Reporte final de “Los peces y el mercurio”

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 501)

Módulo Estadística

Luis Daniel Cano Irigoyen, A00827178

2022-10-26

Resumen

Problemática

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio.

X7 → *MediaMercurio*, X8 → *NumPez*, X9 → *MinMercurio*, X10 → *MaxMercurio*, X11 → *TresMercurio*.

Métodos y técnicas estadísticas

- Análisis de normalidad multivariada.
- Prueba de normalidad de Mardia y prueba de Anderson
- Gráfico de contorno y Gráfica de multinormalidad (Mahalanobis)
- Análisis de Componentes Principales

Principales resultados

- Hay normalidad multivariada para el conjunto de PH y MaxMercurio.
- Con datos estandarizados podemos generar 5 Componentes Principales explicando el 95.69% de la varianza.

Introducción

Problema a resolver

La principal pregunta a resolver que surge es: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?

Importancia del problema

Responderla es importante ya que, al saber qué componente o factor es el principal causante de la concentración del mercurio, este podría ser tratado en las aguas con peces de pesca para reducirlo, y consecutivamente reducir el riesgo de que un pescado con mercurio sea consumido por un humano.

Análisis de los resultados

Datos

En nuestro análisis cambiamos el nombre de las columnas para comprender mejor los datos:

```
ID  Nombre Alcalinidad  PH   Calcio  Clorofila  MediaMercurio  NumPez
MinMercurio  MaxMercurio  TresMercurio  Edad
```

ID y Nombre no son variables numéricas y Edad no es variable continua.
No son explicativas, así que las eliminamos.

1. Análisis de normalidad de las variables continuas para identificar variables normales.

A.

Se realiza la prueba de normalidad de Mardia y prueba de Anderson Darling para identificar las variables que son normales y detectar posible normalidad multivariada de grupos de variables.

Hipótesis:

H_0: Si hay normalidad multivariada

H_a: No hay normalidad multivariada

```
multivariateNormality
             Test        Statistic              p value Result
1 Mardia Skewness  434.33906591642 4.13584083502475e-26     NO
2 Mardia Kurtosis 5.76907272063334  7.9708906142173e-09     NO
3             MVN             <NA>                 <NA>     NO

univariateNormality
              Test      Variable Statistic   p value Normality
1 Anderson-Darling  Alcalinidad     3.6725  <0.001      NO
2 Anderson-Darling       PH         0.3496  0.4611      YES
3 Anderson-Darling    Calcio        4.0510  <0.001      NO
4 Anderson-Darling   Clorofila      5.4286  <0.001      NO
5 Anderson-Darling MediaMercurio    0.9253  0.0174      NO
6 Anderson-Darling    NumPez        8.6943  <0.001      NO
7 Anderson-Darling  MinMercurio     1.9770  <0.001      NO
8 Anderson-Darling  MaxMercurio     0.6585   0.081      YES
9 Anderson-Darling TresMercurio     1.0469  0.0086      NO

Descriptives
               n       Mean    Std.Dev Median  Min    Max  25th  75th
Alcalinidad   53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50
PH            53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40
Calcio        53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60
```

```
Clorofila     53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70
MediaMercurio 53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77
NumPez        53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00
MinMercurio   53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33
MaxMercurio   53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33
TresMercurio  53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70
                     Skew   Kurtosis
Alcalinidad     0.9679170 -0.4705349
PH             -0.2458771 -0.6239638
Calcio          1.3045868  0.6130359
Clorofila       2.4130571  6.1042185
MediaMercurio   0.5986343 -0.6312607
NumPez          2.5808773  6.0089455
MinMercurio     1.0729099  0.4060828
MaxMercurio     0.4645925 -0.6692490
TresMercurio    0.9449951  0.5733500
```

Contamos con una Mardia Skewness de 434.34 y una Mardia Kurtosis de 5.77

Esta función realiza pruebas de normalidad multivariada de sesgo y curtosis y nos da como resultado que No hay normalidad multivariada. De igual manera, utilizando un nivel de significancia de 0.05, podemos ver como los *valores p* de sesgo (4.1e-26) y curtosis (7.9e-09) son menores al nivel de significancia, por lo que rechazamos a H_0 y determinamos que No hay nomralidad multivariada en las variables.

Del Test de univariateNormality de Anderson-Darling encontramos que las variables que son normales son PH y MaxMercurio.

B.

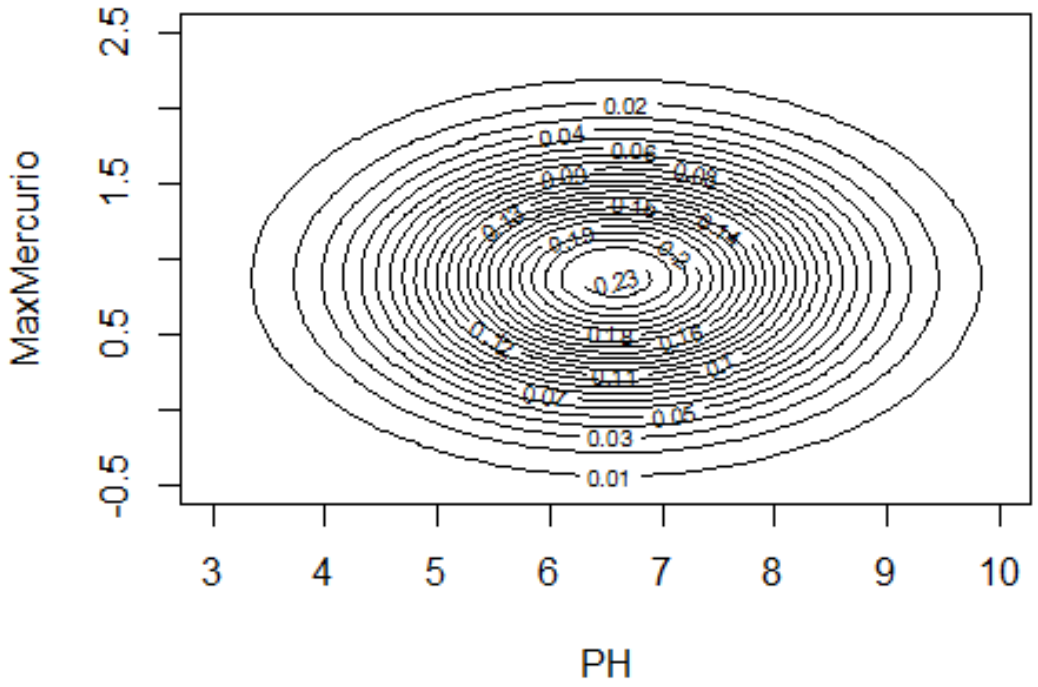
Realizamos las pruebas de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores.

```
multivariateNormality
             Test          Statistic           p value Result
1 Mardia Skewness   6.53855430534145 0.162377302354508    YES
2 Mardia Kurtosis -0.889321233851276 0.373830462900113    YES
3             MVN               <NA>              <NA>    YES

univariateNormality
              Test    Variable Statistic   p value Normality
1 Anderson-Darling     PH         0.3496    0.4611    YES
2 Anderson-Darling MaxMercurio    0.6585    0.0810    YES

Descriptives
             n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew
PH          53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771
MaxMercurio 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
              Kurtosis
PH          -0.6239638
MaxMercurio -0.6692490
```

A diferencia de la prueba de Mardia con todas las variables, al usar solo las normales (PH y MaxMercurio) obtenemos una curtosis entre -1 y 1, lo cual nos dice que hay normalidad. Asimismo, la prueba de normalidad multivariada de sesgo y curtosis realizada nos da como resultado que Si hay normalidad multivariada, y contamos con *valores p* que son mayores al nivel de significancia.

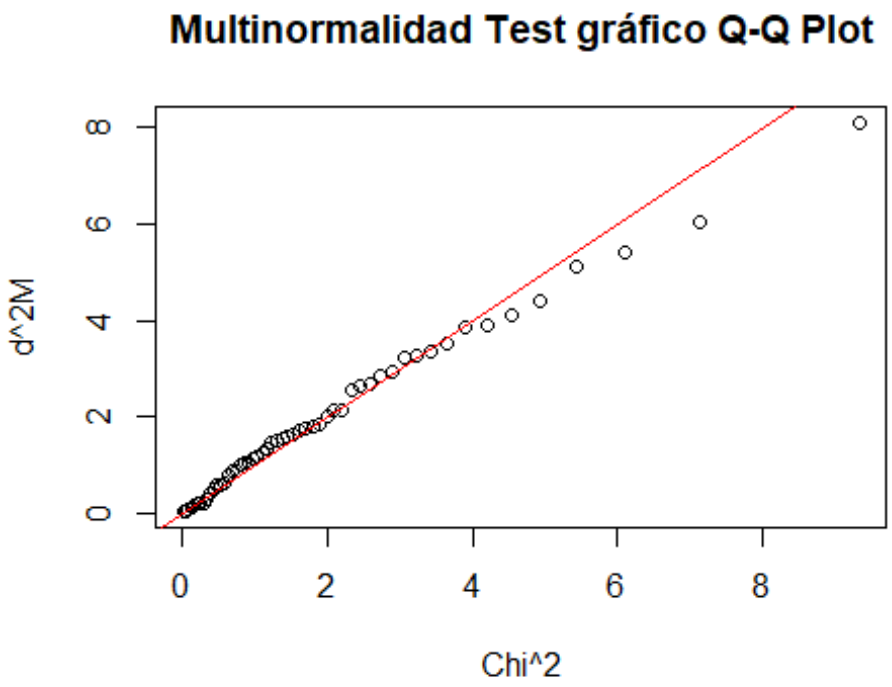
C.

Gráfica de contorno de la normal multivariada obtenida en el inciso B. Realizamos una distribución normal bivariada y con ella creamos el gráfico



D.

Ahora buscamos detectar datos atípicos o influyentes en la normal multivariada encontrada del inciso B. Este gráfico se crea utilizando las medias de PH y MaxMercurio y su matriz de covarianza para determinar la Distancia cuadrada de Mahalanobis. Esta distancia la graficamos contra Chi^2 para crear el gráfico de multinormalidad.

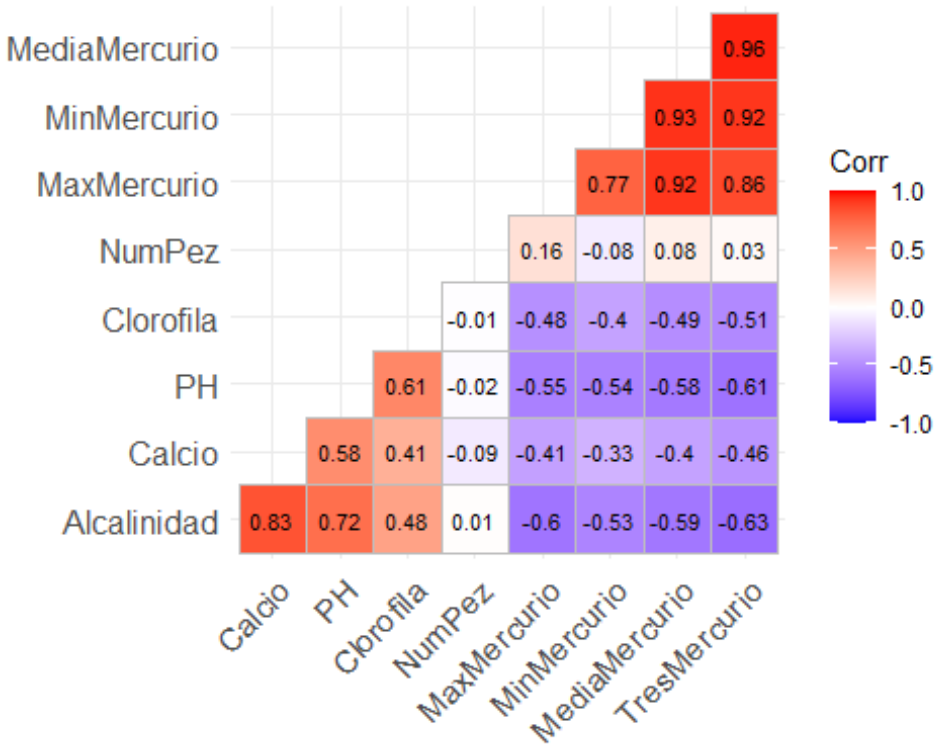


En la gráfica de distancias de Mahalanobis podemos observar que al inicio los puntos se encuentran cerca de la normalidad multivariada, pero conforme avanza la gráfica en las últimas instancias a la derecha se observa una ligera curva hacia abajo. Los últimos 3 puntos podríamos considerarlos como datos atípicos ya que la distancia se aleja bastante.

2. Análisis de componentes principales con la base de datos completa para identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

A.

Para justificar la utilidad de los componentes principales para analizar la base de datos, realizamos una matriz de correlaciones de las variables numéricas:



Fuera de NumPez, tenemos una tabla llena de correlaciones moderadas, altas y muy altas, con muy pocas correlaciones como Calcio y MinMercurio que podríamos considerar como bajas. Hay mucha dependencia entre las variables, podemos determinar que varios pares, como la Alcalinidad y el Calcio, explican casi lo mismo. Por ello si es relevante el usar Componentes Principales en este problema.

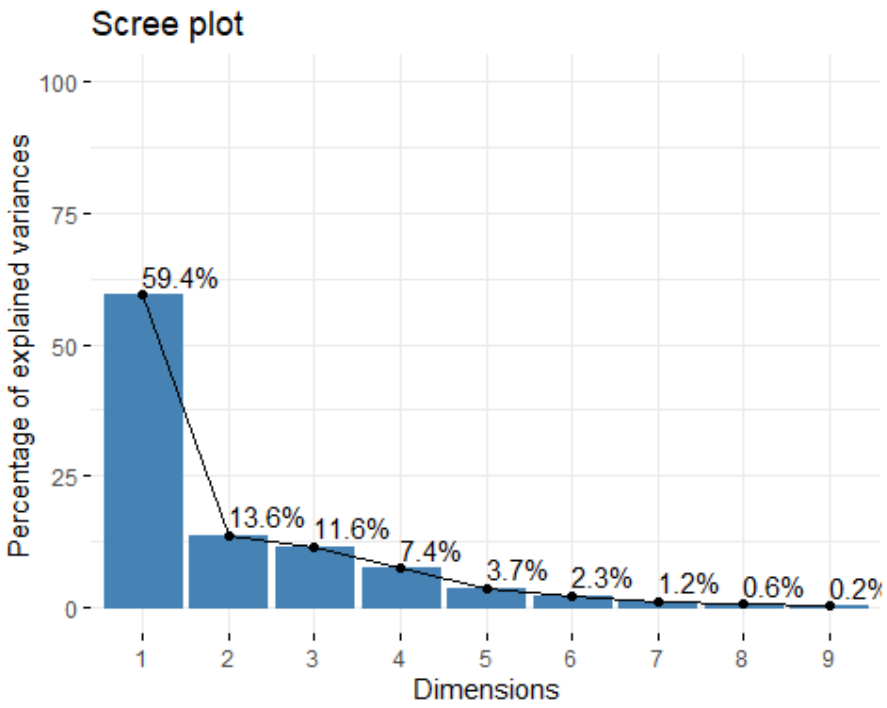
B.

Análisis de componentes principales.

Todo este análisis se realiza con variables estandarizadas, para que todas contribuyan de la misma manera al análisis

```
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 53 individuals, described by 9 variables
```

Nuestro análisis genera 9 componentes, de los cuales podemos ver los siguientes resultados:



Valores eigen:

	eigenvalue	% of variance	cumulative % of variance
comp 1	5.34590819	59.3989799	.39898
comp 2	1.22090789	13.5656432	72.96462
comp 3	1.04253153	11.5836836	84.54831
comp 4	0.66786333	7.4207036	91.96901
comp 5	0.33571266	3.7301407	95.69915
comp 6	0.20893778	2.3215309	98.02068
comp 7	0.10725403	1.1917115	99.21239
comp 8	0.05203127	0.5781252	99.79052
comp 9	0.01885332	0.2094814	100.00000

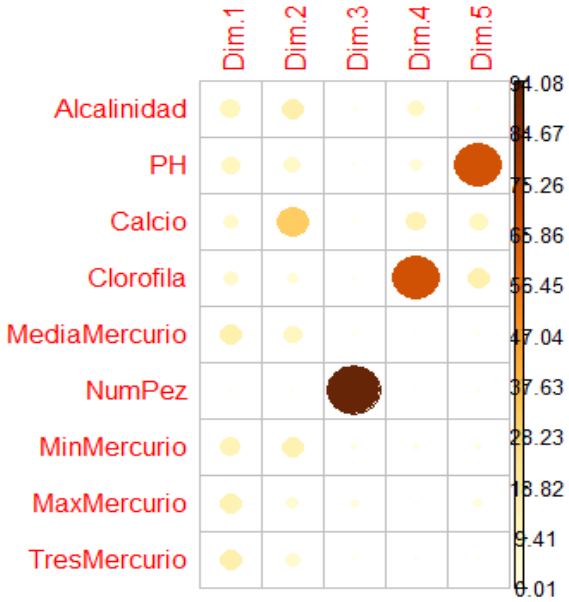
A partir de los resultados de los Componentes Principales, determinamos apropiado que la dimensión de la base sea reducida a los primeros 5 componentes, ya que estos representan más del 95% de la variación total.

Veamos como contribuye cada variable a los 5 Componentes Principales

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Alcalinidad	12.34548788	16.2423951	0.575534992	9.21694334	0.1020241
PH	11.49713128	8.8721570	0.558029911	5.39944549	68.2657403
Calcio	8.01256203	32.4250869	0.089480888	14.00790620	10.7689850
Clorofila	7.91125981	4.6332056	0.377882432	68.98320445	15.5934084
MediaMercurio	15.91287185	10.5746226	0.319004150	0.24802582	0.4276249
NumPez	0.05754606	0.3920636	94.078707352	0.26512448	0.8109000
MinMercurio	13.61982733	14.1730414	1.379131785	1.29984235	1.1163242
MaxMercurio	14.40736248	5.9676906	2.616505311	0.07659795	2.7302433
TresMercurio	16.23595127	6.7197372	0.005723179	0.50290992	0.1847498

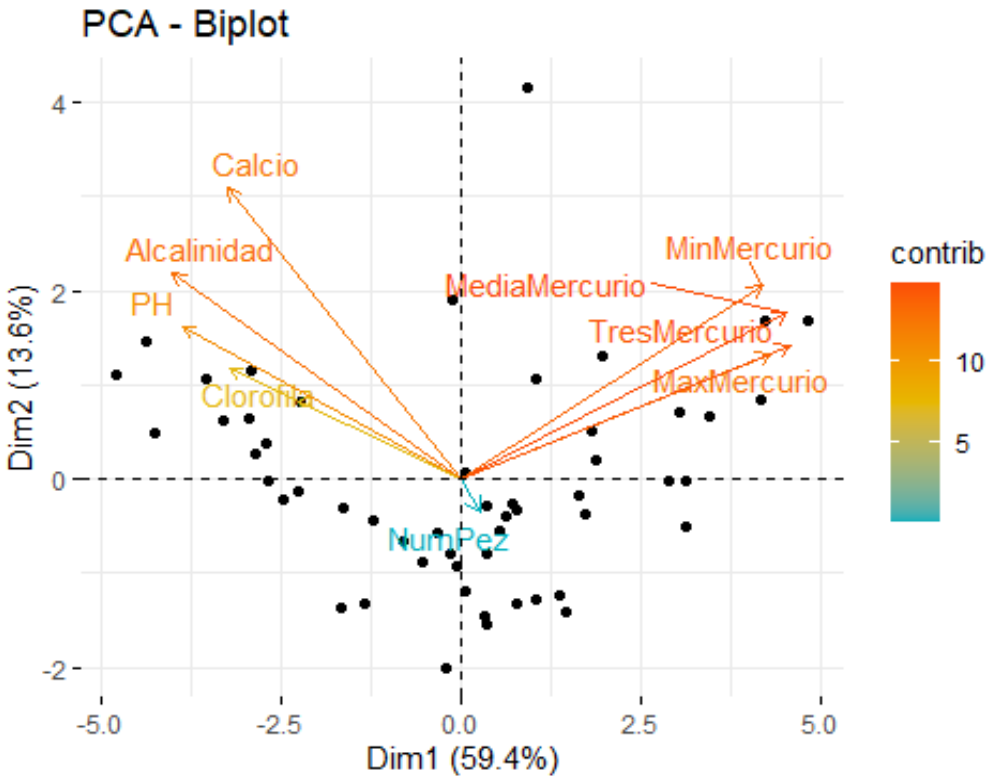
Estos resultados son el porcentaje de contribución de las 9 variables a los 5 componentes.

De igual manera las podemos analizar en una matriz de correlación:



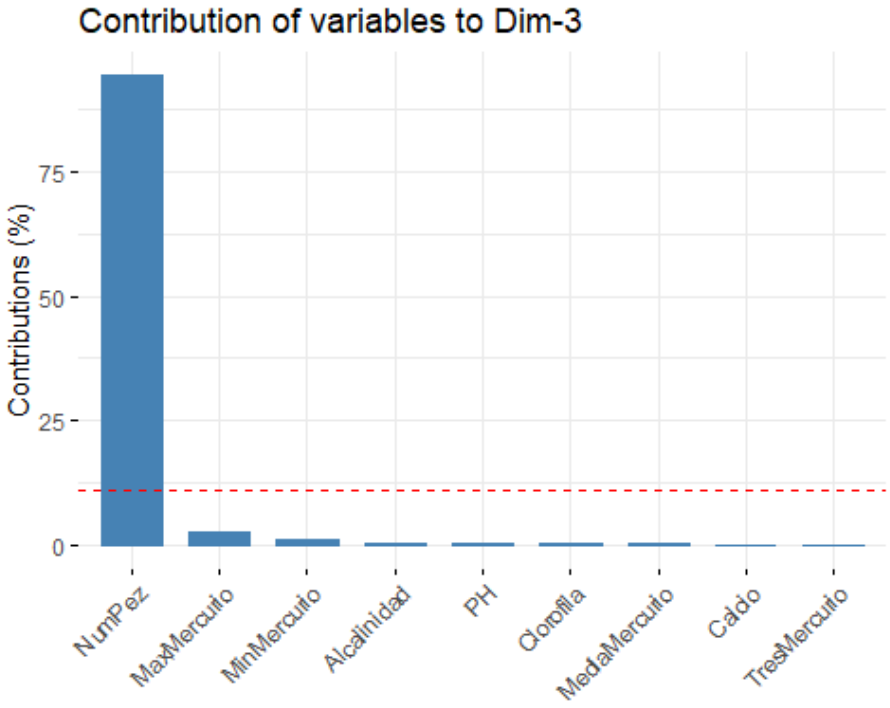
C.

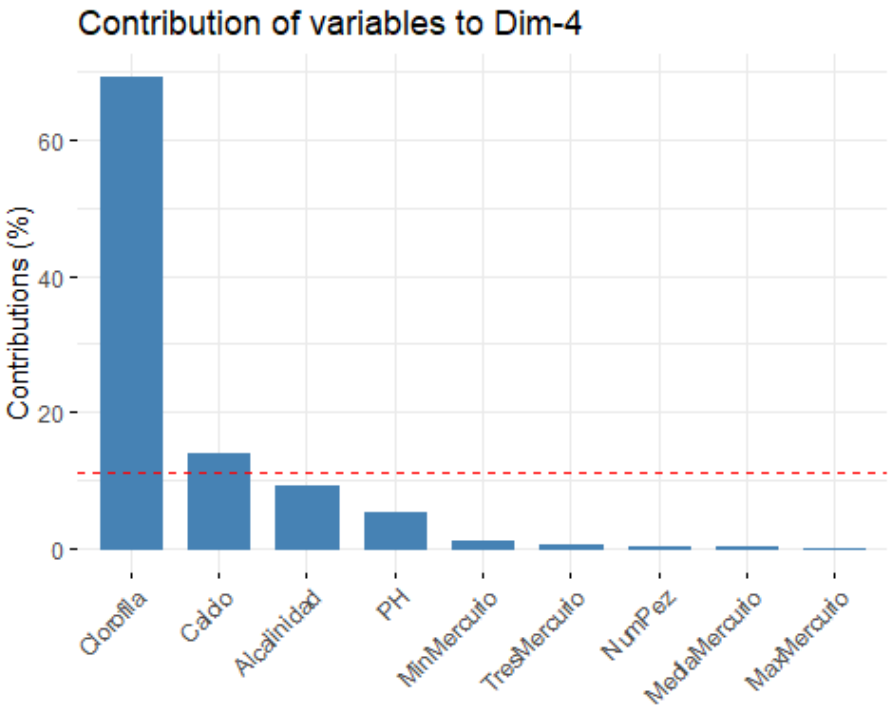
Gráfico de los vectores asociados a las variables y las puntuaciones de las observaciones de los dos primeros componentes.

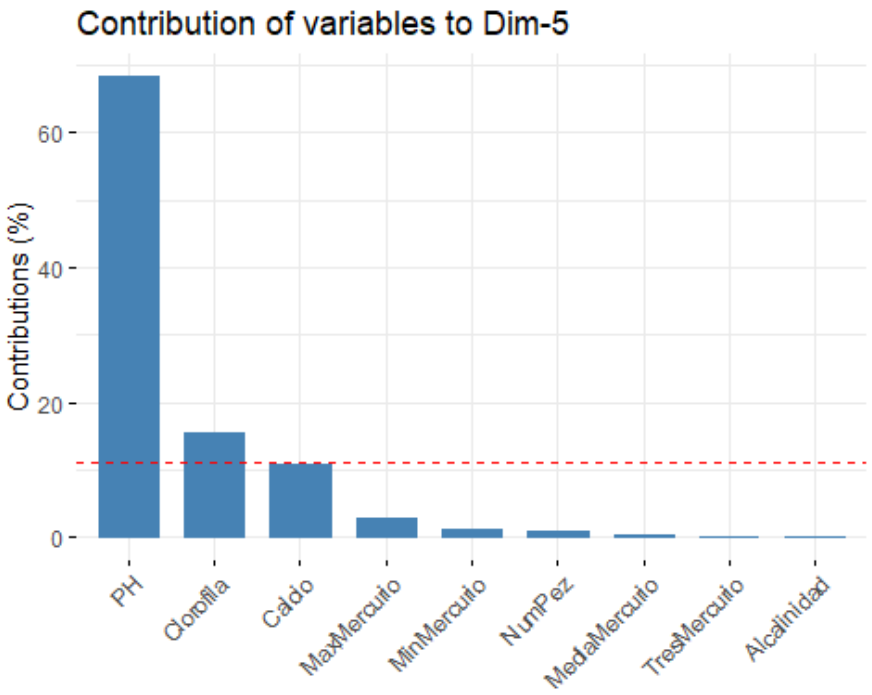


Vemos como todas las variables menos NumPez contribuyen de magnitud similar a la Dimensión 1, y para la Dimensión 2 todas las variables menos NumPez contribuyen con relación positiva, siendo Calcio la que más contribuye a este.

Asimismo, podemos ver las contribuciones de las variables a los demás componentes:







D.

Interpretación de los resultados. Conclusiones del análisis y qué significado tienen los componentes seleccionados en el contexto del problema.

Podemos observar como todas las variables a excepción de NumPez toman parte en explicar los primeros 2 componentes principales, siendo estos los más importantes explicando el 72.96% de la variación, con variables como Alcalinidad, Calcio y MaxMercurio teniendo un ligero mayor impacto en estos 2 componentes.

Sin embargo, los componentes 3, 4 y 5 son mayoritariamente explicados por una sola variable cada uno, siendo NumPez el que contribuye de manera total al componente 3, mayormente Clorofila y poco Calcio para el componente 4, y mayormente PH y poco Clorofila para el componente principal 5.

3. Conclusión general

- **¿De qué forma te ayuda este nuevo análisis a contestar la pregunta principal del estudio: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?**

De este análisis primero encontramos que PH y MaxMercurio son las únicas variables que presentan normalidad multivariada, ya que el conjunto de todas las 9 variables numéricas no presenta una distribución normal multivariada. Esto nos ayuda a determinar que variables podríamos utilizar para buscar generar un mejor modelo que explique los principales factores en el nivel de contaminación por mercurio.
El análisis de componentes principales de igual manera puede ayudar a encontrar una mejor solución al estudio, ya que crear un modelo con los 5 componentes propuestos que explican el 95.69% de la varianza podría ser de gran utilidad.

- **¿En qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas?**

La normalidad encontrada en un grupo de variables detectadas nos ayuda a determinar cómo se distribuyen las variables y qué variables podríamos utilizar para buscar generar un mejor modelo, ya que si no se cuenta con una normalidad multivariada el modelo no es tan confiable.

- **¿Cómo te ayudan los componentes principales a abordar este problema?**

Trabajar con variables que tienen alta correlación entre si llega a ser un problema ya que estas explican cosas similares de la variable dependiente. Cuando realizamos un análisis estadístico deberíamos buscar que las variables independientes tengan correlaciones bajas y expliquen la mayor parte de la varianza total del modelo. Y para esto, en lugar de utilizar las variables en el problema, podemos utilizar un conjunto de componentes principales, los cuales reducen el tamaño de la base, no tienen alta dependencia entre sí, y explican el porcentaje deseado del modelo. En nuestro caso redujimos la base de 9 variables a 5 componentes que explicaran más del 95% de la varianza.

Anexos

Liga al repositorio en Github:

https://github.com/lcanoi/Mercurio_Estadistica