CORRELACION Y REGRESION CON APLICACIONES EN LA HIDROLOGIA

Roger Amisial

CIDIAT

El Centro Interamericano de Desarrollo Integral de Aguas y Tierras (CIDIAT) fue establecido en el año 1964 mediante un acuerdo entre el Gobierno de Venezuela, la Universidad de Los Andes y la Organización de Estados Americanos (OEA).

Después de 10 años de operación y cumplido el proceso de transferencia, el CIDIAT pasó a ser un Centro Venezolano dirigido y administrado conjuntamente por el Gobierno de Venezuela y la Universidad de Los Andes.

Además, para sustentar las actividades del CIDIAT en países miembros de la OEA, se firmó un nuevo acuerdo que rige el denominado "Programa Interamericano" el cual se realiza conjuntamente por el CIDIAT como Institución Venezolana y la Secretaría General de la OEA.

CORRELACION Y REGRESION CON APLICACIONES EN LA HIDROLOGIA

Roger Amisial

Serie Hidrología
Material de Enseñanza
No. H-1

1976
MERIDA - VENEZUELA

CORRELACION Y REGRESION CON APLICACIONES EN LA HIDROLOGIA

La correlación y la regresión tienen muchas aplicaciones en la hidrología. En particular permiten la extensión de registros cortos, la estimación de datos faltantes, la regionalización de la información hidrológica y el diseño de redes hidrometeorológicas.

I. CORRELACION

A. Definición

Se define como la asociación entre dos o más variables aleatorias, que explica sólo parcialmente la variación total de una variable por la variación de otras variables aleatorias involucradas en la ecuación de asociación.

La parte de la variación total que queda sin explicar o sea, la variación no explicada, se debe a errores o a otras variables aleatorias que no han sido tomadas en cuenta en la correlación.

B. Medidas de Correlación

Se necesita un estadístico para medir el grado de asociación correlativa entre las variables bajo consideración. Los estadísticos más utilizados son los coeficientes de correlación y de determinación y la desviación típica de los residuos.

C. Análisis de Correlación

El análisis de correlación consiste en:

a) el cálculo de una medida del grado de correlación
b) la realización de pruebas para determinar si es aceptable el grado de asociación correlativa.

El análisis de correlación está estrechamente relacionado con el análisis de regresión puesto que la fórmula utilizada en el cálculo de la medida de correlación depende del modelo de regresión adoptado. Cuando se selecciona un modelo lineal simple, se habla de correlación lineal simple.

D. Correlación Lineal Simple

En lo que sigue se presentan los estadísticos más utilizados como medida de la asociación correlativa.

1. Coeficiente de Correlación

El coeficiente de correlación es el estadístico más comúnmente utilizado para medir el grado de asociación de dos variables linealmente relacionadas.

a. Fórmula

El coeficiente de correlación se define como

$$\rho(x,y) = \frac{\text{cov}(x,y)}{(\text{var } x \cdot \text{var } y)^{1/2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{[E(x-\mu_x)^2 \, E(y-\mu_y)^2]^{1/2}}$$

siendo ρ el coeficiente de correlación poblacional de las variables x e y. En base a muestras de tamaño n sacadas de las poblaciones de las variables se puede calcular r, que es una estimación de $\rho(x,y)$, mediante la fórmula:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{n s_x s_y} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{n s_x s_y}$$

Para valores de n pequeños se suele utilizar la estimación insesgada.

b. Valores y Variación de r

Cuando x e y son independientes r = 0, puesto que en este caso cov(x,y) = 0

Si r = 0, x e y no son correlacionadas linealmente, aunque pueden ser independientes.

Si r = 1 ó r = -1 hay dependencia lineal funcional entre las dos variables, o sea las series x_i e y_i son idénticas o difieren por un factor constante.

Valores de r entre -1 y 1 describen los varios grados de asociación.

2. Coeficiente de Determinación

El coeficiente de determinación, D, se define como el cuadrado del coeficiente de correlación, o sea

$$D = r^2$$

D es una medida del grado en que la varianza es explicada o tomada en cuenta por la regresión lineal. En otros términos una medida de la diferencia entre la varianza de los valores observados y la varianza de los valores calculados mediante la ecuación de regresión. Cuanto mayor es D, menor es la diferencia.

3. Desviación Típica de los Residuos

Los residuos de la regresión lineal de y versus x son $\Delta y_i = y_i - y$, donde y_i son los valores observados mientras que y es el valor calculado mediante la ecuación de regresión para un valor dado $x=x_i$.

La desviación típica de los residuos es idéntica a la desviación típica condicional de y dado x. Se calcula mediante

$$s_{y/x} = \sqrt{\frac{\Sigma(y_i-y)^2}{n}} = \sqrt{\frac{\Sigma[y_i-(a+bx_i)]^2}{n}}$$

También

$$s_{y/x} = s_y\,(1 - r^2)$$

Una estimación insesgada de la desviación típica de los residuos es

$$s_{y/x} = \sqrt{\frac{\Sigma[y_i-(a+bx_i)]^2}{n-2}} = \sqrt{\frac{n-1}{n-2}}\; s_y \sqrt{1-r^2}$$

Cuanto mayor es el valor de $s_{y/x}$, mayor es la dispersión de los puntos alrededor de la línea de regresión.

II. REGRESION

1. Definición

Representa una ecuación matemática expresando una variable aleatoria como siendo relacionada a otra o varias variables relacionadas. No todas las variables necesitan ser aleatorias.

2. Análisis de Regresión

(a) Definición

Se llama así la determinación de modelos de asociación correlativa de 2 o más variables, de tal manera que la mejor predicción de una variable puede ser obtenida a partir de la, o de las otras variables. Los modelos así desarrollados se llaman funciones de regresión o curvas de regresión o regresiones.
El grado de asociación correlativa depende de la función de regresión seleccionada.

(b) Pasos en el Análisis de Regresión

(i) Selección de una función de relación correlativa, simple o múltiple, lineal o no-lineal.

(ii) Estimación de Parámetros que miden el grado de asociación correlativa.

(iii) Prueba de la significación de los estadísticos que miden la asociación correlativa.

(iv) Estimación de los parámetros de la ecuación o función de regresión.

(v) Prueba de la significación de los parámetros de regresión, o determinar o dibujar los límites de confianza alrededor de las funciones de regresión ajustadas.

3. Determinación de los Parámetros

El método más comúnmente utilizado en la determinación de los parámetros de la regresión es el de los míni

mos cuadrados. El principio del método consiste en determinar los parámetros desconocidos con el criterio de que sea mínimo el valor medio del cuadrado de ciertas desviaciones. En el caso de la regresión, las desviaciones cuyos cuadrados deben ser minimizados son las diferencias entre los valores observados de la muestra y los valores calculados a partir de la ecuación de regresión.

4. Regresión Lineal

Si la función de relación correlativa o ecuación de regresión seleccionada es lineal entonces hablamos de un análisis de regresión lineal. Según la variable independiente es una variable no aleatoria o aleatoria se distingue el modelo lineal 1 y el modelo lineal 2.

a. Modelo lineal 1

En este caso la variable aleatoria Y es funcionalmente dependiente de una variable no aleatoria x cuyo valor puede cambiar de un ensayo al otro. El modelo 1 tiene la forma:

$$E(Y)_x = \alpha + \beta x$$

En cada ensayo x tendrá algún valor x_i y el valor esperado de Y_i será $\alpha + \beta x_i$, de modo que:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

donde ε_i es una variable aleatoria como media cero llamada el término del error. ε_i tiene la misma distribución que Y_i con la diferencia de que su media es cero.

A continuación se presentarán ejemplos de variables que se ajustan al modelo lineal 1. La precipitación es una variable aleatoria que, en zonas montañosas, puede depender funcionalmente de la altitud del punto considerado. La escorrentía o caudal en una sección de un río es una variable aleatoria que depende del área de la cuenca que drena en la sección considerada. La resistencia del suelo también es una variable aleatoria que es una

función de la profundidad a que se toma la muestra. En estos ejemplos la altitud, el área de la cuenca y la profundidad son variables no aleatorias.

b. Modelo Lineal 2

En esta clase de modelos tanto la variable dependiente como la independiente son variables aleatorias. El modelo 2 especifica que el valor esperado condicional de Y dado que X = x es una función lineal de x, o sea:

$$E(Y/X=x) = \alpha + \beta x$$

Para $X = x_i$ el valor esperado Y_i de Y es

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

c. Estimación de Parámetros del Modelo Lineal

Para la determinación de los parámetros obtenemos una muestra (x_i, y_i) de tamaño n y suponemos que la regresión poblacional

$$Y = \alpha + \beta X$$

puede ser estimada por la regresión muestral

$$y = a + bx$$

Para $x = x_i$ podemos calcular el valor correspondiente de y que llamaremos el valor estimado o calculado de y. Denotaremos este valor estimado por $\hat{y}_i$ para diferenciarlos del valor observado y_i. Entonces el valor medio del cuadrado de las desviaciones de los valores estimados y observados es

$$D = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

donde

$$\hat{y}_i = a + bx_i$$

Al combinar estas últimas dos ecuaciones se obtiene

$$D = \frac{1}{n} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Para que D sea un mínimo se debe tener:

$$\frac{\partial D}{\partial a} = \frac{1}{n} \sum_{i=1}^{n} - 2(y_i - a - bx_i) = 0$$

$$\frac{\partial D}{\partial b} = \frac{1}{n} \sum_{i=1}^{n} - 2x_i(y_i - a - bx_i) = 0$$

o sea

$$\Sigma y_i - na - b\Sigma x_i = 0$$

$$\Sigma x_i y_i - a\Sigma x_i - b\Sigma x_i^2 = 0$$

Estas expresiones se llaman ecuaciones normales y se pueden resolver para a y b, obteniendo:

$$a = \frac{\Sigma y_i \Sigma x_i^2 - \Sigma x_i \Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$b = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

Manipulando estas expresiones se puede demostrar que

$$b = r \frac{s_y}{s_x} = \frac{s_{x,y}}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$

Ejemplo

Las precipitaciones del mes de enero para dos estaciones X e Y se presentan en la siguiente tabla para siete años

x_i	12	18	24	30	36	42	48
y_i	5.27	5.68	6.25	7.21	8.02	8.71	8.42

1) Calcular los parámetros de la regresión lineal entre X e Y.

2) Hacer una prueba de significación de r utilizando el estadístico

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

3) Para un nivel de significación del 5% determinar los límites de confianza de Y dado que X = 36.

SOLUCION

1)

x_i	y_i	x_i^2	$x_i y_i$	y_i^2
12.00	5.27	144.00	63.24	27.77
18.00	5.68	324.00	102.24	32.26
24.00	6.25	576.00	150.00	39.06
30.00	7.21	900.00	261.30	51.98
36.00	8.02	1296.00	288.72	64.32
42.00	8.71	1764.00	365.82	75.86
48.00	8.42	2304.00	404.16	70.90

$\Sigma x_i = 210$ $\Sigma x_i^2 = 7308$ $\Sigma y_i = 361.15$

$\Sigma y_i = 49.56$ $\Sigma x_i y_i = 1590.48$

$$\bar{x} = \frac{1}{n}\Sigma x_i = \frac{210}{7} = 30$$

$$\bar{y} = \frac{1}{n}\Sigma y_i = \frac{49.56}{7} = 7.09$$

$$s_x = \sqrt{\frac{\Sigma x_i^2}{n} - (\frac{\Sigma x_i}{n})^2} = \sqrt{\frac{7308}{7} - (\frac{210}{7})^2} = \sqrt{144} = 12$$

$$s_y = \sqrt{\frac{\Sigma y_i^2}{n} - (\frac{\Sigma y_i}{n})^2} = \sqrt{\frac{361.15}{7} - (\frac{49.56}{7})^2} = \sqrt{1.47} = 1.21$$

$$s_{x,y} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\,\bar{y}$$

$$s_{x,y} = \frac{1}{7}(1.590.48) - (30)(7.09) = 14.51$$

$$a = \frac{\Sigma y_i \Sigma x_i^2 - \Sigma x_i \Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{(49.56)(7308) - (210)(1590.48)}{7(7308) - (210)^2}$$

$$a = 4.0$$

$$b = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\,\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{7(1590.48) - (210)(49.56)}{7(7308) - (210)^2}$$

$$b = 0.10$$

La ecuación de regresión es

$$y = 4 + 0.10x$$

Hubiéramos obtenido los mismos resultados utilizando las fórmulas

$$b = \frac{s_{x,y}}{s_x^2} = \frac{14.51}{(12)^2} = 0.10$$

$$a = \bar{y} - b\bar{x} = 7.09 - 0.10(30) = 4.07$$

2) $$r = \frac{s_{xy}}{s_x s_y} = \frac{14.51}{(12)(1.21)} = 0.9985$$

H_0: r no es diferente de cero

$$t = \frac{0.9985\sqrt{7-2}}{\sqrt{1-(0.9985)^2}} = 40.78$$

$$\nu = n - 2 = 5$$

$$t_{\frac{\alpha}{2},n-2} = t_{0.025,5} = 2.57$$

Si t está comprendida entre $-t_{0.025,5}$ y $t_{0.0255}$ se acepta H_0
De lo contrario se rechaza H_o.

Puesto que t = 40.78 no está comprendida entre -2.57 y 2.57 rechazamos la hipótesis nula de que r no es diferente de cero.

3)

$$s_{y/x} = s_y\sqrt{1-r^2} = 1.21\sqrt{1-(0.9985)^2} = 0.066$$

$$\bar{y} \pm b(x_i - \bar{x}) - t_{\frac{\alpha}{2},n-2}\frac{s_{y/x}}{\sqrt{n}}\sqrt{1+\frac{(x_i-\bar{x})^2}{s_x^2}} =$$

$$7.09 + 0.10(36 - 30) - 2.57 \times \frac{0.066}{\sqrt{7}}\sqrt{1+\frac{(36-30)^2}{(12)^2}}$$

$$7.69 - 0.0717 = 7.62$$

El otro límite es

$$7.69 + 0.0717 = 7.76$$

5. Prueba de Significación del Coeficiente de Correlación Lineal

Aún cuando el coeficiente de correlación lineal difiere de cero puede ocurrir que las dos variables Y y X no sean correlacionadas linealmente. El valor de r diferente de cero puede ser debido a errores de muestreo, aunque en la realidad $\rho = 0$

Por consiguiente, conviene probar la hipótesis H_0 de que r no sea diferente de cero de manera significativa. Para tal prueba se puede utilizar el estadístico

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

donde r es el coeficiente de correlación lineal calculado en base a n pares de valores de X e Y.

El estadístico t sigue la distribución t de Student con (n-2) grados de libertad.

Se escogerá a priori un nivel de significancia α para la prueba y se obtendrá los valores $t_{\frac{\alpha}{2}, n-2}$ y $t_{1-\frac{\alpha}{2},n-2}$. Se aplicará el criterio siguiente:

Si $t_{\frac{\alpha}{2},n-2} \leq t \leq t_{1-\frac{\alpha}{2},n-2}$ se aceptará la hipótesis H_0 de que r no es $\neq$ de cero.

Si $t < t_{\frac{\alpha}{2},n-2}$ ó $t \geq t_{1-\frac{\alpha}{2},n-2}$ se rechazará la hipótesis H_0.

Esta prueba se aplica sólo si se puede suponer que la distribución probabilística bivariada de las variables Y y X es normal. Se sabe que la mayoría de las variables hidrológicas no son distribuidas normalmente y siguen una distribución sesgada. Por eso que en hidrología es muy común transformar la variable para normalizarla, y poder así utilizar técnicas requiriendo una distribución normal. Las transformaciones más comúnmente empleadas son la logarítmica, la raíz cuadrada o cúbica.

6. Límites de Confianza sobre la Recta de Regresión

Sea la regresión lineal muestral Y_i vs. X_i definida por:

$$Y = a + bX$$

La regresión de las poblaciones es

$$\eta = \alpha + \beta\xi \qquad Y \text{ es la estima de } \eta \text{ para } \xi = X$$

El valor Y es distribuido normalmente alrededor de η con varianza

$$\hat{s}(Y) = \frac{\hat{s}_{Y.X}}{\sqrt{n}}\sqrt{1 + \frac{(X - \bar{X})^2}{\hat{s}_X^2}}$$

$\hat{s}_Y$ es el error típico de una estima Y.

La cantidad

$$t = \left(\frac{\bar{Y} - \eta}{\hat{s}(Y)}\right)^* \text{ (con (n-2) grados de libertad y) se distribuye según la distribución t de Student para n pequeño*.}$$

Los límites de confianza para un nivel de significancia α son:

$$\bar{Y} + b(X_i - \bar{X}) - t_{\frac{\alpha}{2},n-2}\frac{\hat{s}_{Y.X}}{\sqrt{n}}\sqrt{1 + \frac{(X_i - \bar{X})^2}{\hat{s}_X^2}} \leq \eta \leq \bar{Y} + b(X_i - \bar{X}) - t_{1-\frac{\alpha}{2},n-2}\frac{\hat{s}_{Y.X}}{\sqrt{n}}\sqrt{1 + \frac{(X_i - \bar{X})^2}{\hat{s}_X^2}}$$

Para cada valor X_i de X calculamos los límites de confianza para un nivel α. Ploteamos estos dos puntos. Luego juntamos los puntos del límite inferior también los del límite superior para determinar los límites.

III. TIPOS DE ASOCIACION CORRELATIVA EN HIDROLOGIA.

A. Relaciones Causa - Efecto

1. Donde la variable aleatoria, y, está relacionada correlativamente a factores causativa, X_i, que producen o afectan los valores de y.

2. Ejemplo: relación precipitación-escorrentía, porque la precipitación es el factor causativo básico de la escorrentía, con las características geométricas de la cuenca, el suelo, la humedad y los factores climáticos afectando la relación básica de causa-efecto.

B. Relaciones de Causa Común.

1. Relaciones de variables aleatorias que tienen los mismos factores causativos, tal como

   (a) la asociación correlativa del caudal de un río al caudal de un río adyacente.

   (b) la relación de las variables de la precipitación de estaciones pluviométricas adyacentes.

IV. NORMALIZACIÓN DE VARIABLES

Cuando la variable hidrológica tiene una distribución asimétrica, se puede tratar de definir una nueva variable $Z = f(X)$ que tenga una distribución normal y una varianza estabilizada, para poder utilizar una regresión lineal.

1. Transformación Logarítmica

Forma:

$$Z = \log (X - X_0) \qquad \text{con} \qquad X > X_0$$

Ventajas: Simple y muy utilizada
puede escoger $X_0=0$ si $X_i \neq 0$

Desventajas: Si $X_i=0$ se prefiere la transformación siguiente:

2. Transformación de Potencia Fraccionaria

Forma:

$$Z = X^q \qquad \text{con} \quad X \geq 0 \qquad y \quad 0 < q < 1$$

El valor de q puede ser determinado en función de los coeficientes de sesgo y de curtosis de Pearson de la variable X, utilizando la figura 2 (establecida por M. Rosenberg). Esta da sólo un valor preliminar si $n < 50$.

3. Anamorfosis

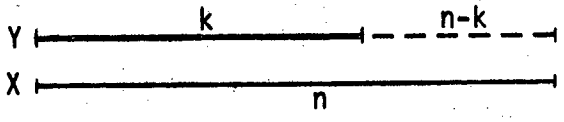
Consideramos la ley de Pearson tipo I F(X) y la ley normal N(Z) por ejemplo. A todo valor X podemos hacer corresponder un valor Z tal que las frecuencias acumuladas sean iguales. Esta operación se llama anamorfósis.

Las transformaciones indicadas en (1) y (2) son sólo unos casos particulares de anamorfosis en donde se puede expresar de manera explícita la relación entre Z y X.

V. APLICACION DE LA REGRESION LINEAL EN LA HIDROLOGIA

A. Extensión de Registros (o Datos Faltantes)

Consideramos dos registros de una variable hidrológica, uno X de longitud n unidades de tiempo (año, por ejemplo) y el otro Y de longitud $k < n$



Se supone que existe alguna correlación entre X e Y y nos interesamos en la media de Y. Supongamos que sean registros de escorrentía anual de dos cuencas vecinas o de dos estaciones sobre el mismo río. Sea $\bar{X}_k$, $\bar{Y}_k$, ${}_k s_x^2$, ${}_k s_y^2$, ${}_k r_{xy}$ los estadísticos calculados en base a los k años comunes. Sea $\bar{X}_n$, ${}_n s_x^2$ los estadísticos calculados en base a los n años disponibles para X.

La ecuación de regresión lineal obtenida por el método de los mínimos cuadrados puede tomar la forma siguiente:

$$(Y - \bar{Y}_k) = {}_k r_{xy} \frac{{}_k s_y}{{}_k s_x} (X - \bar{X}_k)$$

En esta relación $Y=Y_k$ corresponde a $X=X_k$. De modo que si dejamos $X=\bar{X}_n$ estimaremos $\bar{Y}_n$ o sea:

$$(\bar{Y}_n - \bar{Y}_k) = {}_k r_{xy} \frac{{}_k s_y}{{}_k s_x} (\bar{X}_n - \bar{X}_k)$$

$$\bar{Y}_n = \bar{Y}_k + {}_k r_{xy} \frac{{}_k s_y}{{}_k s_x} (\bar{X}_n - \bar{X}_k)$$

De una manera similar estimamos ${}_n s_y^2$

$$ {}_n s_y^2 = {}_k s_y^2 + {}_k r_{xy}^2 \frac{{}_k s_y^2}{{}_k s_x^2} ({}_n s_x^2 - {}_k s_x^2)$$

El valor de ${}_n r_{xy}$ se estima como:

$$ {}_n r_{xy} = {}_k r_{xy} \frac{{}_k s_y}{{}_k s_x} \frac{{}_n s_x}{{}_n s_y} $$

B. Longitud Efectiva de los Registros Extendidos por Correlación

1. Fórmula de Veron

Se calcula en base a consideraciones sobre ganancia de información sobre los parámetros.

El contenido de información de cualquier estadístico $\hat{\theta}$ estimado de un muestreo se define como el inverso de la varianza del estadístico

$$ I = \frac{1}{\text{var } \theta} = \frac{1}{\sigma^2 (\theta)} $$

Para la media $\bar{Y}$: $I = \frac{1}{\text{var } \bar{Y}}$

De modo que la eficiencia de las nuevas estimaciones $\bar{Y}_n$ y ${}_n s_y$, con respecto a las estimaciones $\bar{Y}_k$ y ${}_k s_y$, se medirá al comparar el contenido de información de cada grupo de estimación. La eficiencia relativa de $\bar{Y}_k$ y de $\bar{Y}_n$ se define por

$$ E = \frac{\text{var } \bar{Y}_n}{\text{var } \bar{Y}_k} $$

Veron (Roche: "Hydrologie de Surface") estimó E (calculando las varianzas) como:

$$ E = 1 + (1 - \frac{k}{n}) \left[\frac{1 - (k-2)r^2}{k - 3} \right] $$

Como $\text{var } \bar{Y}_k = \frac{1}{k} \text{var} Y$

y $\text{var } \bar{Y}_n = \frac{1}{n_e} \text{var } Y$

$$ \text{var } \bar{Y}_n = E \cdot \text{var } Y_k $$

$$ \therefore \quad \text{var } \bar{Y}_n = E \cdot \frac{\text{var } Y}{k} $$

Esta varianza corresponde a una $\text{var}\bar{Y}_{ne}$ calculada en base a una longitud efectiva n_e de registro o sea:

$$\frac{1}{n_e}\,\text{var}\,Y = \frac{E}{k}\,\text{var}\,Y$$

$$\therefore \quad n_e = \frac{k}{E}$$

Para decidir si aceptamos una extensión calculamos n_e. Si $n_e > k$ aceptamos la extensión . Si $n_e < k$ rechazamos la extensión.

Ejemplos

1) Para $r = 0.50$ $\quad k = 5$ años $\quad$ y $\quad n = 20$

$$E = 1 + \left(1 - \frac{5}{20}\right)\left[\frac{1 - (5-2)(0.5)^2}{5-3}\right] = 1.09$$

$$n_e = \frac{k}{E} = \frac{5}{1.09} = 4.57 \text{ años} < 5 \text{ No hay ganancia de información.}$$

2) Para $r = 0.75 \longrightarrow E = 0.74$

$$n_e = \frac{5}{0.74} = 6.74 \text{ años}$$

Ganancia: $G = n_e - k = 6.74 - 5 = 1.74$ años

3) Para $r=0.85$ $\quad E = 0.562$

$$n_e = \frac{5}{0.562} = 8.89 \text{ años} \qquad G = 8.89 - 5 = 3.89 \text{ años}$$

4) Para $r = 0.90$ $\quad E = 0.46$

$$n_e = \frac{5}{0.46} = 10.78 \text{ años} \qquad G = 10.78 - 5 = 5.78 \text{ años}$$

2. Fórmula de Langbein

Langbein desarrolló la siguiente fórmula para n_e:

$$n_e = \frac{n}{1 + \frac{n-k}{k-2}(1-r^2)}$$

Ejemplo:

Supongamos un coeficiente de correlación de 0.50 obtenido en base a un período de registro k=5años. Se desea extender el registro de 5 años a 20 años por correlación con la estación base con registro largo. Entonces:

$$n_e = \frac{n}{1 + \frac{n-k}{k-2}(1-r^2)} = \frac{5+15}{1 + \frac{15}{5-2}(1-0.25)}$$

$$n_e = \frac{20}{1 + 5(0.75)} = \frac{20}{1 + 3.75} = \frac{20}{4.75} = 4.2$$

Como $n_e = 4.2 < k = 5$ no hay ganancia de información.

Para	$r = 0.60$	$n_e = 4.75 > k = 5$	$G < 0$ no ganancia
Para	$r = 0.70$	$n_e = 5.63 > k = 5$	$G = 0.63$
Para	$r = 0.75$	$n_e = 6.25 > k = 5$	$G = 1.25$ años
Para	$r = 0.85$	$n_e = 8.3 > k = 5$	$G = 3.3$ años
Para	$r = 0.90$	$n_e = 10.26 > k = 5$	$G = 5.26$ años

Ciertos autores recomiendan un $r > 0.6$ para una correlación aceptable.

C. Aplicación de la Correlación entre Estaciones al Diseño de Redes Hidrométricas.

1. Fórmula de Langbein

La existencia de un coeficiente de correlación aceptable entre los registros de dos estaciones, indica que para conocer la hidrología de una región, no se necesita un número muy

grande de estaciones. Sin embargo, hay que determinar en cada estación de un río la función de regresión existente con las estaciones vecinas. De modo que podemos adoptar el esquema siguiente para la red:

- un número B de estaciones principales o de primer orden o base, observadas sin interrupción durante un período indefinido: constituyen la red base.
- un número S de estaciones secundarias o de segundo orden, observadas durante algunos años y luego desplazadas hacia otro sitio.

Cada estación secundaria es estudiada en correlación con una estación base. El número k de años de observación necesarios para alcanzar una precisión dada, depende de la correlación r entre las 2 estaciones y del número de años de registro en la estación base, según Langbein, puesto que

$$n_e = \frac{n}{1 + \frac{n-k}{k-2}(1-r^2)}$$

El coeficiente de correlación varía en principio de manera inversa con la distancia entre los centros de gravedad de las cuencas de drenaje de cada estación. Pero varía también con la superficie de las cuencas consideradas. Langbein sugiere que se puede aproximar $(1 + r^2)$ por

$$\frac{0.4\ c^2\ A}{B}$$

donde

c = incremento de desviación típica de los valores calculados a partir de la correlación, por km de distancia entre los centros de gravedad de las cuencas.

A ≡ Superficie total de la región

B ≡ número de estaciones base

La solución del problema consiste en maximizar la ganancia G de información donde

$$G = n_e - k$$

Se supone que el recursos financiero es limitado y que se dispone sólo de un presupuesto que permite operar sólo n estaciones cada año, durante un período dado.
Para que G sea un máximo, sus derivadas con respecto a las variables B y k deben ser iguales a cero. Se tiene

$$G = n_e - k$$

$$G = \frac{n\,B(k-2)}{(k-2)\,B + 0.4c^2\,A(n-k)} - k$$

$$\frac{\partial G}{\partial B} = \frac{n(k-2)[(k-2)B + 0.4c^2\,A(n-k)] - n\,B(k-2)^2}{[(k-2)B + 0.4c^2\,A(n-k)]^2} = 0$$

$\therefore$

$$0.4c^2\,A(n-k)\,n(k-2) = 0 \quad \text{para} \begin{cases} k = 2 & \therefore \quad G = -2 \text{ mínimo} \\ o & \\ k = n & \therefore \quad G = 0 \text{ mínimo} \end{cases}$$

no importa B.

$$\frac{\partial G}{\partial k} = \frac{nB[(k-2)B + 0.4c^2A(n-k)] - nB(k-2)(B - 0.4c^2A)}{[(k-2)B + 0.4c^2A(n-k)]^2} - 1 = 0$$

$\therefore$

$$0.4c^2AnB(n-k) + 0.4c^2AnB(k-2) - [(k-2)B + 0.4c^2A(n-k)]^2 = 0$$

$$0.4c^2AnB(n-2) = [(k-2)B + 0.4c^2A(n-k)]^2$$

$$\sqrt{0.4c^2ABn(n-2)} = Bk - 2B + 0.4c^2An - 0.4c^2Ak$$

$$k = \frac{2B - 0.4c^2An + \sqrt{0.4c^2ABn(n-2)}}{B - 0.4c^2A}$$

En vista de que el método analítico, tomando la derivada $\frac{\partial G}{\partial B}$, no permite llegar a un máximo para G, utilizaremos un método por tanteo.

Con un valor de B escogido se calcula k de la ecuación (2). Reemplazando k y B por sus valores en (1) obtenemos el valor correspondiente de G. Se repite el mismo procedimiento y se escoge el conjunto de valores B y k que dan el mayor valor de G y permiten medir el mayor número de ríos.

Ejemplo

Se dispone de fondos para la operación de sólo 20 estaciones hidrométricas al año en una región de 50000 km^2 de superficie. Diseñar una red de estaciones hidrométricas para una vida útil de 30 años, en base al enfoque desarrollado por Langbein.

Solución

La solución debe satisfacer la condición de que el número máximo de ríos deben ser medidos durante el horizonte de 30 años. A tal fin se instalarán B estaciones base o permanentes y S estaciones secundarias, debiendo éstas últimas ser removidas e instaladas en otro sitio al cabo de k años. Luego los registros de k años de longitud serán extendidos a 30 años por correlación con las estaciones base.

El objetivo es determinar los valores de B, S y k que maximizan a la vez la ganancia de información y el número de ríos o puntos aforados teniendo en cuenta las limitaciones de fondos, el grado de correlación de las estaciones de la región y la vida útil del proyecto.

Los datos del problema son

$$A = 50000 \text{ km}^2$$
$$n = 30 \text{ años}$$
$$B + S = 20 \text{ estaciones}$$

Para c se adoptará el valor de 0.01. El parámetro c puede ser

determinado para una región dada mediante un análisis de correlación de las estaciones existentes en la región. En caso de insuficiencia de estaciones, se puede utilizar el valor de c obtenido para otra región similar a la zona bajo estudio.

Las fórmulas a emplear son:

$$S = 20 - B$$

$$k = \frac{2B - 0.4c^2An + \sqrt{0.4c^2An(n-2)B}}{B - 0.4c^2A}$$

$$G = \frac{nB(k-2)}{(k-2)B + 0.4c^2A(n-k)} - k$$

El número de ríos o puntos aforados durante la vida útil del proyecto es:

$$NR = B + \frac{n}{k} S$$

Reemplazando los términos conocidos por sus respectivos valores, se obtiene:

$$k = \frac{2B + 40.^{9878}\sqrt{B} - 60}{B - 2}$$

$$G = \frac{30\ B(k-2)}{B(k-2) + 2(30 - k)} - k$$

$$NR = B + \frac{30}{k} S$$

Para valores de B inferiores o iguales a 20 se calculan los valores de k que maximizan G y también los valores correspondientes de NR. Por ejemplo: para B = 3 se tiene S = 20 - 3 = 17 y

$$k = \frac{2(3) + 40.^{9878}\sqrt{3} - 60}{3 - 2} = 16.^{99} \text{ años}$$

$$G = \frac{30(3)(16.^{99} - 2)}{3(16.^{99} - 2) + 2(30 - 16.^{99})} - 16.^{99} = 19.01 - 16.^{99} = 2.02 \text{ años}$$

$$S = 20 - B = 20 - 3 = 17 \text{ estaciones}$$

$$NR = 3 + \frac{30}{16.^{99}} 17 = 33.^{02} \text{ ríos aforados}$$

En la siguiente tabla se presentan los resultados de tales cálculos:

Nº de estaciones base B (1)	Nº de estaciones secundarias S (2)	Período de operación de las S k (3)	Ganancia de información G (4)	Nº de ríos aforados
3	17	$16.^{99}$	$2.^{02}$	$33.^{02}$
4	16	$14.^{99}$	$4.^{02}$	$36.^{03}$
5	15	$13.^{88}$	$5.^{57}$	$37.^{44}$
6	14	$13.^{10}$	$6.^{80}$	$38.^{06}$
7 *	13	$12.^{49}$	$7.^{82}$	$38.^{23}$
8	12	$11.^{99}$	$8.^{69}$	$38.^{03}$
9	11	$11.^{57}$	$9.^{44}$	$37.^{58}$
10	10	$11.^{20}$	$10.^{10}$	$36.^{78}$
11	9	$10.^{88}$	$10.^{68}$	$35.^{81}$
12	8	$10.^{60}$	$11.^{64}$	$34.^{65}$
13	7	$10.^{34}$	$11.^{68}$	$33.^{30}$
14	6	$10.^{11}$	$12.^{11}$	$31.^{80}$
15	5	$9.^{90}$	$12.^{50}$	$30.^{15}$
16	4	$9.^{71}$	$12.^{86}$	$28.^{36}$
17	3	$9.^{53}$	$13.^{20}$	$26.^{44}$
18	2	$9.^{37}$	$13.^{51}$	$24.^{40}$
19	1	$9.^{22}$	$13.^{80}$	$22.^{26}$
20	0	0	0	$20.^{00}$

* El número de ríos aforados es máximo para B = 7, S = 13 y k = 12.49. En otros términos, el diseño óptimo de la red constará de 7 estaciones base y 13 estaciones secundarias operando éstas últimas durante 12 años y medio en cada sitio.

$NE = B + S$ número de estaciones a operar por año

$$NR = B + \frac{n}{k}(NE - B)$$

$$NR = B + \frac{n(NE-B)(B - 0.4c^2A)}{2B - 0.4c^2An + \sqrt{0.4c^2An(n-2)B}}$$

$$\frac{\partial NR}{\partial B} = 1 + \frac{n[(NE-B)-(B-0.4c^2A)]\,\text{denom.} - [2+\sqrt{0.4c^2An(n-2)}B^{-1/2}]\,\text{numer.}}{(\text{denom.})^2}$$

$$\frac{\partial NR}{\partial B} = 1 + \frac{n(NE+0.4c^2A-2B)\ \text{denomin.} - [2+\sqrt{0.4c^2An(n-2)}B^{1/2}]\ \text{numer.}}{(\text{denom.})^2}$$

2. Fórmula de Veron

La longitud efectiva n_e de un registro extendido por correlación es

$$n_e = \frac{k}{E}$$

donde k es el período corto o longitud de los datos observados y E la eficiencia relativa de $\bar{Y}_k$ con respecto a $\bar{Y}_n$

$$E = 1 + (1 - \frac{k}{n})\left[\frac{1 - (k-2)r^2}{k - 3}\right]$$

La ganancia de información es

$$G = n_e - k = \frac{k}{E} - k = k\left\{\frac{1}{1+(1-\frac{k}{n})\left[\frac{1-(k-2)r^2}{k-3}\right]} - 1\right\}$$

$$G = \frac{k(k - 3)}{k - 3 + (1-\frac{k}{n})\left[1 - (k-2)\ r^2\right]} - k$$

Si utilizamos la estimación de Langbein para $(1 - r^2)$:

$$1 - r^2 = \frac{0.4\ c^2A}{B}$$

$$r^2 = 1 - \frac{0.4c^2A}{B}$$

Entonces

$$G = \frac{k(k-3)}{k-3 + (1-\frac{k}{n})\left[1 - (k-2)(1 - \frac{0.4c^2A}{B})\right]} - k$$

$$\frac{\partial G}{\partial k} = \frac{(2k-3)[\text{denominador}] - \{1-(-\frac{1}{n})[1-(k-2)(1-\frac{0.4c^2A}{B})] + (1-\frac{0.4c^2A}{B})(1-\frac{k}{n})\}}{[\text{denominador}]^2} - 1 = 0$$

$$\frac{\partial G}{\partial B} = \frac{+ k(k-3)(1-\frac{k}{n})(\frac{0.4c^2A}{B^2})(k-2)}{(\text{denominador})^2} = 0$$

VI. APLICACIONES DE LA CORRELACION Y REGRESION EN HIDROLOGIA.

Una de las aplicaciones más comunes de la correlación y regresión en la hidrología, es la extensión de registros cortos de precipitación, escorrentía y otras variables hidrometeorológicas e hidrológicas. En este curso nos limitaremos a la extensión de registros de caudales. Entonces discutiremos los puntos siguientes:

(1) Extensión por correlación con escorrentía de estación adyacente de registro más largo.

(2) Extensión por correlación con registros de precipitación de estación ubicada en la cuenca de drenaje.

(3) Extensión por correlación con registros más largos de escorrentía y precipitación.

(4) Extensión de registros de picos de crecida por correlación con precipitación.

A. Extensión de Registros de Caudales por Correlación con Registros de Precipitación.

i) Justificación

1. Existe una relación de causa-efecto entre precipitación y

escorrentía.

2. Registros de precipitación generalmente disponibles y más largo, proveyendo así una base para extensión.

3. Este tema es una parte del tema general de la Relación Precipitación - Escorrentía.

ii) Datos

1. Sería ideal utilizar datos de precipitación para la cuenca de drenaje de la estación hidrométrica, calculados a partir del método de Thiessen. Sin embargo es más sencillo y práctico utilizar los datos de una sola estación pluviométrica.

2. Se utiliza también, en la correlación, la precipitación previa al período considerado como variable independiente. Y se toma en cuenta tanto la precipitación del período considerado como la precipitación previa.

B. Correlación de Caudales Anuales con Precipitación Anual.

1. Precipitación Efectiva P_e

a) Uso

En caso de datos anuales se correlaciona el caudal con la precipitación efectiva sola. Para prueba de consistencia también se hace uso de la precipitación efectiva.

b) Definición

Es la proporción de la precipitación del año n considerado, P_n, y la proporción de la precipitación de los años previos, que contribuyen a la escorrentía del año considerado.

En general para el año i

$$P_{e_i} = aP_i + bP_{i-1} + cP_{i-2} + \ldots .$$

c) Ajuste

1º La determinación de a, b, c, ... podría hacerse mediante el método de los mínimos cuadrados, sin embargo se pierden muchos grados de libertad para cada parámetro a, b, c, ... que se estima. Se podría reducir la pérdida de grados de libertad al considerar que a, b, c, ... varían uniformemente.

2º El ajuste se hace por el uso de la correlación de orden, empleando el procedimiento siguiente (tabla 3):

(i) Se tabula los datos de caudales y se les asigna un número de orden: al valor máximo se le asigna 1, etc. ... Si dos valores son iguales se les asigna el promedio de los números de orden que les hubiera sido asignados si fuesen diferentes el uno del otro (ej. $\frac{4+5}{2} = 4.5$).

(ii) Se hace igual para la precipitación correspondiente.

(iii) Se calcula la diferencia entre los números de orden de la precipitación y los del caudal correspondiente, y se eleva cada diferencia de orden al cuadrado.

(iv) Se calcula la suma de los cuadrados de las diferencias.

(v) Se asume una serie de valores de a, b, c, ... y se calcula la precipitación efectiva. Por ejemplo:

$$a = 0.8$$
$$b = 0.2$$
$$c = 0$$
$$\therefore \quad P_{ei} = 0.8P_i + 0.2P_{i-1}$$

(vi) Se asigna números de orden a los valores P_{ei} y se calcula la suma de los cuadrados de las diferencias

de orden de los caudales y de la precipitación efectiva.

(vii) Se repite el procedimiento hasta la suma de los cuadrados de las diferencias de número de orden sea mínima.

(viii) Se adoptan los valores de a, b, c, ... que dan la suma mínima.

(ix) Se desarrolla la correlación entre la P_e y el caudal analíticamente o gráficamente.

C. Correlación de Caudal Mensual con Precipitación Mensual.

A. Variables Independientes:

1. Precipitación del mes considerado

2. Indice de precipitación previa:

Para el mes número i:

$$IPP_i = b_1 P_{i-1} + b_2 P_{i-2} + b_3 P_{i-3} + \dots + b_n P_{i-n}$$

$$IPP_i = \sum_{j=1}^{n} b_j P_{i-j}$$

Se hace $b_j = k^j$ donde $k < 1$

$\therefore$ $IPP_i = k P_{i-1} + k^2 P_{i-2} + k^3 P_{i-3} + \dots$

Para datos mensuales se toma $k = 1/2$

$\therefore$ $IPP_i = \frac{1}{2} P_{i-1} + \frac{1}{4} P_{i-2} + \frac{1}{8} P_{i-3} + \frac{1}{16} P_{i-4}$

3. Mes del calendario.

B. Correlación Coaxial

1. Primera Aproximación

a. Graficar X_1 (precipitación previa) contra Y caudal observado e identificar cada punto por el número correspondiente al mes (Figura 1, gráfico de la izquierda).

b. A la derecha de la Figura 1, plotear el caudal observado Y (abcisa) contra el caudal calculado obtenido del gráfico de la izquierda, entrando con los valores correspondientes de- X_1 y X_2.

Identificar cada punto ploteado por el valor correspondiente de la precipitación (X_3).

2. Segunda aproximación

En la segunda aproximación se parte de la gráfica de derecha y - de X_1, para dibujar de nuevo los puntos de la gráfica de izquierda y así se dibujan de nuevo las curvas izquierdas marcadas de - X_2. Luego, a partir de estas nuevas curvas, se dibujan de nuevo las curvas de derecha.

3. Se puede hacer una tercera aproximación, pero dos aproximaciones son generalmente suficientes.

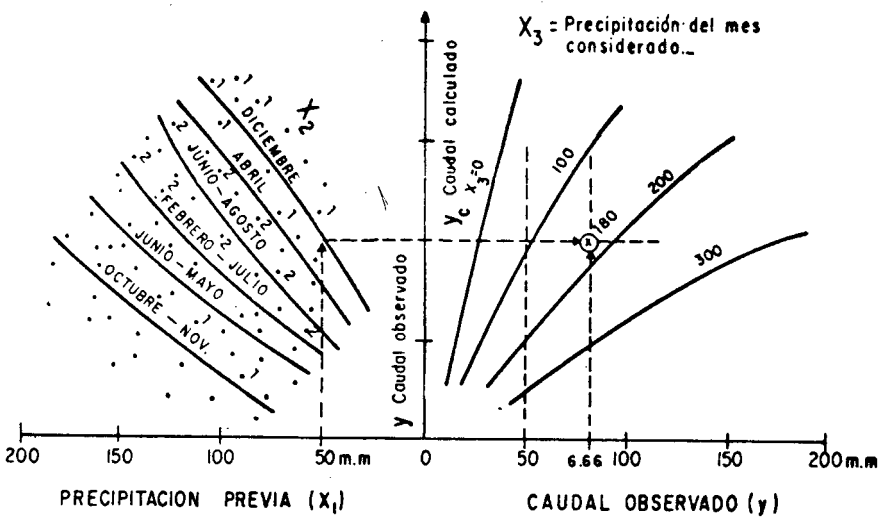


Figura. 1.__ Correlación coaxial, primera aproximación.

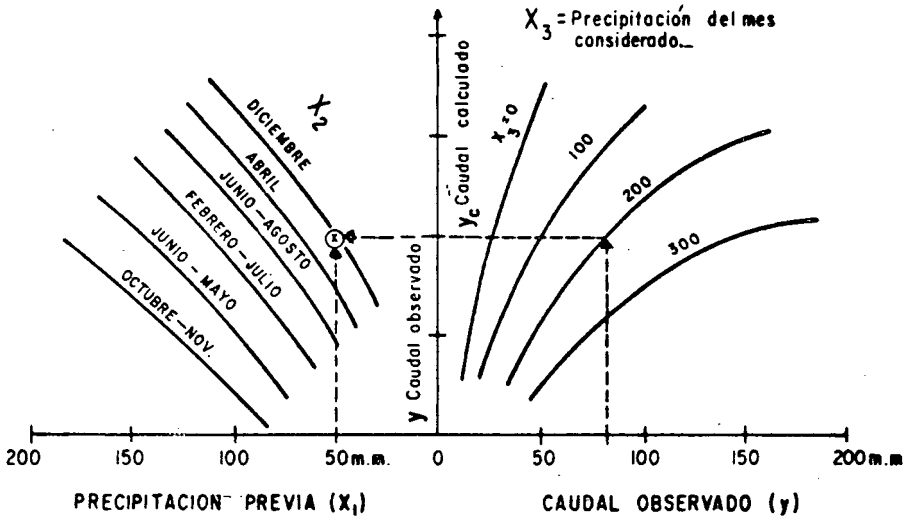


Figura. 2.

Este trabajo se imprimió en el Taller de Reproducciones del CIDIAT en el mes de septiembre de 1981. Se reprodujeron 200 ejemplares.