

Análisis multivariado de datos para determinar los principales factores que influyen en el nivel de contaminación en los peces de los lagos de Florida

Samuel Méndez Villegas - A01652277¹

¹Módulo 5: Estadística avanzada para la ciencia de datos

²Inteligencia artificial avanzada para la ciencia de datos II. Grupo 502

2 de diciembre del 2022

Resumen: El mercurio (Hg) es uno de los contaminantes que más afecta la salud de las personas dado que afectan principalmente el sistema nervioso central. Dicho elemento puede entrar en el cuerpo humano a través de la ingesta de alimentos, por lo que es de suma importancia conocer los niveles de concentración de mercurio en lagos dulces donde habitan peces que suelen ser de consumo. En este reporte se continúa con el análisis realizado con anterioridad. Sin embargo, en este nuevo trabajo se tiene un enfoque de análisis multivariado de datos, es decir se utilizan herramientas estadísticas más avanzadas como análisis de componentes principales. El objetivo del trabajo es de igual forma determinar cuáles son los principales factores que influyen en el nivel de contaminación de los peces de los lagos de Florida. Adicionalmente, se buscó resolver algunas interrogantes que se derivan de la problemática principal. Como resultados, se obtuvo que los factores con más influencia en la presencia de mercurio son la alcalinidad el PH, y la máxima concentración de mercurio en el grupo de peces.

Introducción

El mercurio es un elemento natural que toma forma líquida cuando está en temperatura ambiente, el cual es altamente dañino para la salud humana, pues afecta directamente el sistema nervioso central, el sistema inmunológico, el cerebro, el corazón y los riñones [1]. De igual forma, la problemática con este elemento, es que cuando es liberado de ciertas rocas de forma natural, como lo son explosiones volcánicas, termina en la atmósfera y en el agua. La terminación de este elemento alrededor de la atmósfera también se debe a la actividad humana, a través de la quema de madera, petróleo y combustible. Lo anterior provoca grandes concentraciones de mercurio en cuerpos de agua, como lo son lagos de agua dulce, en donde se encuentran bancos de peces cuyo tejido muscular es contaminado con mercurio [2]. Estos peces con alto nivel de mercurio posteriormente son producto de la pesca, terminando así siendo consumidos por las personas y afectando su salud.

Es por lo anterior que en este reporte se realiza un análisis estadístico multivariado con la herramienta RStudio a una base de datos obtenida de un estudio reciente de los lagos de Florida. El set de datos contiene información sobre los factores que influyen en el nivel de concentración por mercurio en los lagos. Las variables que se estudiarán se describen a continuación:

- X_1 = número de indentificación
- X_2 = nombre del lago
- X_3 = alcalinidad (mg/l de carbonato de calcio)
- X_4 = PH
- X_5 = calcio (mg/l)
- X_6 = clorofila (mg/l)
- X_7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago
- X_8 = número de peces estudiados en el lago
- X_9 = mínimo de la concentración de mercurio en cada grupo de peces
- X_{10} = máximo de la concentración de mercurio en cada grupo de peces
- X_{11} = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
- X_{12} = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Utilizando herramientas estadísticas avanzadas como lo son análisis multivariados y análisis de componentes principales se buscará determinar cuáles son los principales factores que influyen en el nivel de contaminación en los peces de los lagos de florida. Adicionalmente, se buscará resolver algunas interrogantes que se derivan de la problemática como lo son: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida? ¿en qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas? ¿cómo te ayudan los componentes principales a abordar este problema?

Análisis de los resultados

Exploración de la base de datos

Como primer paso para cualquier análisis estadístico, es importante indagar en la base de datos, es decir identificar aquellas variables relevantes para la problemática que se quiere resolver, así como clasificarlas según su naturaleza.

Esta primer exploración de los datos ya fue elaborada en el reporte anterior a éste, por lo que se omitirán varios detalles acerca de éste. Sin embargo, es importante recordar algunas cosas importantes como la correlación entre variables, o cómo están clasificadas según su naturaleza. Recordemos que el data set cuenta con 53 registros y 12 variables, las cuales se clasifican de la siguiente forma:

Variables cualitativas

- Número de identificación (ordinal)
- Nombre del lago (nominal)
- Indicador de la edad de los peces (0: jóvenes; 1: maduros) (nominal)

Variables cuantitativas

- Alcalinidad (mg/l de carbonato de calcio) (escala de razón)
- PH (escala de razón)
- Calcio (mg/l)
- Clorofila (mg/l)
- Concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago (escala de razón)
- Número de peces estudiados en el lago (escala de razón)
- Mínimo de la concentración de mercurio en cada grupo de peces (escala de razón)
- Máximo de la concentración de mercurio en cada grupo de peces (escala de razón)
- Estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) (escala de razón)

Una vez ya clasificadas las variables según su naturaleza, se pasará a realizar algunos análisis.

Análisis de normalidad

La primer técnica que se implementará, es un análisis de normalidad. Este análisis se aplicará a las variables numéricas con el objetivo de identificar qué variables son normales. Para esto, se siguieron una serie de pasos.

Matriz de varianzas y covarianzas y matriz de correlación

Primeramente se obtendrá tanto la matriz de varianzas y covarianzas como la matriz de correlaciones. De igual forma, se obtendrá los vectores de media de cada una de las variables cuantitativas. En las figuras 1 y 2 se puede observar estos resultados:

```

El vector de medias es:
37.53019 6.590566 22.20189 23.11698 0.5271698 13.0566 0.2798113 0.8745283 0.5132075 0.8113208

La matriz de varianzas y covarianzas es:
      X3      X4      X5      X6      X7      X8      X9
X3 1459.509456 35.39971335 793.065711 562.193324 -7.73773984 3.36556604 -4.544071118
X4 35.399713 1.66010160 18.540018 24.159971 -0.25283491 -0.20522496 -0.158097968
X5 793.065711 18.54001814 621.633266 314.949198 -3.40693687 -19.07703193 -1.876788099
X6 562.193324 24.15997097 314.949198 949.645668 -5.16408563 -3.11828737 -2.793996734
X7 -7.737740 -0.25283491 -3.406937 -5.164086 0.11630530 0.23074020 0.071591763
X8 3.365566 -0.20522496 -19.077032 -3.118287 0.23074020 73.28519594 -0.158258345
X9 -4.544071 -0.15809797 -1.876788 -2.793997 0.07159176 -0.15825835 0.051259579
X10 -12.062062 -0.37116800 -5.309432 -7.802021 0.16305729 0.71993106 0.090460486
X11 -8.126195 -0.26746916 -3.922122 -5.286440 0.11080733 0.07481495 0.070485232
X12 -1.432656 0.01933962 -0.020791 -3.444811 0.01464804 0.70319303 0.009002177
      X10      X11      X12
X3 -12.06206241 -8.12619485 -1.432656023
X4 -0.37116800 -0.26746916 0.019339623
X5 -5.30943179 -3.92212155 -0.020791001
X6 -7.80202068 -5.28644013 -3.444811321
X7 0.16305729 0.11080733 0.014648041
X8 0.71993106 0.07481495 0.703193033
X9 0.09046049 0.07048523 0.009002177
X10 0.27253295 0.15203327 0.019332366
X11 0.15203327 0.11473759 0.011962990
X12 0.01933237 0.01196299 0.156023222

```

Figura 1: Vector de medias y matriz de varianzas y covarianzas de los datos.

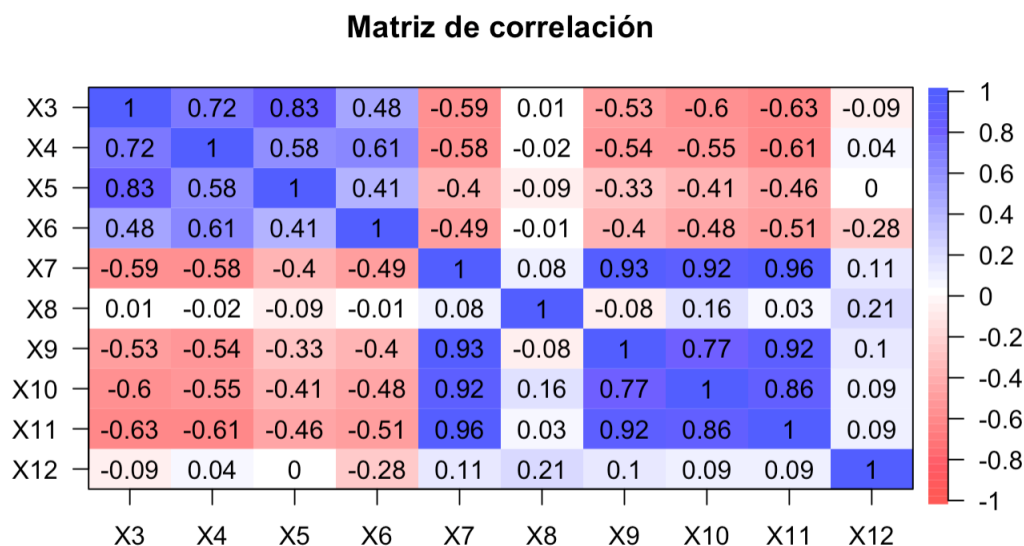


Figura 2: Matriz de correlaciones de los datos.

En la matriz de correlación (2) se puede visualizar con colores, en donde éste indica qué tan correlacionadas están las variables entre ellas. Mientras más cercano a 1 o -1 sea el valor del coeficiente de correlación, más linealmente relacionadas estarán las variables. La matriz de correlaciones se describió a detalle en el trabajo anterior, por lo que a continuación se analizará la matriz de varianzas y covarianzas. Esta matriz como su nombre lo indica, contiene las covarianzas y varianzas entre cada una de las variables. Las varianzas se ubican en la diagonal principal e indican la variabilidad del conjunto de datos respecto a su media. En este contexto, la varianza se podría definir como la covarianza entre la misma variable.

Por otro lado, las covarianzas entre variables se expresan en el resto de la matriz (es una matriz simétrica), en donde se expresa qué tanto varían las variables en forma conjunta respecto a sus medias. Una covarianza positiva indica que mientras el valor de una variable crece, la otra variable de igual forma crece. Por el otro lado, una covarianza negativa indica que mientras una variable aumenta, la otra

disminuye. La covarianza entre las variables nos resulta muy útil para determinar si dos variables son independientes entre sí, pues si la covarianza es distinta de 0, se indica que existe una relación entre las variables y por lo tanto son dependientes. Cabe mencionar que el signo de la covarianza es el que determina de igual forma el signo de la correlación entre variables.

Ya obtenidas las matrices, se pueden graficar todas las variables entre sí con el propósito de ver cómo se comporta cada variable con respecto a las demás. De igual forma si en alguna gráfica se obtiene una nube de puntos similar a una elipse, podríamos estar hablando de una posible distribución normal bivariada.

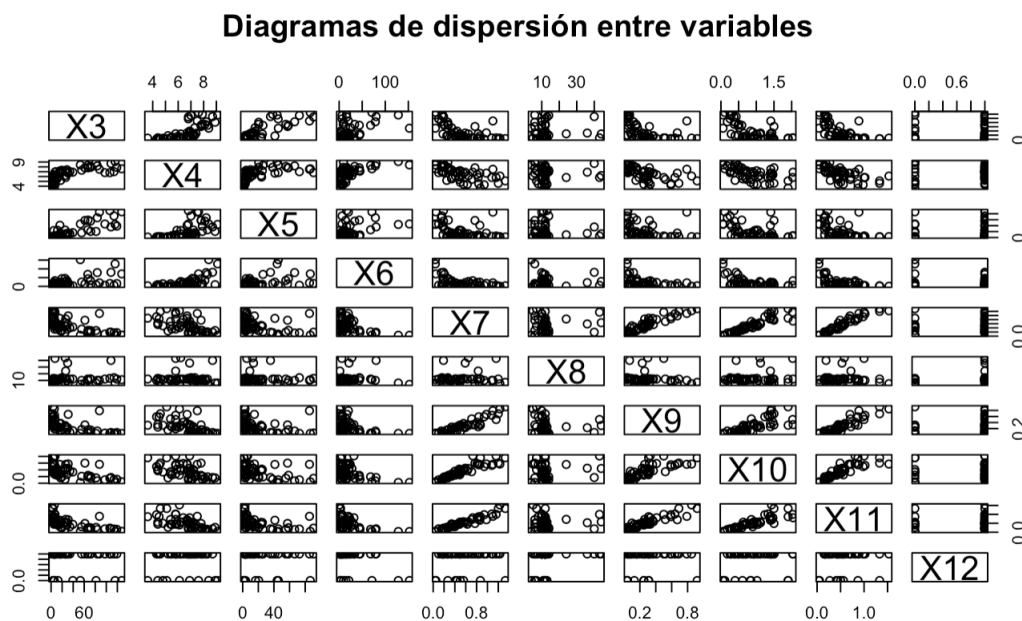


Figura 3: Diagramas de dispersión entre todas las variables.

En el gráfico anterior (figura 3), se muestran los diagramas de dispersión entre todas variables numéricas de la base de datos. Al ser tantas las variables, no se alcanza apreciar del todo bien el comportamiento de las nubes de puntos, sin embargo, se puede observar cierta linealidad entre $X7$, $X9$, $X10$ y $X11$. Esto tiene sentido, pues son las variables que están relacionadas de una forma con la concentración media de mercurio.

Las gráficas anteriores se pueden visualizar de mejor forma graficando por pedazos, u omitiendo algunas variables que se sabe que no es posible formar un comportamiento lineal, como lo es el caso de $X12$, pues esta solo puede tener 2 valores. Por lo tanto, se podría representar de la siguiente forma:

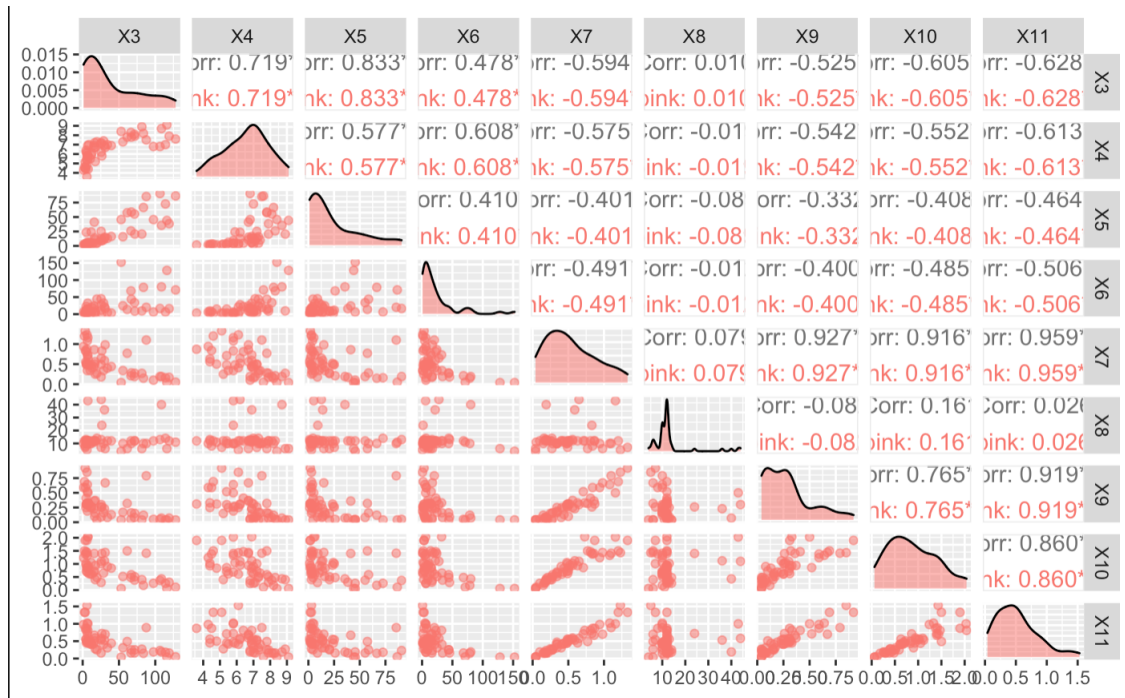


Figura 4: Subconjunto de los diagramas de dispersión.

En la figura 4 se observa con más claridad la linealidad entre las variables antes mencionadas. Como la matriz de gráficos que se forma es simétrica, en la otra mitad se muestra el valor de la correlación entre las variables, el cual coinciden con los valores de la matriz respectiva.

Pruebas de normalidad

Una vez realizados los gráficos, se pasará a realizar las pruebas de normalidad de Mardia y de Anderson Darling con el objetivo de identificar las variables que son normales y detectar posible normalidad multivariada de grupos de variables. A continuación se muestran los resultados de estas pruebas que se aplicaron a todo el conjunto de datos.

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	474.747945136975	8.64265750182786e-21	NO
Mardia Kurtosis	3.59794900484948	0.000320736483631068	NO
MVN	NA	NA	NO
3 rows			

Figura 5: Prueba de normalidad de Mardia sobre el conjunto de datos.

En la figura 5 se pueden observar los resultados de la prueba de normalidad de Mardia, en los cuales se observa que se toman como estadísticos de prueba el sesgo y la kurtosis. Se observa que el p -value del sesgo es alto, al igual que su estadístico correspondiente, indicando así que se tiene una gran variación en los datos. Por lo tanto, no se tiene una distribución simétrica. En el caso de la kurtosis, que simboliza que tan achatada está la campana formada, se tiene un valor pequeño en el p -value el cual debería de ser mayor 0.05, suponiendo que estamos trabajando con una confianza del 95%. Por lo tanto, al tener una

kurtosis relativamente pequeña, se dice que está achatada la campana. De esta forma, se concluye que de esta prueba, se dice que no hay normalidad multivariada entre las 10 variables numéricas analizadas.

	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
1	Anderson-Darling	X3	3.6725	<0.001	NO
2	Anderson-Darling	X4	0.3496	0.4611	YES
3	Anderson-Darling	X5	4.0510	<0.001	NO
4	Anderson-Darling	X6	5.4286	<0.001	NO
5	Anderson-Darling	X7	0.9253	0.0174	NO
6	Anderson-Darling	X8	8.6943	<0.001	NO
7	Anderson-Darling	X9	1.9770	<0.001	NO
8	Anderson-Darling	X10	0.6585	0.081	YES
9	Anderson-Darling	X11	1.0469	0.0086	NO
10	Anderson-Darling	X12	14.3350	<0.001	NO

Figura 6: Prueba de normalidad de Anderson-Darling sobre el conjunto de datos.

El siguiente análisis que se realiza es el de Anderson Darling (6), el cual analiza variable por variable con el objetivo de determinar cuáles variables siguen una distribución normal. En este caso únicamente se obtuvo que $X4$ y $X10$ se comportan de forma normal. Dichas variables corresponden al PH del lago, y al máximo de la concentración de mercurio en cada grupo de peces respectivamente.

	n <int>	Mean <dbl>	Std.Dev <dbl>	Median <dbl>	Min <dbl>	Max <dbl>	25th <dbl>	75th <dbl>	Skew <dbl>
X3	53	37.5301887	38.2035267	19.60	1.20	128.00	6.60	66.50	0.9679170
X4	53	6.5905660	1.2884493	6.80	3.60	9.10	5.80	7.40	-0.2458771
X5	53	22.2018868	24.9325744	12.60	1.10	90.70	3.30	35.60	1.3045868
X6	53	23.1169811	30.8163214	12.80	0.70	152.40	4.60	24.70	2.4130571
X7	53	0.5271698	0.3410356	0.48	0.04	1.33	0.27	0.77	0.5986343
X8	53	13.0566038	8.5606773	12.00	4.00	44.00	10.00	12.00	2.5808773
X9	53	0.2798113	0.2264058	0.25	0.04	0.92	0.09	0.33	1.0729099
X10	53	0.8745283	0.5220469	0.84	0.06	2.04	0.48	1.33	0.4645925
X11	53	0.5132075	0.3387294	0.45	0.04	1.53	0.25	0.70	0.9449951
X12	53	0.8113208	0.3949977	1.00	0.00	1.00	1.00	1.00	-1.5465748

Figura 7: Resumen estadístico de las variables.

Finalmente, la última pestaña que se arroja como resultado muestra un resumen de alguna medidas de tendencia central, dispersión y posición de los datos, así como la kurtosis y el sesgo de cada una.

Sabiendo que no se tiene una normalidad multivariada entre las 10 variables, se volverá a realizar el test, pero ahora utilizando únicamente aquellas variables que se comportan normalmente según la prueba de Anderson-Darling, es decir $X4$ y $X10$. Los resultados se muestran en las siguientes imágenes:

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	6.17538668676458	0.186427564928852	YES
Mardia Kurtosis	-1.12820795824432	0.25923210375991	YES
MVN	NA	NA	YES

Figura 8: Prueba de normalidad de Mardia para las variables normales.

	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
1	Anderson-Darling	X4	0.3496	0.4611	YES
2	Anderson-Darling	X10	0.6585	0.0810	YES

Figura 9: Prueba de normalidad de Anderson Darling para las variables normales.

En este caso, en el test de Mardia (8) se obtuvo que efectivamente las variables tienen un comportamiento binormal, ya que los $p - values$ de la kurtosis y del sesgo son mayores a 0.05. Tenemos que la kurtosis obtenida es negativa, lo que indica que los datos presentan valores atípicos menos extremados a una distribución normal. En cuanto al sesgo, este se nota que es mucho menor que el obtenido con todas las variables, indicando así que no se tiene una gran variación entre los datos. Los resultados de los demás componentes siguen manteniéndose iguales (9). Por lo tanto, se puede decir que la combinación de estas variables forman una distribución normal bivariada, la cual se pasará a graficar a continuación.

Gráficos de normalidad

Primeramente, se grafica el diagrama de dispersión entre las variables normales encontradas, es decir X_4 y X_{10} .

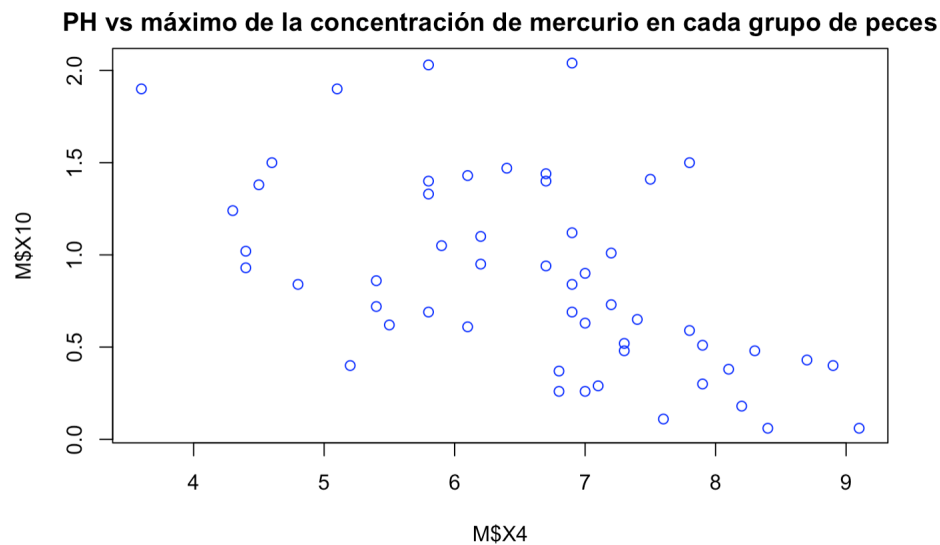


Figura 10: Diagrama de dispersión de X_4 vs X_{10} .

Como se observa, los datos forman más o menos una elipse diagonal, lo cual es sinónimo de una distribución normal bivariada. Dicho gráfico también se puede visualizar en 3D para que se observe de mejor forma la elipse que se forma, así como la kurtosis y el sesgo de la distribución bivariada.

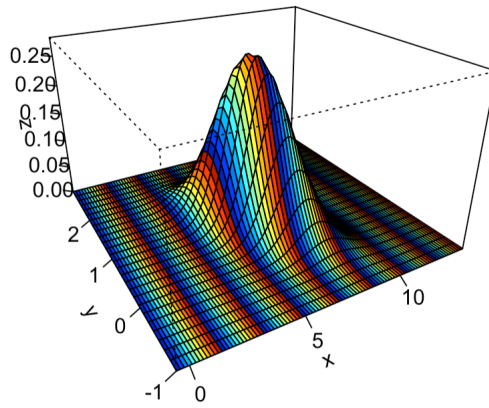


Figura 11: Gráfico en 3D de la distribución normal bivariada entre X_4 y X_{10} .

En el gráfico anterior se puede observar la campana que se forma por la distribución normal bivariada. Como se observa, no se tiene tanto sesgo, pues se muestra como una campana simétrica. Sobre la kurtosis, se nota que se tiene un valor muy pequeño, dado a que la campana no está tan achatada. De igual forma, se puede ver que la campana está colocada diagonalmente, al igual que la nube de puntos que se pudo observar al graficar X_4 contra X_{10} .

Ahora, se pasará a realizar la gráfica de contorno de la normal multivariada con el objetivo de visualizar más claramente la elipse que se forma.

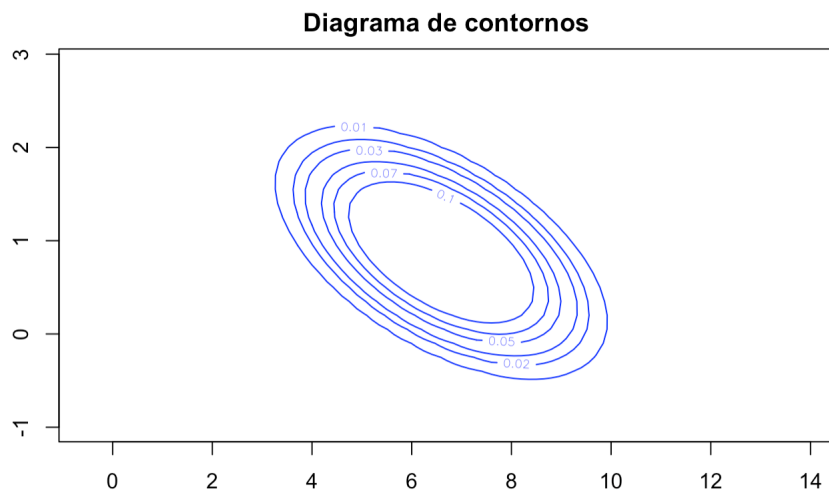


Figura 12: Diagrama de contornos de la distribución bivariada entre X_4 y X_{10} .

La gráfica anterior representan los contornos de la distribución bivariada. Se observa que los contornos tienen forma de elipse, lo que se esperaría de una distribución bivariada. La dirección en la que se expanden las elipses coinciden con los datos del diagrama de dispersión.

Valores atípicos

Finalmente, se pasará a detectar aquellos valores atípicos que se presentan en la distribución. Si observamos de vuelta el diagrama de dispersión entre las variables (10), se nota que se tienen algunos puntos que no están del todo cercanos a los demás, pudiéndose tratar de valores atípicos. Por lo tanto, para

determinar estos datos, se hará uso de la distancia de Mahalanobis y del gráfico QQplot multivariado. En la siguiente imagen se muestran las distancias de Mahalanobis de cada registro:

```
[1] 1.19340732 3.86623312 4.13223576 0.06401297 2.55854258 0.59679099 1.78765285 1.50256004
[9] 1.01802464 1.63445643 1.26957081 0.22408796 0.61327299 5.43003539 2.14558317 2.15128332
[17] 2.68426029 5.11541618 0.78631278 1.84332204 3.88561874 0.04988963 0.88374242 8.09988683
[25] 2.02671417 0.58001802 0.50566324 2.64029211 2.92897033 1.06510339 0.88216107 0.40467236
[33] 6.05908470 3.51592769 1.45168260 1.32485980 1.13164718 2.85893264 0.22445752 3.37787450
[41] 1.54397421 1.72672287 0.28960341 1.76103444 1.60277372 0.09262808 0.19257450 3.22197239
[49] 3.27393627 0.17321589 0.12786732 4.41942776 1.06000856
```

Figura 13: Distancias de Mahalanobis de cada registro.

Con dichas distancias, es posible obtener el QQ plot multivariado que se muestra a continuación:

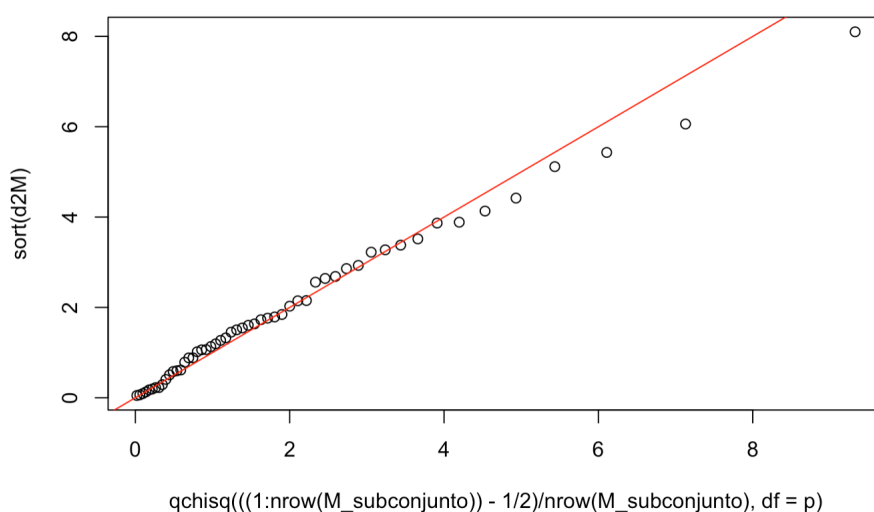


Figura 14: Diagrama QQ-plot.

El diagrama QQ que se observa indica que realmente no se tienen datos atípicos, pues todos los puntos se encuentran realmente cercanos a la recta de color rojo. Los únicos dos posibles datos que podrían tratarse de datos atípicos, son los dos últimos, pues estos se desvían bastante de la recta y son lo que tienen una mayor distancia de Mahalanobis.

Cabe resaltar que los puntos tienen un comportamiento tal que se podría inferir que se cuenta con un leve sesgo hacia la izquierda, es decir, se tiene asimetría negativa.

Análisis de componentes principales (PCA)

Uso adecuado de componentes principales

Ahora, se pasará a realizar un análisis de componentes principales con los datos. Esto es con el objetivo de ver si es posible encontrar nuevas variables (componentes principales) las cuales expliquen una mayor parte de la varianza sin perder tanta información, con tal de reducir la dimensión de la base de datos.

Para este análisis, retomemos un poco la matriz de correlaciones de la figura 2. Como ya se explicó con anterioridad, la matriz de correlaciones indica qué tan linealmente están relacionadas las variables entre sí. Antes de realizar el análisis de componentes principales es importante verificar la correlación, ya

que uno de los supuestos es que las variables estén correlacionadas para formar combinaciones lineales. Si las variables tienen una correlación distinta de 0, quiere decir que se relacionan de una forma, y por lo tanto es válido aplicar componentes principales.

Obteniendo los componentes principales

Ahora se pasará a obtener los componentes principales de los datos a través de la matriz de correlaciones. Para esto, se obtienen los valores y vectores propios de la matriz respectiva.

```
eigen() decomposition
$values
[1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741 0.20673634 0.08682133
[9] 0.05163902 0.01863699

$vectors
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
[1,] -0.35065869 -0.21691594 -0.3472906  0.009131194  0.34050534  0.07547497 -0.33823501
[2,] -0.33700381 -0.21940887 -0.2360975 -0.017242162 -0.39396038  0.73121012 -0.08629646
[3,] -0.28168286 -0.26250672 -0.5113780  0.146950070  0.36205937 -0.31342329  0.34312185
[4,] -0.28334182  0.10195058 -0.2639612 -0.432676049 -0.63093376 -0.44112169  0.13435159
[5,]  0.39830786 -0.12104244 -0.2996635 -0.080630070 -0.03046869  0.07436922 -0.01377825
[6,]  0.02667579 -0.57556151  0.3050633 -0.692854505  0.19646415 -0.05926732 -0.14693148
[7,]  0.36839224 -0.04432459 -0.3876861  0.044658983 -0.13236038 -0.19602465 -0.45674057
[8,]  0.37893835 -0.14237181 -0.2024901 -0.167921215  0.02678086  0.26671839  0.67376588
[9,]  0.40206100 -0.05279514 -0.2562319 -0.042242268 -0.05607416  0.03863899 -0.23387764
[10,] 0.05931430 -0.67421026  0.2294446  0.521815581 -0.37253140 -0.21612970  0.05759514
      [,8]      [,9]      [,10]
[1,]  0.68622998  0.04284021 -0.02239801
[2,] -0.28769221  0.01363551  0.04445261
[3,] -0.45568753 -0.11508339  0.02634676
[4,]  0.19006976 -0.06333133 -0.03982419
[5,] -0.01674789  0.06243320 -0.84827636
[6,] -0.16809481  0.02532023  0.04805976
[7,] -0.18260535  0.53803577  0.35020485
[8,]  0.33602914  0.18844932  0.30445219
[9,]  0.02613406 -0.80648296  0.24018040
[10,] 0.16451240 -0.02782678 -0.01839703
```

Figura 15: Valores y vectores propios de la matriz de correlaciones.

Los vectores propios anteriores se componen de los coeficientes de cada una de las variables para formar una nueva variable. En cada componente, se nota que se tiene un mayor peso en algunas variables, dando así a entender que esas juegan un papel más importante para dicho componente. Por ejemplo, en el primer componente que proviene del primer vector propio, se tiene que las variables X_{11} , X_7 y X_9 tiene mayor relevancia. Cada componente le da una importancia distinta a cada variable y lo que busca es relacionarlas con el fin de darles nosotros mismos un significado al componente.

Una vez realizado lo anterior, es posible calcular el porcentaje de varianza explicada por cada una de las componentes. A continuación se realiza el análisis

```
La variazas explicada de cada componente es:
[1] 0.536122641 0.125426109 0.121668138 0.090943267 0.059141736 0.030314741 0.020673634
[8] 0.008682133 0.005163902 0.001863699
```

Figura 16: Porcentaje de varianza explicada de cada componente principal.

La varianza explicada indica que tanta información de los datos se está explicando en cada componen-

tes. Por ejemplo, el primer componente principal está explicando alrededor del 53 % de la varianza total, Lo que se busca es expresar con el menor número de componentes, el mayor porcentaje de varianza, para así poder realizar la reducción de dimensionalidad sin pérdida de información.

Para poder decidir con qué y cuántos componentes principales nos debemos de quedar, nos apoyamos en algunos gráficos.

Gráficos auxiliares

Primero realizamos dos gráficos que comparan el comportamiento de los dos primeros componentes principales. En el primer gráfico (figura 17), se pueden observar los registros ya transformados, es decir después de haberles aplicado su combinación lineal correspondiente. Como se puede observar, ambos componentes abarcan gran parte de los datos, indicando así que explican gran parte de la varianza, lo cual es bueno cuando se quiere reducir dimensiones. En los ejes de igual forma se ve el porcentaje de varianza explicado de cada componente.

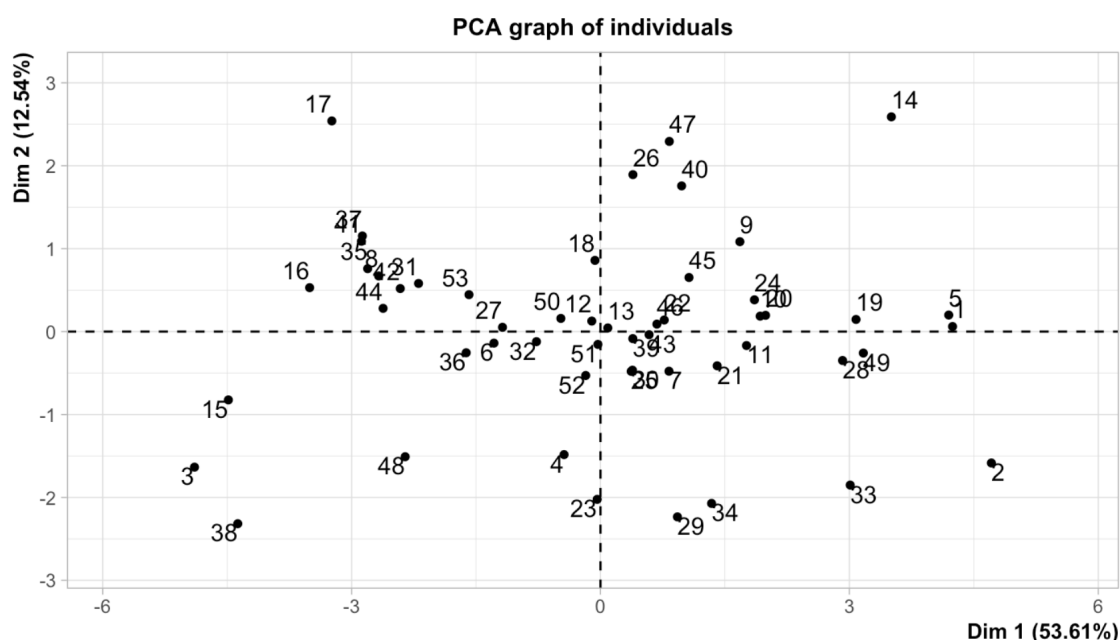


Figura 17: Gráfico de individuos del PCA.

En el segundo diagrama (figura 18) se muestran los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes. Como se observa, la primer componente jala más a X_3 , X_4 , X_7 y X_{11} , mientras que el segundo componente jala más a X_{12} . Sin embargo, de igual forma de ve que la mayoría de vectores tienen una longitud bastante similar, indicando así que estos dos primeros componentes principales están describiendo gran parte de la variación.

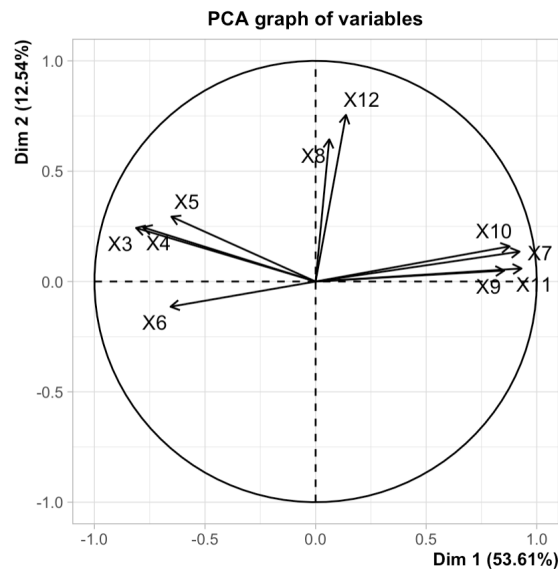


Figura 18: Gráfico de variables del PCA.

Otro gráfico que es de mucho interés, es el de codo con base en la varianza explicada de cada variable. Aquí se muestra qué el primer componente representa el 53 % de la información, mientras que el segundo y tercer componente tiene más o menos el mismo porcentaje. Si nos basamos en el codo de la gráfica, lo mejor sería tomar los tres primeros componentes principales, ya que haciéndolo de esa forma, se reduciría la dimensión original de 10 a 3 y se explicaría el 77 % de la información, lo cual no está nada mal. Los demás componentes no tienen tanta relevancia.

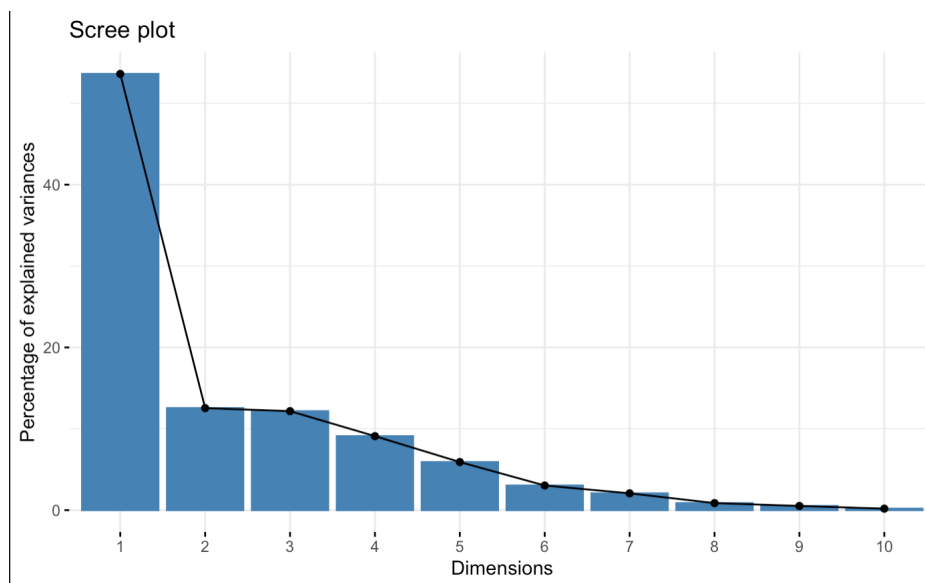


Figura 19: Gráfico de codo del PCA.

Finalmente, como se estará trabajando con los tres primeros componentes, se le debe de dar un significado a cada uno. Por ejemplo, el componente 1 relaciona varias variables, como se muestra en el siguiente gráfico, siendo las más representativas X11, X7, X10, X9, X3 y X4. Viendo las cuatro primeras variables, vemos que están relacionadas con la concentración de mercurio en los lagos, mientras que las otras dos con el pH y la alcalinidad. Esto nos podría decir que tanto la alcalinidad del lago como

su PH son relevantes para la concentración de mercurio, por lo que al componente principal se le podría dar el nombre de "Grado de acidez." similar.

El segundo componente lo describe en su mayoría X_{12} y X_8 , en donde se puede ver que se relaciona la edad de los peces con el número de éstos. Por lo tanto el componente principal se podría interpretar como características de la población", o similar.

Un análisis similar se tendría que hacer para cada uno de los componentes, es decir tratarles de dar un significado dentro del contexto.

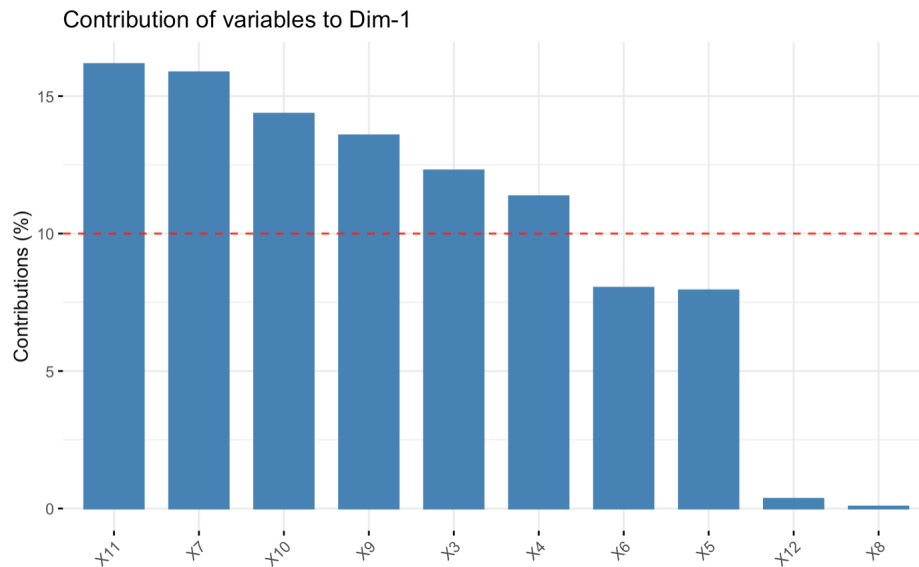


Figura 20: Contribución de las variables al primer componente principal.

Conclusiones

Después de haber realizado tanto el análisis de normalidad, como el análisis de componentes principales, se puede concluir que los principales factores que afectan a la concentración de mercurio en los tejidos de los peces es el de PH, la alcalinidad y el máximo de concentración de mercurio en cada grupo de peces. Esto se debe a que en el análisis de normalidad, fueron algunas de estas variables las que se comportaron normalmente, además de explicar un alto nivel de varianza.

De igual forma, el análisis de normalidad de los datos es muy importante para generar modelos de predicción multivariados. Por ejemplo, el conocer cuáles combinaciones de variables forman una distribución normal bivariada, o multivariada puede facilitar la construcción de modelos como discriminadores, que puedan predecir si el lago tendrá una concentración alta de mercurio, o aplicar otros modelos multivariados para calcular la concentración de mercurio de los lagos.

Finalmente, el análisis de componentes principales resultó crucial para reducir la dimensionalidad del problema. Esto quiere decir que se logró explicar gran parte de la varianza de los datos con muchas menos variables. Esto es de suma relevancia a la hora de construir modelos de aprendizaje de máquina o similares, ya que en muchas ocasiones, el trabajar con pequeñas cantidades de variables, es algo efectivo.

Anexos

Liga a la carpeta de Google Drive que contiene el documento de R en donde se realizó el análisis, y la base de datos utilizada.

<https://drive.google.com/drive/folders/1WBppXACrY2b14a1K-u7ea2hL0fN16qA6?usp=sharing>

Referencias

- [1] N. C. Muñoz. El mercurio como contaminante y factor de riesgo para la salud humana. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1794-44492020000200280
- [2] A. de Protección Ambiental de Estados Unidos. Información básica sobre el mercurio. [Online]. Available: <https://espanol.epa.gov/espanol/informacion-basica-sobre-el-mercurio>