Un modelo para corregir la predicción canónica

Pedro Cervantes-Hernández *

El análisis multivariado es una rama de la estadística que se encarga del estudio simultáneo de diversos tipos de variables (métricas o numéricas y no métricas) (Uriel 1999). Los modelos que incluye esta rama son varios y se clasifican en tres grupos: a) modelos de predicción (*e.g.* Regresión Lineal Múltiple, Correlación Canónica), b) modelos de ordenación y/o clasificación (*e.g.* Análisis de Componentes Principales, Cluster, Análisis de Correspondencia) y c) la fusión de los modelos a y b (*e.g.* Análisis de Discriminante, Escalamiento Multidimensional) (Hair *et al*. 1999).

De los modelos antes mencionados, el que interesa para el desarrollo de este trabajo es el modelo de Correlación Canónica (CC) de Morrison (1967). La CC es un modelo multivariado que se utiliza para predecir simultáneamente dos o más variables dependientes, analizando el efecto que tienen sobre de éstas dos o más variables independientes. El modelo de CC es el siguiente:

$$(1) \qquad Y_1 + \ldots + Y_n = X_1 + X_2 + \ldots + X_n$$

Donde Y_n son el grupo de las variables dependientes y X_n son le grupo de las variables independientes.

Para que el grupo de las variables dependientes pueda ser predecible simultáneamente con base en un mismo grupo de variables independientes, se necesita un nivel de correlación alto y significativo en el grupo de las variables Y_n (Calvo-Gómez 1993, Dallas 2002, Hair *et al*. 1999, Sharon 1999).

Como se observa en la función (1), el modelo de CC está compuesto por n ecuaciones lineales que se consideran independientes u ortogonales (Dallas 2002), esto es, cada ecuación lineal predice una única variable Y_n; sin embargo, al ejecutar las n ecuaciones lineales simultáneamente, el modelo adquiere la cualidad canónica. Para validar la predicción simultánea del grupo Y_n y corroborar el efecto de las variables independientes sobre cada una de éstas, se utilizan los siguientes sellos de garantía estadística:

a) Correlación Canónica (R_{cc}): Nivel de asociación o de correlación entre los grupos de variables dependientes e independientes.
b) Determinación Canónica (R^2_{cc})Porcentaje de varianza explicada del modelo de CC que se asigna a la predicción simultanea del grupo Y_n.
c) Índice de Redundancia (Ir): Nivel de asociación o de correlación entre el grupo de variables dependientes Y_n.

Las variables que se consideran como dependientes en el área de la investigación ecológica y biológica (*e.g.* la abundancia e índices de diversidad), generalmente están alta y significativamente correlacionadas, por lo que la aplicación de un modelo de CC en esta área de la investigación, no representa ningún problema, ya que se cumplen los supuestos antes señalados. Sin embargo, de acuerdo con la práctica, he observado que cuanto mayor es

Universidad del Mar, Instituto de Recursos, Ciudad Universitaria, campus Puerto Ángel, Apdo. Postal 47, Puerto Ángel, Oaxaca, 70902, México. Correo electrónico: pch@angel.umar.mx

la redundancia entre el grupo de las variables Y_n, la predicción de cada una de éstas resulta no confiable, aceptándose en todos los casos la hipótesis alternativa Ha: valor esperado ≠ valor observado.

A través de diversos ensayos en el área de la investigación ecológica y biológica, he corroborado que al aumentar la redundancia en un modelo de CC, por ejemplo, en uno de dos variables dependientes Y_1 y Y_2, las ecuaciones lineales dejan de ser ortogonales y se tornan dependientes una de la otra, ya que al aumentar la redundancia, se requiere de Y_1 para predecir a Y_2 y viceversa, esto es:

$$Y_1 = X_1 + X_2 + \ldots + X_n + Y_2$$

(2)

$$Y_2 = X_1 + X_2 + \ldots + X_n + Y_1$$

En los sistemas ecológico-biológicos, ninguna variable poblacional o ambiental funciona aisladamente, todas están correlacionadas entre sí, dicha correlación entre las diferentes variables, se representa en el modelo de CC (2), al incorporar en cada ecuación lineal el efecto de las demás variables dependientes.

El objetivo de este trabajo es dar a conocer un modelo que permita mejorar la predicción de las variables dependientes en los modelos de CC, cuando en éste, se registre una alta redundancia.

Modelo a desarrollar

Sea f una variable independiente seleccionada para predecir a Y_1 y Y_2 en 3, f se distingue del resto de las variables independientes, ya que la correlación de ésta con respecto al grupo de las variables independientes es baja y su variabilidad anual está condicionada únicamente por el efecto de Y_1 y Y_2. El proceso para corregir la modelación canónica comienza con la selección de una variable independiente clave (en este caso es f), para que a partir de ésta se desarrolle el proceso en cuestión. Esto es:

Sea f la variable seleccionada en el modelo (3):

$$Y_1 = c_{11}f + c_{12}X_{12} + c_{13}X_{13}$$

(3)

$$Y_2 = c_{21}f + c_{22}X_{22} + c_{23}X_{23}$$

Al despejar f en cada una de las ecuaciones lineales en (3), tenemos que:

$$f \quad \frac{Y_1 \quad c_{12}X_{12} \quad c_{13}X_{13}}{c_{11}}$$

(4)

$$f \quad \frac{Y_2 \quad c_{22}X_{22} \quad c_{23}X_{23}}{c_{21}}$$

Las ecuaciones lineales se igualan y se realiza la simplificación algebraica, de manera que:

(5) $$c_{21}(Y_1 - c_{12}X_{12} - c_{13}X_{13}) = c_{11}(Y_2 - c_{22}X_{22} - c_{23}X_{23})$$

El proceso de corrección continúa despejando las variables Y_1 y Y_2 en (5), esto es:

$$Y_1 \quad \frac{c_{11}Y_2 \quad c_{11}c_{22}X_{22} \quad c_{11}c_{23}X_{23} \quad c_{21}c_{12}X_{12} \quad c_{21}c_{13}X_{13}}{c_{21}}$$

(6)

$$Y_2 \quad \frac{c_{21}Y_1 \quad c_{21}c_{12}X_{12} \quad c_{21}c_{13}X_{13} \quad c_{11}c_{22}X_{22} \quad c_{11}c_{23}X_{23}}{c_{11}}$$

Y_1 y Y_2 en (6), se sustituyen en (4) para generar dos nuevos coeficientes (f: Y_1) y (f: Y_2). Estos nuevos coeficientes se sustituyen por f en (3), para obtener el modelo de CC corregido, el cual es:

$$\hat{Y}_1 \quad c_{11}(\ : Y_2) \quad c_{12}X_{12} \quad c_{13}X_{13}$$

(7)

$$\hat{Y}_2 \quad c_{21}(\ : Y_1) \quad c_{22}X_{22} \quad c_{23}X_{23}$$

Donde:

Y_1 y Y_2 son el grupo esperado de las variables dependientes considerando la corrección.

f es la variable independiente seleccionada para realizar la corrección.

$(f: Y_1)$ es el coeficiente que designa el efecto de Y_2 sobre Y_1 en términos de f.

$(f: Y_2)$ es el coeficiente que designa el efecto de Y_1 sobre Y_2 en términos de f.

X_{ij} es la j-enésima variable independiente contenida en la función canónica i

c_{ij} es la j-enésima carga canónica correspondiente a la variable X_{jj} contenida en la función canónica i.

La selección de la variable independiente con la cual se realiza todo el proceso de corrección del modelo CC, esta en función de los criterios ecológico–biológicos que cada investigador consideré según los objetivos y planteamientos de sus investigaciones.

Estudio de caso

Se utilizó un modelo de CC para predecir retrospectivamente, la abundancia poblacional del camarón café *Farfantepenaeus aztecus* (Ives, 1891) de junio 1982 a septiembre 1983, en la región Tamaulipas-norte de Veracruz. El modelo considera la predicción de dos índices poblacionales denominados: abundancia de reclutas (FR) y abundancia de reproductores (DA). Conjuntamente, se analiza el efecto del el esfuerzo de pesca (f_i), la precipitación pluvial (Pp), la temperatura superficial del mar (TSM) y clorofila *a* más feofitínas (PIG) sobre los valores esperados de ambos índices poblacionales.

Después de correr el modelo de CC con base en un programa de cómputo especializado, los resultados que se obtienen se muestran en la Tabla I. De acuerdo con ésta última, el modelo de CC generó dos ecuaciones lineales, éstas se

Tabla I. Resultados del modelo de CC para el intervalo de junio de 1982 a septiembre de 1983.

Variables dependientes	Ecuación lineal 1	Ecuación lineal 2
DA	0.9879	0.1545
FR	0.5538	0.8326
Variables independientes		
i	-0.5143	0.6404
Pp	-0.3808	0.5799
TSM	-0.6410	0.6545
PIG	0.6082	-0.4008

identificaron con base en el valor de carga canónica que resultó para cada índice poblacional. DA generó una carga canónica significativa de 0.9879 (ecuación lineal 1) y FR de 0.8326 (ecuación lineal 2) (R^2_{cc}= 0.5668 y p< 0.05).

El efecto o carga canónica de cada una de las variables independientes sobre los índices FR y DA se interpreta por columna en la Tabla I. De esta manera, en DA fue afectado de manera proporcional en un 60.82% por el PIG e inversamente proporcional por fi (51.43%), Pp (38.08%) y TSM (64.10%). FR fue afectado de manera inversamente proporcional en un 40.08% por el PIG y directamente proporcional por fi (64.04%), Pp (57.99%) y TSM (65.45%).

Con base en los resultados descritos en la Tabla I, las ecuaciones lineales del modelo de CC son las siguientes:

$$FR \quad 0.6404\ i \quad 0.5799Pp \quad 0.6545TSM \quad 0.4008PIG \tag{8}$$

$$DA \quad 0.5143\ i \quad 0.3808Pp \quad 0.6410TSM \quad 0.6082PIG$$

Donde:

FR es el índice esperado asociado a la fuerza del reclutamiento (10^6 organismos).

DA es el índice esperado asociado a la densidad de adultos reproductores (10^6 organismos).

fi es el índice observado asociado al esfuerzo de pesca (número de viajes).

Pp es el índice observado asociado a la precipitación pluvial (mm).

TSM es el índice observado asociado a la temperatura superficial del mar (°C).

PIG es el índice observado asociado a la concentración de Clorofila *a* más feofitínas (mg m^{-3}).

0.6404, 0.5799, 0.6545 y –0.4008 son las respectivas cargas canónicas asociadas a los índices fi, Pp, TSM y PIG en la ecuación lineal 1

-0.5143, -0.3808, -0.6410 y 0.6082 son las respectivas cargas canónicas asociadas a los índices fi, Pp, TSM y PIG en la ecuación lineal 2

Al aplicar las ecuaciones lineales del modelo de CC (8), la predicción o valor esperado (esp) que se obtiene para los índices FR y DA, en comparación con los valores observados (obs) se resume en la Tabla II.

Al confrontar estadísticamente estos resultados con base en la distribución de ji-cuadrada (x^2_t), se tiene que en el caso del índice *FR*, se acepta la hipótesis alternativa (x^2_t =179.19, x^2_α =29.99, gl= 15 y p=0.03595), mientras que en el caso del índice *DA*, la hipótesis que se acepta es la nula (x^2_t =10.15, x^2_α = 24.99, gl = 15 y p = 1.2069).

Con base en lo anterior, se sugiere que el modelo de CC descrito en (8) resultó inviable para predecir simultáneamente a los índices FR y DA. La redundancia obtenida entre ambos índices, resultó de 0.7866. Al respecto, Gracia (1991) y Cervantes-Hernández (1999) indicaron que los índices *FR* y *DA* se encuentran altamente correlacionados a través del ciclo de vida de los camarones de la familia *Peneidae*; esto es, los adultos reproductores mediante el proceso de reproducción generan las nuevas cohortes anuales de reclutamiento, que sustituyen a las cohortes más viejas en la población natural. De manera que para predecir la abundancia *FR*,

Tabla II. Predicción retrospectiva de los índices FR y DA en el intervalo de junio de 1982 a septiembre de 1983 (10^6 organismos).

Mes	FR obs	FR esp	DA obs	DA esp
junio	13.10	18.11	3.66	6.35
julio	9.79	10.74	3.59	5.50
agosto	6.89	3.66	2.96	4.87
septiembre	6.06	33.56	2.17	4.54
octubre	6.30	31.42	1.79	3.99
noviembre	6.17	11.58	1.91	4.38
diciembre	5.41	3.16	1.99	4.18
enero	4.56	3.11	1.44	2.80
febrero	4.37	20.93	1.47	2.28
marzo	4.89	7.06	1.46	2.45
abril	7.29	2.33	1.36	2.17
mayo	13.13	42.68	1.68	2.47
junio	19.94	22.53	2.71	3.11
julio	21.08	11.59	4.54	3.99
agosto	15.85	37.84	5.90	4.43
septiembre	7.35	75.83	3.10	3.61

es necesario integrar en la predicción la abundancia los adultos reproductores *DA*. Lo anterior no fue considerado al ejecutar el modelo de CC (8), por lo que la predicción de los índices *FR* y *DA*, se realizó únicamente en términos de f_i, *Pp*, *TSM* y *PIG*.

Debido a la alta redundancia detectada entre los índices *FR* y *DA*, la función canónica (8) fue corregida según el modelo antes descrito. Para ello, se consideró el siguiente criterio para seleccionar a la variable independiente clave: la variabilidad anual de la abundancia de los índices *FR* y *DA*, condicionan hasta cierto punto la variabilidad anual del esfuerzo de pesca en el año N_t y N_{t+1}, pero no así a las demás variables ambientales; esto es, los cambios en la abundancia *FR* y *DA* no condicionan la variabilidad espacial de *Pp*, *TSM* y *PIG*. Con base en lo anterior, f_i se seleccionó para desarrollar el proceso de corrección del modelo de CC (8). El proceso es el siguiente:

Al despejar f_i en cada una de las ecuaciones lineales en (8), tenemos que:

$$_i \quad \frac{FR \quad 0.5799Pp \quad 0.6545TSM \quad 0.4008PIG}{0.6404} \qquad (i:FR) \tag{9}$$

$$_i \quad \frac{DA \quad 0.3808Pp \quad 0.6410TSM \quad 0.6082PIG}{0.5143} \qquad (i:DA)$$

Las ecuaciones lineales se igualan y se realiza la simplificación algebraica, de manera que:

$$0.5143\ (FR - 0.5799\ Pp - 0.6545\ TSM + 0.4008\ PIG) = 0.6404\ (-DA - 0.3808\ Pp - 0.6410\ TSM + 0.6082\ PIG) \tag{10}$$

El proceso de corrección continúa despejando las variables *FR* y *DA* en (10), esto es:

$$FR \quad \frac{0.6404DA \quad 0.0543Pp \quad 0.0739TSM \quad 0.1834PIG}{0.5143} \tag{11}$$

$$DA \quad \frac{0.5143FR \quad 0.0543Pp \quad 0.0739TSM \quad 0.1834PIG}{0.6404}$$

FR y DA en (11), se sustituyen en (9) para generar dos nuevos coeficientes (f_i: FR) y (f_i: DA). Estos nuevos coeficientes se sustituyen por (f_i) en (8), para obtener el modelo de CC canónico corregido:

$$\hat{FR} \quad 0.6404(i:DA) \quad 0.5799Pp \quad 0.6545TSM \quad 0.4008PIG \tag{12}$$

$$\hat{DA} \quad 0.5143(i:FR) \quad 0.3808Pp \quad 0.6410TSM \quad 0.6082PIG$$

Donde:

$\hat{FR}$ es el nuevo índice esperado asociado a la fuerza del reclutamiento (10^6)

$\hat{DA}$ es el nuevo índice esperado asociado a la densidad de adultos reproductores (10^6)

f_i es el índice observado asociado al esfuerzo de pesca (número de viajes).

Pp es el índice observado asociado a la precipitación pluvial (mm).

TSM es el índice observado asociado a la temperatura superficial del mar (°C).

PIG es el índice observado asociado a la concentración de Clorofila *a* más feofitínas ($mg\ m^{-3}$).

(f_i: FR) es el coeficiente corregido que designa el efecto de DA sobre FR en términos de f_i.

(f_i: DA) es el coeficiente corregido que designa el efecto de FR sobre DA en términos de f_i.

Al aplicar las ecuaciones lineales del modelo de CC (12), la predicción o valor esperado (esp) que se obtiene para los índices FR y DA, en comparación con los valores observados (obs) se resume en la Tabla III.

Al confrontar estadísticamente estos resultados con base en la distribución de ji-cuadrada (X_i^2), ambos casos culminaron con la aceptación de la hipótesis nula ($FR\text{-}\hat{R}\hat{F}\ / X_i^2$ = 1.8919, X_α^2 = 24.99, gl= 15 y p= 0.999) y ($DA\text{-}\hat{D}\hat{A} / X_i^2 = 4.9521, X_\alpha^2 = 24.99$, gl= 15 y p= 0.992).

Tabla III. Predicción retrospectiva de los nuevos índices $\hat{R}\hat{F}$ y $\hat{D}\hat{A}$ en el intervalo de junio de 1982 a septiembre de 1983 (10^6 organismos).

Mes	FR obs	$\hat{F}\hat{R}$	DA obs	$\hat{D}\hat{A}$
junio	13.10	14.10	3.66	4.66
julio	9.79	10.80	3.59	4.59
agosto	6.89	7.90	2.96	3.96
septiembre	6.06	7.08	2.17	3.18
octubre	6.30	7.32	1.79	2.80
noviembre	6.17	7.17	1.91	2.91
diciembre	5.41	6.41	1.99	2.99
enero	4.56	5.57	1.44	2.44
febrero	4.37	5.38	1.47	2.47
marzo	4.89	5.89	1.46	2.46
abril	7.29	8.29	1.36	2.36
mayo	13.13	14.15	1.68	2.69
junio	19.94	20.95	2.71	3.72
julio	21.08	22.11	4.54	5.57
agosto	15.85	16.87	5.90	6.91
septiembre	7.35	8.37	3.10	4.12

Con base en lo anterior, se sugiere que el modelo de CC (12), resultó viable para predecir simultáneamente a los índices FR y DA. En cada caso la diferencia entre los valores esperados y observados resultó no significativa, para FR ($F= 35.77$, $p< 0.05$) y para DA ($F= 209.91$, $p< 0.05$).

El modelo propuesto no aplica en los casos en los que la redundancia es baja (< 0.50), ya que las ecuaciones lineales conservan su independencia. Es decisión del investigador la utilización de este modelo, ya que existen otras herramientas de análisis multivariado, que permiten mejorar la predicción de múltiples variables dependientes a partir de múltiples variables independientes, tal es el caso de los Modelos de Ecuaciones Estructurales (MEE)

En los MEE, las variables involucradas funcionan simultáneamente como dependientes e independientes, este análisis genera n ecuaciones lineales para predecir a todas y cada una de las variables tipificadas por el modelo como dependientes, mientras que el resto funciona como variables independientes (Hair *et al*. 1999). Las bases teóricas de los MEE pueden consultarse en diversos libros de estadística multivariada y en los manuales electrónicos de los programas de computo especializados.

Agradecimientos

Agradezco a Adolfo Gracia Gasca (ICMyL, UNAM), al Instituto Nacional de la Pesca (INP) y al Servicio Meteorológico del estado de Tamaulipas, México, por permitir la utilización de la base de datos de camarón café y parámetros ambientales para la realización del estudio de caso.

Referencias

Calvo-Gómez F. 1993. Técnicas estadísticas multivariantes. Universidad de Deusto, Bilbao, España, 435 pp.

Cervantes-Hernández, P. 1999. Relaciones stock-reclutamiento del camarón *Farfantepenaeus duorarum* en el Banco de Campeche. Tesis de Maestría, Instituto de Ciencias del Mar y Limnología, UNAM, México, 37 pp.

Dallas, E.J. 2002. Métodos multivariados aplicados al análisis de datos. International Thomsom, London, 566 pp.

Gracia, A. 1991. Spawning stock-recruitment relationship of white shrimp in the southwestern Gulf of Mexico. Trans. Amer. Fish. Soc. 120: 519-527.

Hair, F., E. Anderson, l. Tatham & C. Black. 1999. Multivariate data analysis. Prentice Hall, Englewood Cliffs, New Jersey, 745 pp.

Morrison, D. 1967. Multivariate statistical methods. McGraw-Hill, Nueva York, 555 pp.

Sharon, L. 1999. Muestreo, diseño y análisis. Matemáticas Thomsom, México, 480 pp.

Uriel, E. 1999. Análisis de datos, series temporales y análisis multivariante. Editorial AC, Madrid, España, 433 pp.