Marsden 2/ and Robert T. Davis 3/

Water supply forecasters seek to predict the amount of water that will become available at future dates.

In Washington, a major source of the water supply is precipitation in mountain areas. This supply is tapped at points on the major rivers (Columbia and tributaries) often far removed from the mountain source. Much of the supply forecasting is based on the premise that a large part of the summer flow of the major river system originates from the winter precipitation in the mountains. Since this source, wintertime precipitation in the mountains, can be estimated from a variety of measurements, it is used to predict the summer water supply.

Types of measurements of precipitation (inches of rainfall) are total precipitation (recorded at several locations in the drainage basin), seasonal precipitation (spring, fall, winter), breakdown of precipitation by elevation or temperature zone, and snow water equivalent.

Streamflow records accumulated during the summer months provide measurements of actual summer water supply in past years. From these records, a mathematical relation can be established between measurements of winter mountain precipitation and summer water supply. Once established, this equation might be used to predict the summer water supply each year when the winter precipitation records are available. In the past, such equations have been based on multiple regression analyses of the original measurements. This paper presents a method that includes weighted values of all the independent variables--a regression on principal components.

In principal component analysis, the major effects in the system of independent measurements are summarized to form a smaller set of artificial variables or principal components (P.C.'s). These components are weighted summations of the independent variables. The first principal component is the weighted sum of the independents that will explain the maximum possible amount of variation among the independents. The second principal component explains the largest percent of the remaining variation, after the effect of the first component has been removed. Principal components are therefore calculated in descending order of importance, and each is orthogonal, i.e., completely independent of the others.

The development of the prediction equation by means of regression on principal components may be illustrated by the following example in two dimensions ( only 2 independents). From $x_1$ and $x_2$ a variance-covariance matrix is calculated (for method, see Appendix). The weighted sum of $x_1$ and $x_2$ that can explain the maximum amount of variation in this matrix is the first P.C. The remaining variation is explained by the second P.C.

---

1/ Presented at Western Snow Conference, Lake Tahoe, Nevada, April 16 - 18, 1968

2/ Statistician, U. S. Department of Agriculture, Forest Service, Intermountain Forest & Range Experiment Station, Ogden, Utah 84401; stationed at Forest Sciences Laboratory, Moscow, Idaho, which is maintained in cooperation with the University of Idaho.

3/ Snow Survey Supervisor, Soil Conservation Service, Spokane, Washington.

| Observation No. | $x_1$ | $x_2$ | 1st P.C. scores | 2nd P.C. scores | |
|---|---|---|---|---|---|
| 1 | 1 | 2 | .374 | .316 | 1st P. C. = $.044x_1 + .165x_2$ |
| 2 | 2 | 5 | .913 | .438 | 2nd P. C. = $.704x_1 + .194x_2$ |
| 3 | 3 | 5 | .957 | 1.142 | |
| 4 | 2 | 6 | 1.078 | .244 | |
| 5 | 3 | 6 | 1.122 | .948 | |
| 6 | 2 | 1 | .253 | 1.214 | |
| 7 | 2 | 4 | .748 | .632 | |
| 8 | 2 | 3 | .583 | .826 | |
| 9 | 1 | 1 | .209 | .510 | |

The next step in developing this forecast system is to regress the dependent variable on the principal component scores.

For our example the dependent values and the regression on the first principal component follow:

| Observation number | Dependent variable |
|---|---|
| 1 | 1 |
| 2 | 7 |
| 3 | 6 |
| 4 | 7 |
| 5 | 9 |
| 6 | 1 |
| 7 | 6 |
| 8 | 4 |
| 9 | 1 |

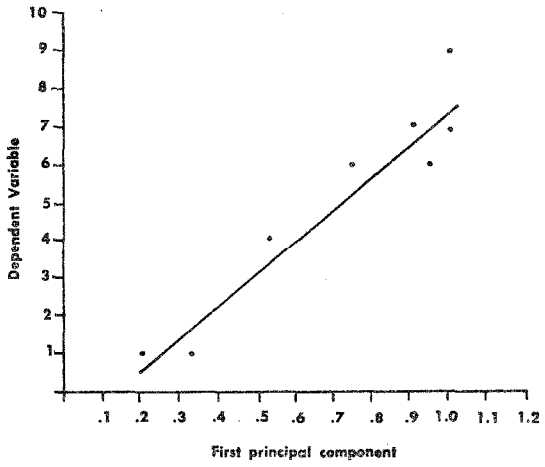
Calculating the regression on the first principal component, we get

$$\hat{y} = -1.133 + 8.369 \quad \text{1st P.C.}$$

Rewriting the equation in the form of the original independents, we get the equation

$$\hat{y} = -1.133 + .368x_1 + 1.381x_2$$

Figure 1.



Now let us apply this method to the water supply in the Yakima River at Cle Elum and Kettle River at Laurier. Local weather variables are represented by total precipitation at each weather station within the respective river basins and are recorded for each of three seasons--fall, winter, and spring. Snow survey data are in inches of water equivalent, recorded on the April 1 survey. For the Kettle River there were five weather stations and four snow survey locations. For the Yakima River there were four weather stations and five snow survey locations. The total number of independent variables measured each year is therefore 19 (3X5 +4) for the Kettle and 17 (3X4 +5) for the Yakima. Data were available for the 21 consecutive years, 1945-1965, on the Yakima, and these years were used in calculating the prediction equation.

The Kettle data could be summarized well by the first 10 principal components, which explained more than 99 percent of total variance among the independents. We used all 17 principal components in the regression analysis of the Yakima River data. Complete analysis of Yakima River data is given in the Appendix.

The following regression equations were computed on the principal components, and nonsignificant coefficients were set equal to zero. The corresponding models for regression on the unaltered independent variables are presented for comparison.

KETTLE

Regression on principal components

$R^2$ = .869 S.E. = 149.08 d.f. =21

Constant = -887.23

| Principal component | Coefficient |
|---|---|
| 1 | -19.9534 |
| 2 | -63.2391 |
| 3 | -73.1850 |
| 4 | -27.7285 |
| 6 | -56.0534 |
| 10 | 81.1782 |

(NOTE: Coefficient times component weight will determine influence of an of an independent observation. Many of the component weights for the Kettle were negative.)

Regression on independents

$R^2$ = .849 S.E. = 153.02 d.f. = 22

Constant = -556.83

| Independent variable | Coefficient |
|---|---|
| 2 | 70.9710 |
| 10 | 107.0263 |
| 15 | 176.7225 |
| 16 | 57.5350 |

YAKIMA

Regression on principal components

$R^2$ = .960 S.E. = 52.70 d.f. = 15

Constant = 228.72

| Principal component | Coefficient |
|---|---|
| 1 | 5.8901 |
| 2 | -2.3157 |
| 3 | 4.9418 |
| 8 | -16.7343 |
| 13 | 84.5690 |

Regression on independents

$R^2$ = .969 S.E. = 46.10 d.f. = 15

Constant = -115.64

| Independent variable | Coefficient |
|---|---|
| 1 | 11.2837 |
| 8 | 8.0651 |
| 12 | 38.2812 |
| 13 | 6.7312 |
| 14 | 6.3192 |

Estimates made from this forecast equation for the years 1945-1965 are given in figure 2 (Appendix).

Why use regression on principal components?

Usually there is correlation among the independent variables. High correlation among the independent variables would result in computation errors or singular matrix if regression on these variables is attempted. These problems can be eliminated either by excluding from the model terms that are highly correlated among themselves or by taking principal components of the original variables. Computing regression on these components does not present the problem of intercorrelation because the components are orthogonal. Each component contains some weighted effect of each of the independents; therefore none of the information is excluded completely prior to regression calculations.

From the comparison of regression on principal components and regression on the independents directly, small differences in precision (standard errors) can be found. The equations from regressions on the independents directly contain only one snow course variable for the Kettle and only one for the Yakima. Water supply predictions would thus depend on only one snow course for the information on snow contribution to water supply. The equations from regressions on principal components contain weighted effect of all snow course measurements.

Principal components should be computed on variables of the same scale or of a common scale, since the measurement of any one variable on a scale different from the others would affect the weights assigned in the calculation of P.C.'s.

Suppose that an inch of snow-water equivalent is not the same as an inch of precipitation; then we would have two different scales of moisture measurement. To avoid any effect on the weighting, independent variables could be treated as two separate groups: snow course measurements and precipitation records.

To illustrate, the Kettle data were so divided, and principal components were calculated for each. Regression analysis was then computed on the resulting two sets of P.C. scores. Nonsignificant coefficients of the regression were set to zero. (The following equation is only an example of a way to handle differences in scale; it is not meant to replace the previous Kettle equation.)

KETTLE

$R^2$ = .794 S.E. = 183.01 d.f. = 21

Constant = -680.40

| Principal component | | Coefficient |
|---|---|---|
| Snow course | 1 | $.34615245X10^{-1}$ |
| Precipitation | 1 | $.15484914X10^{-16}$ |
| | 2 | $-.10026301X10^{-15}$ |
| | 4 | $.24559694X10^{-16}$ |
| | 6 | $-.53816315X10^{-7}$ |

To facilitate use of regression on principal components, the computer program (P COMP) used is available from the authors.

Water supply forecasters need an analytical method to calculate water supply prediction equations. Regression on principal components will generate prediction equations superior to those generated by regression directly on independents. The improvement is achieved by including more information from the set of independent variables and by weighting this information before regression. Fewer degrees of freedom are removed when the regression is calculated. The resulting equation is more stable, for it is dependent on all the independent variables, not on a small subset of them.

## APPENDIX

### Details of Analysis of Yakima

The principal components are calculated from the matrix of independent variables. In particular, the variance-covariance matrix of independent variables is rotated into a set of n vectors, n being the rank of the matrix. These are the latent vectors, which are orthogonal, and together they account for all the variation found among the independent variables. These vectors are arranged in order of the amount of total variation that they explain, their latent root. The largest percent of the total variation is accounted for by the first root and vector.

Runoff can be regressed on these vectors starting with the largest. Only one degree of freedom is associated with each vector. Each vector contains components (weighted scores) of all the original independent variables. The model for prediction of runoff is then derived in two steps.

First, variance-covariance matrix is computed from the original independent variables, namely, precipitation and snow water equivalent.

$$X_{nn} = X'_{np} X_{pn} .$$

In the matrix element, $X_{pn}$, n is the number of original variables and p is the number of observation sets. This matrix X is therefore of rank n, for $n \leq p$.

Example: Data, first observation set (j=1) the $x'_{i1}$ (i = 1,2,...n) would be

| | |
|---|---|
| 1.92 | Fall precipitation Sta. 1 |
| 26.41 | Winter precipitation Sta. 1 |
| --- | |
| --- | |
| --- | |
| 70.6 | Snow water equivalent, April 1, Cayuse Pass |

The principal component analysis will calculate a matrix U to transform the matrix X into an orthogonal matrix Z,

$$Z_{nn} = U'_{nn} X_{nn} .$$

The individual $Z_{ij}$'s are the P.C. scores. The vectors in Z ($Z_{ni}$) are not only orthogonal but are so arranged that $Z_{n1}$ has the maximum variance of all columns in Z. Here $Z_{n2}$ is chosen such that its variance is larger than all other columns except $Z_{n1}$. The matrix is therefore composed of orthogonal vectors arranged in descending order of size of variances.

The vectors in U which define each of the vectors in Z are the latent vectors of the matrix $X_{nn}$.

Example: The largest root of the Yakima data would be used to calculate a new variable $Z_{ij}$:

$$Z_{1,1} = U'_{1,n} X_{1,n}:$$

$$96.30 = .005653(1.92) + .1915(26.41)...+ .6250(70.6).$$

The Z's are used as independent variables in a regression analysis on runoff data.

YAKIMA

Regression model for runoff

on principal components of precipitation and snow water equivalent

| | |
|---|---|
| Standard error of regression | 52.70 |
| Coefficient of determination | .9598 |
| Constant | 288.72 |

| Principal Component | Coefficient |
|---|---|
| 1 | 5.8901 |
| 2 | -2.3157 |
| 3 | 4.9418 |
| 8 | -16.7343 |
| 13 | 84.5690 |

The above model is in terms of the Z's. To put this equation in terms of the original variables, x's, further computations are necessary.

$$\hat{Y}_i = \sum^{k} B_k Z_{ki};\ Z_{ki} = u'_{ki} x_{ki};$$

$$\hat{Y}_i = \sum^{k} B_k (u'_{ki}\, x_{ki});\ = \sum^{k} B_k \sum^{j=1,n} u'_{kj}\, x_{ij}\ ,$$

for k = 1,2,3,8,13.

The model could then be wtitten in terms of x's.

$$\hat{Y}_i = \sum^{j=1,n} x_{ij} \sum^{k} B_k\, u'_{kj}\ .$$

Let $A_j = \sum^{k} B_k u'_{kj}$ .

$$\hat{Y} = \sum^{j=1,n} A_j\, x_{ij}\ .$$

$$A_1 = \sum^{k} B_k\, u'_{k1} = \begin{array}{l} 5.8901\ (.005653) \\ -2.3157\ (.01848) \\ 4.9418\ (-.06853) \\ -16.7434\ (-.03523) \\ 84.5690\ (.07294); \end{array}$$

$$A_1 = \underline{6.4099}.$$

Therefore $\hat{Y}_i = B_0 + \sum^{j=1,n} x_{ij} A_j$ is equavalent to the model of the regression on principal components but is expressed in terms of the original variables, x's.

For the Yakima data the model in terms of A's has been calculated as follows:

| | |
|---|---|
| 228.72 | A(0) = B(0), constant term |
| 6.4099 | A(1), Fall precipitation, Sta. 1 |
| 4.9076 | A(2), Winter " " |
| -14.9383 | A(3), Spring " " |
| .9913 | A(4), Fall precipitation, Sta. 2 |
| 16.7492 | A(5), Winter " " |
| -44.4634 | A(6), Spring " " |
| -8.0534 | A(7), Fall precipitation, Sta. 3 |
| 7.3731 | A(8), Winter " " |
| -2.6743 | A(9), Spring " " |
| .0407 | A(10), Fall precipitation, Sta. 4 |
| -16.3964 | A(11), Winter " " |
| 64.8995 | A(12), Spring " " |
| 5.6907 | A(13) Tunnel Avenue Snow water April 1 |
| 3.6242 | A(14) Olallie Meadow " " " |
| 15.0851 | A(15) Big Boulder Cr. " " " |
| -8.2901 | A(16) Fish Lake " " " |
| .7832 | A(17) Cayuse Pass " " " |

Prediction of water supply can now be made for all past observations using the A's. This could have been done from the relationship Y = BU'X, but since predictions for the future can be handled most easily from the equation Y = AX, these predictions will be made this way also.

Example: Runoff for 1945 predicted by equation

$$\begin{aligned} \hat{Y} = 228.72 &+ 6.4099(1.92) + 4.9076(26.41) - 14.9383(4.07) \\ &+ .9912(1.44) + 16.7492(21.54) - 44.4634(2.80) \\ &- 8.0534(2.66) + 7.3731(31.08) - 2.6743(5.27) \\ &+ .0407(3.27) - 16.3964(35.4) + 64.8995(6.34) \\ &+ 5.6907(16.8) + 3.6242(35.4) + 15.0851(12.6) \\ &- 8.2901(23.2) + .7832(70.6) \end{aligned}$$

$\hat{Y} = 738.41$. Actual runoff was 714.55.

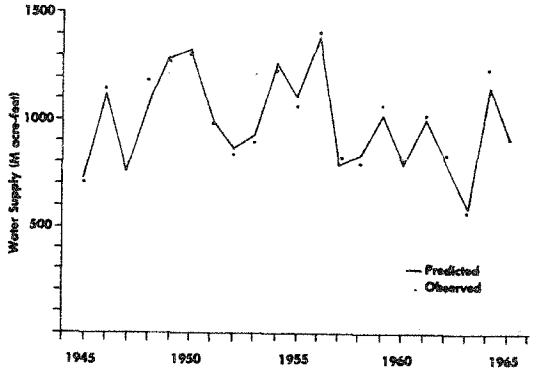



Figure 2.

Observed runoff and predicted runoff are plotted (figure 2) for 1945-1965, the years used in calculating the equation.