Technical Note

Statistical Techniques Used in the VIPER Water Supply Forecasting Software

Purpose

The purpose of this technical note is to provide users of the Visual Interactive Prediction and Estimation Routines (VIPER) water supply forecasting software appropriate background in the statistical procedures contained in the software for developing streamflow forecasting models.

Background

The NRCS Water Supply Forecasting Program, housed at the National Water and Climate Center (NWCC), uses statistical models to make its forecasts of seasonal streamflow volume. Linear regression, in various forms, is the methodology used.

Since the early 1990s, principal components regression has been the standard methodology used by the Program. The specific procedure was developed by Garen (1992), who also wrote software to enable staff hydrologists to develop and calibrate models with this methodology conveniently.

In 2006, a project to develop a new water supply forecasting software environment and process, called VIPER, was started at the NWCC. VIPER is an Excel spreadsheet-based application with data retrieval, visualization, and forecast calibration and execution functions. VIPER supports both principal components regression as well as another method, Z-score regression. This technical note explains and compares the methodologies. Other statistical techniques used in VIPER are explained as well, including searching for optimum combinations of independent variables, searching for optimum time periods covered by selected independent variables, and jackknife testing of models. A user's manual has also been developed and provides a thorough description of the mechanics of operating VIPER.

Principal Components Regression

In multiple linear regression, several independent variables (usually denoted as X) are used to predict a dependent variable (usually denoted as Y). The method of minimizing the sum of squared errors is used to fit an equation of the form:

$$Y = b_0 + \sum_{i=1}^{n} b_i X_i$$

where n is the number of independent variables, and the b's are the coefficients estimated by the least squares algorithm. If the X's are statistically unrelated to each other, that is, if they have minimal correlations among themselves, then a straightforward application of the multiple regression methodology works fine. If, however, the X's are related to each other, that is, if they have significant intercorrelations, then the X's contain redundant information, leading to what is called collinearity (McCuen and Snyder, 1986, chapter 11; Kleinbaum et al., 1988, chapters 11 and 12). If this is the case, standard multiple regression has difficulty in estimating the coefficients (b's), often leading to nonsense values, such as negative coefficients for X's having a positive relationship with Y. If a standard variable selection procedure, such as stepwise regression, is used under these conditions, many of the X's will be rejected even though they have good relationships with Y.

It is preferred, for reasons of physical completeness and model robustness, to use more than just a small subset of the available X's as predictors. To get around the problem of collinearity, two procedures are commonly used: (1) pre-combine the X's into a single composite index or several composite indices of like variable types (e.g., snow water equivalent, precipitation); or (2) principal components regression. The Z-score methodology incorporated into VIPER (described in the next major section) is an example of the first method. Principal components regression is described below.

Principal Components Analysis

Principal components regression is standard regression, but the difference is that, instead of using the X's directly, new variables, called principal components, are used instead. Principal components analysis is a standard multivariate statistical technique discussed in many textbooks and included in most statistical software packages (e.g., McCuen and Snyder, 1986, chapter 11; Johnson and Wichern, 1988, chapter 8).

Principal components are simply linear combinations of the X's. Conceptually, this is similar to the composite index method, except that instead of creating one composite index, there are n principal components.

Each principal component (PC) is a weighted sum of all the X's:

$$PC_1 = \sum_{j=1}^{n} e_{1j} X_j$$

$$PC_2 = \sum_{j=1}^{n} e_{2j} X_j$$

•
•
•

$$PC_n = \sum_{j=1}^{n} e_{nj} X_j$$

where the e's are weights. The set of weights for each PC is called an eigenvector; these eigenvectors are derived from the solution of a matrix equation in the principal components algorithm. The main input to this matrix equation is the correlation matrix of all the X's with each other. The result of the principal components analysis and the construction of PC's is that each PC is statistically unrelated to all of the other PC's. That is, by making this transformation, new variables are constructed that no longer contain collinearity.

This transformation is equivalent to a rotation of axes. This is easily envisioned in two dimensions. Consider as X's the snow water equivalent for a particular time at two sites in a basin. If these are plotted against each other, it is clear that they are closely related (Figure 1). A principal components analysis amounts to a rotation of axes as shown in the figure. By doing this, two new variables (PC's) are created, each of which is a linear combination of the two X's. In this case, most of the variance or information content of the data set is contained in PC_1.

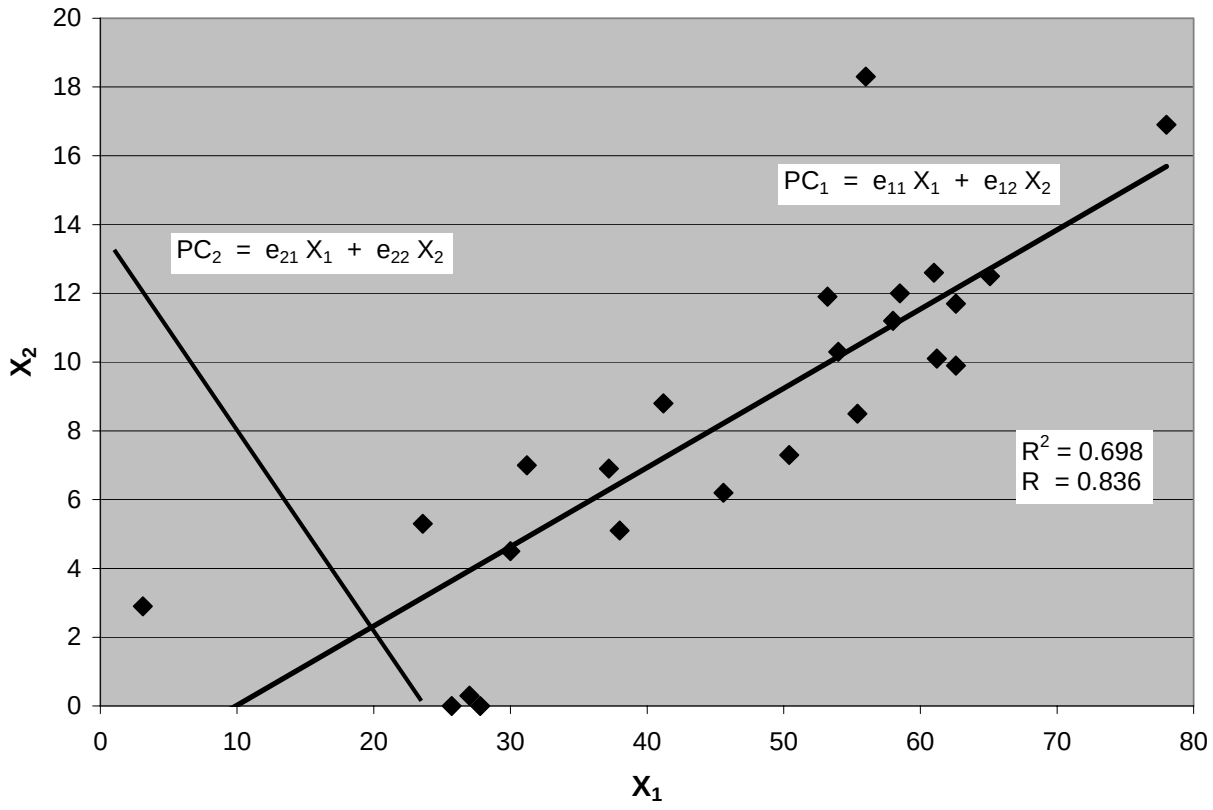

Figure 1. Illustration of principal components as a rotation of axes for two variables.

It should be noted that the weights contained in each eigenvector are based solely on the intercorrelations among the X's and have no knowledge of Y. This is in contrast to the Z-score method, where the weights are based on each X's individual correlation with Y, but there is no knowledge of the intercorrelations among the X's.

Principal components analysis is commonly used as a descriptive tool in situations where there are many intercorrelated X's available, and the analyst wishes to summarize them into a small number of combination variables that relate to some identifiable characteristics. Usually the eigenvector weights on the X's within a given PC will be relatively large on certain variables and noticeably smaller on others. The X's with the higher weights tend to be of like kind. The weights on a different PC will generally be higher on a different kind of X. In this way, the PC's can often be interpreted, and this can be the main purpose of the principal components analysis in some studies, the idea being that a large number of X's can be reduced to a few PC's that still explain most of the variance or information content in the data set.

For example, in water supply forecasting, the snow water equivalent variables are often weighted highly on, say, the first PC, while other variable types, such as fall precipitation or streamflow, are weighted highly on, say, the second PC. While this is not a completely "pure" association, because each PC has at least some weight on every X, these interpretations can often be made.

For water supply forecasting, however, principal components analysis is not done just for descriptive purposes, but rather it is used to prepare new uncorrelated independent variables for developing regression equations. This is described in the next section.

Principal Components Regression

Once the principal components analysis has been done to compute the eigenvectors, and the PC's have been constructed, the data are ready for linear regression. The issue at this point is then to determine how many PC's to include in the regression model. The specific procedure for doing this in VIPER is fully explained in Garen (1992) and is summarized below.

Principal components can be arranged in the order of explained variance in the X data, that is, the first PC (PC_1) explains the highest amount of the variance, the second PC (PC_2) explains the second-highest amount of the variance, etc. PC's are added to the regression model one at a time, beginning with PC_1. The statistical significance of the regression coefficient for PC_1 is tested with a standard t-test, using a user-selected critical t value. If the coefficient passes the t-test, then PC_2 will be added to the model. The statistical significance of its regression coefficient is subjected to the t-test; if it passes, then PC_3 will be tried, and if it fails, the regression model uses only PC_1. PC's are incorporated into the model as long as their regression coefficients are statistically significant.

Once the number of PC's to include has been determined, the regression coefficients and the eigenvector weights are manipulated algebraically to transform the regression results from the PC's back to the original X variables. An additional requirement of the algorithm implemented in VIPER is that the algebraic sign of the coefficients for each X be the same as the sign of its correlation with Y. If not, either a fewer number of PC's are tried or the model is rejected.

Example

Consider a set of twelve X variables to predict the Y variable. These X variables can be described (generically) as follows:

X_1 Snow water equivalent, station 1
X_2 Snow water equivalent, station 2
X_3 Snow water equivalent, station 3
X_4 Snow water equivalent, station 4
X_5 Snow water equivalent, station 5
X_6 Water year to date precipitation, station 1
X_7 Water year to date precipitation, station 2
X_8 Water year to date precipitation, station 3
X_9 Water year to date precipitation, station 4
X_{10} Water year to date precipitation, station 5
X_{11} Antecedent streamflow
X_{12} Climate teleconnection index

Performing a principal components analysis yields the following eigenvectors (these values are from an actual data set):

	PC_1	**PC_2**	**PC_3**	**PC_4**	**PC_5**	**PC_6**	**PC_7**	**PC_8**	**PC_9**	**PC_{10}**	**PC_{11}**	**PC_{12}**
X_1	0.265	0.444	0.004	-0.074	-0.104	0.378	0.074	-0.515	0.126	0.087	0.085	0.522
X_2	0.249	0.325	-0.483	-0.030	0.315	-0.207	-0.538	0.124	0.213	-0.300	0.129	-0.015
X_3	0.335	0.016	-0.178	0.149	-0.314	0.170	0.035	0.699	0.071	0.312	-0.163	0.295
X_4	0.229	0.353	0.456	-0.595	-0.009	0.025	0.003	0.225	0.158	0.145	0.122	-0.385
X_5	0.287	0.332	-0.148	0.120	0.412	-0.110	0.530	0.061	-0.530	-0.008	-0.103	-0.117
X_6	0.339	-0.168	-0.162	-0.106	-0.040	-0.135	-0.178	-0.365	0.056	0.359	-0.662	-0.254
X_7	0.308	-0.329	-0.150	-0.058	-0.015	-0.323	0.034	-0.151	-0.079	0.454	0.656	0.031
X_8	0.317	-0.197	-0.114	0.027	-0.261	0.574	-0.198	-0.060	-0.387	-0.291	0.174	-0.376
X_9	0.304	-0.240	0.299	-0.313	-0.103	-0.346	-0.117	0.037	-0.326	-0.420	-0.151	0.459
X_{10}	0.330	-0.197	-0.197	0.072	-0.129	-0.088	0.528	-0.055	0.572	-0.427	0.018	-0.150
X_{11}	0.235	-0.349	0.351	0.168	0.692	0.344	-0.113	0.074	0.186	0.075	-0.017	0.123
X_{12}	0.232	0.262	0.473	0.675	-0.212	-0.272	-0.215	-0.081	0.000	-0.005	0.055	-0.151
% var.	62.7	15.8	7.8	3.8	3.2	2.7	1.6	1.1	0.7	0.3	0.3	0.2

The "% var." is the percent of variance in the data set explained by each PC. Nearly two thirds of the variance is explained by the first PC alone, and over three fourths is explained by the first and second PC's. If one were using this analysis for a purely descriptive purpose, one could use the first two or three PC's to represent the majority of the information contained in this data set, thereby reducing the number of variables from twelve to two or three. The remaining PC's would then be considered to be insignificant "noise."

The eigenvector weightings in PC_1 are all of a similar magnitude, although the weights for the five snow water equivalent (SWE) variables (X_1 - X_5) tend to be somewhat smaller than the five precipitation variables (X_6 - X_{10}), and the antecedent streamflow (X_{11}) and climate teleconnection (X_{12}) variables are a bit smaller yet. However, PC_1 can be interpreted as a general water availability index, where all twelve variables are indicating the same basic signal. The other PC's

are more mixed in their weightings and are more difficult to interpret clearly, although PC_4 is most strongly associated with the climate teleconnection variable (X_{12}), and PC_5 is most strongly associated with the antecedent streamflow variable (X_{11}).

When these PC's are entered into a regression, one at a time sequentially, to predict Y, it is found that PC_1 is significant, but PC_2 is insignificant. Therefore, the regression model uses only PC_1 as its independent variable. The regression slope and intercept are then transformed back to be in terms of the original X variables. At this point, the algebraic signs of these regression coefficients are tested to ensure that they are the same as the algebraic signs of their correlations with Y. In this case they are, as all X's have a positive correlation with Y, all of the eigenvector weights for PC1 are positive, and the regression coefficient for PC_1 is positive. Had the model included other PC's besides PC_1, however, there would possibly be the opportunity for the model to fail this sign test. If this happened, the procedure would remove PC's one at a time in reverse order (largest to smallest numbered ones) until a model that passed the sign test was obtained. If it is not possible to find a model that passes both the statistical significance test and the sign test, the software reports that no valid model is possible with this combination of X variables.

Z-Score Regression

The Z-score regression methodology is a heuristic technique for combining individual independent variables into a composite index that then becomes the independent variable used in a regression. It relies on standardizing and weighting independent variable components to obtain the composite index. In this regard, it is similar in concept to principal components. In contrast to principal components, however, the weightings used in the Z-score method are based on correlations with the dependent variable, whereas principal component weightings have no knowledge of the dependent variable. Also, the Z-score weightings have no knowledge of the intercorrelations among the independent variables, whereas these intercorrelations are the basis of the principal components weightings.

The Z-score method is particularly useful when dealing with sets of independent variables that are not serially complete (i.e., have missing values) or have varying periods of record. In general, regression methods require that the data for each independent variable be complete for the entire time period being analyzed. If this is not the case, one must either make estimates for the missing values, remove the variables that have missing data from the regression, or restrict the time period used for model calibration to that in which all variables have complete data. The Z-score method is an alternative way of handling missing values without having to make estimates or make these restrictions on variable usage or time period analyzed.

The Z-score method is based on the calculation of a composite index time series using only the data available at each time step. This means that each value of the composite index can be composed of different numbers of independent variables. The assumption, then, is that the composite index so constructed is, to an acceptable degree, a homogeneous index and can be validly used in a regression. Cautions about this assumption are given in a subsequent sub-section.

Z-Score Methodology

The computational steps for Z-score regression are illustrated in Figure 2. The Z-score transform step is simply the common statistical procedure of standardization of a variable in which the mean is subtracted, and the result is divided by the standard deviation:

$$Z = \frac{X - mean}{stdev}$$

This creates a variable (Z) whose mean is 0 and standard deviation is 1. This puts all variables on an "equal footing." The means and standard deviations used in computing Z-scores are calculated from the data set used in calibrating the statistical model (i.e., not from other time periods).

The component index (C) is calculated as a weighted sum of the individual Z-score variables normalized by the sum of the weights. The weights are the coefficient of determination (R^2) of each variable with the dependent variable:

$$C = \frac{\sum_{i=1}^{n} R_i^2 Z_i}{\sum_{i=1}^{n} R_i^2}$$

where n is the number of variables. The purpose of this weighting is to give more emphasis to those variables that are more highly correlated with the dependent variable and less emphasis to those variables whose correlations are lower. Usually, to avoid confounding the index with irrelevant information, a minimum correlation criterion for inclusion of a variable is set by the user -- a common cutoff value of R^2 is 0.09 (R = 0.3), which represents an approximate minimum correlation of reasonable statistical significance for the numbers of observations typically used in developing water supply forecasting models.

Most variables used in water supply forecasting have a positive correlation with the dependent variable. Some variables, however, such as climate teleconnection indices or spring temperature, can have negative correlations. If this is the case, the algebraic sign of the time series is inverted (i.e., time series is multiplied by -1) to give a positive correlation, and the analysis proceeds as usual. This inversion is necessary so that the weighting scheme functions properly.

Notice in Figure 2 that independent variables are grouped by type of data. Typically, these data types include snow water equivalent, precipitation, antecedent streamflow, and climate teleconnection indices, although other types are also possible. If there is more than one type of data, then there are two levels of Z-score transform and component index combination -- one to create a composite index for each data type and one to combine the data type composite indices together to arrive at the final multiple component index that is used as the independent variable in the regression.

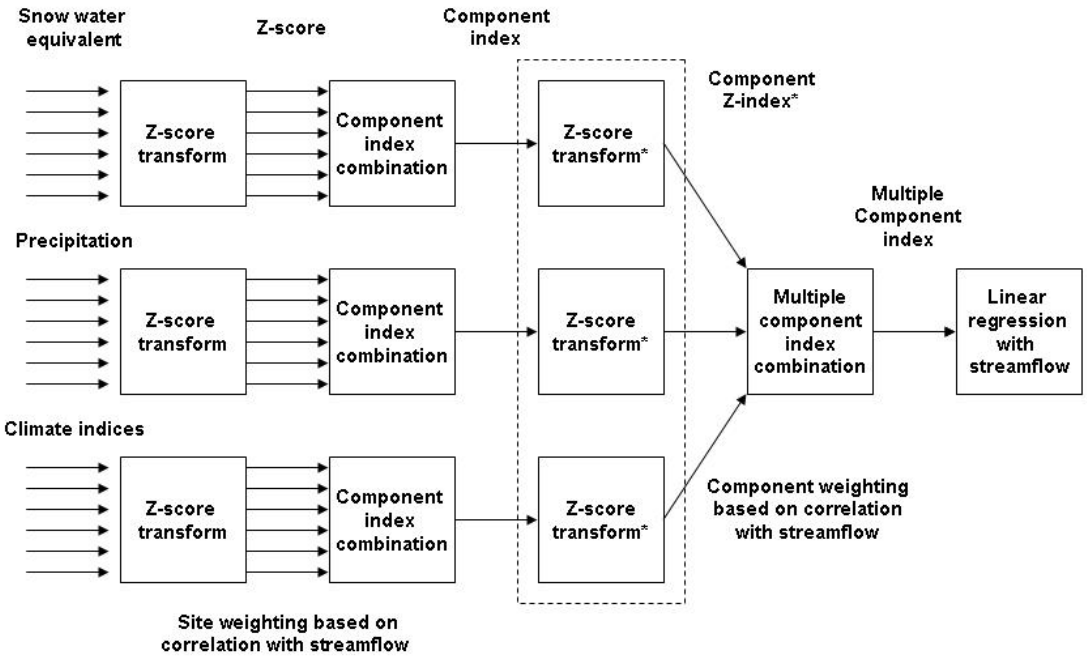

Figure 2. Schematic of Z-score regression process involving several independent variables of three different data types. The steps of Z-score transforms and component index combinations are shown, resulting in a final composite index that is used as the independent variable in the streamflow forecasting model. The Z-score transform step shown in the dotted box and the resultant Component Z-index (all also denoted by asterisks) are not in the algorithm at the time of this writing but are planned to be added.

After the linear regression is performed, the slope and intercept can be algebraically manipulated in conjunction with the component weightings and Z-score transforms to express the final equation in terms of the original X variables used for the calculation of each value of the composite index that was the independent variable in the regression.

Two examples below illustrate the computational procedure.

Simple Example (1 data type, 2 stations)

This example shows the basic calculations without the complication of more than one data type. Example data are given for a short series of years for easy illustration. How missing data are handled is also shown in this example.

Consider the following variables and data:

Y = Streamflow for a series of years

[1975	95
1976	105
1977	83
1978	93
1979	115]

X_1 = SWE, station 1

[1975	10
1976	12
1977	11
1978	missing
1979	12]

X_2 = SWE, station 2

[1975	missing
1976	7
1977	4
1978	8
1979	9]

The first step is to convert each independent variable time series into a Z-score, by subtracting the mean from each value and dividing the result by the standard deviation. For these data:

$\text{Mean}(X_1) = 11.250$
$\text{Stdev}(X_1) = 0.957$

$\text{Mean}(X_2) = 7.000$
$\text{Stdev}(X_2) = 2.160$

This gives the Z-scores as:

Z_1 = Z-score, station 1

[1975	-1.306
1976	0.783
1977	-0.261
1978	missing
1979	0.783]

Z_2 = Z-score, station 2

[1975	missing
1976	0
1977	-1.389
1978	0.463
1979	0.926]

Next, the weights for each station are determined. These are simply the coefficient of determination (R^2, the square of the correlation coefficient) between the station and the dependent variable (Y):

$R^2(Z_1,Y) = 0.420$
[$R(Z_1,Y) = 0.648$]

$R^2(Z_2,Y) = 0.670$
[$R(Z_2,Y) = 0.818$]

Since station 2 has a higher correlation with streamflow than station 1, it receives a higher weighting in the composite index. If values for both sites are available, the composite time series of station 1 and 2 is computed as:

$$C = \frac{R^2(Z_1,Y)Z_1 + R^2(Z_2,Y)Z_2}{R^2(Z_1,Y) + R^2(Z_2,Y)}$$

In 1978, however, station 1 has a missing value, and in 1975, station 2 has a missing value. For these years, the composite index value becomes simply Z_2 and Z_1, respectively. In the more general case, with more than two stations, the composite index would be calculated using all stations with non-missing values.

In this example, the composite time series therefore becomes:

C = Composite index time series

[1975	-1.306
1976	0.302
1977	-0.954
1978	0.463
1979	0.871]

This composite index is then the SWE component index for this example. Since this is the only data type, there are no other component indices to calculate and combine with. This index, then, is the independent variable for the regression. The results of the regression give a slope of 9.153 and an intercept of 99.342, with an R^2 value of 0.505 (R = 0.711).

Extended Example (2 data types, 2 stations apiece)

Two additional stations of a different data type will now be added to the example. The first data type was SWE; this second data type will be considered to be precipitation (e.g., water year precipitation to date). The composite index for the snow water equivalent component will be used again, but the composite index for precipitation now needs to be calculated. Note that in this example two stations have been used for both data types, but in general the number of stations can be different among data types. Also in this example, precipitation will be considered to be for the same station locations as for snow water equivalent, but this is not a requirement.

X_3 = Precipitation, station 1

[1975	23
1976	35
1977	22
1978	missing
1979	40]

Mean(X_3) = 30.000
Stdev(X_3) = 8.907

X_4 = Precipitation, station 2

[1975	missing
1976	45
1977	21
1978	30
1979	45]

Mean(X_4) = 35.250
Stdev(X_4) = 11.84

Z_3 = Z-score, station 1, data type 2

[1975 -0.786
1976 0.561
1977 -0.898
1978 missing
1979 1.123]

Z_4 = Z-score, station 2, data type 2

[1975 missing
1976 0.823
1977 -1.203
1978 -0.443
1979 0.823]

$R^2(Z_3,Y) = 0.893$
[$R(Z_3,Y) = 0.945$]

$R^2(Z_4,Y) = 0.914$
[$R(Z_4,Y) = 0.956$]

The composite index from the previous example will be used below and will now be called C_1, the 1 representing the SWE component. The composite index for the second component, precipitation, is:

C_2 = Composite index time series for precipitation

[1975 -0.786
1976 0.694
1977 -1.052
1978 -0.443
1979 0.971]

The next step is to combine the two component indices. Before doing so, each index needs to be standardized itself, as the standard deviations will be somewhat less than 1 (a result of the summation involved in constructing the composite, related to the Central Limit Theorem of statistics), and the means may not necessarily be 0. From the component index time series:

$Mean(C_1) = -0.125$
$Stdev(C_1) = 0.949$

$Mean(C_2) = -0.123$
$Stdev(C_2) = 0.904$

ZC_1 = Z-score of C_1

[1975 -1.244
1976 0.450
1977 -0.874
1978 0.619
1979 1.049]

ZC_2 = Z-score of C_2

[1975 -0.733
1976 0.904
1977 -1.028
1978 -0.354
1979 1.210]

$R^2(ZC_1,Y) = 0.505$
[R(ZC1,Y) = 0.711]

$R^2(ZC_2,Y) = 0.897$
[R(ZC2,Y) = 0.947]

The multiple component index is calculated in the same manner as for a single component index, that is, as a normalized weighted sum.

MC = Multiple component index

[1975	-0.917
1976	0.740
1977	-0.972
1978	0.003
1979	1.152]

The regression of MC with Y gives a slope of 11.502 and an intercept of 98.200, with an R^2 value of 0.812 (R = 0.901).

Potential Vulnerabilities of Z-Score Regression

By constructing the final composite index using only the data available at each time step (year in these examples), Z-score regression makes provision for the use of non-serially complete data. It also expands the length of the time period covered in the regression analysis to be the union of all of the independent variables instead of the intersection, as with standard regression. This flexibility, however, is also the source of some vulnerabilities that have the potential of introducing inaccuracies into the analysis.

A key assumption of Z-score regression is that independent variables that are combined together into a component index all capture the same signal, and this signal is consistently represented by the variables available at each time step, even though the number of variables may vary. If two or more component indices are combined into a multiple component index, it also assumes that all of the component signals are consistently represented.

In a strict sense, this assumption is not true if different numbers of variables are used to construct the indices. It can, however, be considered to be true to an acceptable level of approximation if the number of missing variables is not great compared with the number of variables used in an index.

What should be avoided, then, is a situation where a component index is missing many of its constituent variables or where an entire component is missing from a multiple component index. One must therefore be careful to avoid these potential pitfalls, both during model calibration and during real-time operations.

Comparison of the Two Regression Methods

It may be validly asked which of the two regression methods is preferable. As a general rule, the standard practice at the NWCC is to use principal components regression. Z-score regression is an alternative methodology available in VIPER that can be used when there are issues with serial completeness (missing data) or varying periods of record of the data.

Because the weighting schemes differ, there will be differences in the regression coefficients for each independent variable between the two methods. Despite this, recent experience with a number of basins has shown that the two methods generally produce similar results in terms of regression R^2 and standard error as well as real-time predictions.

As an illustration, consider the data set given in the example in the Principal Components Regression section above. The table below compares the regression coefficients for each X variable and the regression statistics for both principal components and Z-score models. For the Z-score model, the coefficients are calculated assuming all X variables are available.

Variable	PC regression	Z-score regression
X_1 SWE, station 1	2.914	2.367
X_2 SWE, station 2	3.337	2.528
X_3 SWE, station 3	2.436	2.384
X_4 SWE, station 4	2.273	2.374
X_5 SWE, station 5	2.502	2.375
X_6 Precipitation, station 1	3.343	2.771
X_7 Precipitation, station 2	2.691	1.961
X_8 Precipitation, station 3	2.449	1.499
X_9 Precipitation, station 4	2.974	2.666
X_{10} Precipitation, station 5	2.782	2.072
X_{11} Antecedent streamflow	0.546	0.981
X_{12} Climate teleconnection index	2.470	7.010
Intercept	-79.776	-38.832
R^2	0.821	0.820
R	0.906	0.905
Standard error	62.558	62.607

Note that the regression statistics are nearly identical, but the coefficients differ. The Z-score model coefficients are somewhat smaller than those for the principal components model for SWE and precipitation, while the Z-score model coefficients are larger for antecedent streamflow and the climate teleconnection index. This difference is a direct result of the weighting schemes. The principal components model uses only PC_1, which, as noted previously, gives more weight to the SWE and precipitation variables than X_{11} and X_{12}. Again as noted above, much of the weight for antecedent streamflow and the climate teleconnection index is associated with higher PC's, which are not used in the regression, therefore these two variables receive relatively less weight than with the Z-score model.

This comparison demonstrates the inherent difference between the two methods, given that there is a serially complete data set. If this were not the case, the Z-score method could proceed, but there would be different coefficients for the different situations of missing variables. For the principal components method, the missing values would have to be estimated, a variable would have to be removed from the analysis, or some other adaptation to accommodate the data availability would have to be devised.

Variable Combinations Search

In building statistical models, it can be of assistance to employ a variable selection algorithm to select variables from a list of candidates to optimize model accuracy. Garen (1992) developed such an algorithm, which is implemented in VIPER.

This algorithm is a search procedure that tests combinations of candidate variables in a systematic way to identify variable combinations that result in superior model accuracy (as measured by the standard error). Since testing all possible combinations of candidate variables can be computationally expensive, this procedure tests a subset of combinations by building up models in a logical progression.

The algorithm begins by computing all one variable models and storing the best 30 (or all the models if there are fewer than 30 candidate variables) in a "keep list". Minimizing the standard error (or more correctly, the jackknife standard error; see section below) is used as the optimality criterion. The number 30 was arbitrarily chosen as a compromise between keeping the computations at a reasonable level and giving the algorithm plenty of combinations upon which to build.

In the next iteration of the algorithm, all possible two-variable models built from the 30 stored one-variable models are tested by adding variables from the candidate list to each of the stored models one at a time. The standard errors from the two-variable models along with the 30 one-variable models are sorted, and the best 30 models are retained.

In the third iteration, three-variable models are built from the stored two-variable models, again by adding variables from the candidate list one at a time. The best 30 one-, two-, or three-variable models are then stored.

The algorithm continues its iterations of adding one more variable to the stored models until no more improvements in the standard error occur. At this point, the algorithm terminates, and the user is presented with the list of the best 30 models -- the variables used and the regression statistics.

This search algorithm tends to select for parsimonious models (i.e., ones that do not contain a large number of variables) and does not necessarily find the absolute optimum or all combinations within the range of standard errors in the final list. It does, however, do a good job of identifying the strongest variables and building models that perform well.

Since this algorithm is only a statistical optimization, it is still incumbent upon the user to review these results for physical meaningfulness before selecting any of these models for use. It may be that the user will not want to use any of these models, instead choosing to modify them or choosing an entirely different variable combination, keeping in mind that there are often tradeoffs between statistical optimality and physical meaningfulness. This variable search optimization, then, should be considered to be only a guide to assist in the selection of independent variables to use.

Time Period Search

Since the independent variables used in water supply forecasting have a monthly time step, it is necessary to identify the month or months that contain the most relevant information for predicting the dependent variable. The time period search feature of VIPER is designed to assist with this.

Some types of independent variables can be accumulated or averaged over a period of months to capture a signal into a single aggregated variable. This applies particularly to precipitation, streamflow, and climate teleconnection indices. Aggregating can be helpful in reducing the number of independent variables and in focusing the signal into a single value rather than having it scattered among several monthly values.

The main exception to this is SWE, which is already an accumulated variable. Under certain circumstances, however, it can be helpful to identify which month's SWE is the best predictor of the dependent variable. This applies especially in the spring, when sometimes the SWE from a previous month rather than the current month is a better predictor (e.g., May SWE used in a June forecasting equation). The VIPER time period search routine can assist with identifying such cases.

The time period search algorithm examines variables of like data type together in a group. Each group is examined independently. For all data types except SWE, the algorithm proceeds by computing linear regression models to predict the dependent variable for all combinations of contiguous months for the independent variables within the range specified by the user. Either principal components or Z-score regression can be used, as desired. The range of months tested in each iteration of the algorithm is the same for all variables in the group. The algorithm identifies the model giving the smallest standard error and reports this optimum range of months for the variables in the group.

For SWE, there is no accumulation of months, but the algorithm proceeds as described above, testing individual monthly SWE within the range of months specified by the user. The algorithm reports the SWE month that gives the smallest standard error for all stations together as a group.

As a note of caution, users should be aware that sometimes a monthly range can contain more than one signal. For example, fall precipitation before snow accumulation begins and winter precipitation after snow accumulation has started are two different signals. If a monthly range that spans more than one signal is specified, this should be a conscious decision by the user.

Another item to note is that this time period optimization does not account for interactions among the different data types, as each data type group is evaluated independently. A future enhancement being considered will allow other specified data type groups with fixed time ranges to be included in the models during the time period search for a given data type group. This will allow interactions among data type groups to be considered in determining the optimum time range.

Jackknife Test

Since these models are intended to be used in a forecasting situation, the standard error from the model calibration might be considered to be an overly optimistic expression of the forecast error. To obtain a more realistic evaluation of a model's forecasting potential, a jackknife (also called cross-validation) procedure is used.

The jackknife test for a given combination of independent variables is an iterative procedure of leaving out one observation (typically a year) from the calibration data set, computing the regression coefficients, then using these coefficients with the input data for the withheld observation to make a prediction of the dependent variable. The withheld observation is then returned to the calibration data set, and the next observation is removed. The process is repeated through the entire data set so that when finished, a series of predictions is obtained from models that did not include that observation in the calibration data set. These jackknife predictions are then compared to the observed values, and a jackknife standard error is computed. Generally, the jackknife standard error is a little larger than the standard error from calibration using all observations.

The jackknife standard error is used as the optimality criterion in the variable combinations search algorithm, and it is used in the evaluation of models in which the user specifies the independent variables to be used. The jackknife standard error is used in operational forecasting to compute error bounds around the median forecast.

Nonlinear Procedures

There are two situations in which a linear regression model is not appropriate. The first is when there is a marked nonlinear relationship between the independent and dependent variables, and the second is when the errors around the median predictions from a linear model do not have a normal distribution.

In both of these situations, nonlinear models can generally address the problem. The standard practice at the NWCC, which is implemented in VIPER, is to transform the dependent variable and develop a linear model to predict this transformed Y. The transforms used are square root, cube root, and natural logarithm. The software handles all steps of transforming, computing the linear model to predict the transformed Y, and back-transforming the model results. Note that a

nonlinear model built in this way will have asymmetrical error bounds, being narrower below the median prediction and wider above the median prediction.

Routed Procedures

A so-called "routed procedure" is a statistical model that relates one or more stream locations to another. Typically, upstream points are used to predict a downstream point. Such models are useful when there is a strong relationship between upstream and downstream points, and they can simplify the development of forecasting models by avoiding the full analysis of all of the basic input variables (SWE, precipitation, etc.) as is necessary for a "headwater" basin.

In VIPER, routed procedures are developed in a two-step process. The first step is to develop the relationship between the upstream point(s) and the downstream point using historical streamflow data. If there is more than one upstream point, the streamflows are added together to give a single independent variable. The reasoning behind this is that the upstream flows represent some fraction of the watershed of the downstream point, so the relationship is between the response of this watershed fraction and the entire watershed of the downstream point.

The second step is to estimate the standard error of this model. In real-time forecasting, the independent variable is not an observed value but rather a forecast, which contains uncertainty. This uncertainty needs to be propagated to the prediction at the downstream point.

The standard error for a routed procedure is computed empirically by using the jackknife forecasts from the development of the upstream forecast models as input to the routed model and computing a standard error from the prediction errors at the downstream point. The mechanics of doing this are explained in the VIPER user's manual.

Helper Variables

In the case of missing values of the dependent variable, VIPER allows the user to specify a "helper variable" to be used to estimate the missing data. Typically, the dependent variable is a streamflow volume, and there can be either scattered missing values, or the streamgage was discontinued for some period of time. In this situation, it is often possible to make very good estimates of these missing values using data from a nearby streamgage, either upstream or downstream from the point of interest or in a neighboring basin. This makes it possible to use more years of data in model development.

The estimates are made using a simple linear regression between the helper variable and the dependent variable. The estimates are then used to fill in missing values in the dependent variable. The mechanics of using a helper variable are explained in the VIPER user's manual.

Contact

The contact for this technical note is the Branch Leader, Water and Climate Services, National Water and Climate Center, Portland, Oregon (www.wcc.nrcs.usda.gov).

References

Garen, D. C. (1992). Improved techniques in regression-based streamflow volume forecasting. Journal of Water Resources Planning and Management, 118(6):654-670. (Available on the NRCS National Water and Climate Center's web site under Publications → Professional Publications.)

Kleinbaum, D. G., L. L. Kupper, and K. E. Muller (1988). Applied regression analysis and other multivariable methods. PWS-KENT Publishing Co. (Boston, MA).

Johnson, R. A., and D. W. Wichern (1988). Applied multivariate statistical analysis (2nd edition). Prentice-Hall (Englewood Cliffs, NJ).

McCuen, R. H., and W. M. Snyder (1986). Hydrologic modeling: Statistical methods and applications. Prentice-Hall (Englewood Cliffs, NJ).