

# TRANSFERENCIA DE INFORMACIÓN HIDROLÓGICA MEDIANTE REGRESIÓN LINEAL MÚLTIPLE, CON SELECCIÓN ÓPTIMA DE REGRESORES

## TRANSFERENCE OF HYDROLOGIC INFORMATION THROUGH MULTIPLE LINEAR REGRESSION, WITH BEST PREDICTOR VARIABLES SELECTION

Daniel F. Campos-Aranda<sup>1</sup>

<sup>1</sup> Facultad de Ingeniería de la Universidad Autónoma de San Luis Potosí. Genaro Codina Núm. 240. 78280 San Luis Potosí, San Luis Potosí. (campos\_aranda@hotmail.com).

### RESUMEN

Es necesario contar con registros largos de información hidrológica anual para obtener una imagen más apegada a la realidad de su variabilidad, así como estimaciones confiables de sus propiedades estadísticas. Para obtener tales registros es común buscar fuentes adicionales de datos y técnicas de transferencia. Una técnica es la regresión lineal múltiple, cuya aplicación numérica lleva implícita la selección óptima de los registros largos cercanos (regresores) para buscar que la ampliación del registro corto sea una estimación confiable. Este proceso de selección implica tres análisis: 1) cómo definir las mejores estimaciones, 2) cuáles ecuaciones de regresión investigar, y 3) cuál modelo tiene mejor capacidad predictiva. Para el primer análisis se presentan cuatro criterios basados en las sumas de los cuadrados de los residuos; para el segundo se investigan todas las regresiones posibles porque en los problemas de transferencia de información hidrológica se dispondrá máximo de cinco regresores; para el tercero, seleccionar el mejor modelo predictivo se utiliza el análisis de residuales y la validación cruzada. La aplicación numérica descrita es una ampliación del registro de volúmenes escurridos anuales en la estación hidrométrica Platón Sánchez del sistema del río Tempoal, en la Región Hidrológica No. 26 (Pánuco, México). En este caso se utilizan cuatro regresores que son los registros del resto de las estaciones de aforos de tal sistema. Se concluye que incluso en problemas con multicolinealidad, los criterios de selección y los análisis expuestos conducen a resultados consistentes y permiten obtener las mejores ecuaciones de regresión. La similitud de los resultados alcanzados con los modelos de regresión seleccionados genera confianza en las estimaciones adoptadas.

### ABSTRACT

It is necessary to have long records of annual hydrological data to get a truer picture of their variability, as well as reliable estimates of their statistical properties. To obtain these records it is common to use additional sources of data and transfer techniques. One technique is the multiple linear regression whose numerical application implies the optimum selection of close lengthy records (regressors) to have the extension of short registration be a reliable estimate. This selection process involves three analyses: 1) how to define the best estimates, 2) what regression equations should be investigated, and 3) which model has better predictive ability. For the first analysis four criteria based on the sums of the squares of the residuals are presented; for the second all possible regressions are investigated since in the problems of hydrological information transfer, we will have five regressors at the most; for the third, about selecting the best predictive model, we used the residual analysis and cross-validation. The numerical application described is an extension of the annual runoff volume record in the Platón Sánchez hydrometric station of the Tempoal river system in the 26 Hydrological Region (Pánuco, México). Here we used four regressors that are the records of other gauging stations in such system. We came to the conclusion that even in problems with multicollinearity, the selection criteria and analysis led to consistent results and allowed for the best regression equations. The similarity of the results obtained with the selected regression models generated confidence in the estimates adopted.

**Keywords:** residual mean square, multicollinearity, residual analysis, Durbin-Watson test, cross-validation, Rio Tempoal.

---

\* Autor responsable ♦ Author for correspondence.

Recibido: junio, 2011. Aprobado: octubre, 2011.

Publicado como ARTÍCULO en *Agrociencia* 45: 863-880. 2011.

**Palabras clave:** cuadrado medio de los residuos, multicolinealidad, análisis de residuales, prueba de Durbin-Watson, validación cruzada, Río Tempoal.

## INTRODUCCIÓN

En general, las estimaciones de las características estadísticas de un registro hidrológico de valores anuales son más confiables y consistentes si éste es más amplio, porque al ser más largo es más probable que incluya periodos de años secos y húmedos y no sólo de uno de ellos. Las principales variables en la práctica hidrológica son precipitación, escurrimiento y crecientes, donde el volumen escurrido anual tiene relevancia en todas las estimaciones asociadas con la disponibilidad y el diseño hidrológico de embalses para abastecimiento. La técnica básica para ampliar registros hidrológicos anuales es la regresión lineal, la cual permite la transferencia de información de un sitio a otro. Cuando esta técnica se aplica regionalmente, es decir, se transporta información de varios sitios o registros al de interés, se usa la regresión lineal múltiple y es necesario seleccionar las mejores variables predictivas o registros auxiliares, también llamados regresores.

El objetivo de este estudio fue exponer la técnica de transferencia de información hidrológica de variables anuales, mediante regresión lineal múltiple, para ampliar registros cortos de volúmenes escurridos con base en las series largas cercanas, seleccionando la mejor ecuación de regresión de entre todas las posibles. La formulación matemática se presenta de manera simple al utilizar la solución matricial, se exponen con detalle los criterios de selección y validación, y se desarrolla un ejemplo numérico en el sistema del río Tempoal, de la Región Hidrológica No. 26 (Pánuco, México), para ampliar el registro corto de la estación hidrométrica Platón Sánchez utilizando los cuatro registros largos disponibles en tal sistema.

## MATERIALES Y MÉTODOS

### Regresión lineal múltiple

Esta regresión es útil cuando la variable dependiente ( $y$ ) no está relacionada sólo con otra ( $x$ ), sino que depende de varias, las cuales no están correlacionadas entre sí y tanto  $y$  como todas las otras variables  $x$  proceden de una población Normal multivariada (Gilroy, 1970; Salas *et al.*, 2008). La expresión de este modelo de regresión es:

## INTRODUCTION

In general, estimates of the statistical characteristics of an annual hydrologic record are more reliable and consistent as such record is wider, because as it is longer it is more likely to include periods of dry and wet years and not only one of them. The main variables in hydrological practice are precipitation, runoff and floods, where the annual runoff volume has relevance in all estimates associated with the availability and hydrologic design of reservoirs for supply. The basic technique for extending annual hydrological records is linear regression, which allows the transfer of information from one place to another. When this technique is applied regionally, *i.e.* information from various sites or records is transferred to the site of interest, the multiple linear regression is used and it is necessary to select the best predictive variables or auxiliary registers, also called regressors.

The objective of this study was to put forward the hydrological information transfer technique of annual variables by using multiple linear regression to extend short records of runoff volumes based on the nearby long series, selecting the best regression equation among all possible. The mathematical formulation is presented in a simple way using the matrix solution; the selection and validation criteria are discussed in detail, and a numerical example is developed in the Tempoal river system, Hydrological Region No. 26 (Pánuco, México) to expand the short registration of the hydrometric station Platón Sánchez using the four long records available in that system.

## MATERIALS AND METHODS

### Multiple linear regression

This regression is useful when the dependent variable ( $y$ ) is not only related to another one ( $x$ ), but depends on several others, which are not correlated with each other and  $y$  as well as all the other  $x$  variables come from a Normal multivariate population (Gilroy, 1970, Salas *et al.*, 2008). The expression of this regression model is:

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \dots + a_m \cdot x_m \quad (1)$$

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \dots + a_m \cdot x_m \quad (1)$$

Las ecuaciones normales se obtienen igual que para la recta de regresión lineal, pero ahora la ecuación del error depende de  $x_m$  variables y por tanto se establece igual número de ecuaciones; en forma matricial el sistema es el siguiente (Campos, 2003):

$$\begin{bmatrix} n & \sum_1^n x_{1i} & \sum_1^n x_{2i} & \dots & \sum_1^n x_{mi} \\ \sum_1^n x_{1i} & \sum_1^n x_{1i}^2 & \sum_1^n x_{1i} \cdot x_{2i} & \dots & \sum_1^n x_{1i} \cdot x_{mi} \\ \sum_1^n x_{2i} & \sum_1^n x_{2i} \cdot x_{1i} & \sum_1^n x_{2i}^2 & \dots & \sum_1^n x_{2i} \cdot x_{mi} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \sum_1^n x_{mi} & \sum_1^n x_{mi} \cdot x_{1i} & \sum_1^n x_{mi} \cdot x_{2i} & \dots & \sum_1^n x_{mi}^2 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_1^n y_i \\ \sum_1^n x_{1i} \cdot y_i \\ \sum_1^n x_{2i} \cdot y_i \\ \vdots \\ \sum_1^n x_{mi} \cdot y_i \end{bmatrix} \quad (2)$$

en notación matricial:

$$X \cdot a = B \quad \therefore \quad a = X^{-1} \cdot B \quad (3)$$

Cuando se utiliza este modelo de regresión para transportar información hidrológica desde varios sitios, el problema es seleccionar del grupo de regresores candidatos (registros disponibles), el subconjunto que conviene usar en el modelo. Tal selección implica dos objetivos contrapuestos (Montgomery *et al.*, 2002): 1) que el modelo incluya tantos regresores como sea posible, para que el contenido de información en ellos pueda influir favorablemente en la estimación de  $y$ ; y 2) que el modelo incorpore el menor número posible de regresores porque la varianza de la estimación de  $y$  aumenta con el número de éstos. El proceso de encontrar un modelo que cumpla ambos objetivos se llama selección óptima de regresores y, en general, los diferentes algoritmos para realizarlo conducen a resultados diferentes debido a la presencia de valores atípicos y correlación entre los registros candidatos (McCuen, 1998; Montgomery *et al.*, 2002).

Dos aspectos fundamentales del problema de selección óptima de regresores son la generación de los modelos con subconjuntos y la decisión de si un subconjunto es mejor que otro. Aquí se exponen los cuatro criterios usados para evaluar y comparar ecuaciones de regresión con subconjuntos, y luego cuales modelos revisar.

**Coefficiente de determinación múltiple**

Es quizás la medida más utilizada para medir lo adecuado de un modelo de regresión. Se designa por  $R_p^2$  cuando el modelo tiene un subconjunto de  $p$  términos, es decir,  $p-1$  regresores y un término  $a_0$  de ordenada al origen, y la ecuación es:

The normal equations are obtained in the same way as for the linear regression model, but now the error equation depends on  $x_m$  variables and therefore the same number of equations is provided; in matrix form the system is as follows (Campos, 2003):

in matrix notation:

$$X \cdot a = B \quad \therefore \quad a = X^{-1} \cdot B \quad (3)$$

When using this regression model to transport hydrological information from various sites, the problem is to select from the group of candidate regressors (records available) the subset that should be used in the model. Such selection involves two conflicting objectives (Montgomery *et al.*, 2002): 1) that the model includes as many regressors as possible, so that the information content can favorably influence the estimate of  $y$ ; and 2) that the model incorporates the minimum number of regressors because the  $y$  estimate variance increases with the number of these. The process of finding a model that meets both objectives is called optimal selection of regressors, and in general the different algorithms contributing to develop it lead to different results due to the presence of outliers and the correlation between the candidate records (McCuen, 1998; Montgomery *et al.*, 2002).

Two fundamental aspects of the problem of optimal selection of regressors are the generation of models with subsets and the decision whether a subset is better than another. Here we introduce the four criteria used to evaluate and compare regression equations with subsets, and then the models to be checked.

**Multiple coefficient of determination**

It is perhaps the most commonly used coefficient to measure the adequacy of a regression model. It is called  $R_p^2$  when the model has a subset of  $p$  terms, that is,  $p-1$  regressors and a term  $a_0$  of intercept, and the equation is:

$$R_p^2 = \frac{SC_y - SC_{Res}(p)}{SC_y} = 1 - \frac{\sum_1^n (y_i - \hat{y}_p)^2}{\sum_1^n (y_i - \bar{y})^2} \quad (4)$$

donde  $\hat{y}_p$  es la estimación de la variable  $y_i$  a través de la ecuación de regresión, por ello  $SC_{Res}(p)$  es la suma de cuadrados de los residuos y  $SC_y$  es la varianza total de la variable dependiente cuya media aritmética es  $\bar{y}$ . Siendo  $K$  el número de regresores candidatos, el problema asociado al uso de  $R_p^2$ , es que aumenta conforme lo hace  $p$  y es máximo cuando  $p=K+1$ . Entonces, para aplicar este criterio de selección de modelos se agregan regresores hasta un número en que el siguiente ya no produce un aumento significativo en  $R_p^2$ .

**Coefficiente de determinación múltiple ajustado**

Designado por  $R_{ap}^2$ , no necesariamente aumenta al introducir regresores, sino que al introducir  $s$  regresores,  $R_{a,p+s}^2 > R_{a,p}^2$  si y sólo si, la estadística  $F$  parcial es mayor que 1. Por tanto, este criterio permite seleccionar el subconjunto óptimo a través de su valor máximo (Montgomery *et al.*, 2002) y su fórmula es:

$$R_{a,p}^2 = 1 - \left( \frac{n-1}{n-p} \right) \cdot (1 - R_p^2) \quad (5)$$

**Cuadrado medio de los residuos**

Este criterio tiene un comportamiento de decaimiento que se estabiliza y luego crece, pues en algún punto ( $p$ ) la disminución del numerador no es suficiente para compensar la pérdida de un grado de libertad del denominador. Entonces el subconjunto óptimo será el que define el valor mínimo, cuya expresión es (Montgomery *et al.*, 2002):

$$CM_{Res}(p) = \frac{SC_{Res}(p)}{n-p} = \frac{\sum_1^n (y_i - \hat{y}_p)^2}{n-p} \quad (6)$$

**Estadística  $C_p$  de Mallows**

El valor de  $C_p$  se puede dibujar en una gráfica de  $p$  en las abscisas que incluya la recta a 45° ( $C_p = p$ ). Las ecuaciones de regresión con poco sesgo tendrán valores de  $C_p$  próximos a la recta y aquéllas con sesgo apreciable se apartarán de ésta. Se prefieren los valores menores de  $C_p$ , pues indican menor error total, y la ecuación de  $C_p$  es (Montgomery *et al.*, 2002):

$$R_p^2 = \frac{SC_y - SC_{Res}(p)}{SC_y} = 1 - \frac{\sum_1^n (y_i - \hat{y}_p)^2}{\sum_1^n (y_i - \bar{y})^2} \quad (4)$$

where  $\hat{y}_p$  is the estimate of variable  $y_i$  through the regression equation, so  $SC_{Res}(p)$  is the sum of squared residuals and  $SC_y$  is the total variance of the dependent variable whose mean is  $\bar{y}$ . Since  $K$  is the number of candidate regressors, the problem associated with the use of  $R_p^2$  is that it increases as  $p$  does and is maximum when  $p=K+1$ . So to apply this model selection criterion, regressors are added up to a number in which the next does no longer produce a significant increase in  $R_p^2$ .

**Adjusted multiple coefficient of determination**

Designated by  $R_{ap}^2$ , it does not necessarily increase as more regressors are introduced, but by introducing  $s$  regressors,  $R_{a,p+s}^2 > R_{a,p}^2$  if and only if the partial  $F$  statistic is greater than 1. Therefore, this criterion makes it possible to select the optimal subset through its maximum value (Montgomery *et al.*, 2002) and its formula is:

$$R_{a,p}^2 = 1 - \left( \frac{n-1}{n-p} \right) \cdot (1 - R_p^2) \quad (5)$$

**Residual mean squares**

This criterion has a decay behavior that stabilizes and then grows, as at some point ( $p$ ) the numerator decline is not enough to offset the loss of one degree of freedom of the denominator. Therefore, the optimal subset is the one to define the minimum value, which is expressed (Montgomery *et al.*, 2002) as follows:

$$CM_{Res}(p) = \frac{SC_{Res}(p)}{n-p} = \frac{\sum_1^n (y_i - \hat{y}_p)^2}{n-p} \quad (6)$$

**Mallows's  $C_p$  statistic**

The  $C_p$  value can be drawn on a graph of  $p$  on the abscissa including the line at 45° ( $C_p = p$ ). Regression equations with little bias will have  $C_p$  values close to the line and those with significant bias will depart from it. Smaller values of  $C_p$  are favored as they indicate lower total error, and the  $C_p$  equation (Montgomery *et al.*, 2002) is:

$$C_p = \frac{SC_{Res}(p)}{CM_{Res}(K+1)} - n + 2p \quad (7)$$

Al usar el cuadrado medio del modelo regresional completo como denominador se supone que tiene un sesgo despreciable. Si el modelo completo tiene varios regresores que no contribuyen significativamente, es decir que tienen coeficientes  $a_i$  cercanos a cero, el denominador de la ecuación 7 estará sobreestimado y los  $C_p$  serán pequeños. En tales casos se puede usar el mínimo cuadrado medio obtenido, sin importar que regresores lo originaron; ello conducirá a un  $C_p=p$  para tal modelo regresional.

### Ecuaciones de regresión con subconjuntos

Hay dos procedimientos de análisis de los diferentes modelos de regresión que se pueden formar con los subconjuntos de variables regresoras candidatas; el primero consiste en procesar todas las regresiones posibles y el segundo en realizar una regresión por segmentos. Cuando se analizan todas las regresiones posibles se busca el mejor modelo según uno o varios criterios de selección, entre ecuaciones que tienen un regresor candidato, dos regresores, o más. Ya que el término de ordenada al origen ( $a_0$ ) se incluye en todas las regresiones y como hay  $K$  regresores candidatos, entonces habrá  $2^K$  ecuaciones por estimar y examinar; por ejemplo si  $K=4$ , hay  $2^4=16$  ecuaciones posibles, en cambio si  $K=8$ , hay  $2^8=256$  regresiones por analizar. Este procedimiento se vuelve impráctico para  $K > 5$ . En la regresión por segmentos se evalúa sólo una pequeña cantidad de ecuaciones, agregando o eliminando regresores uno por uno. Hay diversos algoritmos de este procedimiento, por ejemplo, selección hacia adelante, eliminación hacia atrás y sus combinaciones (McCuen, 1998).

Debido a que en la transferencia de información hidrológica difícilmente se dispone de cinco registros aledaños o regresores candidatos, el procedimiento sugerido para el análisis de las regresiones por subconjuntos, es el de procesar todas las ecuaciones posibles, las cuales se indican en el Cuadro 1.

### Validación de los modelos seleccionados

Cuando los regresores usados son series cronológicas se debe verificar que sus residuos no estén autocorrelacionados, pues ello implica violar una de las hipótesis básicas de la regresión lineal: sus errores tienen media cero, varianza constante y no están correlacionados. Tal verificación se realiza mediante gráficas de residuales, cuyo comportamiento indica si se debe detectar autocorrelación positiva o negativa; en el primer caso los residuos se agrupan según su signo y en el segundo cambian demasiado de signo. Esto se verifica mediante la prueba de Durbin-Watson (Makrindakis *et al.*, 1983; Montgomery *et al.*, 2002), cuya hipótesis establece que los errores ( $e$ ) los genera un proceso

$$C_p = \frac{SC_{Res}(p)}{CM_{Res}(K+1)} - n + 2p \quad (7)$$

By using the mean squared of the full regression model as denominator it is assumed the bias is negligible. If the full model has several regressors that do not contribute significantly, meaning that they have  $a_i$  coefficients close to zero, the denominator of equation 7 will be overestimated and  $C_p$  will be small. In such cases you can use the minimum mean square obtained, regardless of the regressors that gave rise to it; this will lead to a  $C_p=p$  for such regression model.

### Regression equations with subsets

There are two methods of analysis of the different regression models that can be formed with subsets of candidate regressor variables. The first is to process all possible regressions and the second to perform a stepwise regression. When analyzing all possible regressions, the aim is to look for the best model according to one or more selection criteria, among equations that have a candidate regressor, two regressors, and so on. Since the intercept term ( $a_0$ ) is included in all regressions and since there are  $K$  candidate regressors, then there will be  $2^k$  equations to estimate and examine; for example, if  $K=4$ ,  $2^4=16$  equations are possible; however if  $K=8$ , there are  $2^8=256$  regressions to analyze. This procedure becomes impractical for  $K > 5$ . In the stepwise regression, only a small number of equations by adding or deleting regressors one by one is evaluated. There are several algorithms of this procedure, for example, forward selection, backward elimination and their combinations (McCuen, 1998).

Because in the transfer of hydrological data hardly five records or candidate regressors are available, the procedure suggested for the analysis of subset regressions is to process all possible equations, which are listed in Table 1.

### Validation of selected models

When the regressors used are time series it is necessary to verify that their residuals are not autocorrelated, because it implies a violation of the basic assumptions of linear regression: their errors have zero mean, constant variance and are not correlated. Such verification is performed using graphs of residuals, whose behavior indicates whether to detect positive or negative autocorrelation. In the first case the residuals are classified according to their sign and in the second they change sign far too much. This is verified by using the Durbin-Watson test (Makrindakis *et al.*, 1983, Montgomery *et al.*, 2002), whose hypothesis is that errors ( $e$ ) are generated by a first-order autoregressive process. When looking for positive autocorrelation the following statistic is used:

**Cuadro 1. Ecuaciones por subconjuntos que se deben analizar, según el número de regresores candidatos ( $K$ ).  
Table 1. Subset equations to be analyzed according to the number of candidate regressors ( $K$ ).**

No. Ec.	$K = 3$			$K = 4$			$K = 5$		
	$k$	$p$	Regresores	$k$	$p$	Regresores	$k$	$p$	Regresores
1	0	1	–	0	1	–	0	1	–
2	1	2	$x_1$	1	2	$x_1$	1	2	$x_1$
3	1	2	$x_2$	1	2	$x_2$	1	2	$x_2$
4	1	2	$x_3$	1	2	$x_3$	1	2	$x_3$
5	2	3	$x_1, x_2$	1	2	$x_4$	1	2	$x_4$
6	2	3	$x_1, x_3$	2	3	$x_1, x_2$	1	2	$x_5$
7	2	3	$x_2, x_3$	2	3	$x_1, x_3$	2	3	$x_1, x_2$
8	3	4	$x_1, x_2, x_3$	2	3	$x_1, x_4$	2	3	$x_1, x_3$
9				2	3	$x_2, x_3$	2	3	$x_1, x_4$
10				2	3	$x_2, x_4$	2	3	$x_1, x_5$
11				2	3	$x_3, x_4$	2	3	$x_2, x_3$
12				3	4	$x_1, x_2, x_3$	2	3	$x_2, x_4$
13				3	4	$x_1, x_2, x_4$	2	3	$x_2, x_5$
14				3	4	$x_1, x_3, x_4$	2	3	$x_3, x_4$
15				3	4	$x_2, x_3, x_4$	2	3	$x_3, x_5$
16				4	5	$x_1, x_2, x_3, x_4$	2	3	$x_4, x_5$
17							3	4	$x_1, x_2, x_3$
18							3	4	$x_1, x_2, x_4$
19							3	4	$x_1, x_2, x_5$
20							3	4	$x_1, x_3, x_4$
21							3	4	$x_1, x_3, x_5$
22							3	4	$x_1, x_4, x_5$
23							3	4	$x_2, x_3, x_4$
24							3	4	$x_2, x_3, x_5$
25							3	4	$x_2, x_4, x_5$
26							3	4	$x_3, x_4, x_5$
27							4	5	$x_1, x_2, x_3, x_4$
28							4	5	$x_1, x_2, x_3, x_5$
29							4	5	$x_1, x_2, x_4, x_5$
30							4	5	$x_1, x_3, x_4, x_5$
31							4	5	$x_2, x_3, x_4, x_5$
32							5	6	$x_1, x_2, x_3, x_4, x_5$

autorregresivo de primer orden. Cuando se busca autocorrelación positiva se usa el siguiente estadístico:

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \tag{8}$$

siendo

$$e_t = y_t - \hat{y}_t \tag{9}$$

donde,  $n$  es el número de datos,  $y_t$  es la variable dependiente y  $\hat{y}_t$  su estimación mediante el modelo de regresión lineal múltiple

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \tag{8}$$

being

$$e_t = y_t - \hat{y}_t \tag{9}$$

where  $n$  is the number of data,  $y_t$  is the dependent variable and  $\hat{y}_t$  its estimate through the multiple linear regression model under test. The null hypothesis ( $H_0$ ) states that there is no autocorrelation, and alternative ( $H_1$ ) that there is. The Durbin-Watson tabulation sets two limits ( $d_L$  and  $d_U$ ) according to  $n$ ,

que se está probando. La hipótesis nula ( $H_0$ ) establece que no existe autocorrelación y la alternativa ( $H_1$ ) que sí. La tabulación de Durbin-Watson establece dos límites ( $d_L$  y  $d_U$ ) según  $n$ , número de regresores ( $K$ ) y nivel de significancia  $\alpha$  de la prueba ( $\alpha=5\%$ , comúnmente). La regla de decisión es:

Si  $d < d_L$  rechazar  $H_0$   
 Si  $d > d_U$  no rechazar  $H_0$   
 Si  $d_L \leq d \leq d_U$  la prueba no es concluyente.

Cuando la prueba se emplea para detectar autocorrelación negativa se emplea el estadístico  $4-d$ , usando como límites  $4-d_L$  y  $4-d_U$  y la misma regla de decisión.

Ya verificado que la autocorrelación de los residuos no existe o es aceptable, se busca el mejor modelo de acuerdo a la capacidad predictiva usando la técnica de validación cruzada, que se describe en la aplicación numérica expuesta.

Otro aspecto importante relacionado con el empleo de regresores que son series cronológicas, es su correlación entre sus elementos, lo cual conduce a la multicolinealidad y sus consecuencias. Ello se detecta a través de la matriz de coeficientes de correlación lineal entre regresores y se cuantifica con base en los factores de inflación de la varianza. Estos tópicos serán expuestos en la aplicación numérica.

### Descripción del sistema de río Tempoal

Al río Tempoal lo forman los ríos Hules y Calabozo, aforados por las estaciones Los Hules y Terrerillos, cuyas cuencas de drenaje inician en la frontera del bajo río Pánuco (Región Hidrológica No. 26 parcial), en los estados de Hidalgo y Veracruz (20° 30' N). El río Tempoal tiene un recorrido de sur a norte y es uno de los colectores más importantes del Río Moctezuma, al cual se une por margen derecha en el poblado El Higo, Veracruz. Antes de la estación hidrométrica Tempoal, última del sistema, llega por margen izquierda el río San Pedro aforado en la estación El Cardón. Finalmente, cerca del poblado de Platón Sánchez, Veracruz, está la estación hidrométrica del mismo nombre sobre el río Tempoal. En la Figura 1 se muestra la ubicación y morfología del sistema del Río Tempoal, con base en la cual se adoptó el siguiente orden de regresores:  $x_1$ =Tempoal,  $x_2$ =Terrerillos,  $x_3$ =Los Hules, y  $x_4$ =El Cardón, para tomar en cuenta su posible relación física o de causa-efecto, con la estación Platón Sánchez.

### Información hidrométrica procesada

En el Cuadro 2 se muestran los cinco registros disponibles de volúmenes escurridos anuales en millones de  $m^3$  ( $Mm^3$ ), en las estaciones hidrométricas del sistema del río Tempoal. Tales

number of regressors ( $K$ ) and level of significance  $\alpha$  of the test ( $\alpha=5\%$ , usually). The decision rule is:

If  $d < d_L$  reject  $H_0$   
 If  $d > d_U$  no rejection of  $H_0$   
 If  $d_L \leq d \leq d_U$  the test is inconclusive.

When the test is done to detect negative autocorrelation, the statistic  $4-d$  is used, with limits  $4-d_L$  and  $4-d_U$ , and the same decision rule.

After verifying that the residual autocorrelation does not exist or is acceptable, the best model according to predictive capability is searched using the cross-validation technique, described in the numerical application presented.

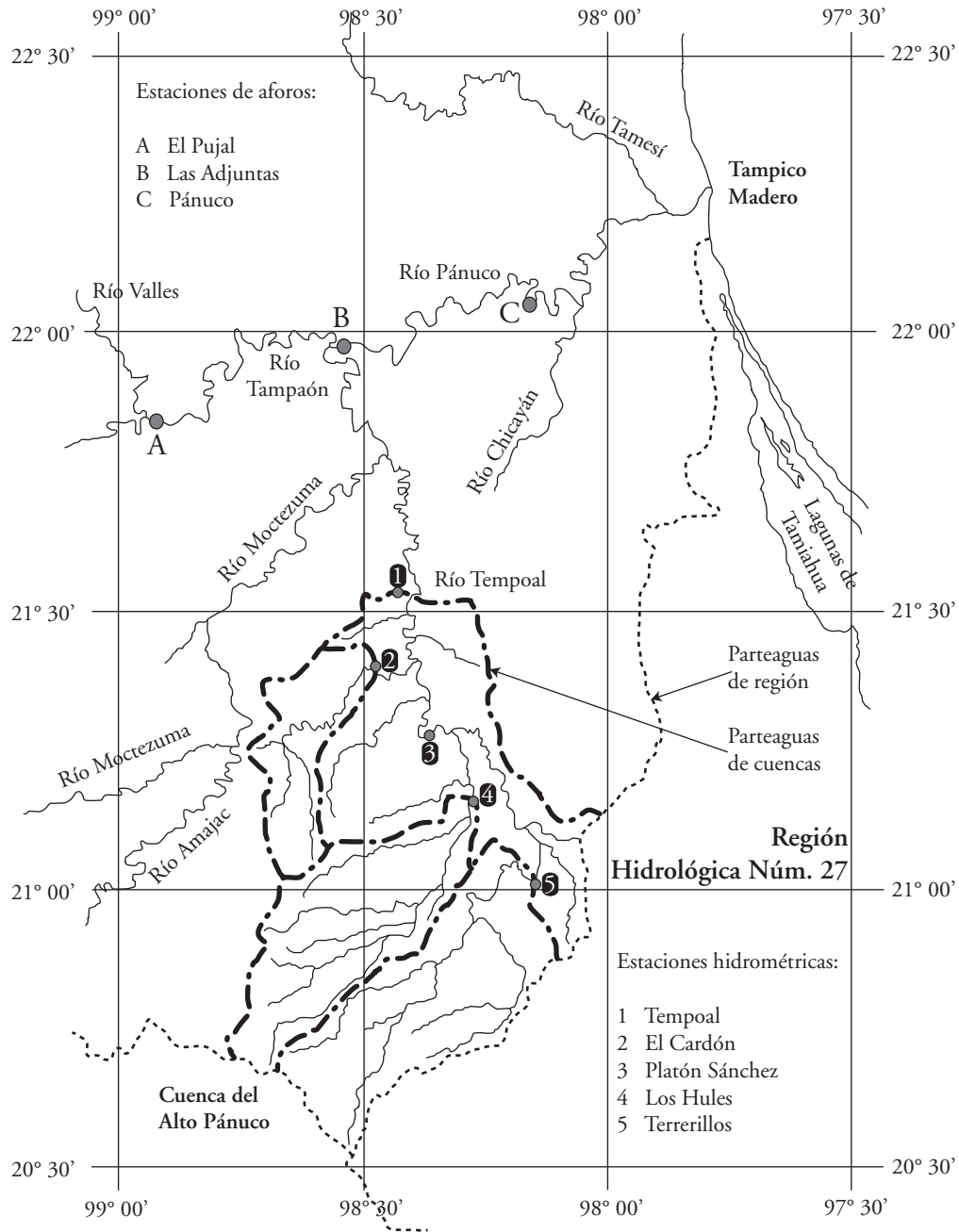
Another important aspect related to the use of time series regressors is the correlation between their elements, leading to multicollinearity and its consequences. This is detected through the matrix of coefficients of linear correlation between regressors and quantified based on the variance inflation factors. These topics will be presented in the numerical application.

### Tempoal river system description

River Tempoal is formed by the rivers Hules and Calabozo, gauged by stations Los Hules and Terrerillos, whose drainage basins begin in the border of the lower Pánuco river (partial Hydrological Region No. 26), in the states of Hidalgo and Veracruz (20° 30' N). River Tempoal has a south-north course and is a major collector of the Moctezuma river to which it joins from the right bank in the town of El Higo, Veracruz. Prior to the Tempoal hydrometric station - the last of the system - the San Pedro river flows on the left bank, gauged by the station El Cardón. Finally, near the town of Platón Sánchez, Veracruz, there is the hydrometric station of the same name on the river Tempoal. Figure 1 shows the location and morphology of the Tempoal river system, based on which the following order of regressors was adopted:  $x_1$ =Tempoal,  $x_2$ =Terrerillos,  $x_3$ =Los Hules, and  $x_4$ =El Cardón, taking into account its possible physical or cause and effect relationship with the Platón Sánchez station.

### Processed hydrometric information

Table 2 shows the five available records of annual runoff volumes in millions of  $m^3$  ( $Mm^3$ ) in the hydrometric stations of the Tempoal river system. Such records are from the BANDAS system (IMTA, 2002) and are presented by increasing order of sizes of drained basins, whose values are: 609, 1269, 1493, 4700 and 5275  $km^2$  for the stations El Cardón, Los Hules, Terrerillos,



**Figura 1. Localización y morfología del sistema del río Tempoal.**  
**Figure 1. Location and morphology of the Tempoal river system.**

registros proceden del sistema BANDAS (IMTA, 2002) y están expuestos por orden creciente de tamaños de cuenca drenada, cuyo valores son: 609, 1269, 1493, 4700 y 5275 km<sup>2</sup>, para las estaciones El Cardón, Los Hules, Terrerillos, Platón Sánchez y Tempoal. Las claves respectivas en tal sistema son: 26286, 26277, 26289, 26433 y 26248.

Platón Sánchez and Tempoal. The respective keys in such a system are: 26286, 26277, 26289, 26433 and 26248.

Table 2 shows that the common period is defined by the Platón Sánchez station, in the period from 1979 to 2002, 24 years, and the feasible expansion for such registration will be of 18 years in the period from 1961 to 1978. Due to missing



En el Cuadro 2 se observa que el periodo común lo define la estación Platón Sánchez, en el lapso de 1979 a 2002, con 24 años y la ampliación factible para tal registro será de 18 años en el periodo 1961 a 1978. Debido a datos faltantes en las estaciones

data in the stations El Cardón, Los Hules and Terrerillos, the common period is reduced to 18 years since it was not considered appropriate to estimate the values of the missing years in order to avoid errors induced by using not real data.

**Cuadro 2. Volúmenes escurridos anuales (Mm<sup>3</sup>) en las estaciones hidrométricas del sistema del río Temporal.**  
**Table 2. Annual runoff volumes (Mm<sup>3</sup>) in the hydrometric stations of the Temporal river system.**

No.	Año	El Cardón	Los Hules	Terrerillos	P. Sánchez	Temporal
18	1961	386.787	1021.195	1211.240	–	3150.302
17	1962	272.019	677.758	628.203	–	1796.844
16	1963	197.102	661.475	668.833	–	1655.044
15	1964	145.934	378.162	324.150	–	1076.755
14	1965	244.780	749.130	865.133	–	2293.958
13	1966	235.217	1011.194	1020.039	–	2786.573
12	1967	548.409	1080.269	1067.163	–	3263.920
11	1968	511.579	945.499	985.655	–	2837.862
10	1969	488.246	1127.562	1336.178	–	3323.340
9	1970	412.411	941.203	944.250	–	2863.385
8	1971	336.091	808.878	1072.324	–	2441.337
7	1972	373.829	950.725	915.385	–	2566.835
6	1973	522.902	1160.392	1243.497	–	3599.619
5	1974	642.511	1312.057	1428.909	–	4296.827
4	1975	570.883	1656.828	1446.878	–	4298.112
3	1976	673.933	1564.284	1868.405	–	4241.779
2	1977	134.540	466.286	521.217	–	1332.365
1	1978	547.106	1376.984	1406.349	–	3688.256
1	1979	284.234	796.465	829.710	2995.751	2103.745
2	1980	227.079	586.212	702.483	1325.674	1586.278
–	<b>1981</b>	<b>728.237</b>	<b>1665.041</b>	–	<b>3666.737</b>	<b>4491.975</b>
3	1982	148.206	359.815	394.620	776.575	880.923
4	1983	271.316	967.822	1190.780	2116.887	2187.518
5	1984	636.325	1832.083	2444.332	4065.671	5057.565
6	1985	361.991	936.464	1390.327	2257.756	2607.572
7	1986	264.761	688.127	891.569	1654.262	1807.878
8	1987	322.006	815.745	1418.647	2018.218	2213.954
9	1988	274.661	729.049	1312.224	1955.431	2325.627
10	1989	288.773	1032.017	958.891	1457.954	1749.932
–	<b>1990</b>	<b>359.339</b>	–	<b>1209.258</b>	<b>2413.499</b>	<b>2680.279</b>
–	<b>1991</b>	<b>713.490</b>	–	<b>1712.530</b>	<b>2682.264</b>	<b>4016.997</b>
11	1992	607.888	1545.657	1929.711	3144.544	4134.539
12	1993	749.586	2230.109	2370.400	4291.278	5629.170
13	1994	305.695	809.173	703.788	1228.833	1634.689
14	1995	343.729	1081.871	843.071	1542.547	1861.508
15	1996	185.781	778.378	594.657	1034.440	1250.085
16	1997	152.708	651.170	520.502	1046.949	1180.939
–	<b>1998</b>	–	<b>950.051</b>	<b>1523.011</b>	<b>2475.934</b>	<b>3314.234</b>
–	<b>1999</b>	–	<b>1294.764</b>	<b>1481.634</b>	<b>2456.724</b>	<b>3206.621</b>
–	<b>2000</b>	–	<b>425.543</b>	<b>839.818</b>	<b>1426.170</b>	<b>1710.830</b>
17	2001	277.771	535.401	805.264	1281.262	1929.465
18	2002	172.370	644.391	683.526	1058.208	1411.568
$\bar{X}$	–	382.570	943.131	1114.745	2098.899	2678.262
$S$	–	182.892	416.849	486.504	983.737	1162.200
$C_v$	–	0.478	0.425	0.436	0.469	0.434
$C_s$	–	0.558	0.960	0.868	0.792	0.604
$C_k$	–	2.315	4.082	3.949	3.168	2.843

El Cardón, Los Hules y Terrerillos, el periodo común se reduce a 18 años, pues no se consideró conveniente estimar los valores de los años faltantes, para evitar inducir errores por emplear datos no reales.

En la parte inferior de la Figura 2 se comparan los registros de volúmenes escurridos anuales de las estaciones hidrométricas El Cardón y Los Hules, y en la porción superior se muestra el relativo a la estación Terrerillos. En la Figura 3 se comparan los registros correspondientes de las estaciones Platón Sánchez y Tempoal. En las cuatro series cronológicas dibujadas no se observa tendencia ni saltos en la media y la variabilidad tampoco cambia, por lo cual tales series probablemente sean estacionarias.

**Verificación de requerimientos estadísticos**

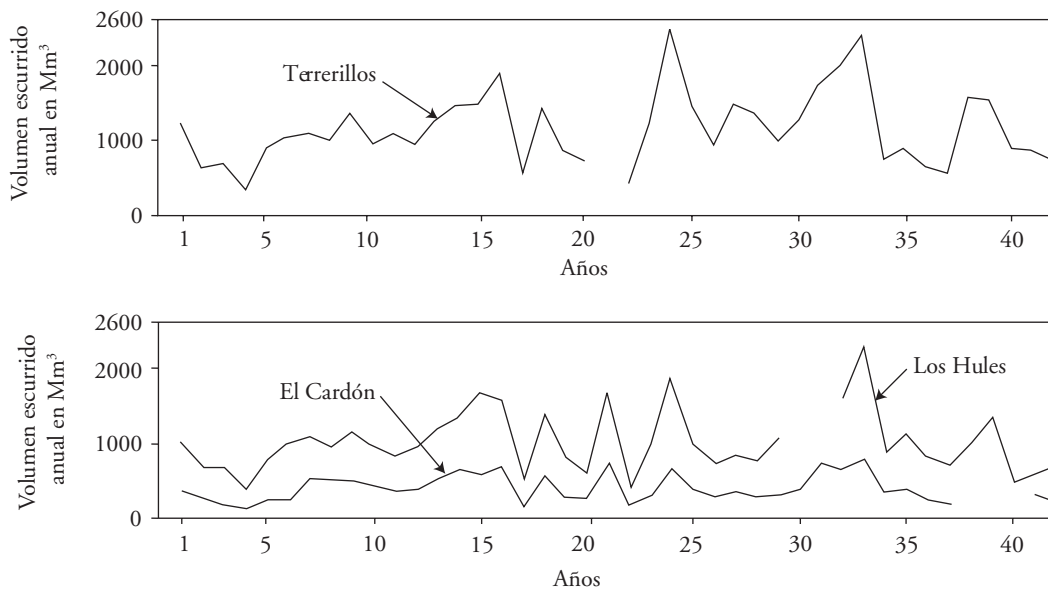
Primeramente se verificó si es factible aceptar que los registros por procesar (Cuadro 2) proceden de poblaciones Normales, lo cual se realizó con la prueba W de Shapiro y Wilk (1965) y se encontró que sólo los registros de Los Hules y El Cardón no proceden de una población Normal. En la parte inferior del Cuadro 2 se muestran los siguientes parámetros estadísticos insesgados de los registros disponibles: media aritmética ( $\bar{X}$ ), desviación estándar ( $S$ ) y coeficientes de variación ( $Cv$ ), asimetría ( $Cs$ ) y curtosis ( $Ck$ ). Se observa que los registros procedentes de poblaciones Normales presentan coeficientes de asimetría y curtosis cercanos a cero y tres, correspondientes a la distribución Normal y que precisamente los registros de El Cardón y Los Hules tienen los valores del  $Ck$  más distantes de tres. Sin embargo, dada la

At the bottom of Figure 2 the records of the annual runoff volumes of the gauging stations El Cardón and Los Hules are contrasted, and the upper portion shows those registered in the Terrerillos station. Figure 3 compares the records of stations Platón Sánchez and Tempoal. The four time series drawn show no trends or jumps in the mean and the variability does not change either, so such series are likely to be stationary.

**Verification of statistical requirements**

First, we verified whether it is feasible to accept that the records to be processed (Table 2) are from Normal populations, which was performed with the W test by Shapiro and Wilk (1965) and found that only the records of Los Hules and El Cardón do not come from a Normal population. Table 2, at the bottom, shows the following unbiased statistical parameters of the records available: arithmetic mean ( $\bar{X}$ ), standard deviation ( $S$ ) and coefficients of variation ( $Cv$ ), skewness ( $Cs$ ) and kurtosis ( $Ck$ ). We observed that the records from Normal populations have coefficients of skewness and kurtosis near zero and three, corresponding to Normal distribution, and precisely the records of El Cardón and Los Hules have  $Ck$  values further distant from three. However, given the similarity between the values of  $Cv$  and  $Cs$  in all records, we did not consider necessary to do some change to work with normalized data.

Specific tests were also applied to search for deterministic components such as persistence, trends and changes in the mean or the variance; and the tests were: serial correlation coefficient



**Figura 2. Registros de volúmenes escurridos anuales en las estaciones hidrométricas Terrerillos, Los Hules y El Cardón. Figure 2. Records of annual runoff volumes in the hydrometric stations of Terrerillos, Los Hules and El Cardón.**

similitud entre los valores de  $C_v$  y  $C_s$  de todos los registros, no se consideró necesario aplicar alguna transformación para trabajar con datos normalizados.

Además se aplicaron pruebas específicas para buscar componentes determinísticas como persistencia, tendencia y cambios en la media o la varianza; las pruebas fueron: coeficiente de correlación serial de orden uno, Kendall, Cramer y Bartlett (WMO, 1971; Ruiz, 1977). Únicamente se encontró que el registro de la estación Terrerillos muestra persistencia ( $r_1 = 0.290$ ).

## RESULTADOS Y DISCUSIÓN

### Detección de regresores colineales

Una consecuencia lógica del uso de registros hidrológicos ubicados dentro de una región homogénea, es que probablemente ellos serán semejantes, es decir, que sus periodos de años secos y húmedos son coincidentes y por tanto mostrarán correlación entre ellos (Figura 2 y 3). Entonces la detección de registros colineales se realiza buscando correlaciones altas ( $r_{xy} > 0.80$ ) en la matriz de coeficientes de correlación lineal (Cuadro 3) y observando sus consecuencias en los coeficientes de los regresores, ya que cuando un regresor está correlacionado con otro, su coeficiente cambiará drásticamente al estar los dos en la ecuación de regresión.

En el Cuadro 3 se observa que todos los regresores usados son colineales, y la mayor correlación fue entre Tempoal ( $x_1$ ) y El Cardón ( $x_4$ ). Como consecuencia, en el Cuadro 4 se observa como los coeficientes de cada regresor (ec. 3) cambian debido a la presencia de otro(s) en la ecuación de regresión. Tales cambios ocurren en magnitud y también en signo; además, los coeficientes de los regresores  $x_1$  y  $x_4$  son los más estables o insensibles a la presencia de otro(s) en la ecuación.

Los factores de inflación de la varianza (VIF, de *variance inflation factors*) constituyen un diagnóstico cuantitativo importante, pues de acuerdo con los ejemplos numéricos de Montgomery *et al.* (2002), cuando exceden a 1000 implican gravísimos problemas de multicolinealidad, menores de 100 problemas aceptables y sin problemas cuando no exceden de 10. La expresión para su estimación práctica es:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (10)$$

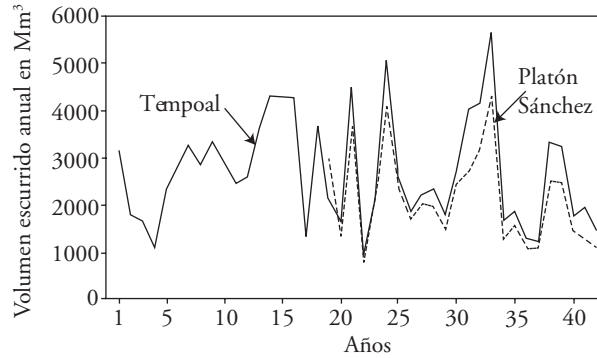


Figura 3. Registros de volúmenes escurridos anuales en las estaciones hidrométricas Tempoal y Platón Sánchez.

Figure 3. Records of annual runoff volumes in the hydrometric stations of Tempoal and Platón Sánchez.

of order one, Kendall, Cramer and Bartlett (WMO, 1971; Ruiz, 1977). We only found that the record of the Terrerillos station shows persistence ( $r_1 = 0.290$ ).

## RESULTS AND DISCUSSION

### Detection of collinear regressors

A logical consequence of the use of hydrological records located within a homogeneous region is that they will be probably similar, *ie.*, their periods of dry and wet years will coincide and thus show correlation between them (Figures 2 and 3). Then the detection of collinear records is performed seeking high correlations ( $r_{xy} > 0.80$ ) in the matrix of linear correlation coefficients (Table 3) and observing their impact on the regressor coefficients, since when a regressor is correlated with another, its coefficient will drastically change as both are in the regression equation.

Table 3 shows that all the regressors used are collinear, and the highest correlation was between Tempoal ( $x_1$ ) and El Cardón ( $x_4$ ). As a result, Table 4 shows how the coefficients of each regressor (Eq. 3) change due to the presence of another (others) in the regression equation. Such changes occur in magnitude and in sign; in addition, the coefficients of regressors  $x_1$  and  $x_4$  are the most stable or insensitive to the presence of another (others) in the equation.

The variance inflation factors (VIF) constitute a significant quantitative diagnosis because according to the numerical examples by Montgomery *et al.*

**Cuadro 3. Matriz de coeficientes de correlación lineal ( $r_{xy}$ ) para los datos del sistema del río Tempoal.**  
**Table 3. Matrix of linear correlation coefficients ( $r_{xy}$ ) for Tempoal river system data.**

Variable dependiente	Regresores				
	Tempoal (x1)	Terrerillos (x2)	Los Hules (x3)	El Cardón (x4)	Platón Sánchez (y)
x1	1.000				
x2	0.970	1.000			
x3	0.942	0.899	1.000		
x4	0.977	0.937	0.953	1.000	
y	0.947	0.920	0.881	0.909	1.000

**Cuadro 4. Coeficientes de regresión de todas las ecuaciones posibles, para los datos del sistema del río Tempoal.**  
**Table 4. Regression coefficients of all possible equations for Tempoal river system data.**

Regresores:	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$
x1	227.9714	0.7496			
x2	206.4563		1.5780		
x3	139.7417			1.9234	
x4	140.6007				5.5697
x1,x2	225.9117	0.7376	0.0268		
x1,x3	261.3518	0.8224		-0.2129	
x1,x4	305.7837	1.0287			-2.2124
x2,x3	107.0753		1.1411	0.6181	
x2,x4	123.6428		0.9613		2.3515
x3,x4	114.1808			0.3665	4.5889
x1,x2,x3	263.0489	0.8307	-0.0164	-0.2157	
x1,x2,x4	315.5988	1.0797	-0.0920		-2.2905
x1,x3,x4	304.1462	1.0270		0.0189	-2.2502
x2,x3,x4	103.4557		0.9497	0.2829	1.6333
x1,x2,x3,x4	314.6669	1.0783	-0.0910	0.0096	-2.3088

donde  $R_j^2$  es el coeficiente de determinación múltiple obtenido haciendo la regresión de  $x_j$  con las demás variables regresoras. Los valores de los VIF para las variables regresoras  $x_1$  (Tempoal),  $x_2$  (Terrerillos),  $x_3$  (Los Hules) y  $x_4$  (El Cardón) fueron 49.53, 18.04, 11.33 y 28.15. Las magnitudes anteriores ratifican los resultados obtenidos con los valores del Cuadro 3 y establecen que es factible proseguir con la selección y validación de modelos.

**Selección de ecuaciones de regresión**

Con base en los resultados del Cuadro 5, se seleccionaron tres ecuaciones de regresión. La primera incluye sólo a  $x_1$  (Tempoal) como regresor y corresponde a los menores valores del cuadrado medio de los residuos (Ecuación 6) y de la estadística de Ma-

(2002), when they exceed 1000 serious problems of multicollinearity emerge; when they are less than 100, problems are acceptable, and no trouble at all if they do not exceed 10. The expression for their estimation is:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{10}$$

where  $R_j^2$  is the coefficient of multiple determination obtained by regressing  $x_j$  with other regressor variables. The VIF values for regressor variables  $x_1$  (Tempoal),  $x_2$  (Terrerillos),  $x_3$  (Los Hules) and  $x_4$  (El Cardón) were 49.53, 18.04, 11.33 and 28.15, respectively. The above figures confirm the results obtained with the values in Table 3 and state that it is

**Cuadro 5. Resultados de los criterios de evaluación de todas las regresiones posibles, en el sistema del río Tempoal.**  
**Table 5. Results of the evaluation criteria of all possible regressions in the Tempoal river system.**

Regresores:	$p$	$SC_{Res}(p)$	$R^2_p$	$R^2_{a,p}$	$CM_{Res}(p)$	$C_p$
ninguno	1	18 282 790.0	–	–	1 075 458.0	141.57
$x_1$	2	1 880 401.0	<b>0.89715</b>	<b>0.89072</b>	<b>117 525.1</b>	<b>2.00</b>
$x_2$	2	2 810 706.0	0.84626	0.83666	175 669.1	9.92
$x_3$	2	4 078 638.0	0.77691	0.76297	254 914.9	20.70
$x_4$	2	3 183 104.0	0.82590	0.81501	198 944.0	13.08
$x_1, x_2$	3	1 880 139.0	0.89716	0.88345	125 342.6	4.00
$x_1, x_3$	3	1 860 841.0	0.89822	0.88465	124 056.1	3.83
$x_1, x_4$	3	1 771 527.0	<b>0.90310</b>	<b>0.89018</b>	<b>118 101.8</b>	<b>3.07</b>
$x_2, x_3$	3	2 529 878.0	0.86163	0.84318	168 658.5	9.53
$x_2, x_4$	3	2 482 138.0	0.86424	0.84613	165 475.9	9.12
$x_3, x_4$	3	3 135 656.0	0.82849	0.80562	209 043.7	14.68
$x_1, x_2, x_3$	4	1 860 748.0	0.89822	0.87641	132 910.6	5.83
$x_1, x_2, x_4$	4	1 768 584.0	<b>0.90327</b>	<b>0.88254</b>	<b>126 327.4</b>	<b>5.05</b>
$x_1, x_3, x_4$	4	1 771 404.0	0.90311	0.88235	126 528.9	5.07
$x_2, x_3, x_4$	4	2 453 973.0	0.86578	0.83701	175 283.8	10.88
$x_1, x_2, x_3, x_4$	5	1 768 552.0	0.90327	0.87350	136 042.5	7.05

llows (Ecuación 7). La segunda con regresores  $x_1$  y  $x_4$  (El Cardón) presenta el coeficiente de determinación múltiple ajustado más alto (Ecuación 5) y el segundo cuadrado medio de los residuos (Ecuación 6) más bajo, en los modelos de dos regresores. La tercera tiene por regresores  $x_1$ ,  $x_2$  (Terrerillos) y  $x_4$ , con el coeficiente de determinación múltiple mayor (Ecuación 4) y el menor cuadrado medio de los residuos en los modelos de tres regresores.

**Análisis de residuales**

En el Cuadro 6 se muestran las estimaciones de la variable dependiente ( $\hat{y}_t$ ) de cada uno de los tres modelos seleccionados y sus residuos en el periodo 1979-2002, así como sus respectivos valores del estadístico  $d$  (Ecuación 8). Para  $n=18$ ,  $\alpha=5.0\%$  y  $K=1, 2$  y  $3$  se obtienen de la tabla de valores límite de Durbin-Watson (Makrindakis *et al.*, 1983):  $d_L=1.16$  y  $d_U=1.39$ ,  $d_L=1.05$  y  $d_U=1.53$ ,  $d_L=0.93$  y  $d_U=1.69$ , por lo cual la primera serie de residuos tiene autocorrelación positiva y para las otras dos la prueba no es concluyente. Lo anterior descarta al modelo del subconjunto  $x_1$ .

Las tres gráficas de residuales (Figura 4) son similares, tienen magnitudes bastante reducidas, excepto el primer residuo, generado por un escurrimiento en Platón Sánchez que es incluso mayor que de Tempoal (Cuadro 2), lo cual es incorrecto. Por tanto, los

feasible to proceed with the selection and validation of models.

**Selection of regression equations**

Based on the results in Table 5, we selected three regression equations. The first includes only  $x_1$  (Tempoal) as a regressor and corresponds to the lowest values of mean square of residuals (Equation 6) and Mallows statistics (Equation 7). The second with  $x_1$  and  $x_4$  regressors (El Cardón) presents the highest adjusted multiple coefficient of determination (Equation 5) in the models of two regressors and the second lowest mean square of the residuals. The third has  $x_1$ ,  $x_2$  (Terrerillos) and  $x_4$  regressors, with the highest coefficient of multiple determination (Equation 4) and the lowest mean square of the residuals in the models of three regressors.

**Analysis of residuals**

Table 6 shows the estimates of the dependent variable ( $\hat{y}_t$ ) of each of the three selected models and their residuals in the period 1979-2002, and their respective values of the statistic  $d$  (Equation 8), for  $n=18$ ,  $\alpha=5.0\%$  and  $K=1, 2$  and  $3$  are obtained from the table of limit values by Durbin-Watson (Makrindakis *et al.*, 1983):  $d_L=1.16$  and  $d_U=1.39$ ,  $d_L=1.05$  and  $d_U=1.53$ ,  $d_L=0.93$  and  $d_U=1.69$ , so the

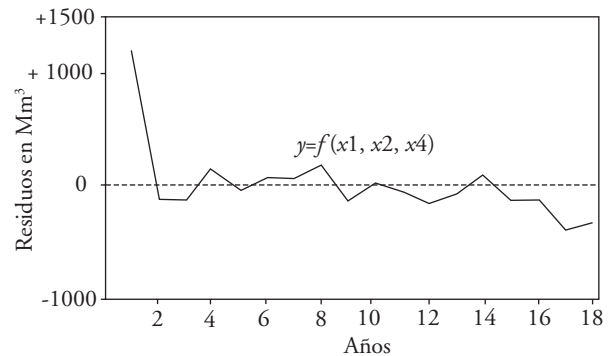
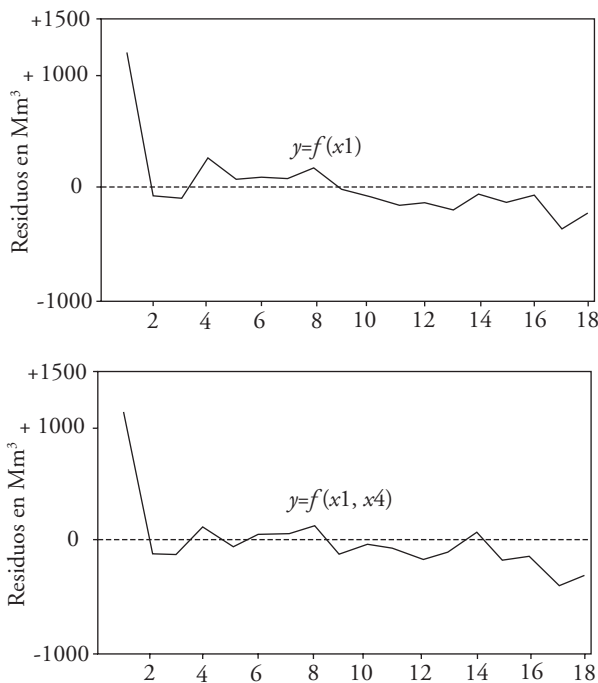
**Cuadro 6. Estimaciones de la variable dependiente ( $\hat{y}_t$ ) obtenidas con cada uno de los tres modelos seleccionados y sus residuos respectivos.**

**Table 6. Estimates of the dependent variable ( $\hat{y}_t$ ) obtained with each of the three selected models and their respective residuals.**

Año	$\hat{y}_t = f(x1)$	Residuo	$\hat{y}_t = f(x1, x4)$	Residuo	$\hat{y}_t = f(x1, x2, x4)$	Residuo
1979	1804.939	1190.812	1841.067	1154.684	1859.641	1136.110
1980	1417.045	-91.371	1435.198	-109.524	1443.550	-117.876
1982	888.311	-111.736	884.098	-107.523	890.960	-114.385
1983	1867.735	249.152	1955.824	161.063	1946.461	170.426
1984	4019.122	46.549	4100.695	-35.024	4093.871	-28.200
1985	2182.608	75.149	2187.324	70.432	2173.944	83.812
1986	1583.157	71.105	1579.791	74.471	1579.105	75.157
1987	1887.551	130.667	1870.872	147.346	1837.935	180.283
1988	1971.261	-15.830	2090.496	-135.065	2076.743	-121.312
1989	1539.721	-81.766	1467.057	-9.103	1455.348	2.606
1992	3327.222	-182.678	3214.093	-69.549	3209.760	-65.216
1993	4447.597	-156.319	4438.127	-146.849	4458.410	-167.132
1994	1453.334	-224.501	1311.069	-82.236	1315.630	-86.797
1995	1623.358	-80.811	1460.251	82.296	1460.595	81.952
1996	1165.035	-130.595	1180.724	-146.284	1185.076	-150.636
1997	1113.203	-66.254	1182.764	-135.815	1192.995	-146.046
2001	1674.298	-393.036	1676.084	-394.822	1688.523	-407.261
2002	1286.083	-227.875	1376.512	-318.304	1381.971	-323.763
Máximo	-	1190.812	-	1154.684	-	1136.110
mínimo	-	-393.036	-	-394.822	-	-407.261
d	-	1.076	-	1.098	-	1.109

tres modelos seleccionados tienen buena capacidad predictiva.

first series of residuals has positive autocorrelation and for the other two the test is inconclusive; this rules out the model of subset x1.



**Figura 4. Gráficas residuales de cada modelo seleccionado. Figure 4. Residual graphs of each selected model.**

**Análisis de validación cruzada**

Cuando los datos de los regresores son series cronológicas, el tiempo es usado para la formación de los datos para estimación y para predicción. El lapso conocido de datos se dividió en dos sub-periodos con nueve valores cada uno. En el Cuadro 7 se muestran los coeficientes de regresión estimados con cada sub-periodo considerado como de estimación y en el Cuadro 8 están las estimaciones y sus correspondientes residuos, para cada subperiodo complementario o de predicción.

En el Cuadro 7 se observan cambios drásticos de un sub-periodo al otro en los coeficientes de los regresores de  $x_2$  y  $x_4$ , además el coeficiente  $a_0$  o constante también cambia bastante. Lo anterior se debe a la presencia de un ciclo húmedo y otro seco en el registro disponible en Platón Sánchez (Figuras 3 y 4).

El análisis de residuales por sub-periodos (Cuadro 8) muestra similitud con los mostrados en el Cuadro 6 y la Figura 4, ya que primero hay residuos positivos y después negativos. Los resultados de Cuadro 8 definen los modelos tercero y segundo como más convenientes por su mejor capacidad predictiva, medida por la menor suma de residuos en cada sub-periodo de predicción; es decir el modelo con subconjunto  $x_1, x_2$  y  $x_4$  y el del subconjunto  $x_1$  y  $x_4$ .

**Validación con datos nuevos**

A través de la Dirección Local San Luis Potosí de la CONAGUA se intentó conseguir las magnitudes del volumen escurrido anual después del año 2002, en las estaciones del sistema del río Tempoal pero

The three residual graphs (Figure 4) are similar, with magnitudes quite small, except the first residual generated by a runoff in Platón Sánchez that is even higher than in Tempoal (Table 2), which is incorrect. Therefore, the three selected models have good predictive ability.

**Cross-validation analysis**

When data from the regressors are time series, time is used for estimation and prediction data. The data known period was divided into two sub-periods, with nine values each. Table 7 lists the estimated regression coefficients for each sub-period considered of estimation, and Table 8 includes the estimates and their corresponding residuals for each complementary or prediction sub-period.

Table 7 shows drastic changes from one sub-period to another in the coefficients of regressors of  $x_2$  and  $x_4$ ; and  $a_0$  or constant coefficient also changes a lot. This is due to the presence of a wet and a dry cycle in the record available in Platón Sánchez (Figures 3 and 4).

The analysis of residuals by sub-periods (Table 8) shows similarity to those shown in Table 6 and Figure 4, because first there are positive residuals and then negative. According to Table 8 results, the third and second models are the most suitable because of their better prediction capacity, measured by the smallest amount of residuals in each sub-period prediction; that is, the model with subset  $x_1, x_2$  and  $x_4$  and that with subset  $x_1$  and  $x_4$ .

**Cuadro 7. Coeficientes de regresión de los tres modelos seleccionados, para los datos del sistema del río Tempoal por subperiodos.**

**Table 7. Regression coefficients of the three selected models for data of the Tempoal river system by subperiods.**

Regresores:	$a_0$	$a_1$	$a_2$	$a_4$	$R_p^2$
Subperiodo 1979-1988					
$x_1$	387.5886	0.7548			0.9087
$x_1, x_4$	428.5020	0.8768		-1.0400	0.9089
$x_1, x_2, x_4$	464.4617	1.4139	-1.1002	-0.9847	0.9270
Subperiodo 1989-2002					
$x_1$	62.9142	0.7468			0.9960
$x_1, x_4$	48.1449	0.6983		0.3699	0.9961
$x_1, x_2, x_4$	25.6819	0.5447	0.4331	0.1489	0.9965

**Cuadro 8. Estimaciones por subperiodos de la variable dependiente ( $\hat{y}_t$ ) obtenidas con cada uno de los tres modelos seleccionados y sus residuos respectivos.**

**Table 8. Subperiod estimates of the dependent variable ( $\hat{y}_t$ ) obtained with each of the three selected models and their respective residuals.**

Año	$\hat{y}_t = f(x1)$	Residuo	$\hat{y}_t = f(x1, x4)$	Residuo	$\hat{y}_t = f(x1, x2, x4)$	Residuo
Subperiodo 1979-1988						
1979	1633.978	1361.773	1622.310	1373.441	1573.297	1422.454
1980	1247.537	78.137	1239.826	85.848	1227.811	97.863
1982	720.782	55.793	718.109	58.466	698.513	78.062
1983	1696.539	420.348	1676.028	440.859	1773.381	343.506
1984	3839.872	225.799	3815.172	250.499	3934.005	131.666
1985	2010.233	247.524	2002.890	254.866	2102.117	155.639
1986	1413.026	241.236	1408.506	245.756	1436.023	218.239
1987	1716.281	301.937	1713.241	304.977	1894.016	124.202
1988	1799.678	155.753	1773.705	181.726	1901.706	53.726
Máximo	-	1361.773	-	1373.441	-	1422.454
Mínimo	-	55.793	-	58.466	-	53.726
Suma	-	3088.300	-	3196.438	-	2625.357
Subperiodo 1989-2002						
1989	1708.432	-250.478	1662.506	-204.552	1599.367	-141.413
1992	3508.325	-363.781	3421.432	-276.888	3588.630	-444.085
1993	4636.468	-345.190	4584.547	-293.269	5077.506	-786.228
1994	1621.447	-392.614	1543.863	-315.030	1700.419	-471.586
1995	1792.649	-250.102	1703.181	-160.634	1830.428	-287.881
1996	1331.149	-296.709	1331.355	-296.915	1394.777	-360.337
1997	1278.958	-232.009	1305.124	-258.175	1411.163	-364.214
2001	1843.943	-562.681	1831.361	-550.099	2033.056	-751.794
2002	1453.036	-394.828	1486.889	-428.681	1538.531	-480.323
Máximo	-	-562.681	-	-550.099	-	-751.794
Mínimo	-	-232.009	-	-160.634	-	-141.413
Suma	-	-3088.392	-	-2784.243	-	-4087.861

no se obtuvo tal información en la estación Platón Sánchez, únicamente en el resto y sólo hasta 2006. Por tanto, no fue posible realizar una validación con datos nuevos.

### Estimaciones finales y su selección

En el Cuadro 9 se muestran los volúmenes escurridos anuales estimados en la estación Platón Sánchez, con cada uno de los dos modelos o ecuaciones de regresión seleccionados, y con sus respectivos parámetros estadísticos. Se observa que las dos estimaciones conducen a registros bastante similares, ya que sus parámetros estadísticos ( $C_v$ ,  $C_s$ ,  $C_k$ ) y valores medios son casi idénticos. Lo anterior genera confianza en la predicción de los valores buscados y se puede adoptar cualquiera de las dos series. Si hay

### Validation with new data

Through the San Luis Potosí CONAGUA local administration, we tried to obtain the annual runoff volume magnitudes after 2002, at the stations of the Tempoal river system, but there was no such information at the Platón Sánchez station, only in the rest and until 2006. Therefore, it was not possible to perform validation with new data.

### Final estimates and their selection

Table 9 presents the annual runoff volumes estimated in the Platón Sánchez station, with each of the two models or regression equations selected, and their respective statistical parameters. It is observed that the two estimates lead to very similar records,



que seleccionar sólo una de ellas, se recomienda la primera por la magnitud menor en su media, lo cual implica cierta garantía en estimaciones asociadas con la disponibilidad.

### CONCLUSIONES

El ejemplo descrito para el sistema del río Tempoal, permitió exponer con detalle los análisis previos y de regresión lineal múltiple realizados con la información hidrológica disponible y sus consecuencias.

Aunque la aplicación numérica expuesta tiene un problema grave de multicolinealidad, lo cual es muy probable que ocurra en todas las aplicaciones prácticas de ampliación de registros de escurrimiento y de lluvia anuales, los criterios expuestos para selección de regresores conducen a resultados consistentes y son una ayuda efectiva en la búsqueda de la mejor ecuación de regresión lineal múltiple.

Cuando los criterios de selección sugieren ecuaciones de regresión diferentes, sus resultados se deben analizar a través del análisis de residuales y las validaciones cruzada y con datos nuevos, para adoptar el modelo candidato más conveniente. Pero la similitud en los resultados de tales modelos, como ocurrió con los datos del sistema del río Tempoal, origina confianza en las estimaciones y en aquéllas adoptadas.

### AGRADECIMIENTOS

Se agradecen los comentarios y sugerencias de los dos árbitros anónimos y del editor asignado, los cuales permitieron completar el trabajo en tópicos no tratados pero relevantes al tema, como: análisis de residuales, factores de inflación de la varianza y validación cruzada.

### LITERATURA CITADA

Campos A., D. F. 2003. Ajuste de curvas. *In*: Introducción a los Métodos Numéricos: Software en Basic y Aplicaciones en Hidrología Superficial. Librería Universitaria Potosina. San Luis Potosí, S.L.P. pp: 93–127.

IMTA (Instituto Mexicano de Tecnología del Agua). 2002. Banco Nacional de Datos de Aguas Superficiales (BANDAS). Secretaría de Medio Ambiente y Recursos Naturales– Comisión Nacional del Agua–IMTA. Jiutepec, Morelos. 8 CD.

Makrindakis, S., S. C. Wheelwright, and V. E. McGee. 1983. Multiple regression. *In*: Forecasting: Methods and Applications. John Wiley & Sons. New York, U.S.A. Second edition. pp: 246–317.

**Cuadro 9. Volúmenes escurridos anuales (Mm<sup>3</sup>) en la estación Platón Sánchez, estimados con cada modelo seleccionado.**

**Table 9. Annual runoff volumes (Mm<sup>3</sup>) in Platón Sánchez station, estimated with each model selected.**

Año	Segundo modelo $y = f(x_1, x_4)$	Tercer modelo $y = f(x_1, x_2, x_4)$
1961	2690.772	2719.610
1962	1552.382	1574.797
1963	1572.259	1589.555
1964	1090.577	1114.088
1965	2124.027	2152.125
1966	2651.937	2691.653
1967	2450.078	2485.343
1968	2093.275	2117.186
1969	2644.308	2662.553
1970	2338.930	2375.697
1971	2073.619	2083.040
1972	2119.228	2146.540
1973	2851.843	2889.999
1974	3304.439	3351.752
1975	3464.230	3515.550
1976	3178.292	3179.910
1977	1378.731	1398.037
1978	2889.475	2915.278
MAX	3464.230	3515.550
mín	1090.577	1114.088
$X$	2359.356	2386.818
$S$	673.056	678.284
$C_v$	0.285	0.284
$C_s$	-0.211	-0.201
$C_k$	2.803	2.808

as their statistical parameters ( $C_v$ ,  $C_s$ ,  $C_k$ ) and mean values are almost identical. This builds confidence in the prediction of the required values and either of the two series can be adopted. If only one of them must be selected, we recommend the first for the lowest amount in its mean, which implies certain guarantee associated with the availability of estimates.

### CONCLUSIONS

The example of the Tempoal river system enabled to set out in detail the preliminary and multiple linear regression analyses that were performed with the available hydrological data and their consequences.

Although the numerical application presented has a serious problem of multicollinearity, which is very likely to occur in all practical applications to extend runoff records and annual rainfall, the criteria put forward for the selection of regressors

- McCuen, R. H. 1998. Stepwise regression. *In: Hydrologic Analysis and Design*. Prentice Hall. New Jersey, U.S.A. pp: 84–87.
- Montgomery, D. C., E. A. Peck, y G. G. Vining. 2002. Selección de variable y construcción del modelo. *In: Introducción al Análisis de Regresión Lineal*. Compañía Editorial Continental. México, D. F. pp. 261–290.
- Gilroy, E. J. 1970. Reliability of a variance estimate obtained from a sample augmented by multivariate regression. *Water Resources Res.* 6: 1595–1600.
- Ruiz M., L. 1977. Condiciones paramétricas del análisis de varianza. *In: Métodos Estadísticos de Investigación*. Instituto Nacional de Estadística. Madrid, España. pp. 233–249.
- Salas, J. D., J. A. Raynal, Z. S. Tarawneh, T. S. Lee, D. Frevert, and T. Fulp. 2008. Extending short record of hydrologic data. *In: Singh, V. P. (ed). Hydrology and Hydraulics*. Water Resources Publications. Highlands Ranch, Colorado, U.S.A. pp: 717–760.
- Shapiro, S. S., and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52: 591–611.
- WMO (World Meteorological Organization). 1971. Standard tests of significance to be recommended in routine analysis of climatic fluctuations. *In: Climatic Change*. Technical Note No. 79. Secretariat of the WMO. Geneva, Switzerland. pp: 58–71.

led to consistent results and can be an effective aid in the search for the best multiple linear regression equation.

When the selection criteria suggest different regression equations, results should be analyzed through the analysis of residuals and cross- and new data validations in order to adopt the most suitable candidate model. But the similarity in the results of such models, as it occurred with the Tempoal river system data, creates confidence in the estimates as well as in those adopted.

—End of the English version—

-----\*-----