Reporte final de "Los peces y el mercurio"

Edgar Antonio Galarza López A00828688

2022-12-04

Resumen

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se analizarán los datos correspondientes a un estudio hecho por investigadores de Florida (USA) para medir la contaminación de mercurio en peces comestibles. Dicho estudio fue realizado en 53 lagos de la zona para examinar posibles factores relacionados al nivel de contaminación por mercurio en peces y características fisicoquímicas del agua. Mediante un análisis de normalidad de las variables presentadas en el problema, así como un análisis de componentes principales se identificarán los factores que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

Introducción

Se considera una problemática de gran importancia debido a que el pescado es una de las principales fuentes de alimentos dentro de la zona de Florida y en muchas otras partes del mundo, y están sumamente expuestos a ser contaminados con metales pesados como el mercurio por su alta presencia en zonas como ríos, lagos y océanos. Actualmente existen diversos estudios relacionados a los riesgos asociados a la ingesta de los contaminantes del pescado, además de distintas regulaciones por parte de los países que dictaminan el límite máximo permitido de concentración de mercurio en pescados y mariscos, variando desde 0,5 hasta 1,5 ppm. Esto considerando los distintos lagos, océanos, mares, así como dependiendo en la variedad del pescado [1].

Dentro del margen de esta problemática, una de las primeras consideraciones tenemos que las mediciones de concentración de mercurio en los peces son en partes por millón. Se midieron variables de interés de los lagos, como el pH con rangos que van desde: 3.6 a 9.1, alcalinidades que van desde: 1.2 a 128 mg/L como CaCO3. Se midió cada una de las edades a partir de la concentración de mercurio de en el tejido muscular de los. Se asumió que los peces absorben mercurio con el tiempo, los peces más viejos tienden a tener concentraciones más altas.

Tomando en cuenta lo mencionado anteriormente, se utilizarán algunas herramientas como el análisis de normalidad y de componentes principales con la finalidad de identificar las variables principales que intervienen en la contaminación por mercurio de los peces para los 53 lagos de la zona de Florida y se responderán ciertas incógnitas tales como sí es que existe una relación directa con la edad o condiciones del lago.

Análisis de los resultados

1. Análisis de normalidad.

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

Se realiza un análisis de normalidad, con el objetivo de analizar cuánto es que difiere la distribución de los datos observados respecto a lo esperado o si tienden a una distribución normal con misma media y desviación atípica.

El test de hipótesis estadística es:

- Ho: los datos siguen una distribución normal.
- Hα: los datos no siguen una distribución normal.

Esto con un nivel de significancia a=0.05, en dónde si p< α, la prueba estadística es significativa, por lo que no existiría normalidad en los datos.

Comenzamos realizando la prueba de normalidad multivariante Mardia con MVN, obteniendo los siguientes resultados:

```
##              Test        Statistic              p value Result
## 1 Mardia Skewness 474.747945136975 8.64265750182764e-21     NO
## 2 Mardia Kurtosis 3.59794900484948 0.000320736483631068     NO
## 3             MVN             <NA>                  <NA>     NO
```

Podemos observar que ninguna de las pruebas indica normalidad multivariante, por lo que los datos no sigue una distribución normal multivariada a un nivel de significación del 0.05: Skewness p(0.4818) <α (8.64e-21), Kurtosis p(0.0003) < α(0.05).

Por lo tanto, se procede a realizar una prueba adicional de normalidad univariante, considerando como hipótesis nula que los datos si proceden de una distribución normal e hipótesis alternativa si no lo hacen El p-value obtenido en las pruebas indica la probabilidad de obtener una distribución como la observada si los datos proceden realmente de una población con una distribución normal.

Utilizando el test de Anderson Darling tenemos que:

```
## $multivariateNormality
##      Test       H     p value MVN
## 1 Royston 122.473 3.25128e-24  NO
##
## $univariateNormality
##                Test             Variable Statistic   p value Normality
## 1  Anderson-Darling     Alcalinidad         3.6725  <0.001      NO
```

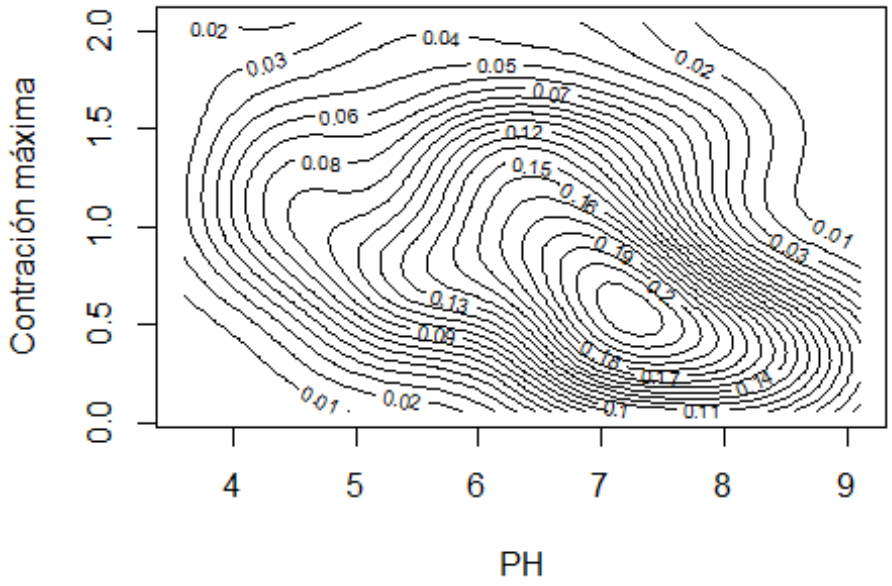
```
## 2  Anderson-Darling          PH              0.3496  0.4611       YES
## 3  Anderson-Darling        Calcio            4.0510  <0.001       NO
## 4  Anderson-Darling      Clorofila           5.4286  <0.001       NO
## 5  Anderson-Darling         CMM              0.9253  0.0174       NO
## 6  Anderson-Darling       N Peces            8.6943  <0.001       NO
## 7  Anderson-Darling Concentración mínima     1.9770  <0.001       NO
## 8  Anderson-Darling  Contración máxima       0.6585   0.081       YES
## 9  Anderson-Darling   Concentración 3a       1.0469  0.0086       NO
## 10 Anderson-Darling         Edad            14.3350  <0.001       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
## CMM                   53  0.5271698  0.3410356   0.48 0.04   1.33  0.27
0.77
## N Peces               53 13.0566038  8.5606773  12.00 4.00  44.00 10.00
12.00
## Concentración mínima 53  0.2798113  0.2264058   0.25 0.04   0.92  0.09
0.33
## Contración máxima     53  0.8745283  0.5220469   0.84 0.06   2.04  0.48
1.33
## Concentración 3a      53  0.5132075  0.3387294   0.45 0.04   1.53  0.25
0.70
## Edad                  53  0.8113208  0.3949977   1.00 0.00   1.00  1.00
1.00
##                            Skew   Kurtosis
## Alcalinidad           0.9679170 -0.4705349
## PH                   -0.2458771 -0.6239638
## Calcio                1.3045868  0.6130359
## Clorofila             2.4130571  6.1042185
## CMM                   0.5986343 -0.6312607
## N Peces               2.5808773  6.0089455
## Concentración mínima  1.0729099  0.4060828
## Contración máxima     0.4645925 -0.6692490
## Concentración 3a      0.9449951  0.5733500
## Edad                 -1.5465748  0.4005116
```

B. Prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores. Interpreta los resultados obtenidos con base en ambas pruebas y en la interpretación del sesgo y la curtosis de cada una de ellas.

```
##     PH Contración máxima
## 1  6.1               1.43
## 2  5.1               1.90
## 3  9.1               0.06
## 4  6.9               0.84
## 5  4.6               1.50
## 6  7.3               0.48
## 7  5.4               0.72
## 8  8.1               0.38
## 9  5.8               1.40
## 10 6.4               1.47
## 11 5.4               0.86
## 12 7.2               0.73
## 13 7.2               1.01
## 14 5.8               2.03
## 15 7.6               0.11
## 16 8.2               0.18
## 17 8.7               0.43
## 18 7.8               1.50
## 19 5.8               1.33
## 20 6.7               1.44
## 21 4.4               0.93
## 22 6.7               0.94
## 23 6.1               0.61
## 24 6.9               2.04
## 25 5.5               0.62
## 26 6.9               1.12
## 27 7.3               0.52
## 28 4.5               1.38
## 29 4.8               0.84
## 30 5.8               0.69
## 31 7.8               0.59
## 32 7.4               0.65
## 33 3.6               1.90
## 34 4.4               1.02
## 35 7.9               0.30
## 36 7.1               0.29
## 37 6.8               0.37
## 38 8.4               0.06
## 39 7.0               0.63
## 40 7.5               1.41
## 41 7.0               0.26
## 42 6.8               0.26
## 43 5.9               1.05
## 44 8.3               0.48
```

```
## 45 6.7              1.40
## 46 6.2              0.95
## 47 6.2              1.10
## 48 8.9              0.40
## 49 4.3              1.24
## 50 7.0              0.90
## 51 6.9              0.69
## 52 5.2              0.40
## 53 7.9              0.51

##              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

## $multivariateNormality
##      Test        H   p value MVN
## 1 Royston 3.924798 0.1210984 YES
##
## $univariateNormality
##               Test          Variable Statistic   p value Normality
## 1 Anderson-Darling        PH            0.3496    0.4611    YES
## 2 Anderson-Darling Contración máxima    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
## Contración máxima 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33
0.4645925
##                      Kurtosis
## PH                 -0.6239638
## Contración máxima -0.6692490
```

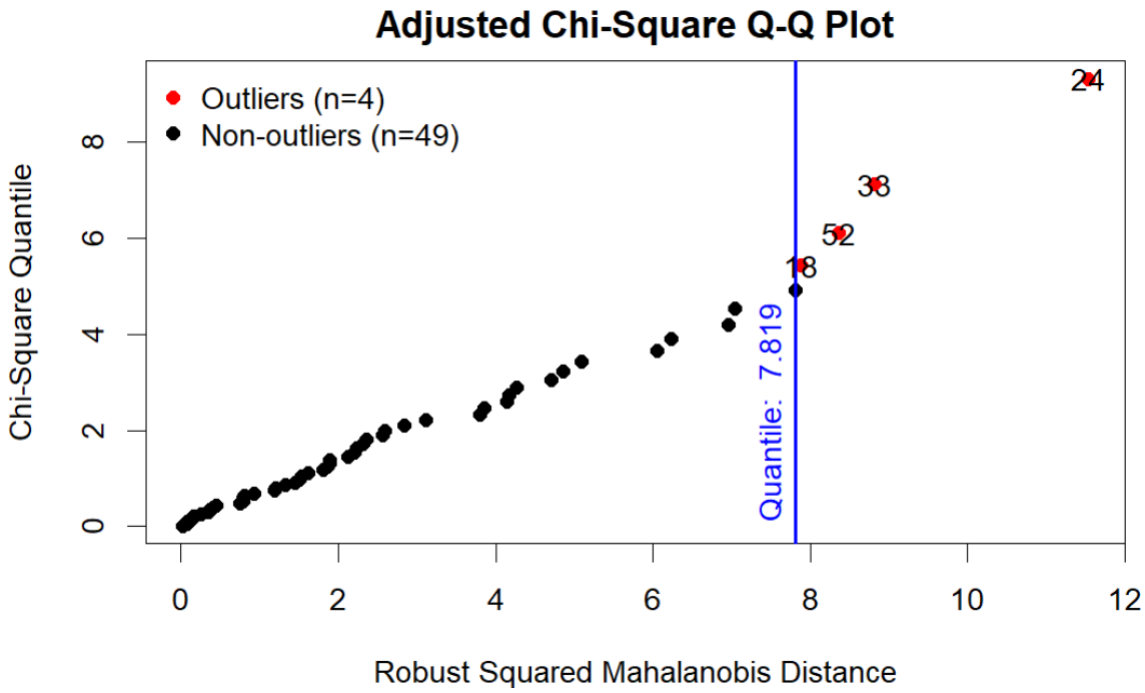
C. Gráfica de contorno de la normal multivariada obtenida en el inciso B.



Como podemos analizar en la gráfica de contorno podemos decir que a cierta medida existe correlación entre el PH y la concentración máxima de mercurio en los peces, esto debido a que dichas líneas de contorno se encuentran alrededor de la diagonal principal. Si dicha correlación fuese 0, las líneas de contorno serían circulares en lugar de tornarse elipsoidales.

```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    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 Contración máxima    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
## Contración máxima 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33
0.4645925
##                      Kurtosis
## PH                -0.6239638
## Contración máxima -0.6692490
```

D. Datos atípicos o influyentes en la normal multivariada encontrada en el inciso B (auxíliate de la distancia de Mahalanobis y del gráfico QQplot multivariado)



Después de realizar el cálculo de la distancia de Mahalanobis ajustada, se encontraron 4 datos atípicos multivariados, por lo que esto será importante considerar antes de realizar una suposición MVN, dado que dicha prueba requiere la ausencia de valores atípicos multivariados.

2. Análisis de componentes principales.

A. Justifique por qué es adecuado el uso de componentes principales para analizar la base (haz uso de la matriz de correlaciones).

El análisis de componentes principales nos sirve mucho en situaciones como estas en las que tenemos una extensa cantidad de variables, por lo que nos ayuda a seleccionar las que poseen una mejor compatibilidad o correlación entre sí.

```
##                      Alcalinidad          PH       Calcio   Clorofila
## Alcalinidad           1.00000000  0.71916568  0.832604192  0.47753085
## PH                    0.71916568  1.00000000  0.577132721  0.60848276
## Calcio                0.83260419  0.57713272  1.000000000  0.40991385
## Clorofila             0.47753085  0.60848276  0.409913846  1.00000000
## CMM                  -0.59389671 -0.57540012 -0.400679584 -0.49137481
## N Peces               0.01029074 -0.01860607 -0.089379013 -0.01182027
## Concentración mínima -0.52535654 -0.54196524 -0.332476229 -0.40045856
```

```
## Contración máxima    -0.60479558 -0.55181523 -0.407916635 -0.48497215
## Concentración 3a     -0.62795845 -0.61284905 -0.464409465 -0.50644193
## Edad                 -0.09493882  0.03800021 -0.002111124 -0.28300234
##                             CMM     N Peces Concentración mínima
## Alcalinidad          -0.59389671  0.01029074          -0.52535654
## PH                   -0.57540012 -0.01860607          -0.54196524
## Calcio               -0.40067958 -0.08937901          -0.33247623
## Clorofila            -0.49137481 -0.01182027          -0.40045856
## CMM                   1.00000000  0.07903426           0.92720506
## N Peces               0.07903426  1.00000000          -0.08165278
## Concentración mínima  0.92720506 -0.08165278           1.00000000
## Contración máxima     0.91586397  0.16109174           0.76535319
## Concentración 3a      0.95921481  0.02580046           0.91908939
## Edad                  0.10873896  0.20795617           0.10066197
##                     Contración máxima Concentración 3a         Edad
## Alcalinidad               -0.60479558      -0.62795845 -0.094938825
## PH                        -0.55181523      -0.61284905  0.038000214
## Calcio                    -0.40791663      -0.46440947 -0.002111124
## Clorofila                 -0.48497215      -0.50644193 -0.283002338
## CMM                        0.91586397       0.95921481  0.108738958
## N Peces                    0.16109174       0.02580046  0.207956171
## Concentración mínima       0.76535319       0.91908939  0.100661967
## Contración máxima          1.00000000       0.85975810  0.093752072
## Concentración 3a           0.85975810       1.00000000  0.089411267
## Edad                       0.09375207       0.08941127  1.000000000
```

B. Realiza el análisis de componentes principales y justifica el número de componentes principales apropiados para reducir la dimensión de la base.

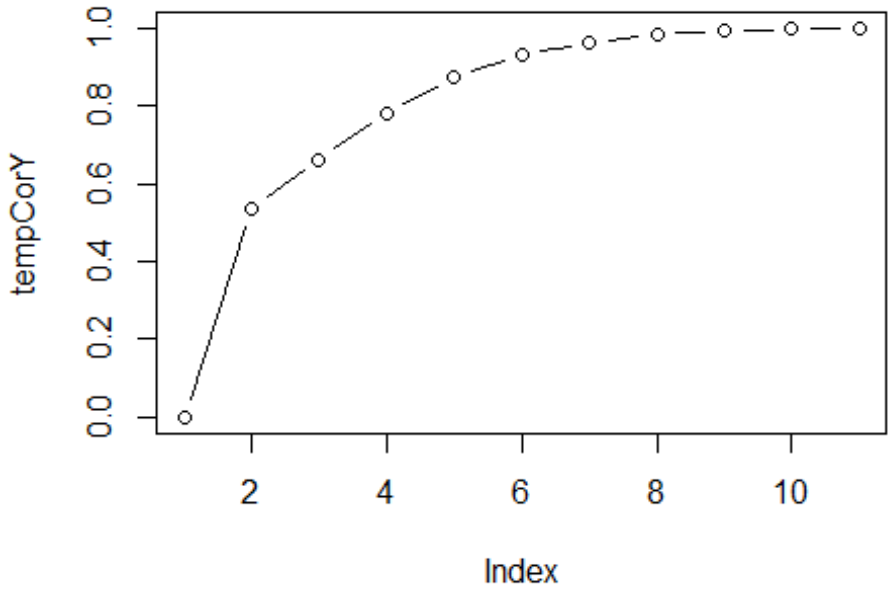
Se debe considerar realizar el análisis de componentes principales utilizando la matriz de correlaciones (R), en donde las componentes principales no dependerán de las unidades de las variables debido a que se encontrarán ya estandarizadas.

```
##
##
## Valores y vectores eigen de la correlación

## eigen() decomposition
## $values
##  [1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741
##  [7] 0.20673634 0.08682133 0.05163902 0.01863699
##
## $vectors
##              [,1]        [,2]       [,3]         [,4]        [,5]
[,6]
##  [1,] -0.35065869 -0.21691594 -0.3472906  0.009131194  0.34050534
0.07547497
##  [2,] -0.33700381 -0.21940887 -0.2360975 -0.017242162 -0.39396038
0.73121012
##  [3,] -0.28168286 -0.26250672 -0.5113780  0.146950070  0.36205937 -
```

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

##
##
## Proporción de varianza explicada

##  [1] 0.536122641 0.125426109 0.121668138 0.090943267 0.059141736
0.030314741
##  [7] 0.020673634 0.008682133 0.005163902 0.001863699

##
##
## Proporción de varianza acumulada

##  [1] 0.5361226 0.6615488 0.7832169 0.8741602 0.9333019 0.9636166
0.9842903
##  [8] 0.9929724 0.9981363 1.0000000
```

Analizando el primer componente se tiene que las variables que más influyen dentro del modelo es la 9 y 5, mientras que para la segunda componente las variables que más influyen son la 6 y la 10, esto debido que explican el mayor porcentaje de variabilidad. Los resultados los podemos verificar con los valores y vectores propios.

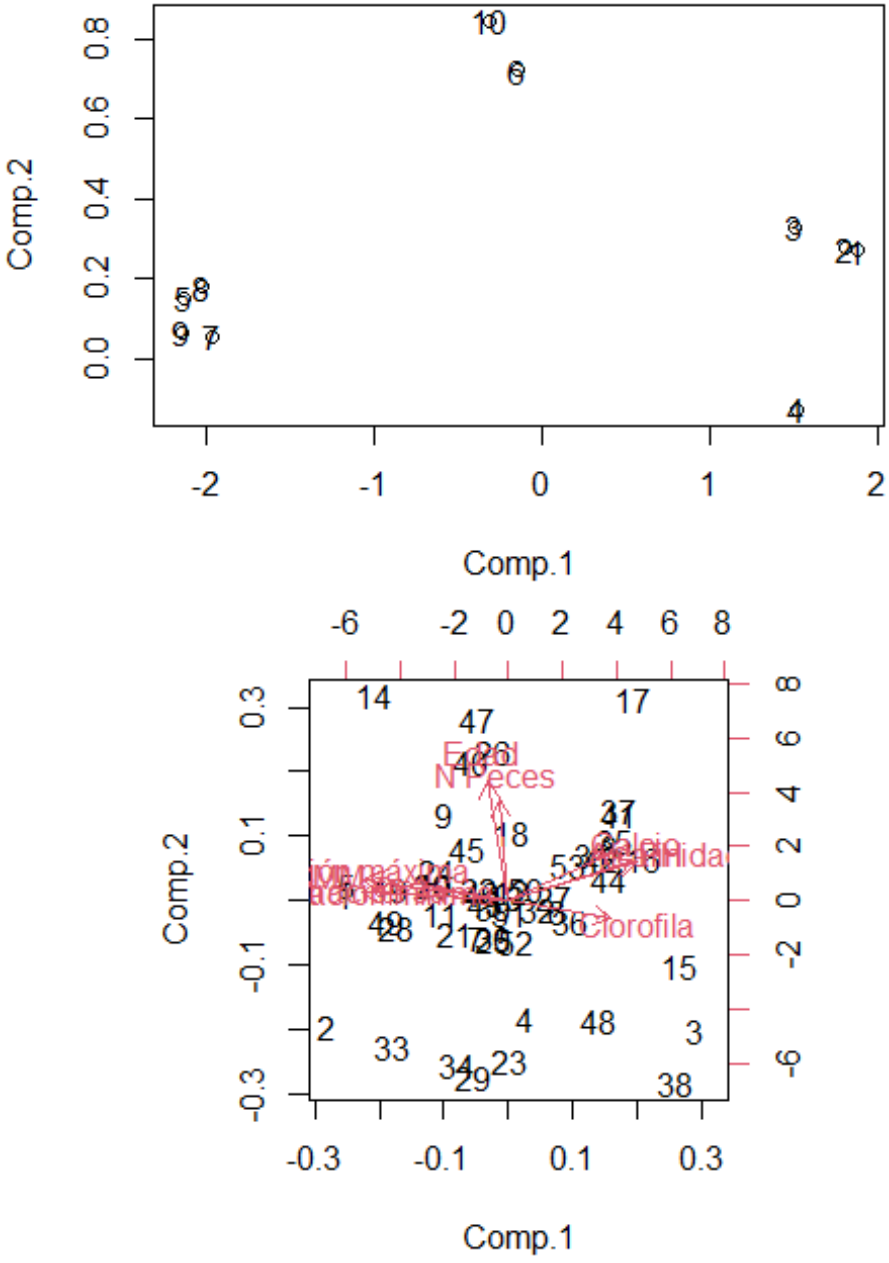
C. Representa en un gráfico los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes

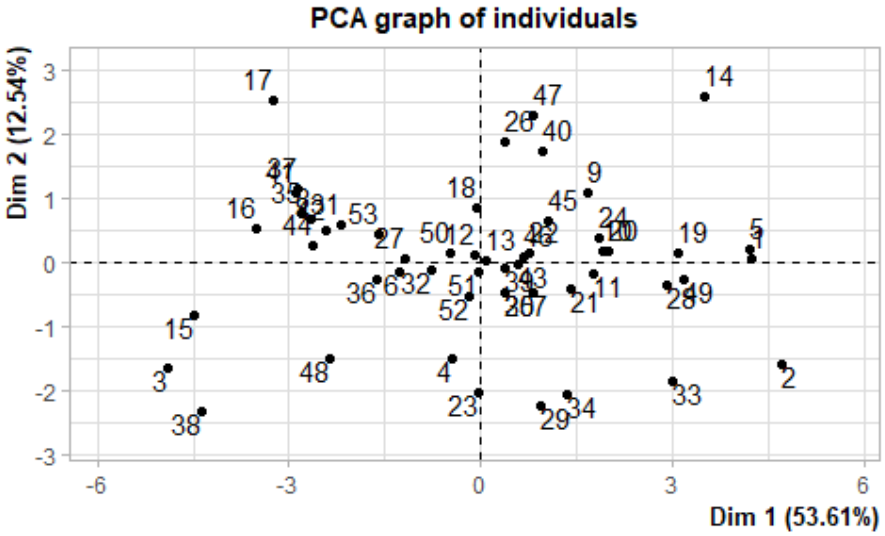


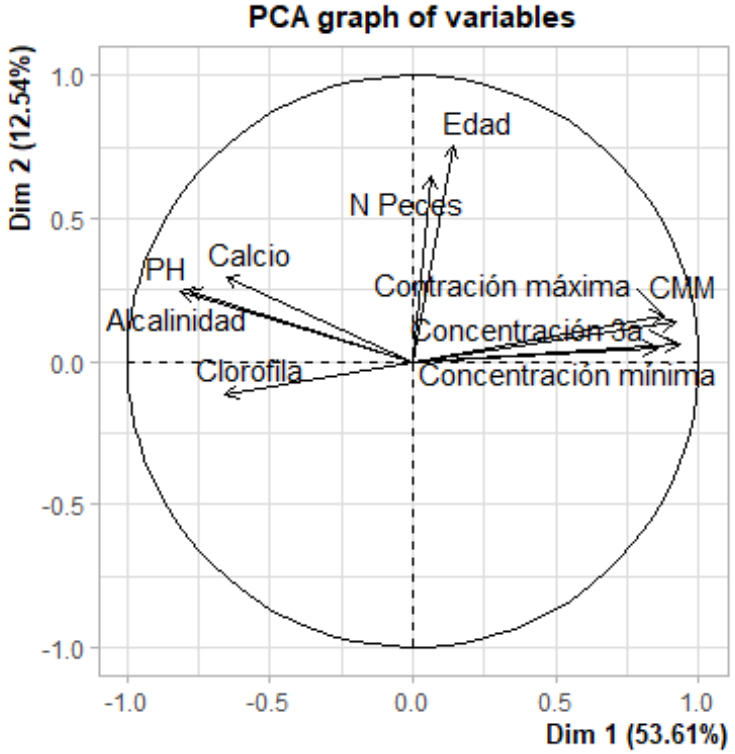
Dentro de las 2 primeras componentes se observa que se explica casi un 60% de la variabilidad de los datos, lo cual indica que es un buen modelo.

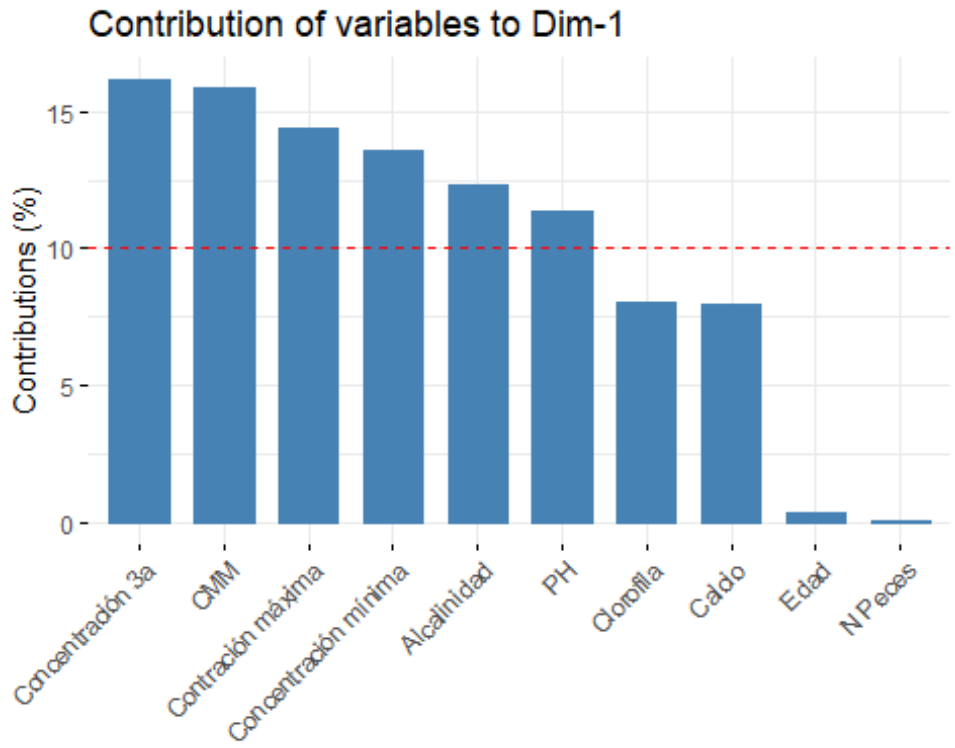
```
## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

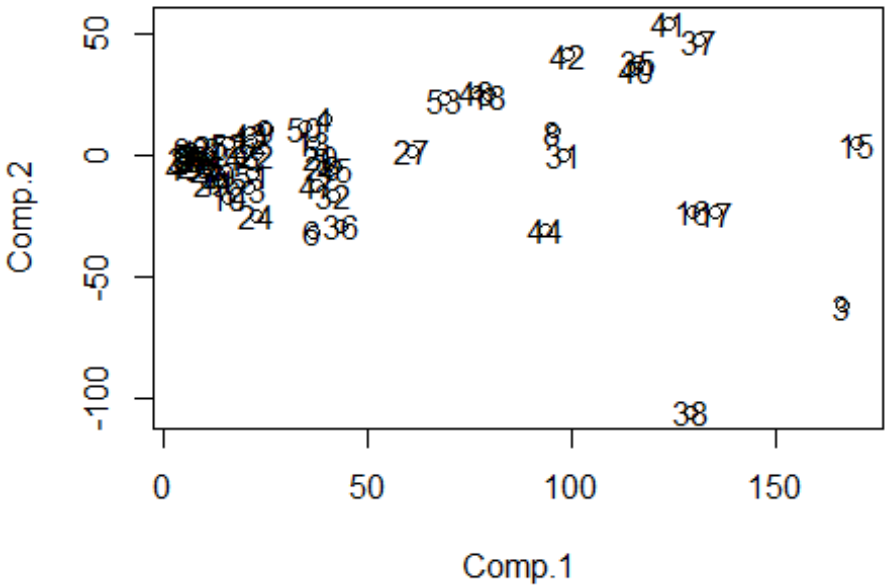
Cor




PCA graph of individuals
Dim 2 (12.54%)
Dim 1 (53.61%)


PCA graph of variables
Dim 2 (12.54%)
Edad
N Peces
PH
Calcio
Contración máxima
CMM
Alcalinidad
Concentración 3a
Clorofila
Concentración mínima
Dim 1 (53.61%)





Según las gráficas obtenidas podemos ver la representación de las principales componentes y el porcentaje de variabilidad explicado por las dos primeras componentes. Se puede verificar la variabilidad de los datos, así como su comportamiento.

Conclusión

Una vez finalizado el análisis se tiene que los principales factores que influyen a las concentraciones dentro de los peces para los lagos de Florida se deben al nivel de alcalinidad y PH dentro del agua. El análisis de normalidad en el grupo de variables resulta ser adecuado para encontrar los datos que mejor se pueden ajustar a un modelo de PCA. Se encontró principalmente que no todas las variables de la base de datos se distribuían normalmente, por lo que al final se hizo el PCA con toda la información, encontrando información suficiente para encontrar lo que más influye en la variabilidad de los datos.

Finalmente sabemos que la alcalinidad y el PH son lo que más influyen a las concentraciones de mercurio de los peces, por lo que ahora se puede brindar información de calidad en caso de que se proponga alertar al gobierno de Florida para controlar dichos niveles y no afecte a la población general.

Referencias bibliográficas

Bartlett, M. S. (1947). Multivariate analysis. *Supplement to the journal of the royal statistical society*, *9*(2), 176-197.

Gurrea, M. (2000). Análisis de componentes principales. *Proyecto e-Math Financiado por la Secretaría de Estado de Educación y Universidades (MECD)*.

Anexos

https://drive.google.com/file/d/1mYPTjA2KE0BEoFmh4itOTEVHX0tZMtuO/view?usp=sharing