TRANSFERENCIA REGIONAL DE INFORMACIÓN HIDROLÓGICA MEDIANTE REGRESIÓN LINEAL MÚLTIPLE DE TIPO RIDGE

REGIONAL TRANSFERENCE OF HYDROLOGIC INFORMATION THROUGH MULTIPLE LINEAR REGRESSION OF RIDGE TYPE

Daniel F. **Campos-Aranda**

Facultad de Ingeniería de la Universidad Autónoma de San Luis Potosí. Genaro Codina Número 240. 78280 San Luis Potosí, San Luis Potosí. (campos_aranda@hotmail.com)

Resumen

Cuando se emplean registros largos de escurrimiento, lluvia o crecientes anuales de una región con respuesta hidrológica similar, para ampliar una serie corta a través de la técnica estadística de regresión lineal múltiple (RLM), es probable que tales registros por su semejanza intrínseca den origen a un problema de multicolinealidad. Tal problema se debe detectar y cuantificar para saber si es aceptable, moderada, fuerte o grave y buscar soluciones alternativas al método de ajuste de la RLM por mínimos cuadrados de los residuos. En este estudio se diagnosticó la multicolinealidad mediante factores de inflación de la variancia e índices de condición, basados en los eigenvalores. Además se presenta como método alternativo el ajuste sesgado de la RLM, conocido como regresión Ridge. Una aplicación numérica en el sistema del río Tempoal, de la Región Hidrológica No. 26 (Pánuco, México), se describió para completar el registro corto de volúmenes escurridos anuales de la estación hidrométrica Platón Sánchez, con base en las otras cuatro estaciones de aforos que cuentan con registros amplios. Se concluye que las principales ventajas de la regresión Ridge son la facilidad de manejo de transferencia con seis o más regresores y la sencillez de su implementación y desarrollo a través de la traza Ridge.

Palabras clave: multicolinealidad, factores de inflación de la varianza, eigenvalores, eigenvectores, índices de condición, traza Ridge, homogeneidad regional, Río Tempoal.

Abstract

When annual long records are used of runoff, rainfall or flooding of a region with similar hydrological response, to amplify short series through the statistical technique of multiple linear regression (MLR), it is likely that those records by reason of their intrinsic similarity will lead to a problem of multicollinearity. This problem should be detected and quantified to know if it is acceptable, moderate, strong or serious and look for alternative solutions to the fitting method of the MLR by least squares of the residuals. In this study a diagnostic was made of multicollinearity through variance inflation factors and condition indices based on the eigenvalues. In addition, the biased fitting of the MLR is presented as an alternative method, known as Ridge regression. A numerical application in the system of the Tempoal river, of Hydrological Region No. 26 (Pánuco, México), was described to complete the short record of runoff volumes of the Platón Sánchez hydrometric station, based on the other four measuring stations that have long records. It is concluded that the principal advantages of Ridge regression are the ease of handling of transference with six or more regressions and the simplicity of its implementation and development by means of the Ridge trace.

Key words: multicollinearity, variance inflation factors, eigenvalues, eigenvectors, condition indices, Ridge trace, regional homogeneity, Tempoal River.

Introducción

La transferencia regional de información hidrológica consiste en la utilización de registros largos de escurrimiento, lluvia o crecientes

Introduction

The regional transference of hydrological information consists of the utilization of annual long records of runoff, rainfall or

* Autor responsable ❖ Author for correspondence.
Recibido: Diciembre, 2012. Aprobado: Abril, 2013.
Publicado como ARTÍCULO en Agrociencia 47: 411-427. 2013.

anuales, para ampliar series cortas disponibles en estaciones hidrométricas o pluviométricas, ubicadas dentro de la misma zona geográfica o de características climáticas y físicas similares. Esta ampliación es conveniente o necesaria porque entre más largo es un registro hidrológico, sus estimaciones estadísticas serán más confiables y exactas. Generalmente la cercanía geográfica no garantiza que los registros, amplio y corto, guarden una relación o correspondencia, sino el hecho de pertenecer a una región que se puede considerar homogénea en sus características de respuesta hidrológica.

Definida la región homogénea y seleccionados los registros amplios disponibles, la transferencia de información para ampliar una serie corta se puede realizar mediante la técnica estadística de regresión lineal múltiple. El hecho de que todos los registros por procesar sean parte de una región con respuesta hidrológica similar, implica que éstos guardarán cierta similitud, es decir, habrá alguna dependencia lineal entre las variables independientes o regresores, generando un problema de multicolinealidad en el análisis de regresión.

La regresión lineal múltiple de tipo Ridge o sesgada es una técnica estadística que permite evitar los problemas que genera la multicolinealidad, en relación con la inestabilidad de los coeficientes de la ecuación de regresión. Zhao *et al.* (1995) la emplearon para identificar hidrogramas unitarios bien formados. Yu y Liong (2007) la aplicaron para obtener pronósticos de series cronológicas hidrológicas y Weimin y Qian (2012) para obtener los parámetros de un modelo hidrológico en función de las características fisiográficas de las cuencas.

El objetivo de este estudio fue exponer como se diagnostica la multicolinealidad y como se desarrolla la regresión lineal múltiple de tipo Ridge para obtener estimaciones confiables de la variable dependiente. Se realiza una aplicación numérica, consistente en ampliar el registro corto de volúmenes escurridos anuales de la estación hidrométrica Platón Sánchez del río Tempoal, con base en las cuatro series amplias de las estaciones de aforos: Tempoal, Terrerillos, Los Hules y El Cardón; el sistema de río Tempoal está dentro de la Región Hidrológica No. 26 (Pánuco, México). Los resultados se contrastan con los obtenidos previamente mediante el método de selección de regresores.

floods, to amplify short series available in hydrometric or pluviometric stations located within the same geographic zone or of similar climatic and physical characteristics. This amplification is convenient or necessary because the longer a hydrological record is, the more exact and reliable its statistics will be. Generally the geographic proximity does not guarantee that the records, long and short, will have a relationship or correspondence, but rather the fact of pertaining to a region that can be considered homogeneous in its characteristics of hydrological response.

Once defined the homogeneous region and the available long records selected, the transference of information to amplify a short series can be made through the statistical technique of multiple linear regression. The fact that all of the records to process are part of a region with similar hydrological response implies that they will have a certain similarity, that is, there will be some linear dependence among the independent variables or regressors, generating a problem of multicollinearity in the analysis of regression.

The Ridge type multiple linear regression or biased regression is a statistical technique that makes it possible to avoid the problems generated by multicollinearity, with respect to the instability of the coefficients of the regression equation. Zhao *et al.* (1995) used it to identify well formed unit hydrographs. Yu and Liong (2007) applied it to obtain predictions of hydrological time series and Weimin and Qian (2012) to obtain the parameters of a hydrological model as a function of the physiographic characteristics of the watersheds.

The objective of the present study was to present how to diagnose multicollinearity and how to develop Ridge type multiple linear regression to obtain reliable estimations of the dependent variable. A numerical application is made, consisting of amplifying the short record of annual runoff volumes of the hydrometric station Platón Sánchez of the Tempoal River, based on the four long series of the measuring stations: Tempoal, Terrerillos, Los Hules and El Cardón; the Tempoal River system is within Hydrological Region No. 26 (Pánuco, México). The results are contrasted with those obtained previously by means of the regressors selection method.

Materiales y Métodos

La regresión lineal múltiple (RLM) y sus problemas

Frecuentemente se puede establecer una relación de tipo lineal entre la variable dependiente (Y) y varias (p) independientes $X1, X2, \ldots, Xp$, que es la generalización o extensión natural de la regresión lineal simple y su expresión es (Ryan, 1998):

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 \ldots + \beta_p Xp + \varepsilon \quad (1)$$

Por tanto, los principios que rigen a la regresión lineal simple y su expresión se aplican a la RLM. Por ejemplo, que tanto Y como las X estén normalmente distribuidas y que los errores ε sean independientes con distribución normal de media cero y misma varianza (σ^2) para cada X. La RLM implica una complejidad real en tres aspectos: 1) selección de cuántas y cuáles variables independientes utilizar; 2) interpretación de los resultados, especialmente de los coeficientes de la regresión (β_i); 3) determinación de cuándo usar un método de ajuste alternativo al de mínimos cuadrados de los residuos.

Para los dos primeros aspectos, un planteamiento correcto de tipo causa-efecto del problema por resolver con la RLM, ayudará a encontrar las mejores variables por utilizar y orientará sobre los valores por esperar en los coeficientes de la regresión. Para la transferencia regional de información hidrológica lo más probable es que se genere una situación de *multicolinealidad*, por la semejanza o correlación de los registros involucrados. Tal problema se debe diagnosticar y resolver, por ejemplo, a través de la regresión tipo Ridge.

Solución de mínimos cuadrados de los residuos

La solución matricial para la RLM, en el caso general de p variables independientes o *regresores* y n observaciones o datos de $Y, X1, X2, \ldots, Xp$, es la siguiente (Ryan, 1998):

$$\mathrm{Y} = \mathrm{X} \cdot \beta + \varepsilon \quad (2)$$

siendo:

Materials and Metods

Multiple linear regression (MLR) and its problems

Frequently a linear relationship can be established between the dependent variable (Y) and various (p) independent variables $X1, X2, \ldots, Xp$, which is the generalization or natural extension of the simple linear regression and its expression is (Ryan, 1998):

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 \ldots + \beta_p Xp + \varepsilon \quad (1)$$

Therefore, the principles that rule over simple linear regression are applied to MLR. For example, that both Y and the Xs are normally distributed and that the errors ε are independent with normal distribution of zero mean and same variance (σ^2) for each X. The MLR implies a real complexity in three aspects: 1) selection of how many and which independent variables to use; 2) interpretation of the results, especially of the coefficients of the regression (β_i); 3) determination of when an alternative fitting method should be used instead of least squares of the residuals.

For the first two aspects, a correct cause-effect type statement of the problem to resolve with MLR will help to find the best variables to use and will point to the values to be expected in the coefficients of the regression. For the regional transference of hydrological information, it is more likely that a situation of *multicollinearity* will be generated, due to the similarity or correlation of the records involved. This problem should be diagnosed and solved, for example, through Ridge type regression.

Solution of least squares of the residuals

The matrix solution for the MRL, in the general case of p independent variables or *regressors* and n observations or data of $Y, X1, X2, \ldots, Xp$, is as follows (Ryan, 1998):

$$\mathrm{Y} = \mathrm{X} \cdot \beta + \varepsilon \quad (2)$$

with:

$$\mathrm{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathrm{X} = \begin{bmatrix} 1 & X1_1 & X2_1 & \cdots & Xp_1 \\ 1 & X1_2 & X2_2 & \cdots & Xp_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X1_n & X2_n & \cdots & Xp_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

El planteamiento de esta solución implica que la sumatoria de uno a *n* de los residuos al cuadrado debe ser minimizada:

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X1_i-\beta_2X2_i-\cdots-\beta_pXp_i\right)^2=0 \quad (3)$$

Entonces, diferenciando el lado derecho de la ecuación anterior con respecto a $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, por separado, se originan las ecuaciones *normales* como función de los parámetros desconocidos. En notación matricial estas ecuaciones son:

$$(X^T\cdot X)\cdot\hat{\beta}=X^T\cdot Y \quad (4)$$

cuya solución es:

$$\hat{\beta}=(X^T\cdot X)^{-1}\cdot(X^T\cdot Y) \quad (5)$$

en la cual, X^T es la matriz transpuesta de X, y $(X^T\cdot X)^{-1}$ indica la matriz inversa de $X^T\cdot X$.

Escalamiento de longitud unitaria de los datos

Sustraer a cada variable independiente o regresor su media aritmética, se conoce como *centrado* de los datos y su ventaja fundamental es que las matrices X involucradas de *n* renglones ahora tienen *p* columnas, ya que la ecuación de RLM es:

$$Y-\overline{Y}=\beta_1\left(X1-\overline{X1}\right)+\beta_2\left(X2-\overline{X2}\right)+\cdots+\beta_p\left(Xp-\overline{Xp}\right) \quad (6)$$

cuyo reacomodo para obtener la ecuación 1 implica que:

$$\beta_0=\overline{Y}-\beta_1\overline{X1}-\beta_2\overline{X2}-\cdots-\beta_p\overline{Xp} \quad (7)$$

El escalamiento de longitud unitaria implica además del centrado, la división entre la raíz cuadrada de la varianza (Montgomery *et al.*, 2002), por lo cual:

$$Wj_i=\frac{Xj_i-\overline{Xj}}{S_j^{1/2}} \text{ con } i=1, 2, 3, \ldots, n,\ j=1, 2, 3, \ldots, p \quad (8)$$

$$Y_i=\frac{Y_i-\overline{Y}}{S_Y^{1/2}} \text{ con } i=1, 2, 3, \ldots, n \quad (9)$$

The approach of this solution implies that the sum of one to *n* of the residuals squared should be minimized:

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X1_i-\beta_2X2_i-\cdots-\beta_pXp_i\right)^2=0 \quad (3)$$

Therefore, differentiating the right side of the above equation with respect to $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, separately, the *normal* equations are formed as a function of the unknown parameters. In matrix notation these equations are:

$$(X^T\cdot X)\cdot\hat{\beta}=X^T\cdot Y \quad (4)$$

whose solution is:

$$\hat{\beta}=(X^T\cdot X)^{-1}\cdot(X^T\cdot Y) \quad (5)$$

in which, X^T is the transposed matrix of X, and $(X^T\cdot X)^{-1}$ indicates the inverse matrix of $X^T\cdot X$.

Scaling of unit length of the data

To subtract from each independent variable or regressor its arithmetic mean is known as *centered* of the data and its advantage is that the X matrices involved of *n* lines now have *p* columns, given that the equation of MLR is:

$$Y-\overline{Y}=\beta_1\left(X1-\overline{X1}\right)+\beta_2\left(X2-\overline{X2}\right)+\cdots+\beta_p\left(Xp-\overline{Xp}\right) \quad (6)$$

whose re-accommodation for obtaining equation 1 implies that:

$$\beta_0=\overline{Y}-\beta_1\overline{X1}-\beta_2\overline{X2}-\cdots-\beta_p\overline{Xp} \quad (7)$$

Scaling to unit length implies along with the centered, the division by the square root of the variance (Montgomery *et al.*, 2002), thus:

$$Wj_i=\frac{Xj_i-\overline{Xj}}{S_j^{1/2}} \text{ with } i=1, 2, 3, \ldots, n,\ j=1, 2, 3, \ldots, p \quad (8)$$

$$Y_i=\frac{Y_i-\overline{Y}}{S_Y^{1/2}} \text{ with } i=1, 2, 3, \ldots, n \quad (9)$$

donde

$$S_j = \sum_{i=1}^{n}\left(Xj_i - \overline{X}j\right)^2 \tag{10}$$

$$S_Y = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \tag{11}$$

El escalamiento de longitud unitaria produce, respecto a la ecuación 4, que la matriz $W^T \cdot W$ sea la matriz de correlación simple entre los regresores *Xj*; además la matriz $W^T \cdot Y$ es la matriz de correlación simple entre cada regresor *Xj* y la variable dependiente *Y*. Este escalamiento y el normal llevan a *coeficientes estandarizados de regresión*, cuya comparación entre ellos define la importancia de cada regresor. Otro escalamiento frecuentemente requerido está asociado con la estabilidad numérica de la matriz inversa de $W^T \cdot W$, pues es común obtenerla planteando esta igualdad $A \cdot A^{-1} = I$; al transformar la matriz A en la matriz identidad I y realizar las mismas operaciones en I, ésta se convierte en la matriz A^{-1} buscada. Cuando la matriz A tiene elementos muy grandes su inversa presentará elementos muy pequeños y entonces los errores por redondeo se vuelven importantes. En tales casos conviene dividir (*escalar*) todos los datos entre una cantidad fija o cociente reductor (COR), antes de aplicar la ecuación 5 y después los resultados de la ecuación 1 se multiplican por el COR.

Diagnóstico de multicolinealidad basada en $W^T \cdot W$ o $(W^T \cdot W)^{-1}$

La manera más simple de descubrir la multicolinealidad es a través de la inspección de la matriz $W^T \cdot W$, cuyos elementos fuera de la diagonal principal corresponden a los coeficientes de correlación simple entre pares de regresores; si hay valores absolutos mayores de 0.80 hay dependencia entre tal pareja. Este método sólo detecta multicolinealidad pero no la cuantifica; en cambio, cuando los factores de inflación de la varianza VIF (*variance inflation factor*) son mayores de 10 implican que los coeficientes de regresión obtenidos con la ecuación 5, no son confiables debido a la multicolinealidad. La expresión de los VIF es (Montgomery *et al.*, 1998; 2002):

$$VIF = \frac{1}{\left(1 - R_j^2\right)} \tag{12}$$

donde R_j^2 es el coeficiente de determinación que resulta de la RLM entre el regresor *Xj* como variable dependiente y el resto

where

$$S_j = \sum_{i=1}^{n}\left(Xj_i - \overline{X}j\right)^2 \tag{10}$$

$$S_Y = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \tag{11}$$

Scaling to unit length produces, with respect to equation 4, that the matrix $W^T \cdot W$ is the matrix of simple correlation between the regressors *Xj*; in addition the matrix $W^T \cdot Y$ is the matrix of simple correlation between each regressor *Xj* and the dependent variable *Y*. This scaling and the normal one lead to *standardized coefficients of regression*, whose comparison among them defines the importance of each regressor. Another scaling that is frequently required is associated with the numerical stability of the inverse matrix of $W^T \cdot W$, as it is commonly obtained proposing this equality $A \cdot A^{-1} = I$; by transforming matrix A in the identity I matrix and carrying out the same operations in I, it converts to the matrix A^{-1} sought. When the A matrix has very large elements its inverse will present very small elements and the errors by rounding become important. In such cases it is convenient to divide (*scale*) all of the data by a fixed amount or reduction quotient (COR), before applying equation 5 and afterwards, the results of equation 1 are multiplied by the COR.

Diagnostic of multicollinearity based on $W^T \cdot W$ or $(W^T \cdot W)^{-1}$

The simplest way to discover multicollinearity is through the inspection of the matrix $W^T \cdot W$, whose elements outside the principal diagonal correspond to the coefficients of simple correlation between pairs of regressors; then if there are absolute values greater than 0.80, there is dependence between this pair. This method only detects multicollinearity but it does not quantify it; however, *variance inflation factors* (VIF) higher than 10 imply that the regression coefficients obtained with equation 5 are not reliable due to the multicollinearity. The expression of the VIFs is as follows (Montgomery *et al.*, 1998; 2002):

$$VIF = \frac{1}{\left(1 - R_j^2\right)} \tag{12}$$

where R_j^2 is the coefficient of determination that results from the MRL between the regressor *Xj* as dependent variable and

$p-1$ como regresores. Los VIF corresponden a la diagonal de la matriz inversa de $W^T \cdot W$.

Diagnóstico de multicolinealidad con base en los eigenvalores de $W^T \cdot W$

Los eigenvalores de la matriz $W^T \cdot W$ se designan por $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p$, se conocen como valores propios, corresponden a las raíces de la ecuación característica $|A-\lambda \cdot I|=0$ de la matriz A, y se obtienen con procedimientos de métodos numéricos, por ejemplo el método de potencias (Carnahan *et al.*, 1969). Si existe una o más dependencias casi lineales en los datos, uno o más de los eigenvalores serán pequeños. El *número de condición* κ de la matriz $W^T \cdot W$ se define así (Montgomery *et al.*, 1998; 2002):

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{13}$$

y representa el espectro de variación de los eigenvalores de la matriz $W^T \cdot W$. En general, cuando κ es menor de 100 prácticamente no existen problemas de multicolinealidad, cuando varía de 100 a 1000 hay multicolinealidad moderada a fuerte y cuando excede a 1000 habrá graves problemas asociados a ésta. Los *índices de condición* κ_j de la matriz $W^T \cdot W$ son:

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j} \quad \text{con} \quad j = 1, 2, 3, \ldots, p \tag{14}$$

Los valores de κ_j definen el número y magnitud de las dependencias lineales que existen en los datos. Además, los eigenvectores asociados a cada eigenvalor permiten establecer numéricamente la dependencia lineal que existe entre los regresores. Esto último se mostrará en la aplicación numérica.

La regresión Ridge

En general, cuando el método de mínimos cuadrados de los residuos se aplica a datos que presentan multicolinealidad, la estimación de los coeficientes de regresión no es confiable, ya que su valor absoluto está exagerado y además es inestable. Las técnicas básicas para combatir la multicolinealidad son tres: 1) obtener más datos, lo cual puede no ser posible y además es probable que los datos nuevos reflejen el comportamiento de los anteriores; 2) reespecificar el modelo, redefiniendo los regresores, por ejemplo, si *X*1, *X*2 y *X*3 son linealmente dependientes, se puede adoptar una función de ellos del tipo $X = (X1 + X2)/X3$, o bien $X = X1*X2*X3$ que preserva el contenido de la información de los regresores originales, pero que reduzca el deterioramiento de los

the subtraction $p-1$ as regressors. The VIFs correspond to the diagonal of the inverse matrix of $W^T \cdot W$.

Diagnostic of multicollinearity based on the eigenvalues of $W^T \cdot W$

The eigenvalues of the matrix $W^T \cdot W$ are designated by $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p$, known as eigenvalues and correspond to the roots of the characteristic equation $|A-\lambda \cdot I|=0$ of the matrix A, are obtained with procedures of numerical methods, for example, the powers method (Carnahan *et al.*, 1969. If there are one or more nearly linear dependences in the data, one or more of the eigenvalues will be small. The condition number κ of the matrix $W^T \cdot W$ is defined as (Montgomery *et al.*, 1998; 2002):

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{13}$$

and represents the spectrum of variation of the eigenvalues of the matrix $W^T \cdot W$. In general, when κ is less than 100, there are practically no problems of multicollinearity, when it varies from 100 to 1000 there is moderate to strong multicollinearity and when it is more than 1000 there will be serious problems associated with it. The *condition indices* κ_j of the matrix $W^T \cdot W$ are:

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j} \quad \text{with} \quad j = 1, 2, 3, \ldots, p \tag{14}$$

The values of κ_j define the number and magnitude of the linear dependencies that exist in the data. Furthermore, the eigenvectors associated with each eigenvalue make it possible to numerically establish the linear dependence that exists between the regressors. The latter will be shown in the numerical application.

Ridge regression

In general, when the method of least squares of the residuals is applied to data that present multicollinearity, the estimation of the coefficients of regression is not reliable, given that its absolute value is exaggerated and is also unstable. The basic techniques to combat multicollinearity are three: 1) obtain more data, which may not be possible and also it is probable that the new data reflect the behavior of the previous data; 2) re-specify the model, redefining the regressors, for example, if *X*1, *X*2 and *X*3 are linearly dependent, a function of them can be developed of the type $X = (X1 + X2)/X3$, or $X = X1*X2*X3$ that preserves the

datos debido a la multicolinealidad; otro método de reespecificación efectivo consiste en la eliminación de una o más variables o regresores, que definitivamente reduce la multicolinealidad pero puede dañar notablemente la capacidad predictiva del modelo; 3) obtener estimaciones sesgadas, como la RLM de tipo Ridge.

El método de mínimos cuadrados de los residuos permite que la estimación $\hat{\beta}$ (ecuación 5) tenga varianza mínima pero la multicolinealidad genera varianza muy grande, por lo cual sus estimaciones son inestables. Suponiendo que se obtiene un estimador sesgado $\hat{\beta}^*$ con mucho menor varianza, se puede aceptar una cantidad pequeña de sesgo en $\hat{\beta}^*$, de manera que el error medio cuadrático de $\hat{\beta}^*$ sea menor que la varianza del estimador insesgado $\hat{\beta}$. La menor varianza del estimador sesgado implica que $\hat{\beta}^*$ es un estimador más estable de β que el insesgado $\hat{\beta}$.

Hay varios procedimientos para obtener estimadores sesgados de los coeficientes de regresión β. Uno de ellos es la regresión Ridge (Hoerl y Kennard, 1970) cuyo nombre se debe a la semejanza de sus operaciones matemáticas con el análisis Ridge usado para describir el comportamiento de superficies de respuesta de segundo orden. El estimador Ridge $\hat{\beta}_R$ se obtiene resolviendo una versión ligeramente modificada de las ecuaciones normales, expuestas como ecuaciones 4 y 5 (Montgomery *et al.*, 1998, 2002):

$$(W^T \cdot W + k \cdot I) \cdot \hat{\beta}_R = W^T \cdot Y \quad (15)$$

por lo cual: $$\hat{\beta}_R = (W^T \cdot W + k \cdot I)^{-1} \cdot (W^T \cdot Y) \quad (16)$$

En esas expresiones la constante $k \geq 0$, o *parámetro de sesgo*, se selecciona durante el proceso de aplicación de la regresión Ridge. El estimador Ridge es una transformación lineal del estimador de mínimos cuadrados de los residuos cuyo sesgo crece al aumentar k, pero al mismo tiempo disminuye su varianza. Con la regresión Ridge se obtiene una estimación estable de sus coeficientes, a cambio de no ser el mejor ajuste a los datos. Por tanto, aunque no hay demostración matemática concluyente, se considera que conduce a ecuaciones de regresión que funcionan mejor para predecir observaciones futuras, comparada con la de mínimos cuadrados de los residuos.

Hoerl y Kennard (1970) sugieren que un valor adecuado de k puede estimarse por inspección de la *traza Ridge*, que es una gráfica de las magnitudes de $\hat{\beta}_R$ dibujados en las ordenadas, contra sus respectivos valores de k en las abscisas. Los valores de k suelen estar en el intervalo de 0 a 1. Si la multicolinealidad es grave, los coeficientes $\hat{\beta}_R$ variarán mucho, pero en un cierto valor de k se estabilizan. Lo fundamental es seleccionar el valor de k más pequeño, donde los $\hat{\beta}_R$ sean estables. Con ello es posible obtener una ecuación de regresión con menor error medio cuadrático que el correspondiente a mínimos cuadrados.

content of the information of the original regressors, but which reduces the deterioration of the data due to the multicollinearity; another effective re-specification method consists of the elimination of one or more variables or regressors, definitely which reduces multicollinearity but it can notably damage the predictive capacity of the model; 3) obtain biased estimations, such as the Ridge type MLR.

The method of least squares of the residuals allows that the estimation $\hat{\beta}$ (equation 5) has minimum variance but the multicollinearity generates very large variance, thus its estimations are unstable. Assuming that a biased estimator $\hat{\beta}$ is obtained with much lower variance, a small amount of bias in $\hat{\beta}$ can be accepted, thus the mean quadratic error of $\hat{\beta}$ is lower than the variance of the unbiased estimator $\hat{\beta}$. The lower variance of the biased estimator implies that $\hat{\beta}$ is a more stable estimator of β than the unbiased $\hat{\beta}$.

There are various procedures to obtain biased estimations of the regression coefficients β. One of them is the Ridge regression (Hoerl and Kennard, 1970), whose name is due to the similarity of its mathematical operations with the Ridge analysis used to describe the behavior of second order response surfaces. The Ridge estimator $\hat{\beta}_R$ is obtained resolving a slightly modified version of the normal equations, presented as equations 4 and 5 (Montgomery *et al.*, 1998; 2002):

$$(W^T \cdot W + k \cdot I) \cdot \hat{\beta}_R = W^T \cdot Y \quad (15)$$

Thus: $$\hat{\beta}_R = (W^T \cdot W + k \cdot I)^{-1} \cdot (W^T \cdot Y) \quad (16)$$

In the above expressions the constant $k \geq 0$, denominated bias parameter, is selected during the process of application of the Ridge regression. The Ridge estimator is a linear transformation of the estimator of least squares of the residuals whose bias increases as k increases, but at the same time its variance decreases. With the Ridge regression a stable estimation of its coefficients is obtained, in exchange for not being the best fit for the data. Therefore, although there is no conclusive mathematical demonstration, it is believed that leads to regression equations that function better for predicting future observations, compared with that of least squares of the residuals.

Hoerl and Kennard (1970) suggest that an adequate value of k can be estimated by inspection of the *Ridge trace*, which is a graph of the magnitudes of $\hat{\beta}_R$ drawn on the ordinates, against their respective values of k in the abscissas. The values of k tend to be in the interval of 0 to 1. If the multicollinearity is serious, the coefficients $\hat{\beta}_R$ will vary greatly, but in a certain value of k they stabilize. The fundamental is to select the smallest value of k, where the $\hat{\beta}_R$ are stable. With this it is possible to obtain a regression equation with a smaller mean square error than the one corresponding to least squares.

Aplicación numérica

El río Tempoal es uno de los afluentes importantes del río Moctezuma, que junto con el río Tampaón forman el río Pánuco, en la Región Hidrológica No. 26 de México. El río Tempoal tiene cinco estaciones hidrométricas: El Cardón (*X*4), Los Hules (*X*3), Terrerillos (*X*2), Tempoal (*X*1) y Platón Sánchez (*Y*). En la Figura 1 se muestra la ubicación y morfología del sistema del río Tempoal. Campos (2011) mostró las características generales de ellas y sus registros disponibles de volumen escurrido anual (en Mm^3), en el periodo de 1961 a 2002, con seis años faltantes en el lapso de 1979 a 2002 (Cuadro 1). También verificó las características estadísticas de los datos y realizó la estimación del registro faltante en Platón Sánchez con base en los otros cuatro, siguiendo el *método de selección de variables*. Ahora se aplicará la regresión Ridge con el mismo propósito de ampliar el registro corto de esta estación de aforos y comparar resultados.

Verificación de la homogeneidad regional

El uso estadístico de la información del Cuadro 1 establece que se requieren dos verificaciones de la homogeneidad regional, la primera para el periodo común de los datos (1979-2002) pues éstos definirán los coeficientes de la ecuación de regresión, y la segunda para el lapso total de datos porque el periodo 1961-1978 permite ampliar del registro corto. La segunda verificación es más difícil de cumplir debido a la diferencia de amplitudes de registros y por ello, sólo de esta comprobación se presentan sus resultados. Antes de usar una prueba de homogeneidad, en este caso la versión corregida del *Test de Langbein* (Fill y Stedinger, 1995; Campos, 2012), se debe verificar la calidad estadística de la información, por ejemplo a través del *Test de Discordancias* aplicado en el periodo común, el cual permite detectar valores anómalos, tendencias determinísticas o cambios en la media, ya que mide lo discordante que es cada registro en sus cocientes *L* con respecto al promedio del grupo considerado como un todo (Hosking y Wallis, 1997; Campos, 2010). En el Cuadro 2 se exponen los resultados del Test de Discordancias, cuyo valor crítico de Discordancia es D_c=1.333, ya que sólo hay cinco estaciones en el grupo. Se observa que ningún registro es discordante con el grupo.

En el Cuadro 3 se muestran los resultados de la aplicación del nuevo Test de Langbein al sistema del río Tempoal, los cuales indican que sus cinco estaciones hidrométricas forman una región homogénea, pues ninguna de ellas queda fuera de sus curvas de control definidas por los periodos de retorno inferior (Tr_{inf}) y superior (Tr_{sup}), en años.

Numerical application

The Tempoal River is one of the most important tributaries of the Moctezuma River, which along with the Tampaón River form the Pánuco River, in Hydrological Region No. 26 of México. The Tempoal River has five hydrometric stations: El Cardón (X4), Los Hules (X3), Terrerillos (X2), Tempoal (X1) and Platón Sánchez (Y). Figure 1 shows the location and morphology of the Tempoal River system. Campos (2011) showed their general characteristics and their available records of annual runoff volume (in Mm^3), in the period of 1961 to 2002, with six years missing in the lapse of 1979 to 2002 (Table 1). He also verified the statistical characteristics of the data and carried out the estimation of the missing record in Platón Sánchez based on the other four, following the *variables selection method*. Now the Ridge regression will be applied with the same purpose of amplifying the short record of this gauging station and comparing results.

Verification of regional homogeneity

The statistical use which for the information of Table 1 establishes that two verifications are required of regional homogeneity, the first for the common period of the data (1979-2002), as these data will define the coefficients of the regression equation, and the second for the total lapse of data because the period 1961-1978 allows the expansion of the short record. The second verification is more difficult to complete due to the difference of amplitudes of records and for this reason, the results are shown only from this verification. Before using a homogeneity test, in this case the corrected version of the *Langbein Test* (Fill and Stedinger, 1995; Campos, 2012), a verification should be made of the statistical quality of the information. For example, using the *Discordances Test* applied in the common period, it is possible to detect anomalous values, deterministic tendencies or changes in the mean, given that it measures the discordance of each record in its *L* quotients with respect to the average of the group considered as a whole (Hosking and Wallis, 1997; Campos, 2010). Table 2 shows the results of the Discordances Test, whose critical value of Discordance is D_c=1.333, given that there are only five stations in the group. It is observed that no record is discordant with the group.

Table 3 shows the results of the application of the new Langbein Test to the Tempoal river system, which indicate that their five hydrometric stations form a homogeneous region, as none of them is outside of the control curves defined by the lower and (Tr_{inf}) higher (Tr_{sup}) return periods, in years.

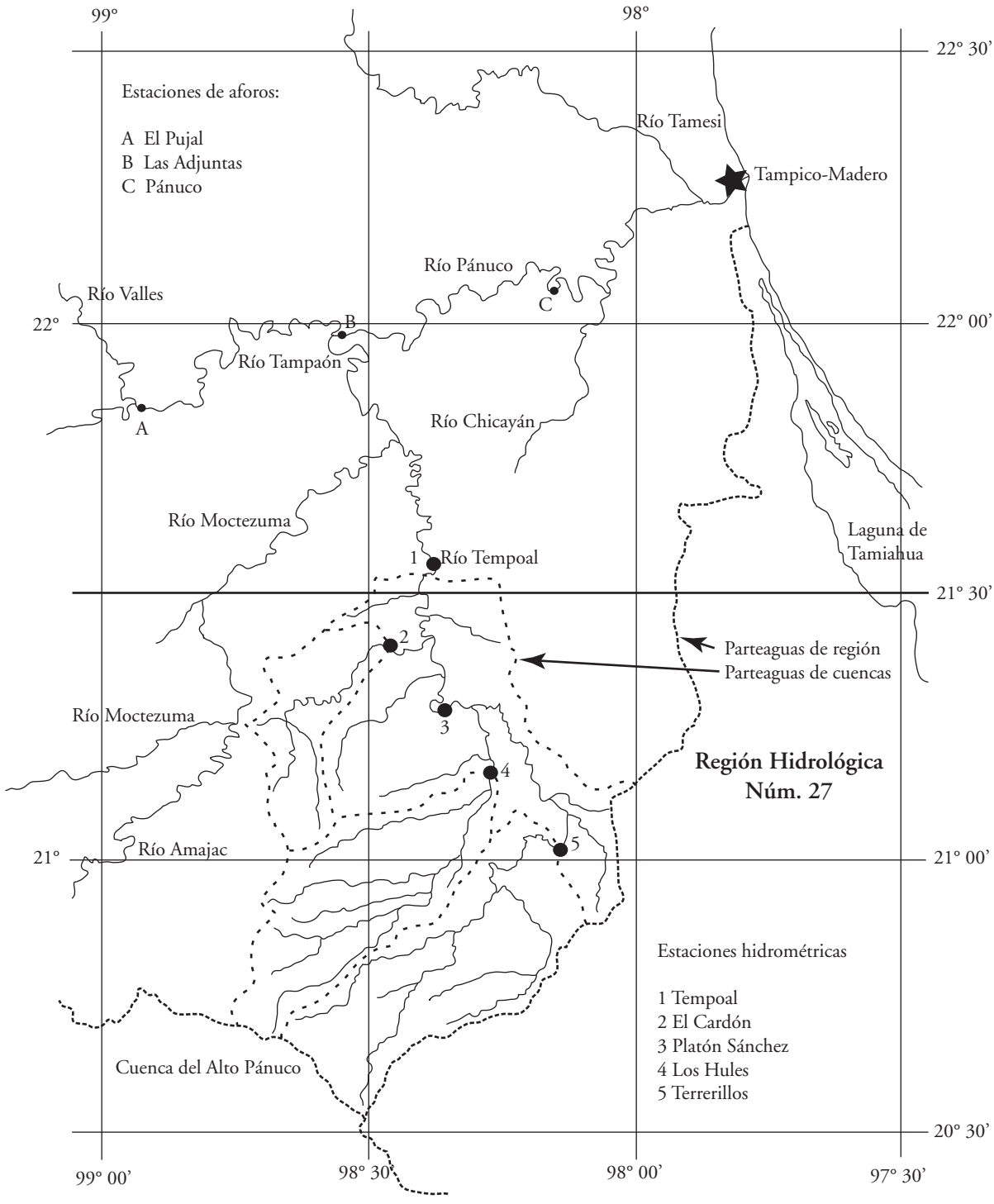

Figura 1. Localización y morfología del sistema del río Tempoal.
Figure 1. Location and morphology of the Tempoal river system.

Cuadro 1. Volúmenes escurridos anuales (Mm³) en las estaciones hidrométricas del sistema del Río Tempoal.
Table 1. Annual runoff volumes (Mm³) in the hydrometric stations of the Tempoal river.

Año	P. Sánchez (*Y*)	Tempoal (*X*1)	Terrerillos (*X*2)	Los Hules (*X*3)	El Cardón (*X*4)
1961	–	3150.302	1211.240	1021.195	386.787
1962	–	1796.844	628.203	677.758	272.019
1963	–	1655.044	668.833	661.475	197.102
1964	–	1076.755	324.150	378.162	145.934
1965	–	2293.958	865.133	749.130	244.780
1966	–	2786.573	1020.039	1011.194	235.217
1967	–	3263.920	1067.163	1080.269	548.409
1968	–	2837.862	985.655	945.499	511.579
1969	–	3323.340	1336.178	1127.562	488.246
1970	–	2863.385	944.250	941.203	412.411
1971	–	2441.337	1072.324	808.878	336.091
1972	–	2566.835	915.385	950.725	373.829
1973	–	3599.619	1243.497	1160.392	522.902
1974	–	4296.827	1428.909	1312.057	642.511
1975	–	4298.112	1446.878	1656.828	570.883
1976	–	4241.779	1868.405	1564.284	673.933
1977	–	1332.365	521.217	466.286	134.540
1978	–	3688.256	1406.349	1376.984	547.106
1979	2995.751	2103.745	829.710	796.465	284.234
1980	1325.674	1586.278	702.483	586.212	227.079
1982	776.575	880.923	394.620	359.815	148.206
1983	2116.887	2187.518	1190.780	967.822	271.316
1984	4065.671	5057.565	2444.332	1832.083	636.325
1985	2257.756	2607.572	1390.327	936.464	361.991
1986	1654.262	1807.878	891.569	688.127	264.761
1987	2018.218	2213.954	1418.647	815.745	322.006
1988	1955.431	2325.627	1312.224	729.049	274.661
1989	1457.954	1749.932	958.891	1032.017	288.773
1992	3144.544	4134.539	1929.711	1545.657	607.888
1993	4291.278	5629.170	2370.400	2230.109	749.586
1994	1228.833	1634.689	703.788	809.173	305.695
1995	1542.547	1861.508	843.071	1081.871	343.729
1996	1034.440	1250.085	594.657	778.378	185.781
1997	1046.949	1180.939	520.502	651.170	152.708
2001	1281.262	1929.465	805.264	535.401	277.771
2002	1058.208	1411.568	683.526	644.391	172.370
$\overline{X}$	1958.458	2585.169	1081.620	969.718	364.421
S	1037.043	1175.534	498.134	410.447	171.522
Cv	0.530	0.455	0.461	0.423	0.471
Cs	1.173	0.795	1.066	1.137	0.610
Ck	3.931	3.219	4.364	4.668	2.489

Resultados y Discusión

Diagnóstico de multicolinealidad

Las matrices $W^T \cdot W$, $W^T \cdot Y$ y $(W^T \cdot W)^{-1}$ obtenidas para los datos de Cuadro 1, procesados lógicamente

Results and Discussion

Diagnostic of multicollinearity

The matrices $W^T \cdot W$, $W^T \cdot Y$ and $(W^T \cdot W)^{-1}$ obtained for the data of Table 1, processed logically

Cuadro 2. Aplicación del Test de Discordancias en el sistema del río Tempoal.
Table 2. Application of the Discordancies Test in the Tempoal river system.

Estación hidrométrica:	Cocientes de momentos *L* muestrales			Discordancia calculada (D_i)
	t_2	t_3	t_4	
Platón Sánchez	0.29031	0.31360	0.13440	1.00
Tempoal	0.28790	0.39201	0.30166	1.30
Terrerillos	0.29684	0.31259	0.15733	0.77
Los Hules	0.25948	0.36578	0.30049	0.96
El Cardón	0.26783	0.34226	0.28231	0.97

Cuadro 3. Resultados de la aplicación de la versión corregida del Test de Langbein en el sistema del río Tempoal.
Table 3. Results of the application of the corrected version of the Langbein Test in the Tempoal river system.

C_v^R=0.4609 Sesgo (C_v^R) = −0.0085

Estación hidrométrica:	Número de datos	Valor medio ($\overline{X}$) en Mm³	Coeficiente de Variación (*Cv*)	Tr_{inf} (años)	Tr_{sup} (años)	$Tr(X)_{10}$ (años)	*k*
Platón Sánchez	18	1958.458	0.5295	6.0	20.2	7.9	–
Tempoal	36	2585.169	0.4547	6.9	15.4	9.9	–
Terrerillos	36	1081.620	0.4605	6.9	15.4	9.7	–
Los Hules	36	969.718	0.4233	6.8	15.5	11.2	–
El Cardón	36	364.421	0.4707	6.9	15.4	9.4	–

con escalamiento unitario son:

with unitary scaling are:

$$\mathrm{W}^{\mathrm{T}}\cdot\mathrm{W}=\begin{array}{c} \\ X1\\ X2\\ X3\\ X4\end{array}\begin{array}{c}\begin{array}{cccc} X1 & X2 & X3 & X4\end{array}\\ \begin{bmatrix} 1.000000 & 0.970261 & 0.942155 & 0.976884\\ 0.970261 & 1.000000 & 0.899202 & 0.936976\\ 0.942155 & 0.899202 & 1.000000 & 0.952889\\ 0.976884 & 0.936976 & 0.952889 & 1.000000\end{bmatrix}\end{array}\qquad \mathrm{W}^{\mathrm{T}}\cdot\mathrm{Y}=\begin{array}{c}\begin{bmatrix} 0.947179\\ 0.919927\\ 0.881427\\ 0.908789\end{bmatrix}\\ \begin{array}{c} X1\\ X2\\ X3\\ X4\end{array}\end{array}$$

$$\left(\mathrm{W}^{\mathrm{T}}\cdot\mathrm{W}\right)^{-1}=\begin{bmatrix} 49.519800 & -22.050200 & -4.551862 & -23.377190\\ -22.050200 & 18.039850 & 1.457870 & 3.248384\\ -4.551862 & 1.457870 & 11.327140 & -7.712855\\ -23.377190 & 3.248384 & -7.712855 & 28.142640\end{bmatrix}$$

La inspección de la matriz $\mathrm{W}^{\mathrm{T}}\cdot\mathrm{W}$ muestra que los cuatro registros están altamente correlacionados, mostrando la mayor relación Tempoal con El Cardón (r_{xy}=0.977) y la menor Terrerillos con Los Hules (r_{xy}=0.899). Lo anterior establece que existe un problema de multicolinealidad con tales datos. En la matriz $\mathrm{W}^{\mathrm{T}}\cdot\mathrm{Y}$ se observa que la mayor correlación entre

The inspection of the matrix $\mathrm{W}^{\mathrm{T}}\cdot\mathrm{W}$ shows that the four records are highly correlated, showing the highest Tempoal relationship with El Cardón (r_{xy}=0.977) and the lowest Terrerillos with Los Hules (r_{xy}=0.899). The above establishes that there is a problem of multicollinearity with this data. In the matrix $\mathrm{W}^{\mathrm{T}}\cdot\mathrm{Y}$ it is observed that the highest

los regresores y el registro de Platón Sánchez es con Tempoal y la menor con Los Hules, esto se reflejará en las variaciones de los coeficientes Ridge ($\hat{\beta}_R$).

El primer renglón de resultados del Cuadro 4 procede de los elementos de la diagonal de la matriz inversa de $W^T \cdot W$, mostrando que existe multicolinealidad pues todos los VIF_j exceden de 10. Sin embargo no exceden de 100, de manera que tales problemas son moderados o aceptables. En el segundo renglón de resultados del Cuadro 4 se exponen los eigenvalores y en el tercero los índices de condición κ_j de la matriz $W^T \cdot W$, los cuales ratifican los resultados anteriores en relación con la multicolinealidad, ya que sólo uno excede en valor absoluto a 100.

Con base en los elementos del cuarto eigenvector se establece la siguiente ecuación relativa a la multicolinealidad presente:

$$0.80988 \cdot X1 - 0.36334 \cdot X2 - 0.01671 \cdot X3 - 0.46021 \cdot X4 = 0 \quad (17)$$

considerando que el coeficiente de $X3$ es cercano a cero se obtiene:

$$0.80988 \cdot X1 = 0.36334 \cdot X2 + 0.46021 \cdot X4 \quad (18)$$

$$X1 = 0.4486 \cdot X2 + 0.5682 \cdot X4 \quad (19)$$

la ecuación anterior establece la relación entre $X1$ y aproximadamente las mitades de $X2$ y $X4$.

correlation between the regressors and the record of Platón Sánchez is with Tempoal and the lowest with Los Hules, which will be reflected in the variations of the Ridge coefficients ($\hat{\beta}_R$).

The first line of results of Table 4 proceeds from the elements of the diagonal of the inverse of $W^T \cdot W$, showing that there is multicollinearity, as all of the VIFs exceed 10. However, they do not exceed 100, thus such problems are moderate or acceptable. In the second line of results of Table 4 the eigenvalues are shown and in the third the indices of condition κ_j of the matrix $W^T \cdot W$, which ratify the above results in relation to the multicollinearity, given that only one exceeds 100 in absolute value.

Based on the elements of the fourth eigenvector, the following equation is established relative to the multicollinearity present:

$$0.80988 \cdot X1 - 0.36334 \cdot X2 - 0.01671 \cdot X3 - 0.46021 \cdot X4 = 0 \quad (17)$$

considering that the coefficient of $X3$ is close to zero the following is obtained:

$$0.80988 \cdot X1 = 0.36334 \cdot X2 + 0.46021 \cdot X4 \quad (18)$$

$$X1 = 0.4486 \cdot X2 + 0.5682 \cdot X4 \quad (19)$$

The above equation establishes the relationship between $X1$ and approximately the halves of $X2$ and X4.

Cuadro 4. Resultados del diagnóstico de multicolinealidad en el sistema del río Tempoal.
Table 4. Results of the diagnostic of multicollinearity in the Tempoal river system.

Indicadores	Regresores			
	$X1$	$X2$	$X3$	$X4$
VIF_j	49.51980	18.03985	11.32714	28.71286
λ_j	3.83960	0.10665	0.03992	−0.01382
$\kappa_j = \frac{\lambda_{max}}{\lambda_j}$	1.000	36.002	96.182	−277.829
Regresores		eigenvectores		
$X1$	0.50655	0.19203	−0.20621	0.80988
$X2$	0.49570	0.66806	0.42378	−0.36334
$X3$	0.49404	−0.69993	0.51566	−0.01671
$X4$	0.50361	−0.16409	−0.71553	−0.46021

Cálculo y análisis de la traza Ridge

Aplicando la ecuación 16 con los valores de *k* indicados en el Cuadro 5 se obtuvieron los coeficientes de regresión tipo Ridge ahí mostrados, cuyos coeficientes de determinación (R^2) respectivos, también se citan en este cuadro. El cálculo de R^2 se realizó haciendo el centrado de los datos y utilizando un COR=1000. Con base en los resultados del Cuadro 5 se ha construido la traza Ridge, mostrada en la Figura 2.

La inspección de la traza Ridge muestra que el coeficiente β_1 es el de mayor variación o cambio con

Calculation and analysis of the Ridge trace

Applying equation 16 with the values of *k* indicated in Table 5, the Ridge regression coefficients that are shown were obtained, whose respective coefficients of determination (R^2) are also shown in this table. The calculation of R^2 was made centered the data and using a COR = 1000. Based on the results of Table 5 the Ridge trace has been constructed, shown in Figure 2.

The inspection of the Ridge trace shows that the coefficient β_1 has the highest variation or change with

Cuadro 5. Coeficientes de regresión tipo Ridge ($\hat{\beta}_R$) obtenidos para los valores del parámetro de sesgo indicado.
Table 5. Coefficients of Ridge regression ($\hat{\beta}_R$) obtained for the indicated bias values.

$\hat{\beta}_R$	Valores del parámetro de sesgo (*k*)											
	0.0000	0.0025	0.0050	0.0075	0.0100	0.0125	0.0150	0.0175	0.020	0.025	0.030	0.035
β_1	1.3625	1.1983	1.0774	0.9845	0.9109	0.8510	0.8014	0.7595	0.7236	0.6654	0.6202	0.5839
β_2	−0.0530	0.0146	0.0635	0.1001	0.1284	0.1508	0.1688	0.1836	0.1958	0.2146	0.2282	0.2382
β_3	0.0044	0.0121	0.0189	0.0250	0.0306	0.0357	0.0405	0.0449	0.0491	0.0568	0.0638	0.0701
β_4	− 0.3767	−0.2864	−0.2201	−0.1693	−0.1292	−0.0967	−0.0698	−0.0472	−0.0278	0.0034	0.0275	0.0468
R^2	0.9033	0.9032	0.9030	0.9028	0.9025	0.9023	0.9021	0.9019	0.9017	0.9013	0.9010	0.9008
$\hat{\beta}_R$	Valores del parámetro de sesgo (*k*)											
	0.040	0.050	0.060	0.070	0.080	0.090	0.100	0.110	0.120	0.130	0.140	0.160
β_1	0.5540	0.5079	0.4736	0.4471	0.4259	0.4085	0.3940	0.3816	0.3708	0.3614	0.3531	0.3391
β_2	0.2458	0.2558	0.2617	0.2651	0.2670	0.2678	0.2680	0.2677	0.2671	0.2662	0.2653	0.2631
β_3	0.0759	0.0860	0.0947	0.1023	0.1089	0.1148	0.1199	0.1246	0.1287	0.1325	0.1359	0.1417
β_4	0.0625	0.0866	0.1043	0.1178	0.1285	0.1371	0.1441	0.1500	0.1550	0.1593	0.1630	0.1690
R^2	0.9005	0.9002	0.8999	0.8996	0.8994	0.8992	0.8990	0.8989	0.8987	0.8986	0.8985	0.8982

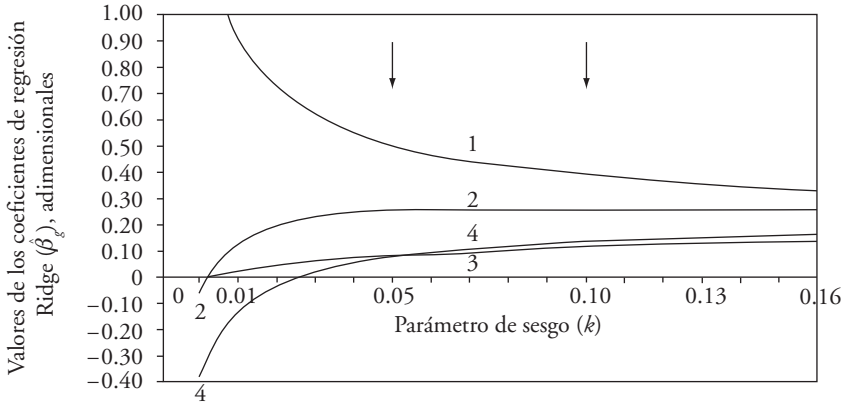

Figura 2. Traza Ridge para los datos del sistema del río Tempoal.
Figure 2. Ridge trace for the data of the Tempoal river system.

k, siguiéndolo el β_4 y el β_2 que incluso cambian de signo; por el contrario, el coeficiente β_3 fue el más estable. Entonces, durante la regresión Ridge el coeficiente que más cambia es el correspondiente al registro de Tempoal y que menos lo hace el del registro de Los Hules. Teniendo en cuenta que el parámetro de sesgo (*k*) debe tener el menor valor posible, cuando ya los coeficientes de regresión Ridge se pueden considerar estabilizados, se seleccionaron dos valores para *k*: 0.050 y 0.100 (Figura 2).

Estimaciones Ridge y su contraste

En el Cuadro 6 se muestran las 18 estimaciones de la variable dependiente $\left(\hat{Y}_i\right)$, es decir del registro histórico en Platón Sánchez Mm^3 para el periodo 1979-2002, así como sus residuos respectivos, realizadas con las regresiones Ridge que emplean *k*=0.050 y 0.100. Los coeficientes de regresión respectivos se muestran en el Cuadro 7 y fueron obtenidos con datos centrados y usando un COR de mil.

En el Cuadro 7 están los resultados del contraste entre los residuos de los mejores modelos de regresión obtenidos a través de selección óptima de regresores (Campos, 2011) y las regresiones Ridge adoptadas. La regresión Ridge origina valores escasamente mayores de los residuos y de la suma de residuos al cuadrado, pero la suma algebraica de sus errores es menor.

Estimaciones Ridge adoptadas

En el Cuadro 8 se muestran los 18 volúmenes escurridos anuales estimados en la estación Platón Sánchez en el periodo de 1968 a 1978, mediante las regresiones Ridge adoptadas, así como sus respectivos parámetros estadísticos.

En la Figura 3 se muestra la comparación entre la segunda serie de volúmenes escurridos anuales estimados con regresión Ridge y los valores adoptados bajo el planteamiento de selección de regresores (Campos, 2011). Ambas series estimadas de volúmenes escurridos anuales presentan el mismo comportamiento, pero la procedente de la regresión Ridge es menor y con valores mínimos más acusados, lo cual origina una media y un coeficiente de variación más parecidos a los datos históricos de Platón Sánchez (Cuadro 1).

k, followed by β_4 and β_2 which even change signs; in contrast, the coefficient β_3 was the most stable. Then, during the Ridge regression the coefficient that changes the most is the one corresponding to the register of Tempoal and the one that changes the least is that of Los Hules. Considering that the parameter of bias (*k*) should have the lowest possible value, when the coefficients of Ridge regression can be considered stabilized, two values for *k* were selected, 0.050 and 0.100 (Figure 2).

Ridge estimations and their contrast

Table 6 shows the 18 estimations of the dependent variable ($\left(\hat{Y}_i\right)$), that is of the historic register of Platón Sánchez Mm^3 for the period 1979-2002, as well as their respective residuals, made with the Ridge regressions that employ *k* = 0.050 and 0.100. The respective coefficients of regression are shown in Table 7 and were obtained with centered data and using a COR of one thousand.

Cuadro 6. Estimaciones de la variable dependiente ($\hat{Y}_i$) en Mm^3 obtenidas con las regresiones Ridge y sus residuos respectivos.

Table 6. Estimations of the dependent variable ($\hat{Y}_i$) in Mm^3 obtained with the Ridge regressions and their respective residuals.

Año	*k* = 0.050	Residuo	*k* = 0.100	Residuo
1979	1817.664	1178.087	1803.750	1192.001
1980	1432.449	−106.775	1425.375	−99.701
1982	897.946	−121.371	894.951	−118.376
1983	1885.840	231.047	1883.035	233.852
1984	4043.281	22.390	4034.190	31.481
1985	2199.088	58.668	2202.787	54.970
1986	1592.529	61.733	1592.221	62.041
1987	1903.194	115.025	1919.363	98.855
1988	2029.024	−73.592	2019.968	−64.537
1989	1495.657	−37.703	1514.250	− 56.296
1992	3299.914	−155.370	3310.448	−165.904
1993	4418.631	−127.353	4408.121	−116.843
1994	1407.141	−178.308	1421.407	−192.574
1995	1551.752	−9.205	1573.395	−30.848
1996	1148.355	−113.915	1149.466	−115.026
1997	1119.689	−72.740	1110.812	−63.863
2001	1702.048	−420.786	1691.108	−409.846
2002	1308.034	−249.826	1297.616	−239.408
Máximo	4418.631	1178.087	4408.121	1192.001
mínimo	897.946	−420.786	894.951	−409.846

Cuadro 7. Indicadores de los residuos obtenidos con los modelos de mínimos cuadrados y con la regresión Ridge.
Table 7. Indicators of the residuals obtained with the models of least squares and with Ridge regression.

Modelo analizado:	Coeficientes de regresión				Valores de los residuos				
	β_0	β_1	β_2	β_3	β_4	Mínimo	Máximo	$\sum_{i=1}^{18}\varepsilon_i$	$\sum_{i=1}^{18}\varepsilon_i^2$
$Y=f(X1)$	227.9714	0.7496	–	–	–	−393.036	1190.812	0.659	$1880.4\cdot10^3$
$Y=f(X1, X4)$	305.7837	1.0287	–	–	−2.2124	−394.822	1154.684	0.193	$1771.5\cdot10^3$
$Y=f(X1, X2, X4)$	315.5988	1.0797	−0.0920	–	−2.2905	−407.261	1136.110	1.723	$1768.6\cdot10^3$
Ridge con k=0.050	0.265019	0.840930	0.025685	−0.107688	−0.534767	−420.786	1178.087	0.006	$1825.2\cdot10^3$
Ridge con k 0.100	0.252525	0.780914	0.070965	−0.089600	−0.278420	−409.846	1192.001	−0.023	$1846.3\cdot10^3$

Comentarios en torno a la hidrología estadística

La regresión lineal múltiple de tipo Ridge permite realizar la ampliación de un registro hidrológico corto, empleando varios cercanos y aledaños que por lógica serán multicolineales. Esta técnica estadística y otras, basadas en el análisis multivariado de series cronológicas, permiten un mejor uso de la información hidrológica disponible. Sin embargo, ningún procedimiento de la Hidrología estadística genera nueva información o suple su escasez.

Por lo anterior, se debe requerir a la Comisión Nacional del Agua (CONAGUA) que restablezca la red de mediciones climatológicas y hidrométricas que había en los años ochenta. Lafragua *et al.* (2006)

Cuadro 8. Volúmenes escurridos anuales (Mm^3) estimados en la estación Platón Sánchez, con base en la regresión Ridge.
Table 8. Annual runoff volumes (Mm^3) estimated in the Platón Sánchez station, based on Ridge regression.

Año	k = 0.050	k = 0.100	Año	k = 0.050	k = 0.100
1961	2363.483	2346.882	1973	2654.376	2649.680
1962	1308.702	1311.298	1974	3165.146	3160.407
1963	1232.319	1225.765	1975	3167.864	3151.737
1964	795.037	789.342	1976	3086.177	3117.260
1965	1739.706	1717.504	1977	1011.652	998.213
1966	2134.833	2092.369	1978	2696.828	2704.310
1967	2362.535	2375.092	MAX	3167.864	3160.407
1968	2036.365	2058.923	mín	795.037	789.342
1969	2446.493	2453.098	$\bar{X}$	2111.401	2108.471
1970	2110.259	2103.911	*Cv*	0.339	0.341
1971	1813.699	1816.522	*Cs*	–0.216	–0.211
1972	1879.747	1880.171	*Ck*	2.764	2.755

Table 7 shows the results of the contrast between the residuals of the best regression models obtained through the optimum selection of regressors (Campos, 2011) and the adopted Ridge regressions. The Ridge regression originates values that are scarcely larger than the residuals and of the sum of residuals squared, but that the algebraic sum of their errors is lower.

Adopted Ridge estimations

Table 8 shows the 18 annual runoff volumes estimated in the station Platón Sánchez in the period of 1968 to 1978, through the adopted Ridge regressions, along with their respective statistical parameters.

Figure 3 shows the comparison between the second series of annual runoff volumes estimated with Ridge regression and the values adopted under the concept of selection of regressors (Campos, 2011). Both estimated series of annual runoff volumes present the same behavior, but the one proceeding from the Ridge regression is lower and with more accused minimum values, which originates a mean and a coefficient of variation that are more similar to the historic data of Platón Sánchez (Table 1).

Comments regarding statistical hydrology

Ridge type multiple linear regression makes it possible to carry out the amplification of a short hydrological record, employing various nearby records which will logically be multicollinear. This statistical technique along with other techniques, based on the multivariate analysis of time series,

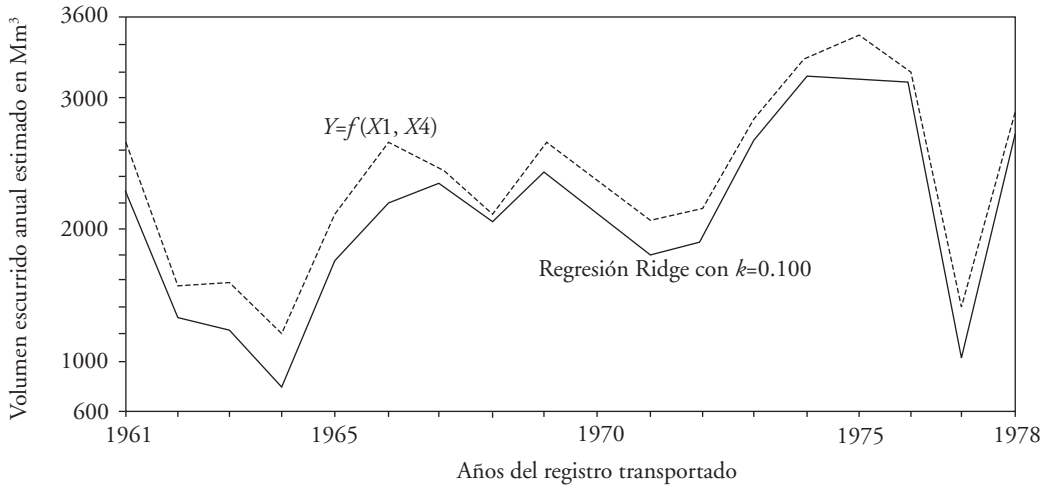

Figura 3. Contraste de estimaciones en la estación hidrométrica Platón Sánchez.
Figure 3. Contrast of estimations in the Platón Sánchez hydrometric station

señalan que la situación respecto a la información hidrológica es verdaderamente crítica, a pesar de ser la base para la estimación de la disponibilidad superficial en las cuencas y en el país. Por ejemplo, en el 2004 había 50% de las estaciones climatológicas existentes en 1980 y las hidrométricas habían disminuido de casi 1,600 en 1980 a sólo 430 en el 2002.

Conclusiones

Es mejor usar parte de la información estadística de todos los regresores, como lo hace la regresión Ridge, que emplear toda la información de algunos regresores y nada de otros, como actúa el método de selección de variables. Además, la regresión Ridge es un procedimiento directo, de fácil implementación dentro de la solución de mínimos cuadrados de los residuos y la interpretación de la traza Ridge no presenta dificultades.

Respecto a la aplicación numérica descrita, problema previamente abordado con eliminación de variables, los resultados de la regresión Ridge son bastante semejantes pero más apegados al coeficiente de variación de los datos disponibles en Platón Sánchez. En problemas con seis o siete registros amplios disponibles, la regresión Ridge será una mejor opción que la inspección de 64 o de 128 posibles modelos obtenidos por mínimos cuadrados de los residuos, como lo establece el esquema de eliminación de variables.

permit a better use of the information. However, none of the procedure of the hydrology statistic generates new information or substitutes its scarcity.

Therefore, the National Water Commission (Comisión Nacional del Agua, CONAGUA) should re-establish the network of climatological and hydrological measurements that existed in the decade of the eighties. Lafragua *et al.* (2006) point out that the situation with respect to the hydrological information is truly critical, even though it is the base for the estimation of the superficial availability in the watersheds and throughout the country. For example, in 2004 there were 50 % of the climatological stations existing in 1980 and the hydrometric stations had been reduced from nearly 1,600 in 1980 to only 430 in 2002.

Conclusions

It is better to use part of the statistical information of all of the regressors, as with the Ridge regression, than to use all of the information of some regressors and none of others, as with the method of the selection of variables. Furthermore, Ridge regression is a direct procedure, of easy implementation within the solution of least squares of the residuals and the interpretation of the Ridge trace does not present difficulties.

With respect to the numerical application described, problem previously dealt with by the

Agradecimientos

Se agradecen las correcciones y sugerencias de los dos árbitros anónimos y del editor asignado, las cuales permitieron completar el trabajo en ciertos aspectos teóricos y cálculos no expuestos, entre los segundos el análisis de homogeneidad regional.

Literatura Citada

Campos A., D. F. 2010. Verificación de la homogeneidad regional mediante tres pruebas estadísticas. Tecnología y Ciencias del Agua I(4): 157-165.

Campos A., D. F. 2011. Transferencia de información hidrológica mediante regresión lineal múltiple, con selección óptima de regresores. Agrociencia 45(8): 863-880.

Campos A., D. F. 2012. Descripción y aplicación de la versión corregida del Test de Langbein para verificar homogeneidad regional. Ingeniería, Investigación y Tecnología XIII(4): 411-416.

Carnahan, B., H. A. Luther, and J. O. Wilkes. 1969. Matrices and related topics. *In:* Applied Numerical Methods. John Wiley & Sons. New York, U.S.A. pp: 210-268.

Fill, H. D., and J. R. Stedinger. 1995. Homogeneity test based upon Gumbel distribution and a critical appraisal of Dalrymple's test. J. Hydrol. 166: 81-105.

Hosking, J. R. M., and J. R. Wallis. 1997. Screening the data. *In:* Regional Frequency Analysis. An Approach Based on L–moments. Cambridge University Press. Cambridge, United Kingdom. pp: 44-53.

Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12: 55-67.

Lafragua C., J., D. González R., y Y. Solís A. 2006. Cantidad y calidad de la información climatológica e hidrométrica para el cálculo de la disponibilidad de agua superficial. *In:* Memoria del XIX Congreso Nacional de Hidráulica, Tema: Hidrología. Cuernavaca, Morelos, México. (En CD).

elimination of variables, the results of the Ridge regression are quite similar, but more closely related to the coefficient of variation of the data available in Platón Sánchez. In problems with six or seven available long records, Ridge regression would be a better option than the inspection of 64 or 128 possible models obtained by least squares of the residuals, as established by the scheme of elimination of variables.

—End of the English version—

Montgomery, D. C., E. A. Peck, y G. G. Vining. 2002. Multicolinealidad. *In:* Introducción al Análisis de Regresión Lineal. Compañía Editorial Continental. México, D. F. pp: 291-342.

Montgomery, D. C., E. A. Peck, and J. R. Simpson. 1998. Multicollinearity and biased estimation in regression. *In:* Wadsworth, H. M. (ed). Handbook of Statistical Methods for Engineers and Scientists. McGraw-Hill, Inc. New York, U.S.A. 2nd ed. pp: 16.3-16.27.

Ryan, T. P. 1998. Linear regression. *In:* Wadsworth, H. M. (ed). Handbook of Statistical Methods for Engineers and Scientists,. McGraw-Hill, Inc. New York, U.S.A. Second edition. pp: 14.1-14.43.

Weimin, B., and L. Qian. 2012. Estimating selected parameters for the XAJ model under multicollinearity among watersheds characteristics. J. Hydrol. Eng. 17(1): 118-128.

Yu, X., and S. Liong. 2007. Forecasting of hydrology time series with ridge regression in feature space. J. Hydrol. 332(3-4): 290-302.

Zhao, B., Y. Tung, and J. Yang. 1995. Estimation of unit hydrograph by ridge least-squares method. J. Irrig. Drain. Eng. 121(3): 253-259.