Los peces y el mercurio: análisis multivariado

Inteligencia Artificial avanzada para la ciencia de datos - TC3007C

Grupo 501

Portafolio de Implementación - Módulo 5

Cristofer Becerra Sánchez - A01638659

***Resumen*—Los metales pesados como el mercurio son altamente tóxicos tanto para los ecosistemas acuáticos como para los seres humanos que se alimentan de ellos. Por consiguiente, es de suma importancia abordar la presencia de estos químicos y su concentración en los lagos de consumo ya que representa un peligro para la salud de las personas. El presente reporte analiza un conjunto de datos de 53 lagos en Florida y se intenta determinar los diferentes factores que influyen en la concentración de mercurio en el tejido de los peces. Se realizó un análisis multivariado de normalidad, seguido de una descomposición de las varianzas de las variables en Componentes Principales. Se encontró que únicamente hay dos variables que siguen una distribución normal, a saber, el pH de los lagos y el nivel máximo de concentración de mercurio encontrado en los lagos; a partir de estas variables se constituye la única normal bivariada del conjunto que cumple con la prueba Henze-Zirkler de normalidad. Finalmente a partir del análisis de componentes principales se encontró que la variación en el conjunto se debe principalmente a la cantidad máxima de mercurio registrado por lago, con relación a la alcalinidad y concentración de calcio en el lago; también la concentración media de mercurio y la concentración mínima de mercurio repercuten en la máxima variación.**

***Index Terms*—Mercurio, Metilmercurio, peces, lagos, ANOVA, prueba de hipótesis**

I. Introducción

El tema de la concentración de metales pesados en ecosistemas acuáticos, en particular en animales como peces y mariscos, cobra relevancia debido a que metales como el mercurio son altamente tóxicos en los seres humanos, por lo que altas concentraciones de estos metales en el cuerpo representan un peligro. Por tanto, vale la pena investigar los factores principales que inciden en la concentración de mercurio en estos ecosistemas.

En el presente documento se analiza un conjunto de datos obtenido de un estudio de 53 lagos en Florida (III), y pretende elucidar sobre los diferentes factores que influyen en la cantidad de mercurio en el tejido de los peces a través de un análisis multivariado; primero, se realizan varias pruebas de normalidad de las variables involucradas, y se buscan comportamientos bivariados o multivariados de los datos; segundo, se realiza un análisis de componentes principales para entender las variables que introducen la mayor variación o intentar agrupar los lagos por diferentes características (variables). Los resultados se presentan a detalle en la siguiente sección (II).

Cristofer Becerra Sánchez, pertence al Tec de Monterrey Campus Monterrey, Monterrey, Nuevo León, Mexico, C.P. 64849

II. Resultados y análisis

II-A. Normalidad Multivariada

Se comenzó el análisis haciendo pruebas de normalidad de Anderson-Darling a cada una de las variables del conjunto de datos. La prueba de Anderson-Darling se define como

$$H_0 : \ f(x) = f_0(x) \tag{1}$$
$$H_a : \ f(x) \neq f_0(x) \tag{2}$$

donde $f(x)$ es la distribución de la variable que se desea probar, y $f_0(x)$, en este caso, es una distribución normal. Ésta prueba emplea un estadístico de prueba $A^2 = -N - S$, donde N es el tamaño de la muestra y S toma la forma

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} \left[\ln F(Y_i) + \ln\left(1 - F(Y_i)\right)\right]$$

donde F es la función de densidad de probabilidad acumulada de $f_0(x)$, es decir, la distribución normal.

Tras iterar sobre las variables del conjunto las pruebas de normalidad arrojaron únicamente dos variables: el pH del lago y el nivel máximo de mercurio registrado para el lago. Para la primera variable se computó un estadístico $A^2 = 0.3496$ y, para un nivel de significancia de $\alpha = 0.05$, su respectivo valor crítico es $A^2_{\text{Crit}} = 0.738$. Por tanto, con dicho nivel de significancia no puede rechazarse la hipótesis nula y se concluye que el pH es una variable que sigue una distribución normal. Es de ayuda complementar este resultado con un gráfico cuantil-cuantil acompañado de un histograma de la distribución en cuestión, como se ilustra en la figura 1

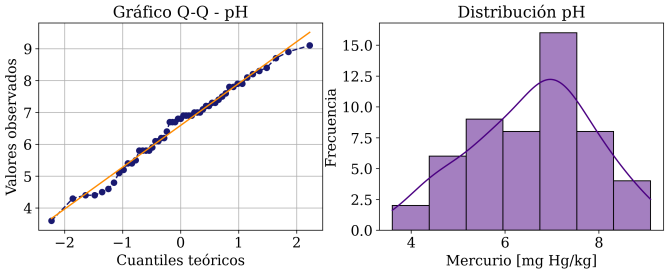

Figura 1. (a) Gráfico cuantil-cuantil de la variable del pH de los lagos. (b) Histograma de la muestra del pH de los lagos de Florida; se ilustra la densidad teórica de la distribución mientras que el eje y cuantifica la frecuencia de los datos en cada intervalo.

Puede resaltarse que el ajuste a la recta teórica es considerablemente bueno, y sólo hay un ligero desfase en la cola derecha, mientras que el mayor desfase sucede cerca de la cola izquierda y un poco al centro de la distribución; esto se traduce al histograma que muestra la clase a la derecha a la central como la más frecuente, en lugar de distribuirse un tanto más a la central.

Para la segunda variable se obtuvo un estadístico $A^2 = 0.6585$ y, utilizando el mismo valor crítico para el mismo nivel de significancia, es imposible rechazar la hipótesis nula y se concluye que el nivel máximo de mercurio registrado en los lagos también sigue una distribución normal. De manera similar, la figura 2 ilustra la distribución y la el ajuste del gráfico Q-Q para apoyar la prueba de hipótesis

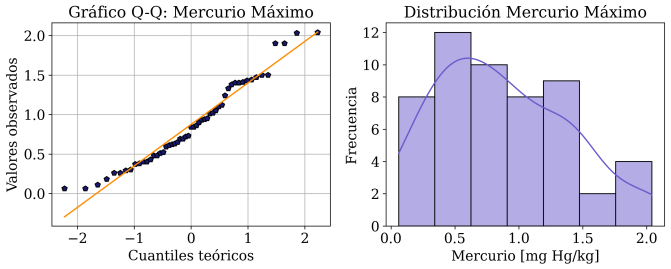

Figura 2. (a) Gráfico cuantil-cuantil de la variable del nivel máximo de mercurio registrado en cada lago. (b) Histograma de la muestra de la misma variable (máximo de mercurio); también se graficó la densidad teórica de la distribución y el eje y también cuantifica la frecuencia por clase.

Se observa que esta variable tiene colas bastante mal ajustadas, lo cual se hace evidente en el histograma; puede decirse que esta distribución, a pesar de haber tenido una prueba de normalidad exitosa, parece tener un moderado sesgo a la derecha con $\alpha_3 = 0.4921$ y un tanto platicúrtica con valor de $\alpha_4 = -0.5149$. Quizá esto pueda deberse a la naturaleza de las mediciones, es decir, el hecho de que sólo se registren valores positivos o 0.

Tras identificar las variables que siguen una distribución normal en una dimensión, es práctico simplificar el proceso y concluir que la única distribución normal multivariada se compone por las únicas distribuciones normales univariadas; es decir, el pH y el máximo de mercurio. No obstante, se optó por realizar las pruebas de normalidad multivariada de todas las combinaciones posibles de variables. En efecto, al aplicar pruebas de normalidad multivariada de Henze-Zirkler de manera exhaustiva en las variables del sistema se encontró sólo una normal multivariada compuesta por el pH de los lagos y el nivel máximo de mercurio –las variables normales encontradas con antelación–. La prueba arrojó un estadístico de prueba $HZ = 0.7696$ con un respectivo p-value $p = 0.1025$, por lo cual, con un nivel de significancia de $\alpha = 0.05$, se concluye que la(s) muestra(s) sigue(n) una distribución normal bivariada. Esta distribución se plasma en la figura 3.

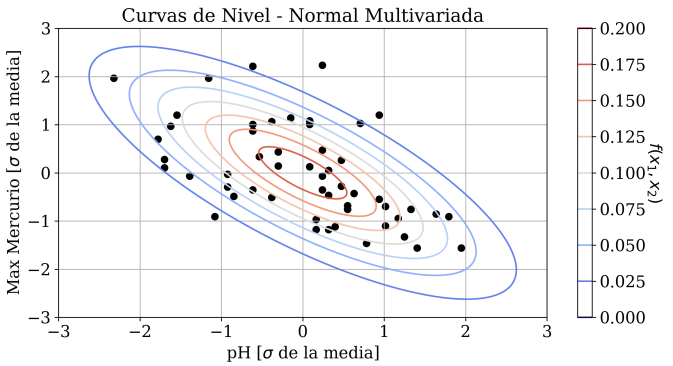

Figura 3. Única distribución normal multivariada del conjunto. El eje x representa a la variable de pH estandarizada, X_1, mientras que el eje y representa la variable del valor máximo de mercurio registrado en los lagos, X_2. Además, se grafican las curvas de nivel de la función de densidad de probabilidad correspondiente.

Una vez computada la distribución normal bivariada, se desea realizar un análisis de intervalos de confianza de los datos que la conforman. Dado un vector $\mathbf{X} = [X_1, X_2]$ de variables aleatorias que conforman una distribución normal bivariada ($p = 2$), se define la distancia de Mahalanobis como

$$d^2 = (\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu) = K$$

donde K es una variable aleatoria y μ es el vector de medias. Entonces, la superficie sobre la cual $K = cte$ representa una elipse centrada en el vector de medias; el volumen debajo de esta superficie representa el intervalo de confianza de los datos para una probabilidad dada. La forma cuadrática que representa la variable aleatoria K es aquella de una distribución chi-cuadrada con p grados de libertad,

$$P\{K \leq \chi_2^2(\alpha)\} = 1 - \alpha$$

en donde $\chi_2^2(\alpha)$ es el percentil $100(1-\alpha)$ de la distribución; en consecuencia, $1 - \alpha$ es el nivel de confianza que un valor dado de $\mathbf{X}$ se encuentre dentro de la elipse. La presente implementación toma en cuenta dos intervalos de confianza, a saber, el 68 % y el 95 %; así,

$$\chi_2^2(0.32) = 2.279 \qquad \chi_2^2(0.05) = 5.991.$$

Luego, se define

$$[X_1 - \mu_1, X_2 - \mu_2] \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \leq \chi_2^2(0.05)$$

donde $\Sigma^{-1} = TDT^{-1}$ para la cual $T = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ y

$$D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

son los eigenvectores y matriz de eigenvalores respectivamente. Ahora, si se define un vector ω,

$$\begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} = T^{-1} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}$$

y sabiendo que $T^T = T^{-1}$, el cuadrado de la diferencia se expresa como

$$[\omega_1, \omega_2] D^{-1} \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} \leq \chi^2_2(0.05)$$

se arriba a la expresión del elipse que representa el intervalo de confianza de la distribución normal bivariada [1].

$$\frac{\omega_1^2}{\chi^2_2(0.05)\lambda_1} + \frac{\omega_2^2}{\chi^2_2(0.05)\lambda_2} \leq 1 \tag{3}$$

En la figura 4 se pueden apreciar ambos intervalos de confianza que representan la ecuación 3 para $\alpha = 0.32, 0.05$.

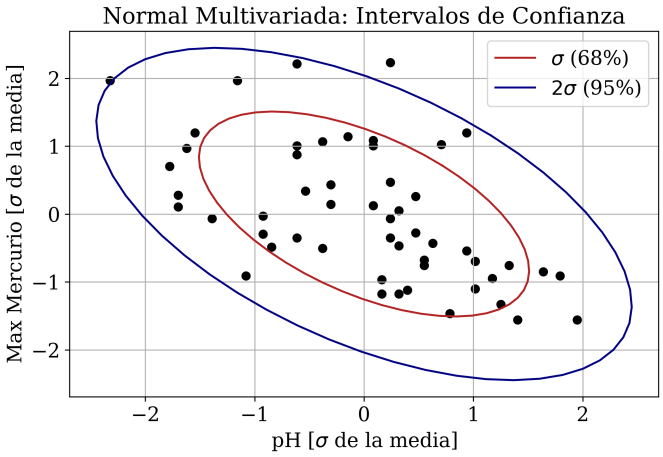

Figura 4. Única distribución normal multivariada del conjunto con los elipses que representan los intervalos de confianza dados.

II-B. Análisis de Componentes Principales

Sea $\mathbf{x}$ un vecor de p variables aleatorias, y sea $\boldsymbol{\alpha}_1'\mathbf{x}$ una función lineal de la forma

$$\boldsymbol{\alpha}_1'\mathbf{x} = \sum_{j=0}^{p} \alpha_{ij} x_j.$$

Ahora, sea $\boldsymbol{\alpha}_2'\mathbf{x}$ otra función lineal sin correlación con $\boldsymbol{\alpha}_1'\mathbf{x}$; si se maximiza $var[\boldsymbol{\alpha}_1'\mathbf{x}]$ manteniendo la nula correlación con $\boldsymbol{\alpha}_2'\mathbf{x}$, y esta, a su vez, tiene una varianza maximizada, se puede seguir con estas condiciones hasta $\boldsymbol{\alpha}_k'\mathbf{x}$. Se dice entonces que $\boldsymbol{\alpha}_2'\mathbf{x}$ es la k-ésima Componente Principal de $\mathbf{x}$. Se desea retener una gran cantidad de la varianza expresada por el conjunto para m componentes principales tal que $m \ll p$ [2].

Ahora, es posible realizar este análisis al obtener los eigenvalores y eigenvectores de la matriz de correlación de las variables numéricas del conjunto (ver figura 5).

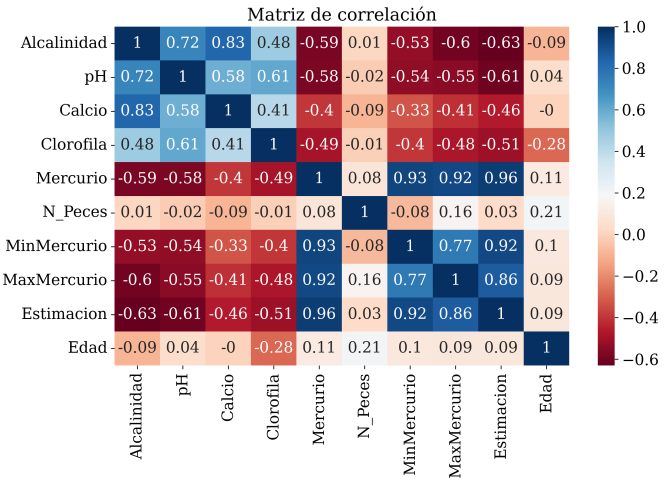

Figura 5. .

Con este análisis se observa que las primeras 4 componentes principales explican el 87.416 % de la variación en el conjunto. Dado que, en términos de dimensionalidad, es cierto que $m \ll p$ para $m = 4$ y $p = 11$, se considera exitosa esta reducción en la complejidad. La figura 6 ilustra los coeficientes de las primeras 4 componentes seleccionadas.

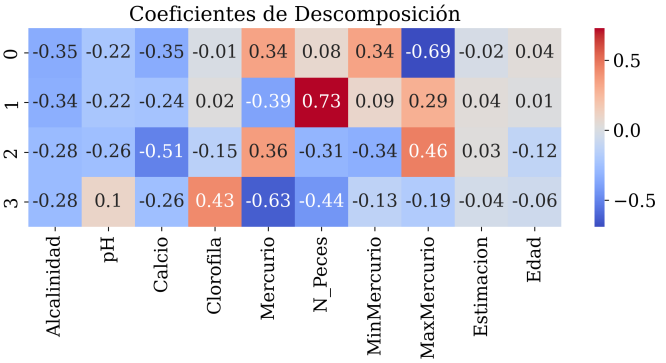

Figura 6. Coeficientes de la descomposición de la varianza del conjunto; es decir, los eigenvectores de la matriz de correlación.

Ahora, utilizando los eigenvectores extraídos de la matriz de correlación, es decir, la nueva base, es posible obtener las variables resultantes. Cabe mencionar que para este procedimiento se estandarizó cada variable antes de realizar el cambio de base. Puede visualizarse las primeras dos componentes principales encontradas, que comprenden el 66.15 % de la variación del conjunto; dicha visualización se aprecia en la figura 7.

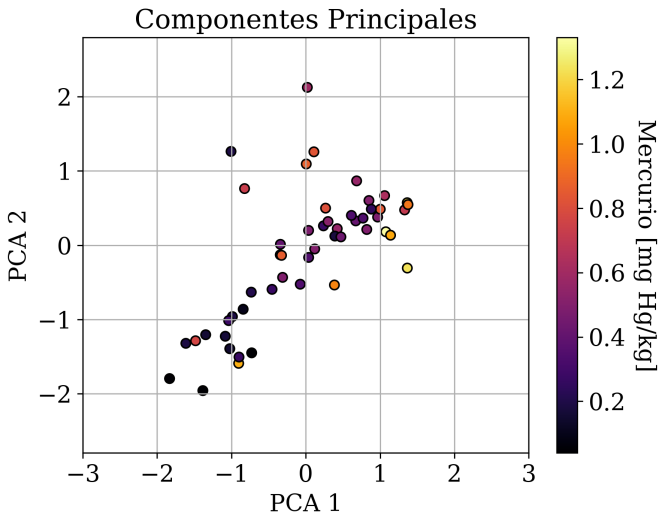

Figura 7. Diagrama de dispersión de las priemras dos componentes principales calculadas. Se agrega una dimensión de color que representa la concentración de mercurio promedio registrada en cada lago (observación).

Puede notarse una especie de tendencia proporcional entre la segunda componente *PCA 2* y la concentración de mercurio; sin embargo, hay observaciones con concentraciones de mercurio inusualmente altas, por lo cual, la correlación entre ambas no es particularmente buena.

Adicionalmente se grafica un diagrama de pares para las primeras cuatro componentes principales extraídas (ver figura 8). Ésta visualización también se tiene la dimensión agregada de la concentración de mercurio como color de las observaciones.

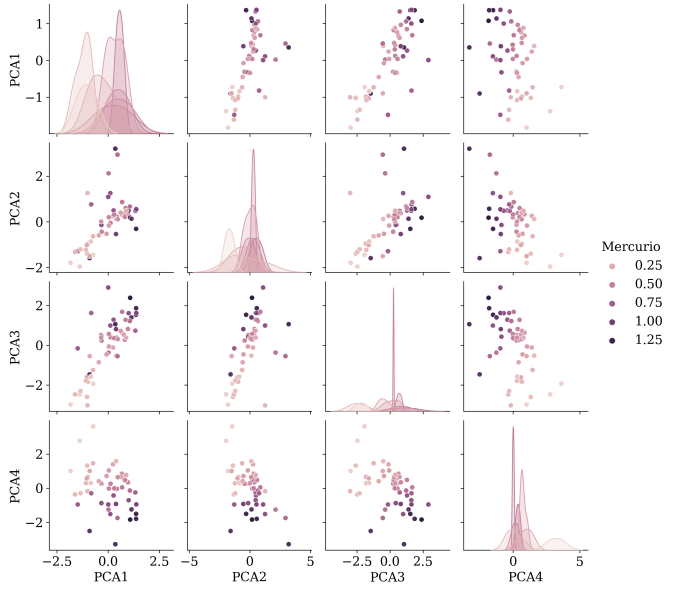

Figura 8. Diagrama de pares de las primeras 4 componentes principales calculadas.

Se resalta una considerable correlación entre PCA1 y PCA2 ($r_{1,2} = 0.6$), pero una baja correlación entre las mismas y la concentración media de mercurio, m ($r_{1,m} = 0.63$, $r_{2,m} = 0.44$); estas correlaciones se ilustran en la figura 9. Esto sugiere que la relación entre la concentración media de mercurio en los lagos depende principalmente de la variación del nivel máximo de mercurio en los lagos con respecto a la alcalinidad del lago, su pH y su concentración de calcio –a juzgar por los coeficientes de la primera componente–. Además, tiene sentido que la segunda fuente de variación, que depende principalmente del número de peces, no tenga una fuerte correlación con la concentración de mercurio, ya que su correlación inicial es sumamente pequeña.

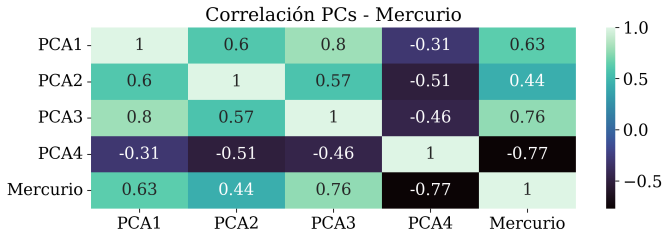

Figura 9. Diagrama de pares de las primeras 4 componentes principales calculadas.

Es notable también que la mayoría de las componentes presentan una correlación entre ligera y razonable con respecto a otras componentes y la concentración media de mercurio en los lagos. Resaltan los casos de las componentes 3 y 4, que tienen unas correlaciones de $r_{3,m} = 0.76$ y $r_{4,m} = -0.77$ respectivamente; por eso es posible notar que, a pesar de que no tienen una correlación particularmente fuerte entre sí ($r_{3,4} = -0.46$), la dimensión agregada en color aumenta conforme PCA4 disminuye, y aumenta junto con la PCA3.

III. Conclusión

En conclusión, del análisis anterior se pueden extraer varias ideas. Primero, que el análisis multivariado de los datos permite encontrar distribuciones normales multivariadas sobre las cuales se puede realizar un cálculo de intervalos de confianza que permite una limpieza más formal de los datos. Segundo, el análisis de componentes principales permite agrupar los datos con respecto a sus mayores fuentes de variación, lo cual puede resultar en un agrupamiento de datos similares o en el fortalecimiento o aislamiento de ciertas características comunes entre observaciones que surgen de las relaciones entre variables.

Anexo

Enlace al repositorio de GitHub, *Mercurio*, con todos los archivos del proyecto (Jupyter Notebook, Notebook en formato .py, la base de datos utilizada, y el presente documento PDF): https://github.com/crisb-7/MercurioPCA

Referencias

[1] Erten, O. & Deutsch, C.V. (2020). Combination of Multivariate Gaussian Distributions through Error Ellipses. In J.L. Deutsch (Ed.), Geostatistics Lessons. Retrieved from http://geostatisticslessons.com/lessons/errorellipses.

[2] Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. Springer-Verlag. doi: https://doi.org/10.1007/b98835.