Nota técnica

Ampliación de registros de volumen escurrido anual con base en información regional y regresión de tipo Ridge

• Daniel Francisco Campos-Aranda •
Profesor jubilado de la Universidad Autónoma de San Luis Potosí, México
*Autor de correspondencia

Resumen

Campos-Aranda, D. F. (julio-agosto, 2014). Ampliación de registros de volumen escurrido anual con base en información regional y regresión de tipo Ridge. *Tecnología y Ciencias del Agua*, *5*(4), 173-185.

En general, la planeación, diseño y manejo de las obras de infraestructura hidráulica se realiza con base en los registros históricos disponibles de crecientes, escurrimientos y lluvias anuales. Conforme tales registros abarcan más años, sus estimaciones hidrológicas tienen una mayor exactitud. Por lo anterior, siempre es necesario ampliar los registros cortos (*Y*), por ejemplo a través de la *regresión lineal múltiple* (RLM), la cual utiliza la información *regional* disponible. El establecimiento de una RLM tiene varias dificultades, quizá la más importante en el transporte de información hidrológica sea la presencia de correlación entre los registros auxiliares o variables predictivas (X_i), lo cual da origen a un problema de *multicolinealidad*. En este trabajo se expone con detalle el diagnóstico cuantitativo de tal problema por medio de los factores de inflación de la varianza y de los eigenvalores de la matriz $\mathbf{X}' \cdot \mathbf{X}$. También se describe ampliamente la RLM de tipo Ridge o sesgada como estrategia para minimizar los efectos de la multicolinealidad, buscando su parámetro de sesgo con base en la traza Ridge. Se detalla una aplicación numérica para ampliar el registro de volúmenes escurridos anuales en la estación hidrométrica Santa Isabel de la cuenca del Alto Río Grijalva, utilizando cuatro registros amplios cercanos. Por último se formulan las conclusiones, las cuales destacan las ventajas del uso de la RLM de tipo Ridge.

Palabras clave: multicolinealidad, factores de inflación de la varianza, eigenvalores, eigenvectores, alto río Grijalva.

Abstract

Campos-Aranda, D. F. (July-August, 2014). Extension of Annual Runoff Volume Records Based on Regional Information and Ridge Regression. Water Technology and Sciences *(in Spanish),* 5*(4), 173-185.*

The planning, design and management of water infrastructure are typically based on available historical records of annual floods, runoff and rainfall. The more years covered by these records the more accurate the hydrological estimates. Therefore, it is always necessary to expand short records (Y), for example, through multiple linear regression (MLR), which uses available regional information. Establishing a MLR has several difficulties, perhaps the most important regarding the transport of hydrological information is the presence of correlation between the auxiliary or predictor variables (X_i), which gives rise to a problem of multicollinearity. In this work, the quantitative evaluation of multicollinearity is presented in detail through variance inflation factors and eigenvalues for the $\mathbf{X}' \cdot \mathbf{X}$ matrix. In addition, the biased or Ridge MLR is described extensively as a strategy to minimize the effects of multicollinearity, seeking its biasing parameter based on the Ridge trace. A numerical application is presented in detail, which expands the annual runoff volume records in the Santa Isabel gauging station in the upper Grijalva River using four broad records nearby. Lastly, several conclusions are formulated which highlight the advantages of using the Ridge MLR.

***Keywords**: Multicollinearity, variance inflation factors, eigenvalues, eigenvectors, upper Grijalva River.*

Introducción

En general, la planeación, diseño, operación y revisión de las obras hidráulicas se realiza con base en los registros históricos disponibles de datos hidrológicos, principalmente crecientes, escurrimientos y lluvias *anuales*. Al contar sólo con registros cortos, la confianza en sus estimaciones estadísticas es baja y por ello se debe buscar información adicional y técnicas de ampliación de las series disponibles (Salas *et al*., 2008). Para el caso específico de los escurrimientos anuales, la fuente más común de datos adicionales son los registros largos de las estaciones hidrométricas cercanas y la técnica estadística más utilizada para el llamado

transporte regional de información hidrológica es la regresión lineal múltiple (RLM).

Bajo este enfoque, el registro corto (Y) debe tener un periodo común de información con las series largas (X_i, *regresores*) y guardar una cierta dependencia o correlación con ellas. Obtenida y validada su ecuación, se pueden obtener las estimaciones que amplían el registro corto, con base en los valores observados en los regresores. Esta técnica estadística implica una complejidad real tan sólo en los tres aspectos siguientes (Ryan, 1998): (1) selección de cuántos y cuáles registros amplios e independientes utilizar; (2) interpretación de los resultados, en especial de los coeficientes de la regresión (β_i), y (3) determinación de cuándo un método de ajuste, alternativo al de mínimos cuadrados de los residuos, debe ser utilizado.

Como el registro corto debe estar correlacionado con los auxiliares o circunvecinos, resulta lógico esperar que también éstos muestren cierta dependencia entre sí, pues además de ser cercanos guardan correlación con la variable dependiente. La correlación entre los regresores implica que alguna parte de la información estadística contenida en cada uno también está presente en alguna de las otras $i - 1$ variables independientes (Haan, 1977). Esta situación genera un problema de *multicolinealidad* debido a la semejanza o correlación existente entre los registros involucrados. Tal problema se debe diagnosticar y resolver, por ejemplo, a través de la regresión tipo Ridge (Montgomery *et al*., 1998; 2002).

Los tres *objetivos* básicos de este trabajo son: (1) describir la teoría estadística relativa a la RLM y su ajuste por mínimos cuadrados de los residuos; (2) explicar los conceptos y el diagnóstico cuantitativo de la multicolinealidad, y (3) exponer y aplicar la RLM de tipo Ridge o sesgada, como método eficiente para contrarrestar la dependencia lineal entre los regresores. Se realiza una aplicación numérica con cinco estaciones hidrométricas de la cuenca del Alto Río Grijalva para ampliar el registro corto de volúmenes escurridos anuales en la estación Santa Isabel. Se comparan los resultados con los obtenidos previamente, mediante el enfoque de selección exhaustiva de variables predictivas.

Resumen de la teoría operativa

Regresión Lineal Múltiple (RLM) y su ajuste

Con frecuencia se puede establecer una relación de tipo lineal entre la variable dependiente (Y) y varias (p) independientes X_1, X_2,..., X_p, que es la generalización o extensión natural de la regresión lineal simple; su expresión es (Ryan, 1998):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \cdots + \beta_p X_p + \varepsilon \qquad (1)$$

Debido a esto último, los principios que rigen a la regresión lineal se aplican a la RLM; por ejemplo, que tanto Y como las X_i estén normalmente distribuidas, y que los errores ε sean independientes también con distribución normal de media cero y misma varianza (σ^2) para cada X_i. Por lo general, la estimación de los coeficientes de la regresión (β_i) se realiza mediante el llamado ajuste de mínimos cuadrados de los residuos. Tal solución matricial para la RLM, en el caso general de p variables independientes o *regresores* y n observaciones o datos de Y, X_1, X_2,..., X_p, es la siguiente (Ryan, 1998):

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2)$$

Siendo:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

$$=\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_{1i}-\beta_2X_{2i}-\cdots-\beta_pX_{p\,i}\right)^2=0 \quad (3)$$

Entonces, diferenciando el lado derecho de la ecuación anterior con respecto a β_0, β_1, β_2,..., β_p, por separado, se originan las ecuaciones llamadas *normales*, función de los parámetros desconocidos. En notación matricial, estas ecuaciones son:

$$\left(\mathbf{X}'\cdot\mathbf{X}\right)\cdot\hat{\beta}=\mathbf{X}'\cdot\mathbf{Y} \quad (4)$$

cuya solución es:

$$\hat{\beta}=\left(\mathbf{X}'\cdot\mathbf{X}\right)^{-1}\cdot\left(\mathbf{X}'\cdot\mathbf{Y}\right) \quad (5)$$

en la cual **X′** es la matriz transpuesta de **X** y $(\mathbf{X}'\cdot\mathbf{X})^{-1}$ indica la matriz inversa de **X′ · X.**

Coeficiente de determinación múltiple

Designado por R^2 es probablemente el estadístico más utilizado para medir lo adecuado de un modelo de regresión; indica cuánta de la varianza de Y la explica el modelo; por ello su expresión es (Hirsch *et al.*, 1993):

$$R^2=\frac{SC_Y-SC_{Res}}{SC_Y}=1-\frac{\sum_{i}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{1}^{n}\left(Y_i-\overline{Y}\right)^2} \quad (6)$$

en la cual $\hat{Y}_i$ es la estimación de la variable Y_i a través de la ecuación de regresión; por ello SC_{Res} es la suma de cuadrados de los residuos y SC_Y es la varianza total de la variable dependiente, cuya media aritmética es $\overline{Y}$.

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 tiene como ventaja fundamental que las matrices **X** involucradas de n renglones ahora tienen p columnas, ya que la ecuación de RLM es:

$$Y-\overline{Y}=\beta_1\left(X_1-\overline{X_1}\right)+\beta_2\left(X_2-\overline{X_2}\right)$$

$$+\cdots+\beta_p\left(X_p-\overline{X_p}\right) \quad (7)$$

cuyo reacomodo para obtener la ecuación (1) implica que:

$$\beta_0=\overline{Y}-\beta_1\overline{X_1}-\beta_2\overline{X_2}-\cdots-\beta_p\overline{X_p} \quad (8)$$

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:

$$E_{j\,i}=\frac{X_{j\,i}-\overline{X_j}}{S_j^{1/2}}\ \text{con}\ i=1,2,3,\ldots,n,\ j=1,2,3,\ldots,p \quad (9)$$

$$Y_i=\frac{Y_i-\overline{Y}}{S_Y^{1/2}}\ \text{con}\ \ i=1,2,3,\ldots,n \quad (10)$$

Donde:

$$S_j=\sum_{i=1}^{n}\left(X_{j\,i}-\overline{X_j}\right)^2 \quad (11)$$

$$S_Y=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2 \quad (12)$$

El escalamiento de longitud unitaria produce, en relación con la ecuación (4), que la matriz **E′ · E** sea una matriz de correlación simple entre los regresores X_j; además, la matriz **E′ · Y** es ahora una matriz de correlación simple entre cada regresor X_j y la variable dependiente Y. Este escalamiento y el normal conducen a *coeficientes estandarizados de regresión*, cuya comparación entre ellos define la importancia de cada regresor.

Otro escalamiento que se requiere con frecuencia está asociado con la estabilidad numérica de la matriz inversa de $\mathbf{E}' \cdot \mathbf{E}$, pues es común obtenerla planteando esta igualdad $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}$; al transformar la matriz **A** en la matriz identidad **I** y realizar las mismas operaciones en **I**, ésta se convierte en la matriz $\mathbf{A}^{-1}$ 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.

Multicolinealidad: definición y soluciones

Como ya se indicó, en el caso de una ampliación de un registro hidrológico con base en la información *regional* disponible, el conjunto de datos siempre mostrará un cierto grado de multicolinealidad, a menos que las columnas de la matriz **X** sean ortogonales, es decir que $\mathbf{X}' \cdot \mathbf{X}$ sea una matriz diagonal, lo cual sólo sucederá en un experimento diseñado (Montgomery *et al.*, 1998; 2002). Siendo $\mathbf{X}_j$ la *j*-ésima columna de la matriz **X**, la *multicolinealidad* se define de manera formal como la dependencia lineal entre tales columnas, es decir, que existe un conjunto de constantes $t_1, t_2, ..., t_p$, no todas cero, tales que:

$$\sum_{j=1}^{p} t_j \cdot \mathbf{X}_j = 0 \qquad (13)$$

Si la ecuación anterior es exactamente válida para un subconjunto de las columnas de **X**, el rango de la matriz $\mathbf{X}' \cdot \mathbf{X}$ es menor que *p* y entonces no existe $(\mathbf{X}' \cdot \mathbf{X})^{-1}$. Cuando la ecuación (13) es válida, sólo aproximadamente existe multicolinealidad; es decir, que la matriz $\mathbf{X}' \cdot \mathbf{X}$ presenta un cierto grado de deterioramiento. En general, cuando se aplica el método de mínimos cuadrados de los residuos 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 las tres siguientes (Ryan, 1998; Montgomery *et al.*, 1998):

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. Re-especificar 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 $X = (X_1 + X_2)/X_3$, o bien $X = X_1 \cdot X_2 \cdot X_3$, que preserva el contenido de la información de los regresores originales, pero que reduce el deterioramiento de los datos debido a la multicolinealidad. Otro método de re-especificación muy efectivo consiste en la eliminación de una o más variables o regresores, esto de manera definitiva reduce la multicolinealidad, pero puede dañar notablemente la capacidad predictiva del modelo.
3. Obtener estimaciones sesgadas, como la RLM de tipo Ridge.

Diagnóstico cuantitativo de la multicolinealidad con base en $(\mathbf{E}' \cdot \mathbf{E})^{-1}$

La manera más simple de descubrir la multicolinealidad es a través de la inspección de la matriz $\mathbf{E}' \cdot \mathbf{E}$, cuyos elementos fuera de la diagonal principal corresponden a los coeficientes de correlación simple entre pares de regresores; entonces, si existen valores absolutos mayores de 0.80, se tiene dependencia entre tal pareja. Este método sólo detecta la multicolinealidad, pero no la cuantifica; en cambio, cuando los factores de inflación de la varianza VIF de *Variance Inflation Factor* son mayores que 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):

$$\text{VIF}_j = C_{jj} = \frac{1}{\left(1 - R_j^2\right)} \qquad (14)$$

donde R_j^2 es el coeficiente de determinación que resulta de la RLM entre el regresor X_j como variable dependiente y el resto $p - 1$ como regresores. Los VIF_j corresponden a la diagonal principal de la matriz inversa de $\mathbf{E}' \cdot \mathbf{E}$.

Diagnóstico cuantitativo de la multicolinealidad con base en los eigenvalores de E′ · E

Los eigenvalores de la matriz $\mathbf{E}' \cdot \mathbf{E}$ se designan por λ_1, λ_2, λ_3,..., λ_p; también se conocen como valores propios y corresponden a las raíces de la ecuación característica $\left|\mathbf{A} - \lambda \cdot \mathbf{I}\right| = 0$ de la matriz **A**. 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 $\mathbf{E}' \cdot \mathbf{E}$ se define como (Montgomery *et al.*, 1998; 2002):

$$\kappa = \frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}} \qquad (15)$$

y representa el espectro de variación de los eigenvalores de la matriz $\mathbf{E}' \cdot \mathbf{E}$. En general, cuando κ es menor que 100, prácticamente no existen problemas de multicolinealidad; cuando varía de 100 a 1 000 se tiene multicolinealidad de moderada a fuerte, y cuando excede de 1 000 seguramente se tendrán graves problemas asociados con ésta. Los *índices de condición* κ_j de la matriz $\mathbf{E}' \cdot \mathbf{E}$ son:

$$\kappa_j = \frac{\lambda_{\text{máx}}}{\lambda_j} \quad \text{con} \quad j = 1,2,3,..., p \qquad (16)$$

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

La regresión Ridge

El método de mínimos cuadrados de los residuos asegura 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 puede obtener un estimador sesgado $\hat{\beta}^*$ que tenga mucho menor varianza, entonces 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 $\hat{\beta}$.

Se han desarrollado varios procedimientos para obtener estimadores sesgados de los coeficientes de regresión β. Uno de ellos es la *regresión Ridge* o de cresta, que fue propuesta a comienzos de la década de los años setenta por Hoerl y Kennard (1970), y que debe su nombre a la semejanza de sus operaciones matemáticas con el análisis Ridge empleado 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); ésta es (Montgomery *et al.*, 1998; 2002):

$$\left(\mathbf{E}' \cdot \mathbf{E} + k \cdot \mathbf{I}\right) \cdot \hat{\beta}_R = \mathbf{E}' \cdot \mathbf{Y} \qquad (17)$$

por lo cual:

$$\hat{\beta}_R = \left(\mathbf{E}' \cdot \mathbf{E} + k \cdot \mathbf{I}\right)^{-1} \cdot \left(\mathbf{E}' \cdot \mathbf{Y}\right) \qquad (18)$$

en las expresiones anteriores, la constante $k \geq 0$, denominada *parámetro de sesgo*, se selecciona durante el proceso de aplicación de la regresión Ridge. En realidad, 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. Debido a esto último se cree, pues no hay demostración matemática concluyente, que conduce a ecuaciones de regresión que funcionan mejor para predecir observaciones futuras, en comparación con la de mínimos cuadrados de los residuos.

Hoerl y Kennard (1970) sugirieron 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 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. La idea fundamental es seleccionar el valor de *k* más pequeño, donde los $\hat{\beta}_R$ ya sean estables. Con ello es posible que se obtenga una ecuación de regresión con menor error medio cuadrático que el correspondiente a mínimos cuadrados.

Aplicación numérica

Datos en el Alto Río Grijalva

En la figura 1 se muestra la ubicación, dentro de la Región Hidrológica 30 (Ríos Grijalva y Usumacinta), de cinco cuencas pertenecientes a las estaciones hidrométricas Santa Isabel, La Escalera, El Boquerón II, Las Flores II y Santa María, cuyos datos generales se tienen en el cuadro 1. El planteamiento general de esta aplicación numérica consiste en ampliar el registro corto de Santa Isabel a través de RLM de tipo Ridge, empleando los otros cuatro registros largos. Esta estimación ya fue realizada (Campos, 2012), con base en el método de selección de variables predictivas.

Campos (2012) recopiló en el sistema *BANDAS* (IMTA, 2002) los datos disponibles en las estaciones citadas, correspondientes al volumen escurrido anual en millones de metros cúbicos (Mm^3); también estimó valores mensuales perdidos para los años incompletos y dedujo las magnitudes anuales faltantes en el periodo común, definido de 1956 a 1973. Además, estableció el periodo de ampliación de 1974 a 1994. Tales datos se presentan en el cuadro 2. Finalmente, probó que los datos no tuvieran componentes determinísticas y verificó su procedencia de una distribución normal, con base en el test de Shapiro y Wilk (Shapiro, 1998).

Diagnóstico de la multicolinealidad

Las matrices $\mathbf{E'} \cdot \mathbf{E}$, $\mathbf{E'} \cdot \mathbf{Y}$ y $(\mathbf{E'} \cdot \mathbf{E})^{-1}$ obtenidas para los datos del cuadro 1, procesados lógicamente con escalamiento unitario y subrutinas de multiplicación e inversión de matrices elaboradas *ex professo* son:

$$\mathbf{E'} \cdot \mathbf{E} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 \end{matrix} \\ \begin{bmatrix} 1.000000 & 0.706983 & 0.522109 & 0.409604 \\ 0.706983 & 1.000000 & 0.867273 & 0.759131 \\ 0.522109 & 0.867273 & 1.000000 & 0.911521 \\ 0.409604 & 0.759131 & 0.911521 & 1.000000 \end{bmatrix} \end{array} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{matrix}$$

$$\mathbf{E'} \cdot \mathbf{Y} = \begin{bmatrix} 0.742823 \\ 0.651777 \\ 0.576905 \\ 0.365996 \end{bmatrix} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{matrix}$$

$$\left(\mathbf{E'} \cdot \mathbf{E}\right)^{-1} = \begin{bmatrix} 2.175362 & -2.174041 & 0.340291 & 0.449162 \\ -2.174041 & 6.304874 & -4.623234 & 0.318447 \\ 0.340291 & -4.623233 & 10.405540 & -6.114622 \\ 0.449162 & 0.318447 & -6.114621 & 6.147887 \end{bmatrix}$$

La inspección de la matriz $\mathbf{E'} \cdot \mathbf{E}$ muestra que únicamente existen dos correlaciones importantes: la mayor (r_{xy} = 0.9115), entre X_3 y X_4, y la menor (r_{xy} = 0.8673), entre X_2 y X_3. Por lo anterior, se detecta un problema de multicolinealidad en los datos, pero quizás sea aceptable o moderada. En relación con el vector $\mathbf{E'} \cdot \mathbf{Y}$, ninguna correlación es importante y éstas disminuyen conforme las estaciones hidrométricas están más alejadas de la estación Santa Isabel (ver figura 1).

El diagnóstico cuantitativo de la multicolinealidad se tiene en el cuadro 3, cuyo primer renglón de resultados corresponde a los valores de los factores de inflación de la varianza (VIF_j) y son los elementos de la diagonal principal de la matriz inversa de $\mathbf{E'} \cdot \mathbf{E}$. Como la magnitud mayor de los VIF_j escasamente excede de 10, se encuentra multicolinealidad aceptable. Por otra parte, como ninguno de los índices

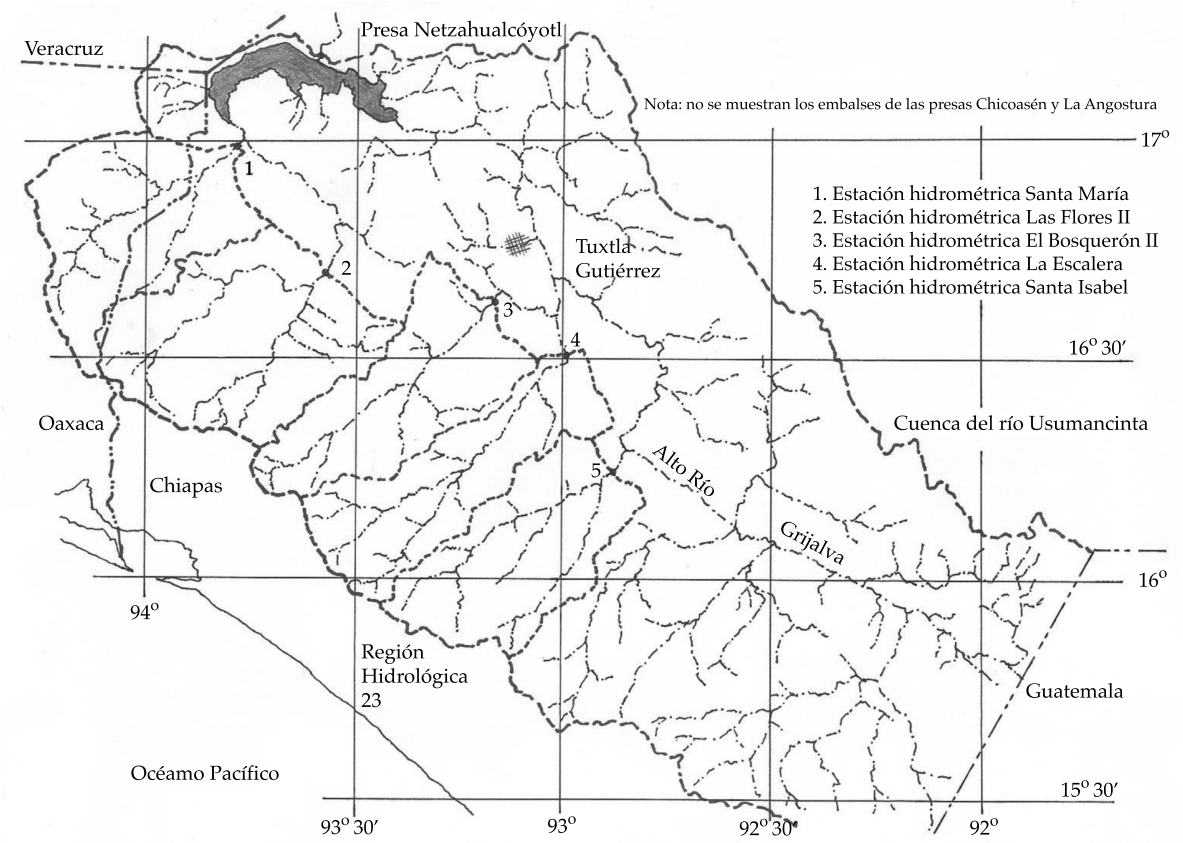

Figura 1. Localización geográfica de las cinco estaciones hidrométricas procesadas del Alto Río Grijalva.

Cuadro 1. Características generales de las estaciones hidrométricas utilizadas de la cuenca del Alto Río Grijalva.

Nombre	Clave*	Río aforado	Latitud N	Long. WG	Área de cuenca (km²)	Registro (años faltantes)
Santa Isabel	30053	El Dorado	16° 16′	92° 53′	1 873	1956-1973 (0)
La Escalera	30041	Santo Domingo	16° 32′	92° 57′	1 808	1954-2002 (4)
El Boquerón II	30020	Suchiapa	16° 40′	93° 09′	1 870	1949-2002 (6)
Las Flores II	30072	Zoyatenco	16° 42′	93° 33′	2 551	1962-2002 (5)
Santa María	30071	Encajonado	16° 57′	93° 46′	1 958	1962-2001(13)

*Según sistema *BANDAS*.

de condición (κ_j) excede de 100, entonces los problemas asociados con la multicolinealidad no serán serios, lo cual ratifica la conclusión anterior.

Con base en los elementos del cuarto eigenvector, que corresponde al menor de los eigenvalores, se establece la siguiente ecuación (13), relativa a la multicolinealidad presente:

$$-0.0629 \cdot X_1 + 0.3877 \cdot X_2 - 0.7862 \cdot X_3 + 0.4771 \cdot X_4 = 0 \quad (19)$$

considerando que el coeficiente de X_1 es cercano a cero se obtiene:

$$0.3877 \cdot X_2 + 0.4771 \cdot X_4 \cong 0.7862 \cdot X_3 \quad (20)$$

Cuadro 2. Volúmenes escurridos anuales (Mm^3) y sus parámetros estadísticos en las estaciones hidrométricas indicadas de la cuenca del Alto Río Grijalva.

Núm.	Año	Santa Isabel (Y)	La Escalera (X_1)	El Boquerón II (X_2)	Las Flores II (X_3)	Santa María (X_4)
1	1956	1 113.119	524.303	700.426	617.200	1 348.100
2	1957	846.957	343.245	327.830	195.700	829.200
3	1958	1 134.394	759.834	576.979	477.500	1 176.100
4	1959	748.138	538.022	456.217	340.900	1 008.000
5	1960	1 259.572	831.507	544.870	441.200	1 131.400
6	1961	876.884	555.208	340.309	209.800	846.500
7	1962	1 345.812	700.956	594.985	528.749	780.873
8	1963	1 151.548	793.521	519.857	509.515	1 062.852
9	1964	1 103.385	809.059	572.255	428.682	1 040.491
10	1965	1 227.758	619.464	424.201	311.510	926.283
11	1966	671.762	612.066	543.392	400.731	1 234.455
12	1967	629.364	302.969	337.493	319.296	820.282
13	1968	863.049	443.917	327.265	245.051	886.820
14	1969	1 071.681	872.190	702.579	654.717	1 475.765
15	1970	1 182.934	636.918	674.119	971.609	2 071.694
16	1971	1 131.230	737.992	707.531	499.456	1 156.651
17	1972	627.775	313.800	287.889	244.434	823.366
18	1973	1 183.812	701.963	800.585	800.408	1 714.570
1	1974	–	270.474	431.411	447.160	1 185.743
2	1975	–	285.090	429.059	258.044	1 006.232
3	1976	–	219.347	280.568	127.823	852.063
4	1977	–	272.813	253.843	95.508	710.200
5	1978	–	563.517	609.759	288.524	1 018.654
6	1979	–	286.202	371.625	318.484	1 124.668
7	1980	–	590.997	552.401	1 312.039	2 438.163
8	1981	–	844.403	729.838	659.665	1 516.715
9	1982	–	662.449	442.376	190.252	954.137
10	1983	–	436.859	545.409	395.663	874.889
11	1984	–	600.143	572.296	522.289	1 250.051
12	1985	–	434.315	392.148	190.942	736.901
13	1986	–	409.003	424.557	234.945	794.653
14	1987	–	298.749	334.747	143.007	764.274
15	1988	–	618.826	671.978	731.723	1 318.376
16	1989	–	601.612	738.056	662.633	988.181
17	1990	–	393.349	437.495	253.685	791.806
18	1991	–	177.277	257.679	137.659	751.560
19	1992	–	612.984	432.906	162.104	843.562
20	1993	–	414.408	504.291	375.818	1 096.850
21	1994	–	153.414	230.450	73.578	657.381
Máximo		1 345.812	872.190	800.585	1 312.09	2 438.163
Mínimo		627.775	153.414	230.450	73.578	657.381
$\overline{X}$		1 009.399	519.055	489.274	404.564	1 077.14
S		229.553	202.140	154.990	257.809	369.465
Cv		0.227	0.389	0.317	0.637	0.343
Cs		–0.509	–0.040	0.179	1.447	1.953
Ck		2.377	2.141	2.240	6.014	7.703
r_1		–0.196	0.216	–0.103	0.229	0.175

Cuadro 3. Resultados del diagnóstico de multicolinealidad para los datos del Alto Río Grijalva.

Indicadores	Regresores			
	X_1	X_2	X_3	X_4
VIF_j	2.17536	6.30487	10.40554	6.14789
λ_j	3.11744	0.67478	0.14690	0.06089
$\kappa_j = \lambda_{máx}/\lambda_j$	1.000	4.620	21.222	51.198
Regresores	**Eigenvectores**			
X_1	0.4096	0.8201	0.3946	–0.0629
X_2	0.5377	0.1170	–0.7395	0.3877
X_3	0.5380	–0.2964	–0.0679	–0.7862
X_4	0.5036	–0.4753	0.5411	0.4771

$$X_3 \cong 0.4931 \cdot X_2 + 0.6068 \cdot X_4 \qquad (21)$$

la ecuación anterior establece la relación aproximada entre X_3 con X_2 y X_4.

Cálculo y análisis de la traza Ridge

La aplicación de la ecuación (18), con base en un programa de cómputo elaborado *ex professo*, el cual utiliza los valores de k indicados en el cuadro 4, condujo a los coeficientes de regresión tipo Ridge ahí concentrados, cuyos coeficientes de determinación (R^2) correspondientes también se muestran en este cuadro. El cálculo de R^2 se realizó haciendo el centrado de los datos y utilizando un COR = 500. A partir de los resultados del cuadro 4 se ha construido la traza Ridge, mostrada en la figura 2.

El estudio de la traza Ridge muestra que sólo el coeficiente de regresión de la variable X_1, es decir, de la estación hidrométrica La Escalera es estable; en cambio, los relativos a las estaciones Las Flores II (X_3) y Santa María (X_4) varían bastante y de manera similar; por último, el de El Boquerón II (X_2) fluctúa menos, pero incluso cambia de signo. Con el objeto de establecer el menor valor para el parámetro de sesgo (k), se acepta que en la traza Ridge sus coeficientes ya están estables en 0.25 y más apropiadamente en 0.35.

Estimaciones Ridge y su contraste

En el cuadro 5 se exponen las 18 estimaciones de la variable dependiente ($\hat{Y}_j$), esto es, el registro histórico en la estación hidrométrica Santa Isabel en el periodo 1956-1973, así como sus residuos correspondientes, realizadas con las regresiones Ridge, que emplean k = 0.250 y 0.350. Los coeficientes de regresión respectivos se muestran en el cuadro 6 y fueron obtenidos con datos centrados y usando un COR de 500.

En el cuadro 6 se han concentrado los resultados del contraste entre los residuos de los dos mejores modelos de regresión obtenidos a través de selección de variables predictivas (Campos, 2012) y las regresiones Ridge adoptadas. Se observa que la regresión Ridge origina valores ligeramente mayores de los residuos negativos y escasamente menores de los residuos positivos; la suma de residuos al cuadrado es mayor, pues no es el mejor ajuste a los datos, pero la suma algebraica de sus errores es menor.

Estimaciones Ridge adoptadas

Finalmente, en el cuadro 7 se presentan los 21 volúmenes escurridos anuales estimados en la estación hidrométrica Santa Isabel para el periodo de 1974 a 1994, 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 variables predictivas (Campos, 2012). Se observa que ambas series estimadas de volúmenes escurridos anuales presentan el

Cuadro 4. Coeficientes de regresión tipo Ridge ($\hat{\beta}_R$) obtenidos para los valores del parámetro de sesgo indicado.

$\hat{\beta}_R$	**Valores del parámetro de sesgo (*k*)**							
	0.000	**0.005**	**0.010**	**0.020**	**0.030**	**0.040**	**0.050**	**0.075**
β_1	0.5596	0.5533	0.5476	0.5378	0.5294	0.5219	0.5152	0.5003
β_2	-0.0562	-0.0266	-0.0018	0.0376	0.0671	0.0898	0.1077	0.1386
β_3	1.0045	0.9334	0.8724	0.7731	0.6958	0.6340	0.5835	0.4903
β_4	-0.7363	-0.6878	-0.6457	-0.5760	-0.5204	-0.4749	-0.4369	-0.3638
R^2	0.6892	0.6891	0.6890	0.6885	0.6878	0.6870	0.6861	0.6834
$\hat{\beta}_R$	**Valores del parámetro de sesgo (*k*)**							
	0.100	**0.120**	**0.150**	**0.180**	**0.200**	**0.250**	**0.300**	**0.350**
β_1	0.4873	0.4778	0.4646	0.4523	0.4446	0.4267	0.4105	0.3963
β_2	0.1576	0.1678	0.1781	0.1845	0.1874	0.1915	0.1930	0.1929
β_3	0.4267	0.3887	0.3457	0.3138	0.2968	0.2641	0.2409	0.2236
β_4	-0.3110	-0.2781	-0.2388	-0.2079	-0.1907	-0.1560	-0.1294	-0.1084
R^2	0.6805	0.6782	0.6746	0.6711	0.6689	0.6633	0.6580	0.6530

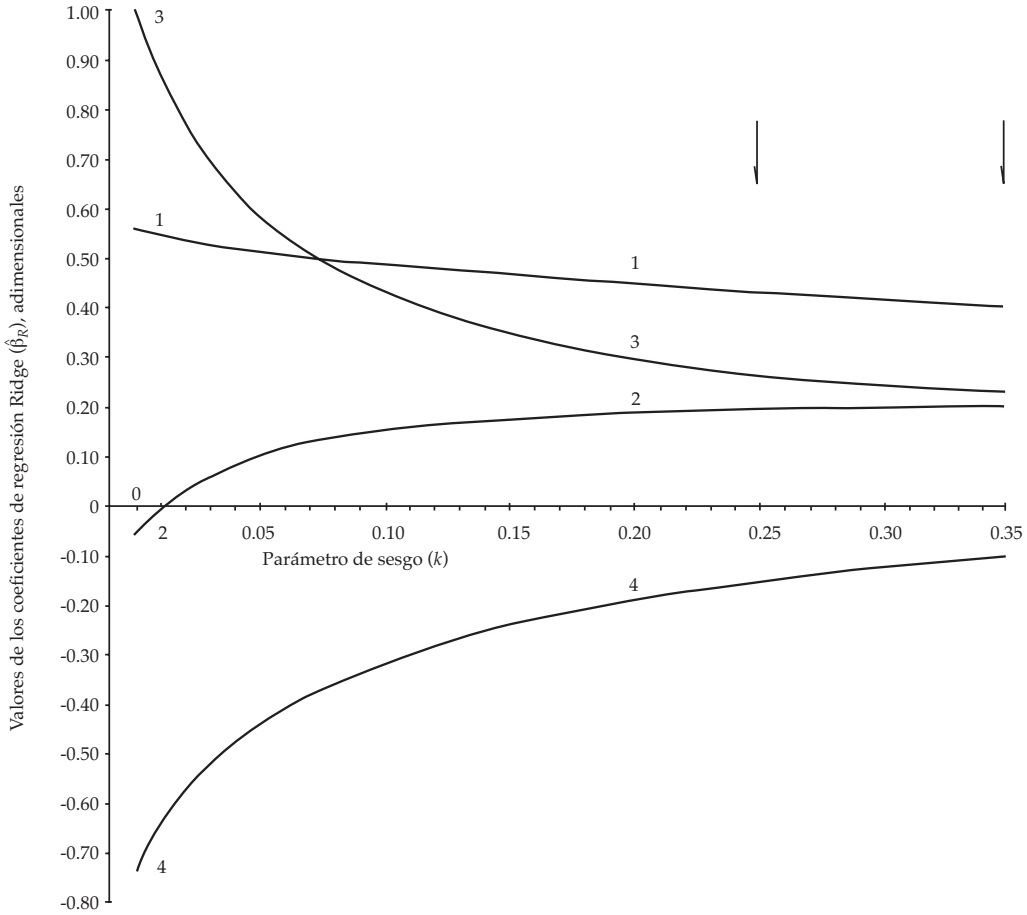

Figura 2. Traza Ridge para los datos del Alto Río Grijalva.

Cuadro 5. Estimaciones (Mm^3) de la variable dependiente ($\hat{Y}_i$) obtenidas con las regresiones Ridge y sus residuos correspondientes.

Año	$k = 0.250$	Residuo	$k = 0.350$	Residuo
1956	1 025.148	87.971	1 027.709	85.410
1957	730.486	116.471	740.300	106.657
1958	1 108.431	25.963	1 106.384	28.010
1959	913.045	-164.907	917.222	-169.084
1960	1 136.477	123.095	1 132.984	126.588
1961	866.418	10.466	871.102	5.782
1962	1 214.719	131.093	1194.033	151.779
1963	1 167.052	-15.504	1 155.790	-4.242
1964	1 146.003	-42.618	1 140.809	-37.424
1965	961.314	266.444	961.738	266.020
1966	949.722	-277.960	958.822	-287.060
1967	783.116	-153.752	784.173	-154.809
1968	805.278	57.771	811.278	51.771
1969	1 225.672	-153.990	1 222.444	-150.763
1970	1 099.002	83.932	1 101.088	81.846
1971	1 140.574	-9.344	1 139.339	-8.109
1972	734.440	-106.665	740.087	-112.312
1973	1 162.270	21.542	1 163.875	19.937
Máximo	1 225.672	266.444	1 222.444	266.020
Mínimo	730.486	-277.960	730.300	-287.060

Cuadro 6. Indicadores de los residuos obtenidos con los mejores modelos de mínimos cuadrados y con la regresión Ridge.

Modelo analizado	Coeficientes de regresión					Valores de los residuos			
	β_0	β_1	β_2	β_3	β_4	Mínimo	Máximo	$\sum_{i=1}^{18} e_i$	$\sum_{i=1}^{18} e_i^2$
$Y = f(X_1, X_3, X_4)$	648.9886	0.6939	–	1.0636	-0.4884	-225.249	269.998	0.298	278 905.4
$Y = f(X_1, X_2, X_3, X_4)$	657.9566	0.7188	-0.0818	1.1091	-0.4903	-220.896	267.891	0.347	278 457.1
Ridge con $k = 0.250$	1.128022	0.61220	0.20869	0.58922	-0.27422	-277.960	266.444	0.006	301 594.6
Ridge con $k = 0.350$	1.111032	0.58847	0.23013	0.51196	-0.23256	-287.060	266.020	-0.003	310 879.8

mismo comportamiento, pero la procedente de la regresión Ridge es mayor y con los valores máximos más pequeños, lo cual origina un valor medio ($\overline{X}$ = 862.3 Mm^3) y un coeficiente de variación (Cv = 0.223) más parecidos a los de los datos históricos de Santa Isabel ($\overline{X}$ = 1 009.4 Mm^3 y Cv = 0.227), en comparación con los obtenidos mediante selección de regresores ($\overline{X}$ = 831.1 Mm^3 y Cv = 0.293).

Conclusiones

La regresión Ridge es un procedimiento directo, de fácil implementación dentro de la solución de mínimos cuadrados de los residuos (ecuaciones (5) y (18)) y la interpretación y uso de la traza Ridge no presenta ninguna dificultad.

En relación con los problemas de ajuste que origina la multicolinealidad, existe consenso

Cuadro 7. Volúmenes escurridos anuales (Mm^3) estimados en la estación Santa Isabel mediante la regresión Ridge.

Año	$k = 0.250$	$k = 0.350$	Año	$k = 0.250$	$k = 0.350$
1974	757.946	767.132	1988	1 152.709	1 142.331
1975	704.198	720.120	1989	1 205.796	1 188.828
1976	598.509	616.445	1990	828.465	833.405
1977	645.525	658.206	1991	601.333	614.829
1978	926.913	938.268	1992	893.815	902.676
1979	696.028	710.956	1993	843.611	852.755
1980	1 145.583	1 135.113	1994	596.110	583.616
1981	1 206.035	1 205.375	Máx	1 206.035	1 205.375
1982	912.337	922.659	Mín	569.110	583.616
1983	938.495	937.210	$\overline{X}$	856.160	862.282
1984	1 015.802	1 017.064	S	200.618	192.259
1985	822.168	827.724	Cv	0.234	0.223
1986	823.526	829.383	Cs	0.420	0.396
1987	691.446	703.830	Ck	2.580	2.566

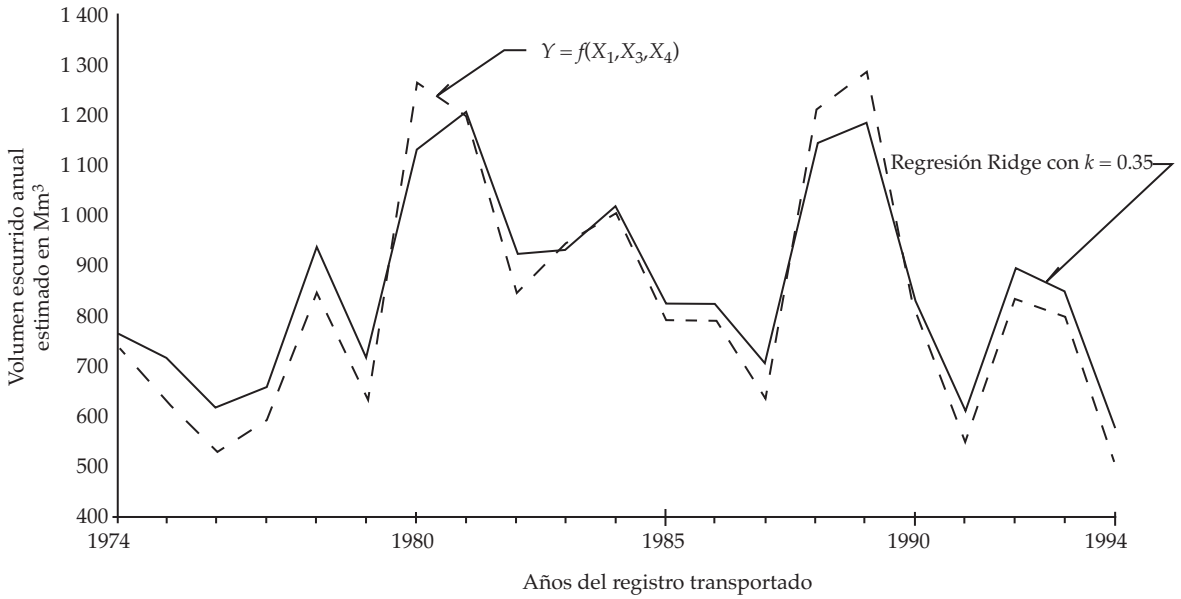

Figura 3. Contraste de estimaciones en la estación hidrométrica Santa Isabel del Alto Río Grijalva.

para recomendar que es mejor usar algo 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 predictivas.

Respecto a la aplicación numérica descrita, problema previamente abordado con eliminación de variables, los resultados de la regresión Ridge son bastante semejantes (ver figura 3), pero más apegados a los parámetros estadísticos históricos de la estación Santa Isabel (cuadro 2).

Finalmente, en problemas con seis o siete registros amplios disponibles, caso común al transportar registros de lluvia anual, la

regresión Ridge será una mejor opción que la inspección de 64 o 128 posibles modelos obtenidos por mínimos cuadrados de los residuos, como lo establece el esquema de eliminación de variables predictivas.

Recibido: 26/12/12
Aceptado: 09/01/14

Referencias

Campos-Aranda, D. F. (2012). *Ampliación de registros de volumen escurrido anual a través de Regresión Lineal Múltiple, con selección de variables predictivas*. XXII Congreso Nacional de Hidráulica. Tema: Aprovechamiento Integral de Cuencas, Ponencia 29, del 7 al 9 de noviembre, Acapulco, Guerrero.

Carnahan, B., Luther, H. A., & Wilkes, J. O. (1969). Matrices and Related Topics. Chapter 4. In *Applied Numerical Methods* (pp. 210-268). New York: John Wiley & Sons.

Haan, C. T. (1977). Multivariate Analysis. Chapter 12. In *Statistical Methods in Hydrology* (pp. 236-262). Ames, USA: The Iowa State University Press.

Hirsch, R. M., Helsel, D. R., Cohn, T. A., & Gilroy, E. J. (1993). Statistical Analysis of Hydrologic Data. Chapter 17. In D. R. Maidment (Ed.), *Handbook of Hydrology* (pp. 17.1-17.55). New York: McGraw-Hill, Inc.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55-67.

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

Montgomery, D. C., Peck, E. A., & Simpson, J. R. (1998). Multicollinearity and Biased Estimation in Regression. Chapter 16. In H. M. Wadsworth (Ed.), *Handbook of Statistical Methods for Engineers and Scientists* (pp. 291-342). Second edition. New York: McGraw-Hill, Inc.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2002). Multicolinealidad. Capítulo 10. En *Introducción al análisis de regresión lineal* (pp. 291-342). México, DF: Compañía Editorial Continental.

Ryan, T. P. (1998). Linear Regression. Chapter 14. In H. M. Wadsworth (Ed.), *Handbook of Statistical Methods for Engineers and Scientists* (pp. 14.1–14.43). New York: McGraw-Hill, Inc.

Salas, J. D., Raynal, J. A., Tarawneh, Z. S., Lee, T. S., Frevert, D., & Fulp, T. (2008). Extending Short Record of Hydrologic Data. Chapter 20. In V. P. Singh (Ed.), *Hydrology and Hydraulics* (pp. 717-760). Highlands Ranch, USA: Water Resources Publications.

Shapiro, S. S. (1998). Selection, Fitting and Testing Statistical Models. Chapter 6. In H. M. Wadsworth (Ed.), *Handbook of Statistical Methods for Engineers and Scientists* (pp. 6.1-6.35). Second edition. New York: McGraw-Hill, Inc.

Dirección del autor

Dr. Daniel Francisco Campos Aranda

Profesor jubilado de la Universidad Autónoma de San Luis Potosí
Genaro Codina 240, Colonia Jardines del Estadio
78280 San Luis Potosí, San Luis Potosí, México
Teléfono: +52 (444) 8151 431
campos_aranda@hotmail.com