Los peces y el mercurio: Procesamiento de datos multivariados

Carlos David Contreras Chacon

2022-12-02

Resumen

En el presente reporte se realizó un análisis de las variables involucradas en un estudio reciente de variables medidas en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. Se realizaron pruebas de hipótesis para evaluar si el mercurio se podía predecir con las mediciones en los peces y para elegir un modelo adecuado se utilizó la prueba de ANOVA y un análisis multivariable. Luego del análisis se concluyó que a las variables que realmente afecta el nivel de mercurio son la clorofila y el calcio, a menor cantidad de ellos mayor nivel de mercurio medido.

Introducción

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. Las variables que se midieron se encuentran en la siguiente base de datos:

- X1 = número de identificación
- X2 = nombre del lago
- X3 = alcalinidad (mg/l de carbonato de calcio)
- X4 = PH
- X5 = calcio (mg/l)
- X6 = clorofila (mg/l)
- X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago
- X8 = número de peces estudiados en el lago
- X9 = mínimo de la concentración de mercurio en cada grupo de peces
- X10 = máximo de la concentración de mercurio en cada grupo de peces
- X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
- X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

En el reporte presentamos un análisis análisis de normalidad de las variables continuas para identificar la normalidad de las mismas, asi como un análisis de componentes principales utilizando todos los datos para identificar que factores son los mas importantes en el problema.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Análisis de resultados

Análisis de normalidad

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.

Realizaremos un análisis de normalidad de las variables con la prueba de Mardia, que nos ayuda a analizar como cambian las distribuciones de los datos con respecto a lo que esperaríamos con una distribución normal estándar.

La prueba de hipótesis es: Ho: los datos siguen una distribución normal Ha: los datos no siguen una distribución normal

Trabajaremos con una significancia de a=0.05, sabemos que si p<a, la prueba estadística es significativa, por lo que no habría normalidad en los datos.

Realizamos la prueba de normalidad multivariable Mardia con MVN:

```
##              Test        Statistic              p value Result
## 1 Mardia Skewness 502.667343452414  3.6277693977554e-24     NO
## 2 Mardia Kurtosis 4.83254138772002 1.34801075923896e-06     NO
## 3             MVN             <NA>                 <NA>     NO
```

Observamos que ninguna de las pruebas indica normalidad multivariante, así que no hay distribución normal multivariada con un nivel de significancia de 0.05.

Ahora realizaremos una prueba adicional de normalidad por cada variable: Ho: Los datos si proceden de una distribución normal Ha: No lo hacen

Usando ahora el test de Anderson Darling:

```
## $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  Mercurio      0.9253  0.0174      NO
## 6  Anderson-Darling  NumPeces      8.6943  <0.001      NO
## 7  Anderson-Darling MercurioMin    1.9770  <0.001      NO
## 8  Anderson-Darling MercurioMax    0.6585   0.081      YES
## 9  Anderson-Darling MercurioEst    1.0469  0.0086      NO
## 10 Anderson-Darling  EdadPeces    14.3350  <0.001      NO
##
## $Descriptives
##              n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew
## Alcalinidad 53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170
## PH          53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771
## Calcio      53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868
## Clorofila   53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571
## Mercurio    53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343
## NumPeces    53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773
## MercurioMin 53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33  1.0729099
## MercurioMax 53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33  0.4645925
## MercurioEst 53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70  0.9449951
## EdadPeces   53  0.8113208  0.3949977   1.00 0.00   1.00  1.00  1.00 -1.5465748
##               Kurtosis
## Alcalinidad -0.4705349
## PH          -0.6239638
## Calcio       0.6130359
## Clorofila    6.1042185
## Mercurio    -0.6312607
## NumPeces     6.0089455
## MercurioMin  0.4060828
## MercurioMax -0.6692490
## MercurioEst  0.5733500
## EdadPeces    0.4005116
```

Viendo los resultados de la prueba, tenemos que el PH y que el Mercurio Max si se comportan como una variable normal.

Realiza la 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.

Realizamos la prueba de normalidad multivariable Mardia con MVN ahora solo con las variables que pasaron la prueba de normalidad:

```
##              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
```

Observamos que ambas de las pruebas indican normalidad multivariante, así que si con un nivel de significancia de 0.05 hay distribución normal multivariada.

Ahora realizaremos una prueba adicional de normalidad por cada variable: Ho: Los datos si proceden de una distribución normal Ha: No lo hacen

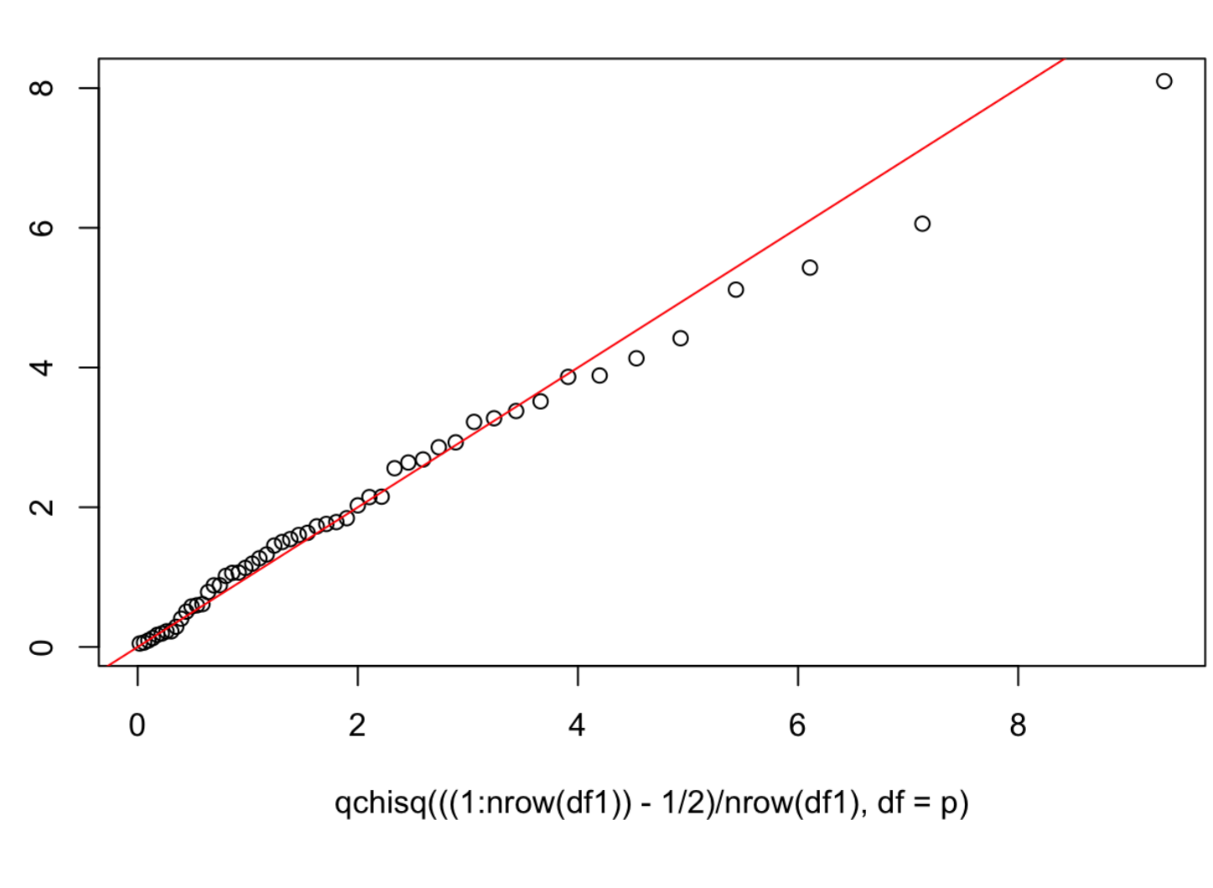
Usando ahora el test de Anderson Darling:

```
## $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 MercurioMax    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
## MercurioMax 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
##               Kurtosis
## PH          -0.6239638
## MercurioMax -0.6692490
```

Los resultados son los mismos que en la anterior vez, pero ahora comprobamos que se comportan tanto de manera independiente como de manera conjunta como una distribución normal

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

Detección datos atípicos o influyentes en la normal multivariada encontrada en el inciso B



Luego de hacer la prueba vemos que la mayoría de los datos caen en una buena posición y aunque hay algunos que se alejan ligeramente de la linea objetivo, son mas bien pocos.

Análisis de componentes principales

El uso de componentes principales para analizar la base

Es importante que podamos buscar utilizar las variables con mayor importancia dentro de los datos, ya que al utilizar todas podemos introducir ruido y hacer un análisis mucho peor. Utilizaremos la matriz de correlaciones para evaluar como se afectan unas variables a las otras y viendo las de mayor correlación podemos disminuir el conjunto de variables.

```
##             Alcalinidad          PH       Calcio   Clorofila    Mercurio
## Alcalinidad  1.00000000  0.71916568  0.832604192  0.47753085 -0.59389671
## PH           0.71916568  1.00000000  0.577132721  0.60848276 -0.57540012
## Calcio       0.83260419  0.57713272  1.000000000  0.40991385 -0.40067958
## Clorofila    0.47753085  0.60848276  0.409913846  1.00000000 -0.49137481
## Mercurio    -0.59389671 -0.57540012 -0.400679584 -0.49137481  1.00000000
## NumPeces     0.01029074 -0.01860607 -0.089379013 -0.01182027  0.07903426
## MercurioMin -0.52535654 -0.54196524 -0.332476229 -0.40045856  0.92720506
## MercurioMax -0.60479558 -0.55181523 -0.407916635 -0.48497215  0.91586397
## MercurioEst -0.62795845 -0.61284905 -0.464409465 -0.50644193  0.95921481
## EdadPeces   -0.09493882  0.03800021 -0.002111124 -0.28300234  0.10873896
##                NumPeces MercurioMin MercurioMax MercurioEst    EdadPeces
## Alcalinidad  0.01029074 -0.52535654 -0.60479558 -0.62795845 -0.094938825
## PH          -0.01860607 -0.54196524 -0.55181523 -0.61284905  0.038000214
## Calcio      -0.08937901 -0.33247623 -0.40791663 -0.46440947 -0.002111124
## Clorofila   -0.01182027 -0.40045856 -0.48497215 -0.50644193 -0.283002338
## Mercurio     0.07903426  0.92720506  0.91586397  0.95921481  0.108738958
## NumPeces     1.00000000 -0.08165278  0.16109174  0.02580046  0.207956171
## MercurioMin -0.08165278  1.00000000  0.76535319  0.91908939  0.100661967
## MercurioMax  0.16109174  0.76535319  1.00000000  0.85975810  0.093752072
## MercurioEst  0.02580046  0.91908939  0.85975810  1.00000000  0.089411267
## EdadPeces    0.20795617  0.10066197  0.09375207  0.08941127  1.000000000
```

Análisis de componentes principales

A continuación calcularemos que variables influyen mas en las respectivas componentes, utilizando los valores propios de la matriz de correlaciones

```
## 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
```

```
##             Alcalinidad          PH     Calcio  Clorofila    Mercurio
## Alcalinidad 1459.509456 35.39971335 793.065711 562.193324 -7.73773984
## PH            35.399713  1.66010160  18.540018  24.159971 -0.25283491
## Calcio       793.065711 18.54001814 621.633266 314.949198 -3.40693687
## Clorofila    562.193324 24.15997097 314.949198 949.645668 -5.16408563
## Mercurio      -7.737740 -0.25283491  -3.406937  -5.164086  0.11630530
## NumPeces       3.365566 -0.20522496 -19.077032  -3.118287  0.23074020
## MercurioMin   -4.544071 -0.15809797  -1.876788  -2.793997  0.07159176
## MercurioMax  -12.062062 -0.37116800  -5.309432  -7.802021  0.16305729
## MercurioEst   -8.126195 -0.26746916  -3.922122  -5.286440  0.11080733
## EdadPeces     -1.432656  0.01933962  -0.020791  -3.444811  0.01464804
##                NumPeces  MercurioMin  MercurioMax MercurioEst    EdadPeces
## Alcalinidad  3.36556604 -4.544071118 -12.06206241 -8.12619485 -1.432656023
## PH          -0.20522496 -0.158097968  -0.37116800 -0.26746916  0.019339623
## Calcio     -19.07703193 -1.876788099  -5.30943179 -3.92212155 -0.020791001
## Clorofila   -3.11828737 -2.793996734  -7.80202068 -5.28644013 -3.444811321
## Mercurio     0.23074020  0.071591763   0.16305729  0.11080733  0.014648041
## NumPeces    73.28519594 -0.158258345   0.71993106  0.07481495  0.703193033
## MercurioMin -0.15825835  0.051259579   0.09046049  0.07048523  0.009002177
## MercurioMax  0.71993106  0.090460486   0.27253295  0.15203327  0.019332366
## MercurioEst  0.07481495  0.070485232   0.15203327  0.11473759  0.011962990
## EdadPeces    0.70319303  0.009002177   0.01933237  0.01196299  0.156023222
```

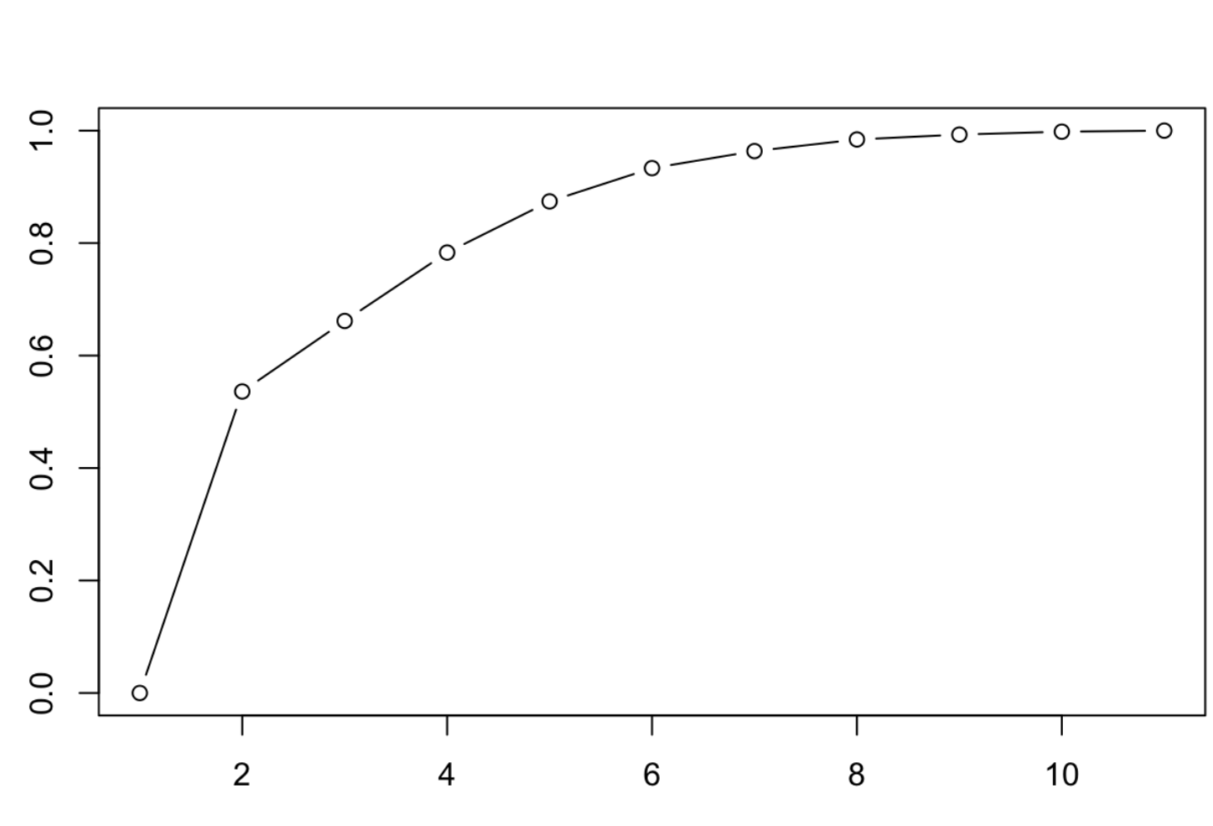
```
## eigen() decomposition
## $values
##  [1] 2.256462e+03 6.326380e+02 1.473944e+02 6.887516e+01 6.737500e-01
##  [6] 2.533448e-01 1.124379e-01 2.884345e-02 4.589790e-03 1.908078e-03
##
## $vectors
##                [,1]          [,2]         [,3]          [,4]         [,5]
##  [1,]  0.770051693 -0.3595624308  0.512212889 -1.211960e-01 -0.022379095
##  [2,]  0.020607395  0.0064777176  0.013445136 -7.815087e-03  0.970898339
##  [3,]  0.459103405 -0.2606043150 -0.824602375  2.029292e-01  0.005136564
##  [4,]  0.442396618  0.8959523487 -0.034896812  8.682740e-03 -0.014101339
##  [5,] -0.004349946 -0.0015153985 -0.006280655  6.263085e-03 -0.070098841
##  [6,] -0.003461490  0.0017191525  0.236827940  9.713807e-01  0.003620863
##  [7,] -0.002482186 -0.0006039064 -0.004911645 -6.911159e-05 -0.061772060
##  [8,] -0.006732314 -0.0020103109 -0.009275841  1.487778e-02 -0.076192576
##  [9,] -0.004611177 -0.0012561924 -0.004970382  3.212703e-03 -0.079787785
## [10,] -0.001169630 -0.0040547980 -0.002920752  1.197391e-02  0.188987340
##               [,6]         [,7]         [,8]          [,9]         [,10]
##  [1,] -0.011296594  0.007055454 -0.002545190 -1.581653e-05  0.0001951853
##  [2,] -0.152551489 -0.173065657  0.057214145  3.485511e-03 -0.0117752811
##  [3,]  0.009060983 -0.006678314  0.002454478 -9.773667e-04 -0.0002013405
##  [4,] -0.003742677  0.006784090 -0.001424066 -5.372560e-04  0.0004978075
##  [5,] -0.470036339  0.066615441  0.279756491  3.236927e-01  0.7658810638
##  [6,]  0.009949302 -0.011307227  0.008472817  9.632211e-04 -0.0020987371
##  [7,] -0.292267405  0.110635778  0.446372646  5.796271e-01 -0.6027245954
##  [8,] -0.693362599 -0.035769112 -0.693360284 -3.142457e-04 -0.1762334539
##  [9,] -0.432229792  0.082492810  0.471915643 -7.474356e-01 -0.1363231520
## [10,]  0.049765942  0.972120176 -0.125200363 -2.410429e-02  0.0190692867
```

También podemos calcular la proporción de varianza explicada por cada componente

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

Gráfico los vectores asociados a las variables y las puntuaciones de las observaciones

Igualmente podemos podemos ver las puntuaciones de los componentes principales y su nivel de explicabilidad de manera gráfica para una mejor comprensión.

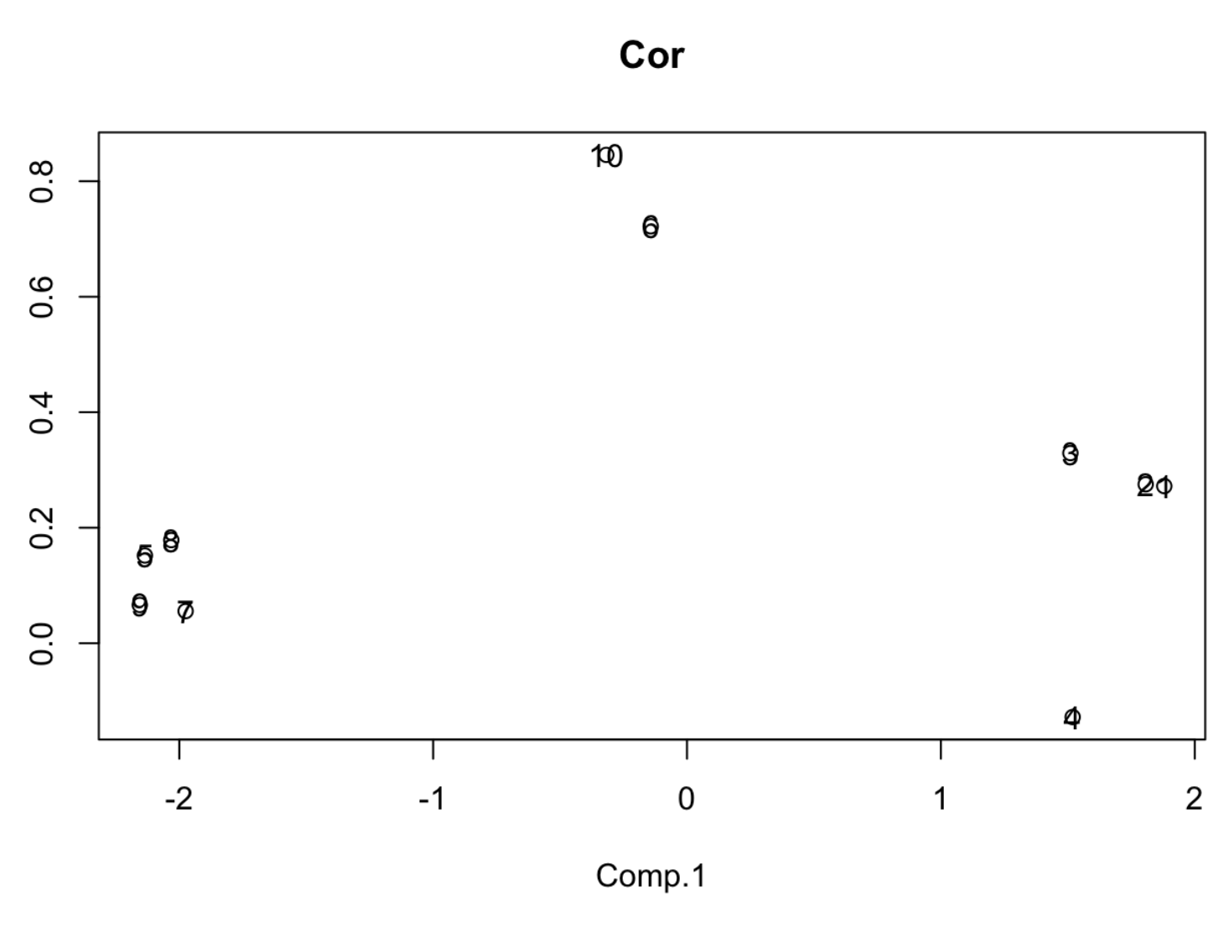


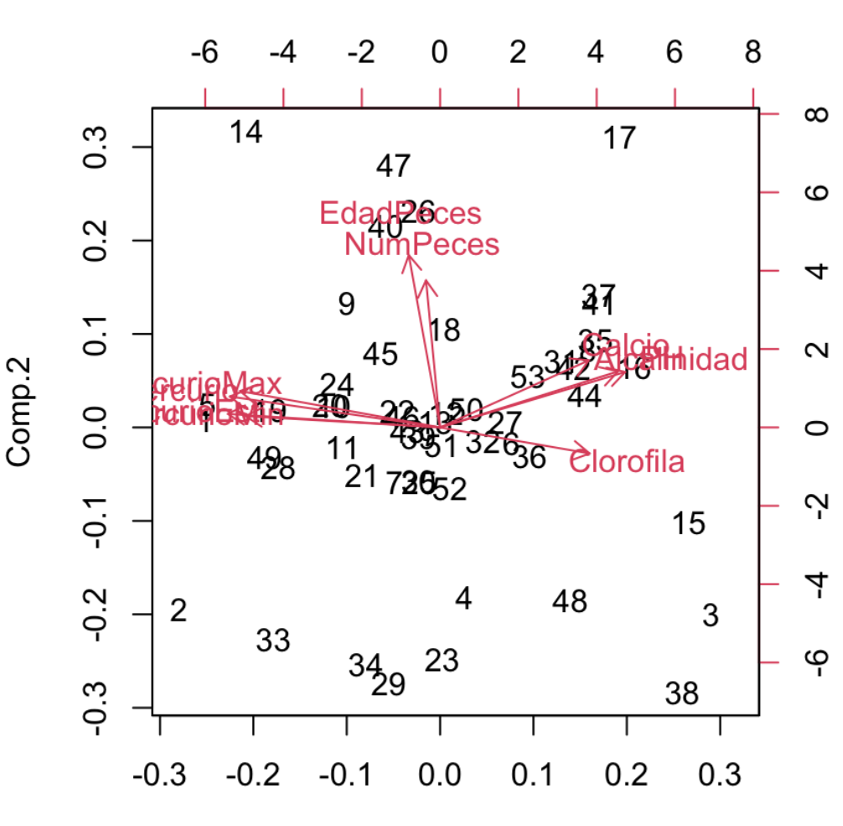
Viendo la gráfica observamos que para la 6a componente ya se tiene mas del 0.9 explicado pero para las primeras dos solo hay aproximadamente un 55% de nivel.

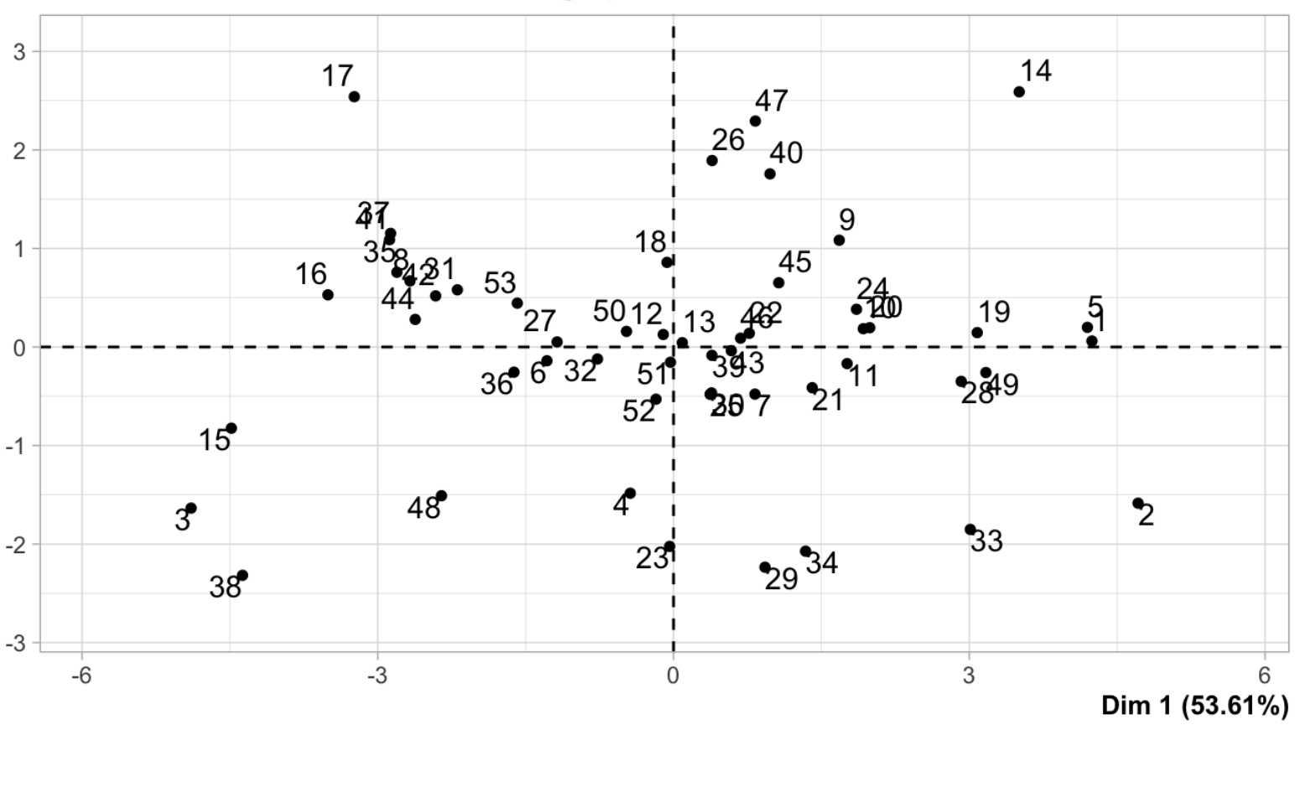
Interpretación de resultados

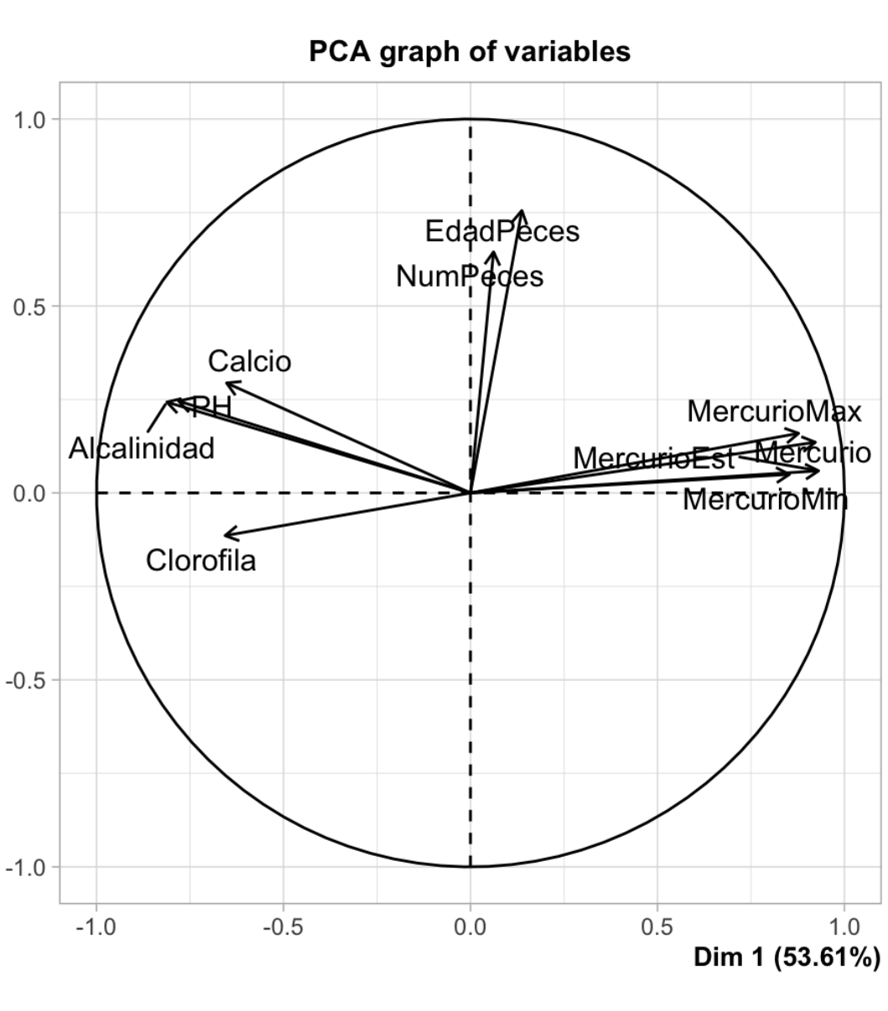
```
## Loading required package: ggplot2
```

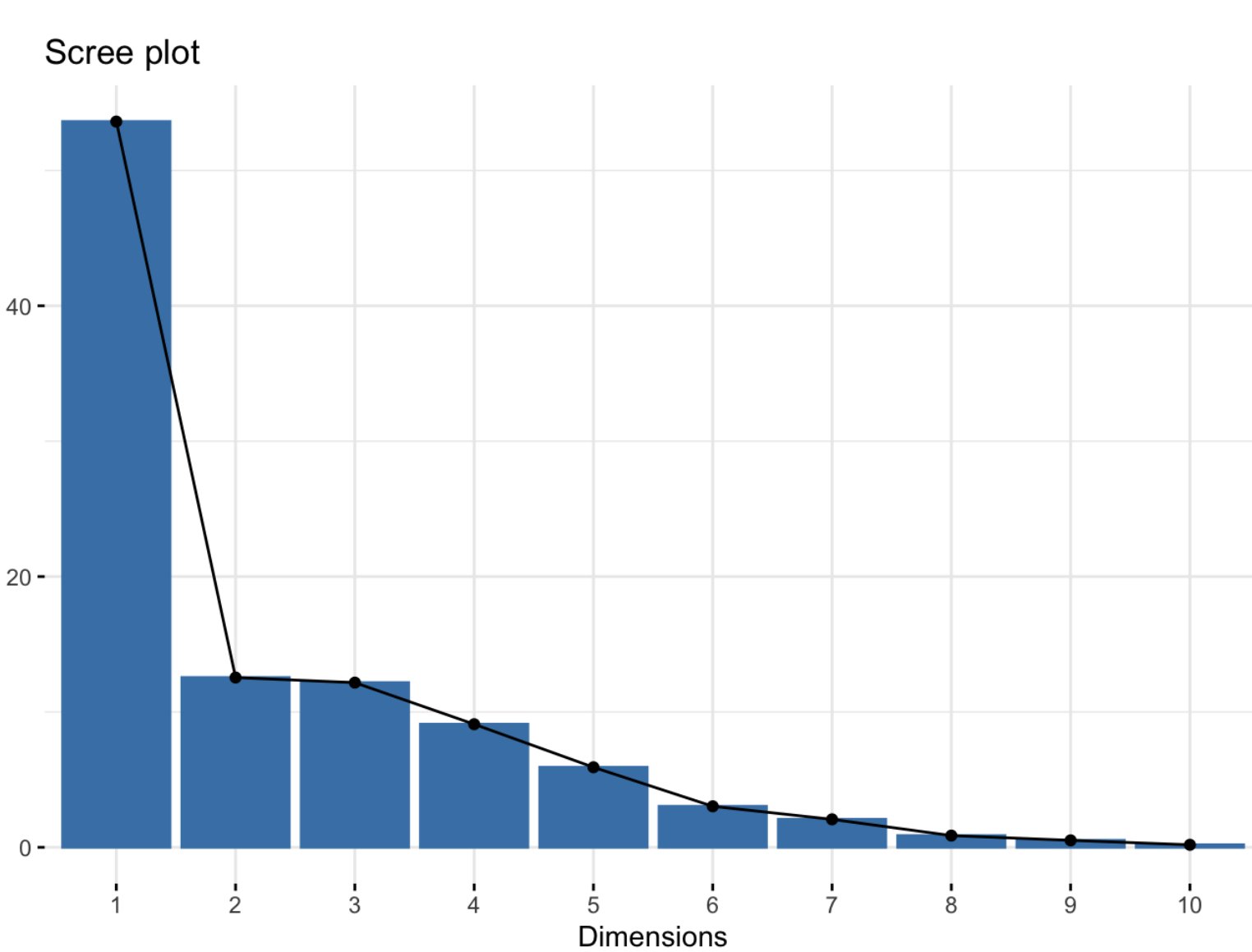
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

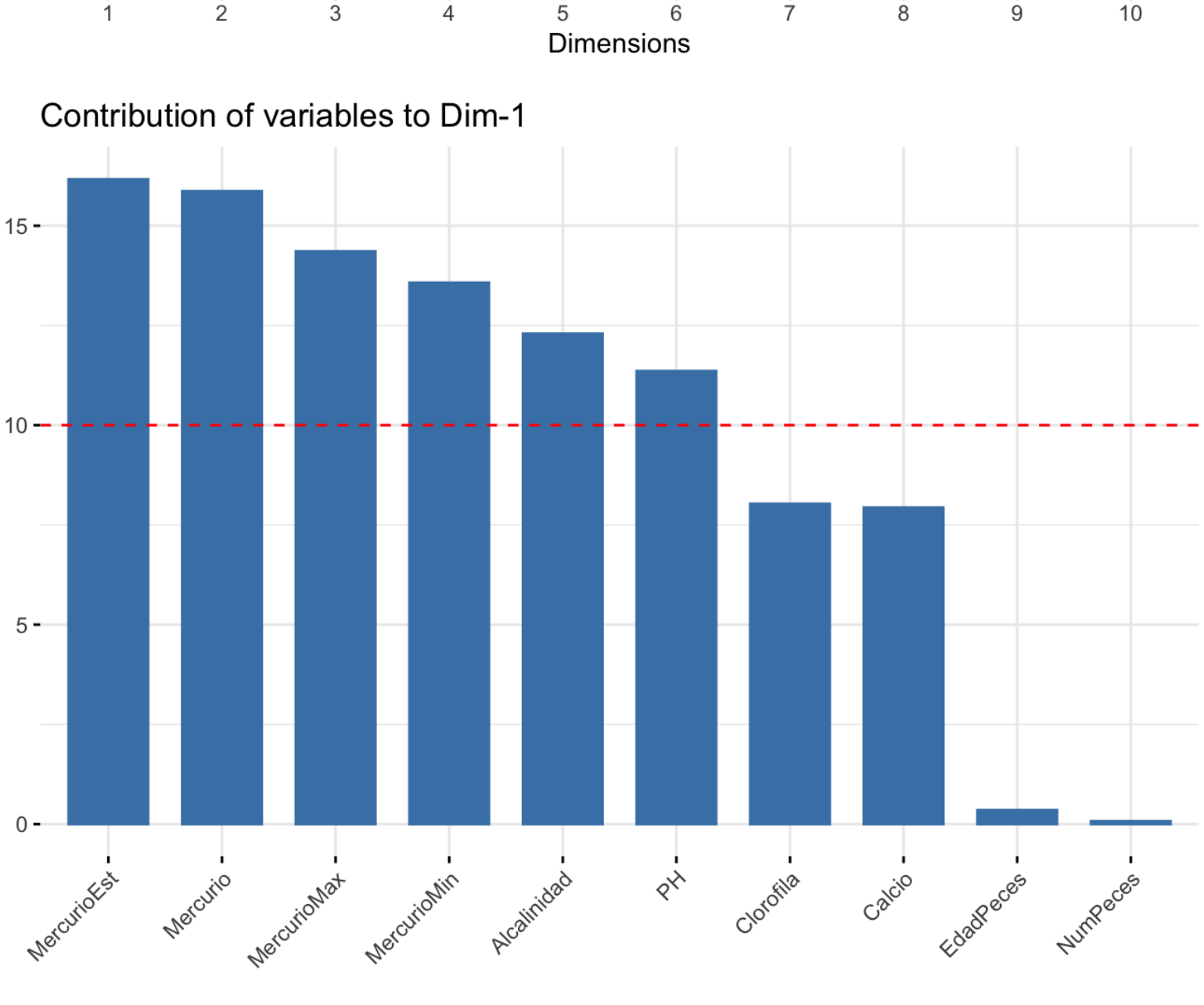












Conclusiones

Luego de analizar las variables de las que se compone nuestro dataset, pudimos ver en las pruebas de normalidad que solo el PH y el MercurioMax se comportaban como normales, por lo que son variables que en caso de utilizarse pueden ser estadisticamente modeladas con un grado mayor de confiabilidad como normales. Vimos también que estas mismas variables en las contribuciones a los componentes principales eran importantes y relevantes. Pero de la misma manera vimos que al menos utilizando solo los primeros dos componentes principales no llegabamos a un nivel de explicabilidad justificable como para concentrarnos en esos dos.