Momento de Retroalimentación: Módulo 1 Construcción de un modelo estadístico base (Portafolio Implementación)

Roel De la Rosa - A01197595

13/9/2022

Durante este reporte lo que se busca es analizar un conjunto de datos que contienen información sobre los niveles de contaminación que se encuentran en peces en distintos lagos. Se busca encontrar y justificar relaciones entre las distintas variables para poder encontrar, en este caso, cuales son las variables que más afectan el nivel de contaminación de mercurio en la carne de los peces. Para hacerlo se utilizaron métodos estadisticos tales como las pruebas de normalidad de Mardia y de Anderson-Darling, para saber el comportamiento de las variables, además de análisis de correlación y de componentes principales para entender mejor las relaciones entre las variables. Se concluyo que la Alcalinidad, el PH, el Calcio y la Clorofila son los factores más relevantes.

Algunas de las preguntas que se quieren responder son las siguientes: ¿Cómo se distribuyen las variables que se analizaran? ¿Hay relaciones entre las distintas variables? ¿Es necesario escalar los datos para obtener mejores análisis?

Considero que este tipo de estudios son necesarios, pues por ejemplo, el mercurio es uno de los metales más tóxicos con los que una persona puede ingerir. Es necesario que se hagan estudios para entender como es que ciertas variables pueden afectar el nivel de contaminación de mercurio en los animales que más se consumen para evitar que alguien consuma material contaminado y termine con efectos adversos a su salud.

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

```
## Warning: package 'reshape2' was built under R version 4.1.3
```

```
## Warning: package 'psych' was built under R version 4.1.3
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
## Warning: package 'Hmisc' was built under R version 4.1.3
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following object is masked from 'package:psych':
##
##     describe
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## Warning: package 'clusterSim' was built under R version 4.1.3
```

```
## Warning: package 'factoextra' was built under R version 4.1.3
```

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

```
## Warning: package 'mvnormalTest' was built under R version 4.1.3
```

```
##
## Attaching package: 'mvnormalTest'
```

```
## The following object is masked from 'package:psych':
##
##     mardia
```

```
## Warning: package 'MVN' was built under R version 4.1.3
```

```
## Warning: package 'RVAideMemoire' was built under R version 4.1.3
```

```
## *** Package RVAideMemoire v 0.9-81-2 ***
```

Leemos los datos

```
##   X1            X2    X3  X4   X5    X6   X7 X8   X9  X10  X11 X12
## 1  1    Alligator   5.9 6.1  3.0   0.7 1.23  5 0.85 1.43 1.53   1
## 2  2        Annie   3.5 5.1  1.9   3.2 1.33  7 0.92 1.90 1.33   0
## 3  3       Apopka 116.0 9.1 44.1 128.3 0.04  6 0.04 0.06 0.04   0
## 4  4 Blue Cypress  39.4 6.9 16.4   3.5 0.44 12 0.13 0.84 0.44   0
## 5  5        Brick   2.5 4.6  2.9   1.8 1.20 12 0.69 1.50 1.33   1
## 6  6       Bryant  19.6 7.3  4.5  44.1 0.27 14 0.04 0.48 0.25   1
```

```
## 'data.frame':    53 obs. of  12 variables:
##  $ X1 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X2 : chr  "Alligator" "Annie" "Apopka" "Blue Cypress" ...
##  $ X3 : num  5.9 3.5 116 39.4 2.5 19.6 5.2 71.4 26.4 4.8 ...
##  $ X4 : num  6.1 5.1 9.1 6.9 4.6 7.3 5.4 8.1 5.8 6.4 ...
##  $ X5 : num  3 1.9 44.1 16.4 2.9 4.5 2.8 55.2 9.2 4.6 ...
##  $ X6 : num  0.7 3.2 128.3 3.5 1.8 ...
##  $ X7 : num  1.23 1.33 0.04 0.44 1.2 0.27 0.48 0.19 0.83 0.81 ...
##  $ X8 : int  5 7 6 12 12 14 10 12 24 12 ...
##  $ X9 : num  0.85 0.92 0.04 0.13 0.69 0.04 0.3 0.08 0.26 0.41 ...
##  $ X10: num  1.43 1.9 0.06 0.84 1.5 0.48 0.72 0.38 1.4 1.47 ...
##  $ X11: num  1.53 1.33 0.04 0.44 1.33 0.25 0.45 0.16 0.72 0.81 ...
##  $ X12: int  1 0 0 0 1 1 1 1 1 1 ...
```

Transformación de los datos

Renombro las variables para poderlas interpretar de mejor manera.

X1 = número de indentificación 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 muscualar 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)

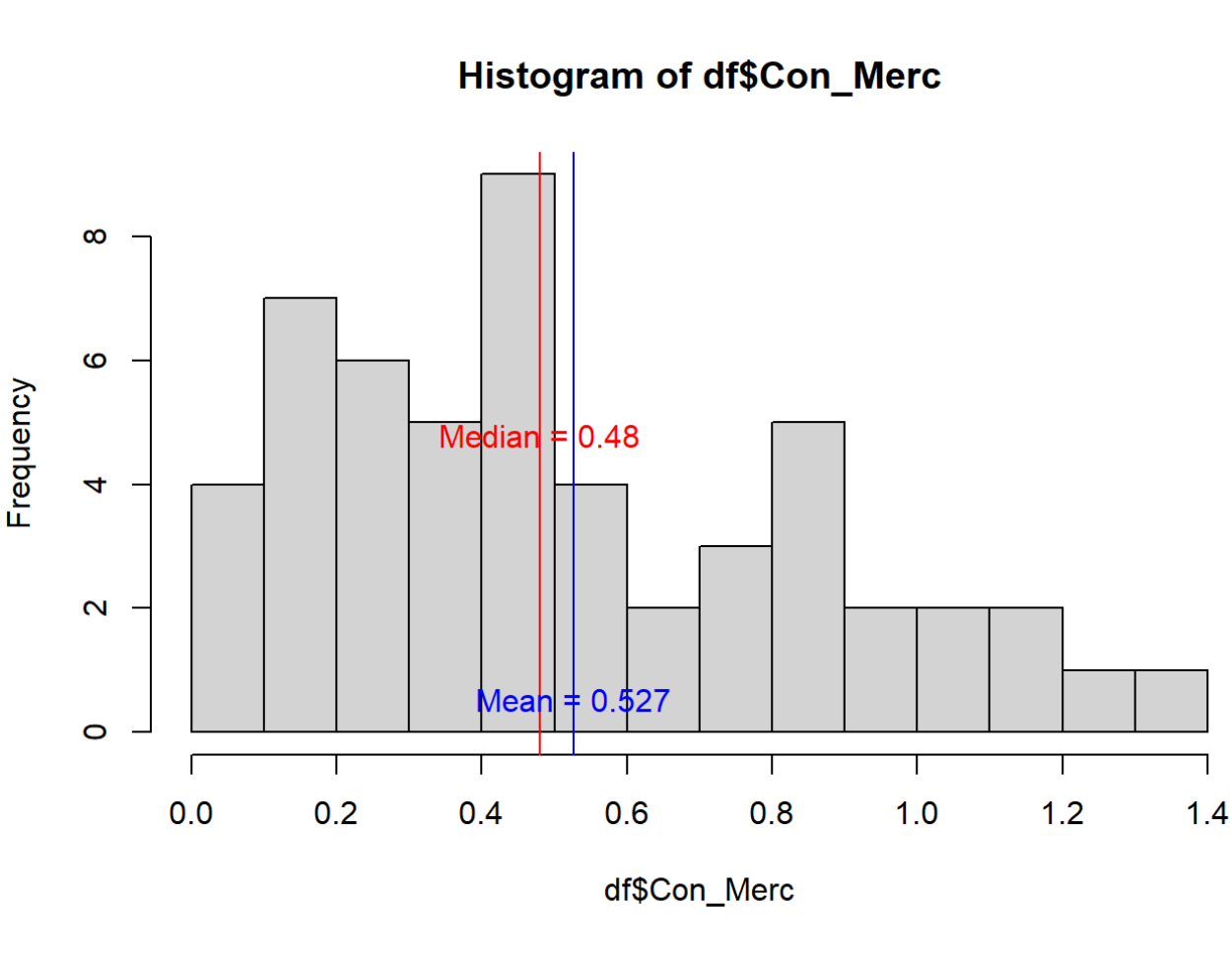
```
##           Lago Alcalinidad  PH Calcio Clorofila Con_Merc Peces_est Min_merc
## 1    Alligator         5.9 6.1    3.0       0.7     1.23         5     0.85
## 2        Annie         3.5 5.1    1.9       3.2     1.33         7     0.92
## 3       Apopka       116.0 9.1   44.1     128.3     0.04         6     0.04
## 4 Blue Cypress        39.4 6.9   16.4       3.5     0.44        12     0.13
## 5        Brick         2.5 4.6    2.9       1.8     1.20        12     0.69
## 6       Bryant        19.6 7.3    4.5      44.1     0.27        14     0.04
##   Max_merc Est_conc Edad_peces
## 1     1.43     1.53          1
## 2     1.90     1.33          0
## 3     0.06     0.04          0
## 4     0.84     0.44          0
## 5     1.50     1.33          1
## 6     0.48     0.25          1
```

```
##  [1] "Alligator"         "Annie"             "Apopka"
##  [4] "Blue Cypress"      "Brick"             "Bryant"
##  [7] "Cherry"            "Crescent"          "Deer Point"
## [10] "Dias"              "Dorr"              "Down"
## [13] "Eaton"             "East Tohopekaliga" "Farm-13"
## [16] "George"            "Griffin"           "Harney"
## [19] "Hart"              "Hatchineha"        "Iamonia"
## [22] "Istokpoga"         "Jackson"           "Josephine"
## [25] "Kingsley"          "Kissimmee"         "Lochloosa"
## [28] "Louisa"            "Miccasukee"        "Minneola"
## [31] "Monroe"            "Newmans"           "Ocean Pond"
## [34] "Ocheese Pond"      "Okeechobee"        "Orange"
## [37] "Panasoffkee"       "Parker"            "Placid"
## [40] "Puzzle"            "Rodman"            "Rousseau"
## [43] "Sampson"           "Shipp"             "Talquin"
## [46] "Tarpon"            "Tohopekaliga"      "Trafford"
## [49] "Trout"             "Tsala Apopka"      "Weir"
## [52] "Wildcat"           "Yale"
```

```
## [1] 53
```

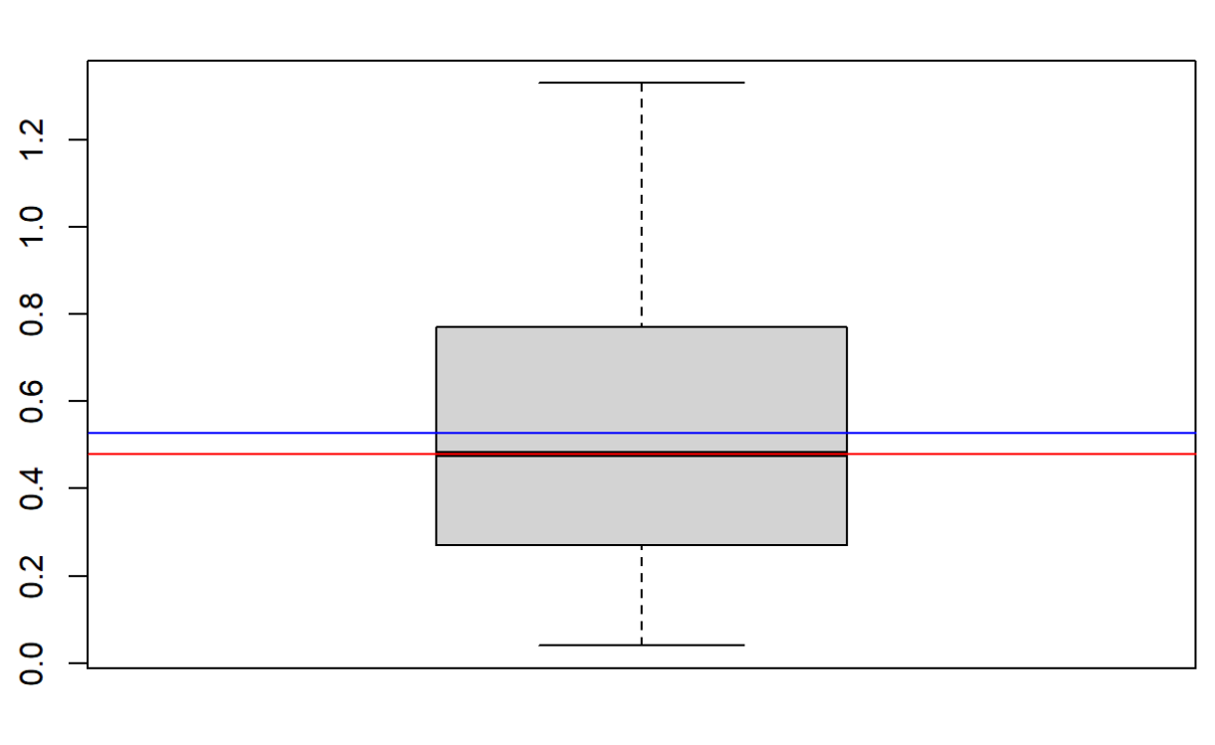
Hay 53 registros y se tienen 53 diferentes lagos, por lo que sabemos que esta columna realmente no nos aporta mucho.

Distribución de la concentración de mercurio



Se puede observar que hay bastantes casos en los que la concentración de mercurio por kg de pez es mayor a 0.5, lo cual incumple el Reglamento 34687-MAG y los reglamentos internacionales CE 1881/2006 y Codex Standard 193-1995.

De hecho, la mediana es 0.48 y la moda es de 0.527. Esto se encuentra incluso ya pasando los límites de lo establecido.



Normalidad Multivariada

##Prueba de Normalidad de Mardia

```
## $mv.test
##           Test Statistic p-value Result
## 1     Skewness  502.6673       0     NO
## 2     Kurtosis    5.8768       0     NO
## 3 MV Normality      <NA>    <NA>     NO
##
## $uv.shapiro
##             W      p-value UV.Normality
## Alcalinidad 0.8203 0       No
## PH          0.981  0.5552  Yes
## Calcio      0.7913 0       No
## Clorofila   0.6817 0       No
## Con_Merc    0.9421 0.0125  No
## Peces_est   0.583  0       No
## Min_merc    0.877  1e-04   No
## Max_merc    0.9555 0.0467  No
## Est_conc    0.9258 0.0028  No
## Edad_peces  0.4774 0       No
```

Se puede observar que tras este test de Normalidad de Mardia, la única variable que parece que se comporta con normalidad es el PH, pues tiene un p-value mayor a 0.05. Aunque el máximo de mercurio y la concentración de mercurio podrían ser considerados normales si el criterio del p-value fuera ser mayor a 0.01

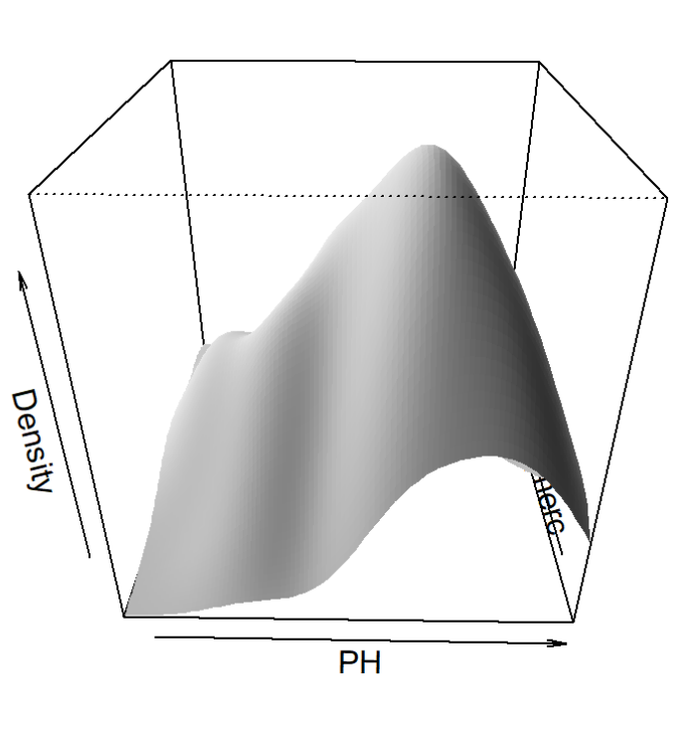
##Prueba de Anderson-Darling

```
##               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  Con_Merc      0.9253  0.0174      NO
## 6  Anderson-Darling  Peces_est     8.6943  <0.001      NO
## 7  Anderson-Darling  Min_merc      1.9770  <0.001      NO
## 8  Anderson-Darling  Max_merc      0.6585   0.081      YES
## 9  Anderson-Darling  Est_conc      1.0469  0.0086      NO
## 10 Anderson-Darling Edad_peces    14.3350  <0.001      NO
```

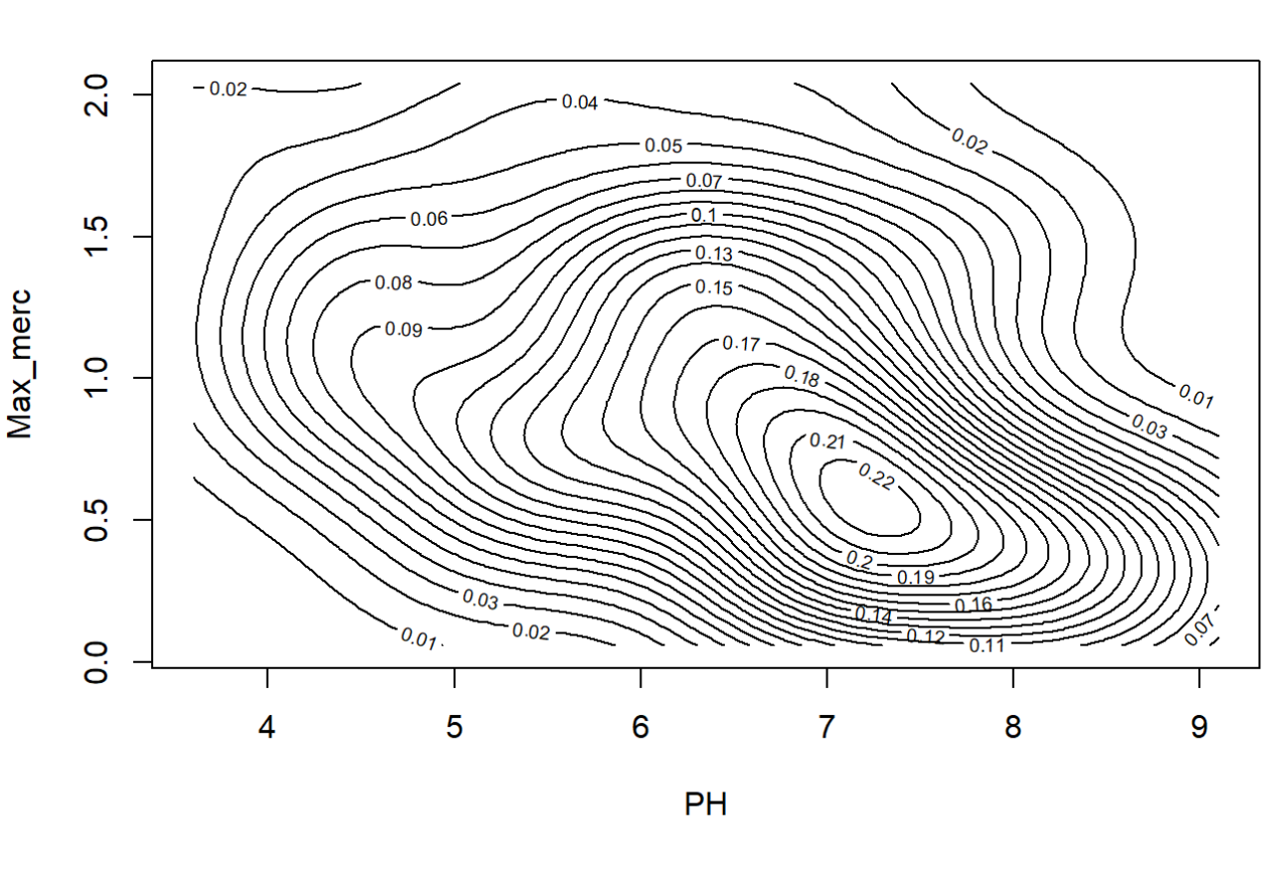
Aquí podemos observar que en tras la prueba de Anderson Darling, las únicas variables que parece que muestran normalidad son el Ph(el cual vimos en la prueba de Mardia) y el máximo de mercurio.

```
##           Test Statistic p-value Result
## 1     Skewness  502.6673       0     NO
## 2     Kurtosis    5.8768       0     NO
## 3 MV Normality      <NA>    <NA>     NO
```

A partir de la prueba de Mardia, podemos saber que no se tiene normalidad multivariada pues los p-values del sesgo y de la kurtosis son ambos prácticamente 0.

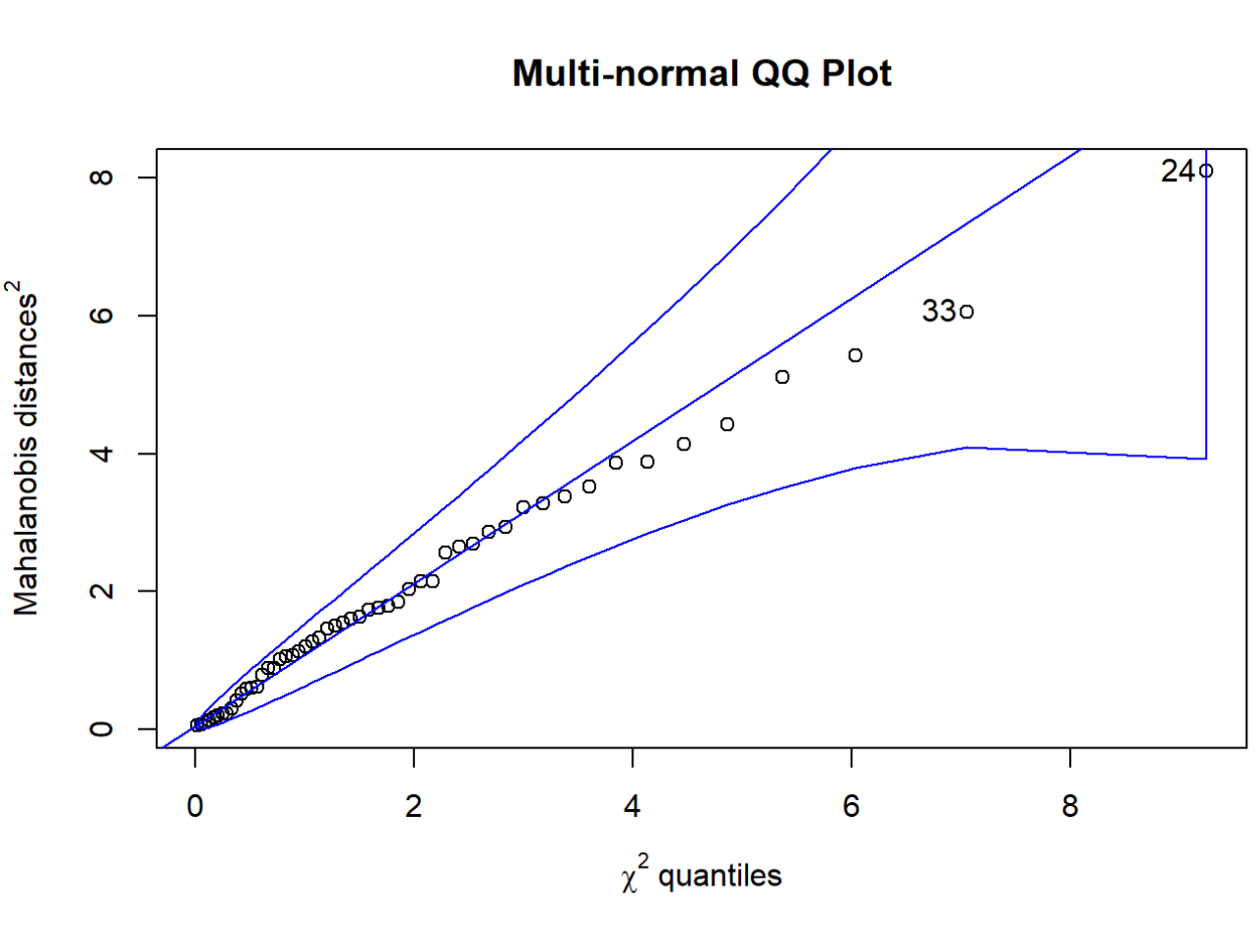


```
## $multivariateNormality
##              Test          Statistic           p value Result
## 1 Mardia Skewness   6.53855430534145 0.162377302354508    YES
## 2 Mardia Kurtosis -0.889321233851276 0.373830462900113    YES
## 3             MVN               <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    PH        0.3496    0.4611    YES
## 2 Anderson-Darling Max_merc     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
## Max_merc 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925
##            Kurtosis
## PH       -0.6239638
## Max_merc -0.6692490
```



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

Outliers



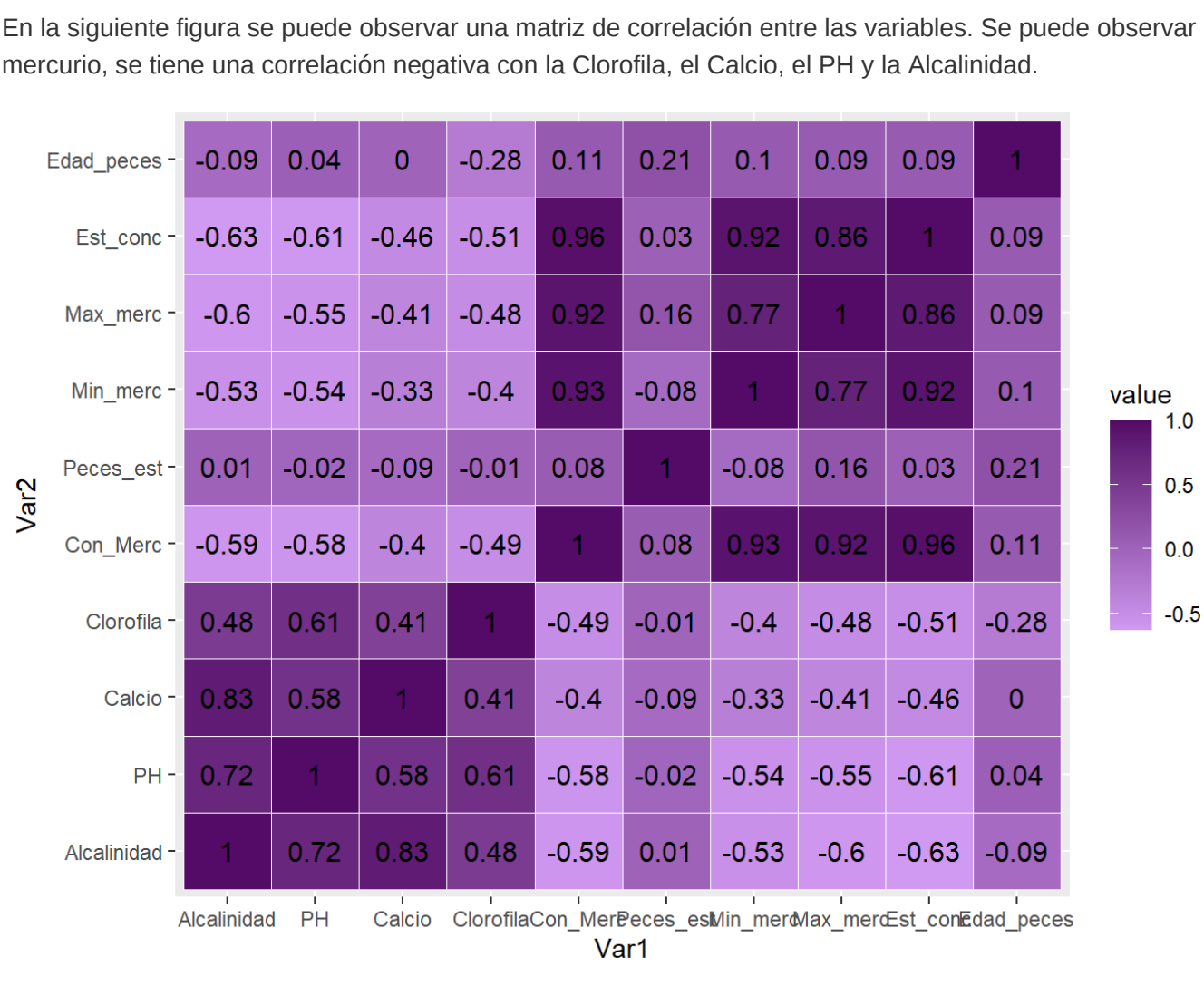
```
## [1] 24 33
```

Análisis de Correlación

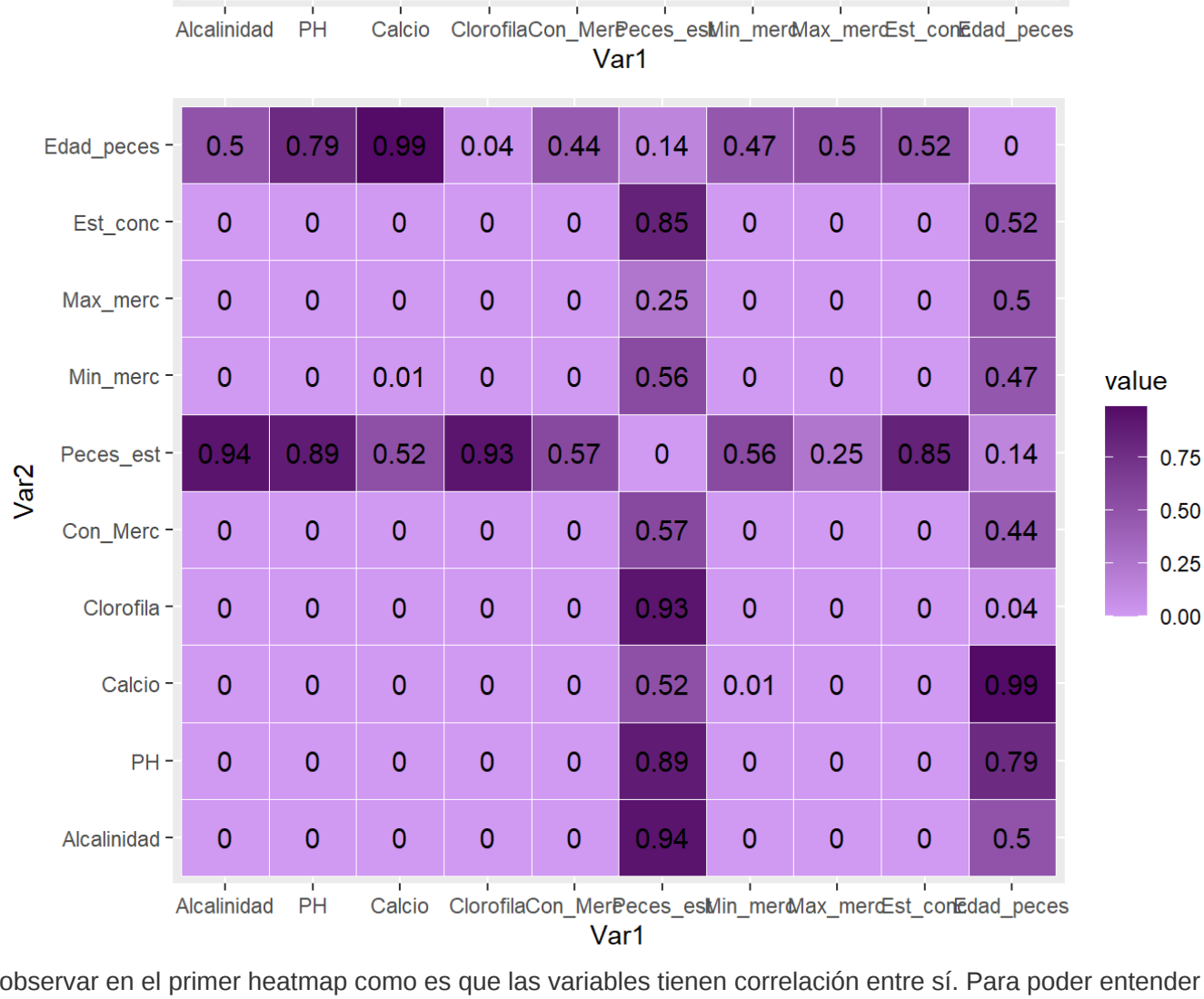
Para conocer los factores que más pueden influir hacemos un análisis de la correlación entre las variables a examinar.

Matriz de Correlación

En la siguiente figura se puede observar una matriz de correlación entre las variables. Se puede observar que, para la concentración de mercurio, se tiene una correlación negativa con la Clorofila, el Calcio, el PH y la Alcalinidad.



Var2	Alcalinidad	PH	Calcio	Clorofila	Con_Merc	Peces_est	Min_merc	Max_merc	Est_conc	Edad_peces
Edad_peces	-0.09	0.04	0	-0.28	0.11	0.21	0.1	0.09	0.09	1
Est_conc	-0.63	-0.61	-0.46	-0.51	0.96	0.03	0.92	0.86	1	0.09
Max_merc	-0.6	-0.55	-0.41	-0.48	0.92	0.16	0.77	1	0.86	0.09
Min_merc	-0.53	-0.54	-0.33	-0.4	0.93	-0.08	1	0.77	0.92	0.1
Peces_est	0.01	-0.02	-0.09	-0.01	0.08	1	-0.08	0.16	0.03	0.21
Con_Merc	-0.59	-0.58	-0.4	-0.49	1	0.08	0.93	0.92	0.96	0.11
Clorofila	0.48	0.61	0.41	1	-0.49	-0.01	-0.4	-0.48	-0.51	-0.28
Calcio	0.83	0.58	1	0.41	-0.4	-0.09	-0.33	-0.41	-0.46	0
PH	0.72	1	0.58	0.61	-0.58	-0.02	-0.54	-0.55	-0.61	0.04
Alcalinidad	1	0.72	0.83	0.48	-0.59	0.01	-0.53	-0.6	-0.63	-0.09



Var2	Alcalinidad	PH	Calcio	Clorofila	Con_Merc	Peces_est	Min_merc	Max_merc	Est_conc	Edad_peces
Edad_peces	0.5	0.79	0.99	0.04	0.44	0.14	0.47	0.5	0.52	0
Est_conc	0	0	0	0	0	0.85	0	0	0	0.52
Max_merc	0	0	0	0	0	0.25	0	0	0	0.5
Min_merc	0	0	0.01	0	0	0.56	0	0	0	0.47
Peces_est	0.94	0.89	0.52	0.93	0.57	0	0.56	0.25	0.85	0.14
Con_Merc	0	0	0	0	0	0.57	0	0	0	0.44
Clorofila	0	0	0	0	0	0.93	0	0	0	0.04
Calcio	0	0	0	0	0	0.52	0.01	0	0	0.99
PH	0	0	0	0	0	0.89	0	0	0	0.79
Alcalinidad	0	0	0	0	0	0.94	0	0	0	0.5

A partir de lo anterior se puede observar en el primer heatmap como es que las variables tienen correlación entre sí. Para poder entender mejor la relación entre las variables se han calulado los p-values entre las correlaciones de todas las variables en el segundo heatmap. Después se han usado pruebas de hipótesis para ver si las variables tienen independencia o asociación entre sí.

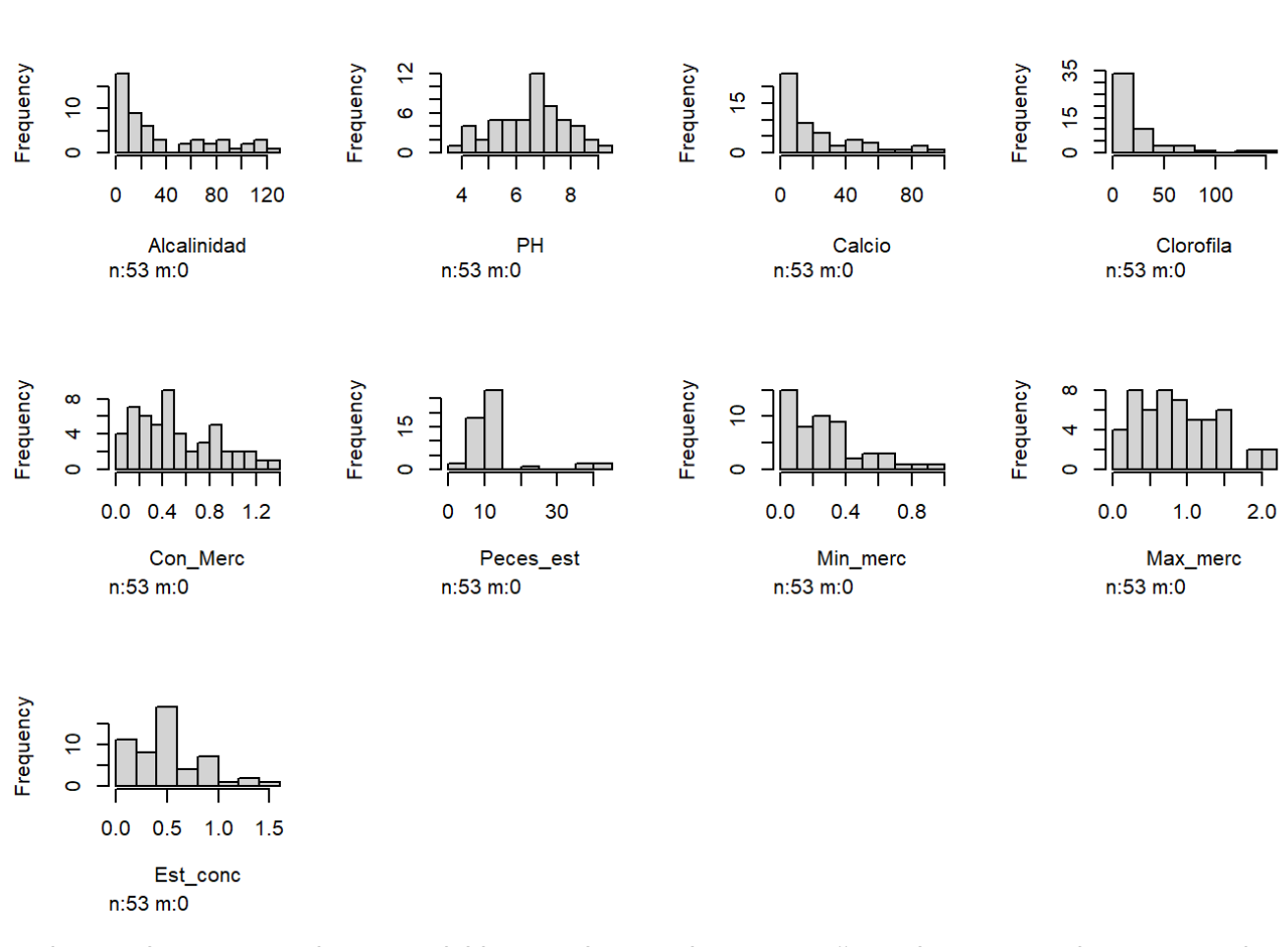
$H_0 : \rho = 0$ El caso de independencia entre variables $H_1 : \rho \neq 0$ El caso de asociación entre variables

Regla de decisión: Rechazar H_0 si el p-value < 0.05.

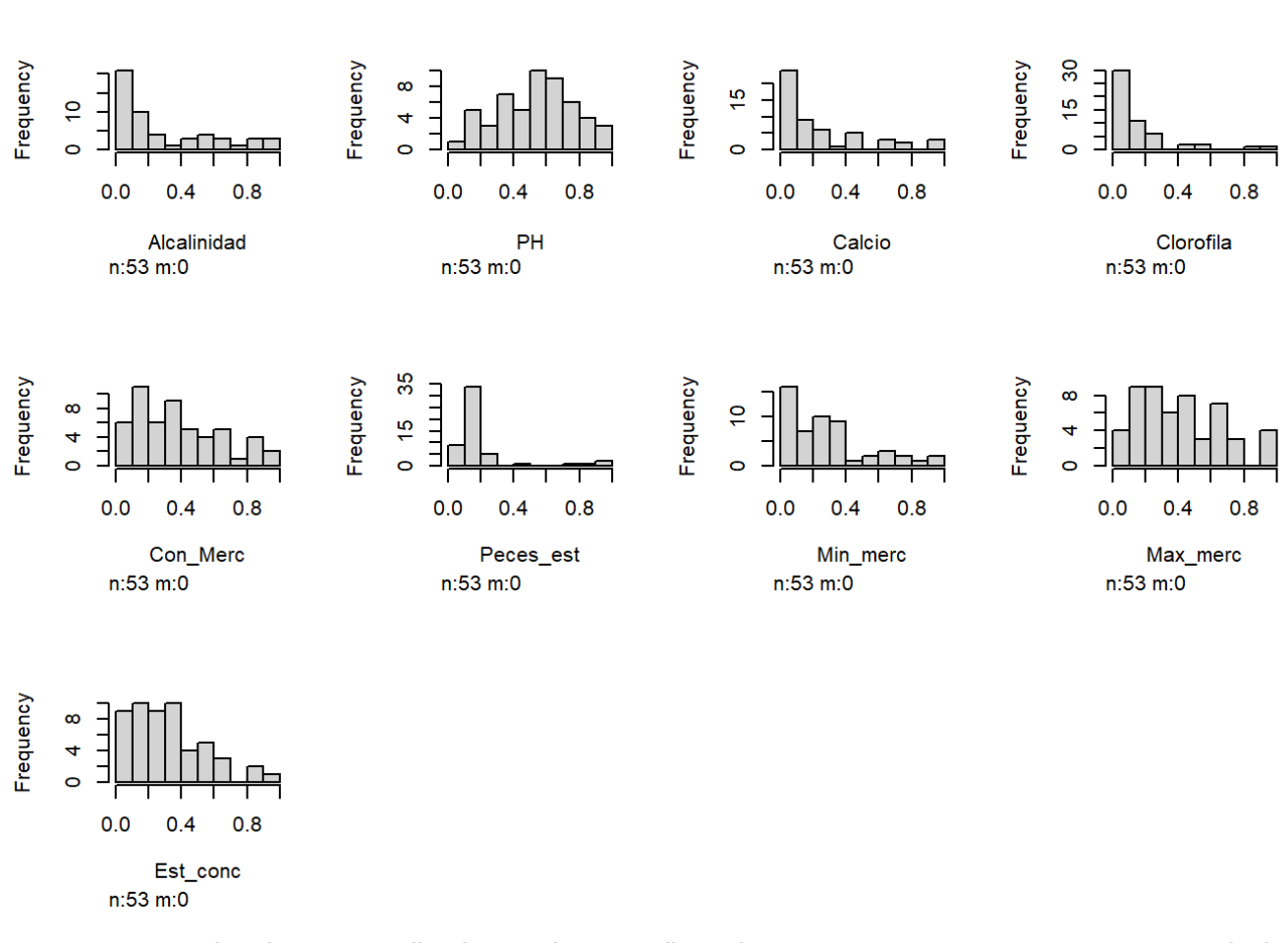
Dado que queremos ver que variables afectan a la concentración de mercurio, podemos observar que Alcalinidad, PH, Calcio, Clorofila, el mínimo de mercurio, el máximo de mercurio y la estimación de mercurio tienen un p-value menor a 0.05, por lo que se rechaza H_0 y se llega a la conclusión que estas variables son las que pueden afectar a la concentración de mercurio.

Análisis de Componentes Principales

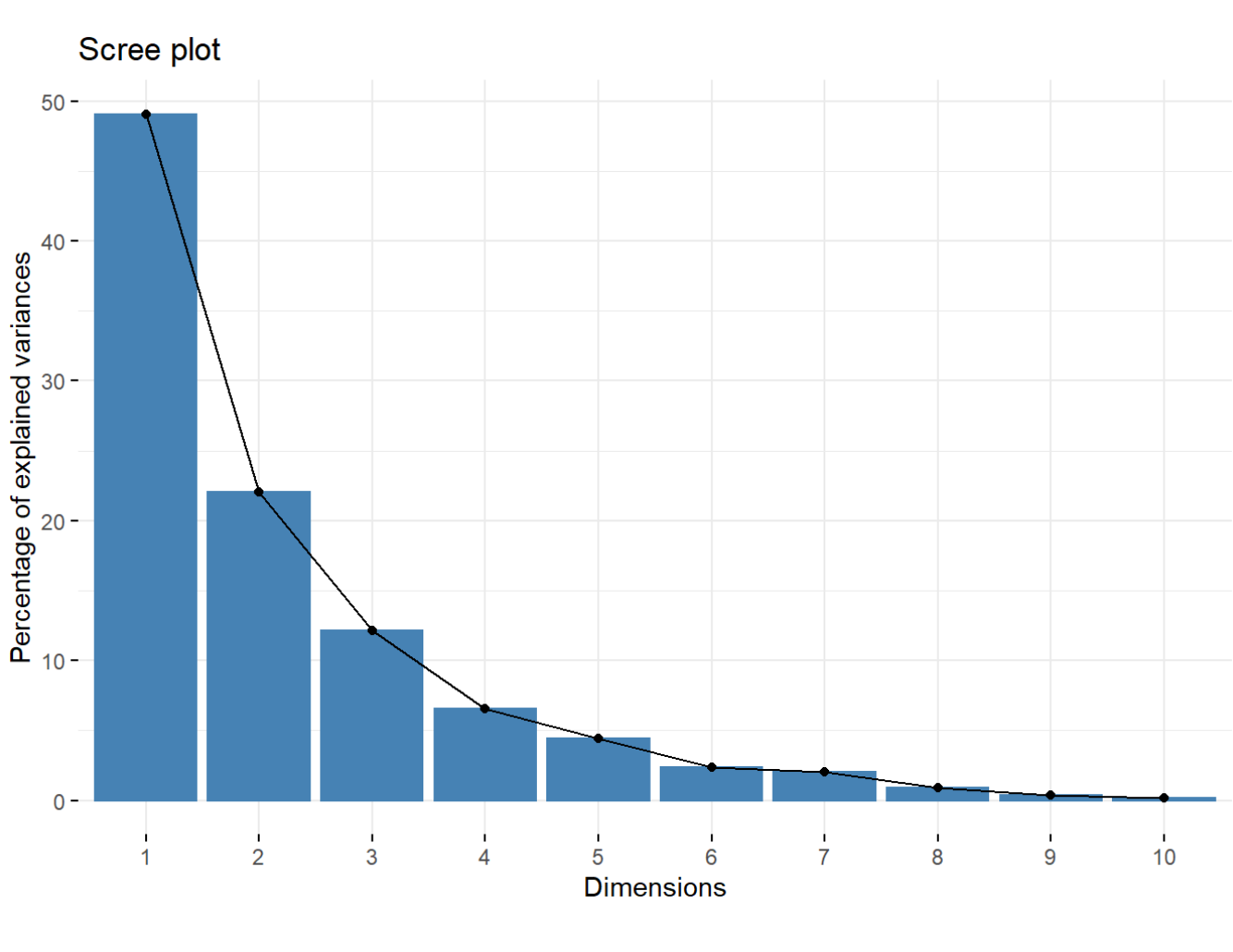
Antes de hacer el PCA debemos de saber si es necesario escalar los datos. Para poder saber esto primero vamos a ver las distribuciones de las variables.

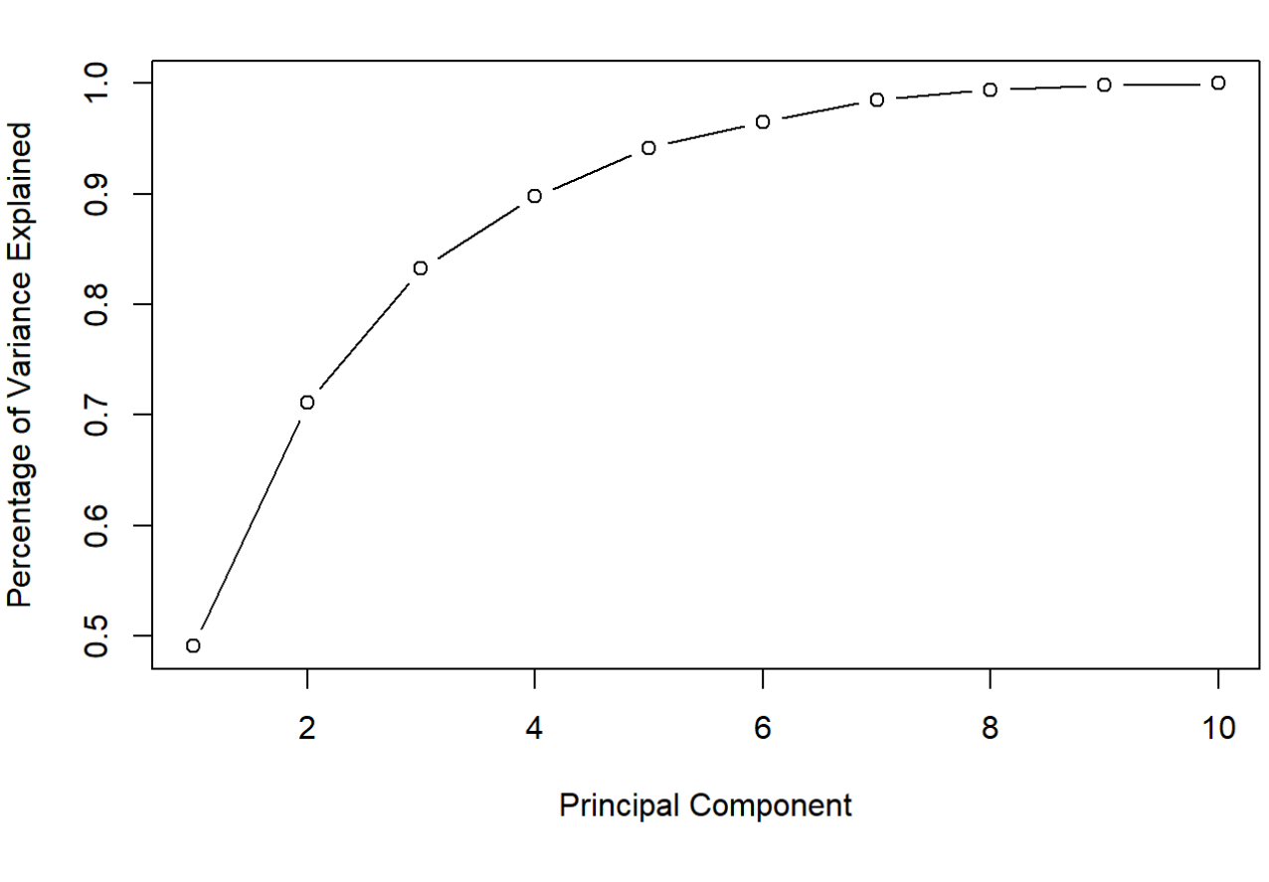


Podemos observar que algunas variables que tienen valores pequeños, mientras que algunas otras tienen valores relativamente mucho más grandes, por ello vamos a buscar normalizar los datos para que cuando se realize el PCA no se tenga algún sesgo hacia las variables con valores más altos.



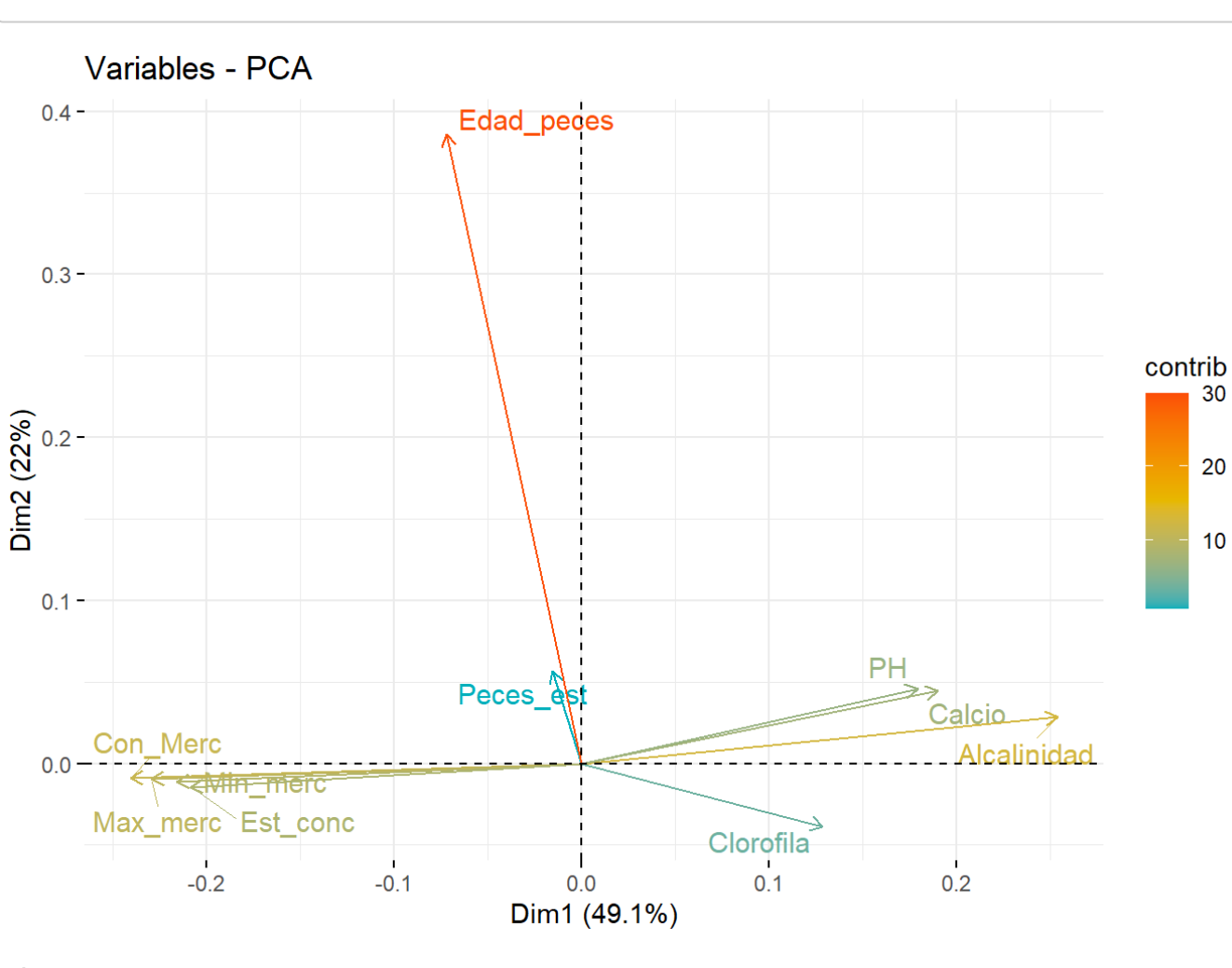
Ya que tenemos los datos normalizados, podemos aplicar el PCA para ver cuantos componentes principales representan la mayoría de la variabilidad de los datos y que variables son las que tienen mayor relevancia en esos componentes.





Cómo se puede observar con 4 componentes principales se tiene cerca del 90% de la variabilidad de los datos. Ahora solo falta observar esos componentes principales.

```
## Standard deviations (1, .., p=10):
##  [1] 0.59581904 0.39917379 0.29659562 0.21724577 0.17846501 0.13020889
##  [7] 0.12054968 0.08034046 0.05387504 0.03552301
##
## Rotation (n x k) = (10 x 10):
##                     PC1         PC2         PC3          PC4         PC5
## Alcalinidad  0.42599010  0.07110301 -0.45380835 -0.092834387  0.22948361
## PH           0.30061298  0.11482867 -0.17870490 -0.069014050 -0.55950798
## Calcio       0.31893493  0.11188130 -0.59403512  0.099361373  0.28519334
## Clorofila    0.21582602 -0.09664014 -0.08869493 -0.142529907 -0.70038086
## Con_Merc    -0.40348016 -0.02205194 -0.34761399 -0.065785424 -0.07823426
## Peces_est   -0.02601474  0.14242070  0.04076418 -0.928309977  0.12533319
## Min_merc    -0.36228310 -0.02807425 -0.37522624  0.161153945 -0.16959234
## Max_merc    -0.38487657 -0.02403028 -0.27289223 -0.221215975 -0.01713438
## Est_conc    -0.34987337 -0.03625495 -0.24119831  0.008891634 -0.06986001
## Edad_peces  -0.12083256  0.96770451  0.07375163  0.123927733 -0.08157767
##                     PC6         PC7         PC8         PC9        PC10
## Alcalinidad -0.17189310  0.43307746  0.57319435 -0.01376188  0.02023250
## PH          -0.63015912  0.03565135 -0.38101324 -0.02411459 -0.05073958
## Calcio       0.29582901 -0.42882768 -0.40366493  0.08098410 -0.02767055
## Clorofila    0.54136410 -0.15098977  0.31542473  0.08653852  0.05010209
## Con_Merc    -0.06531699  0.04565994 -0.03511790  0.01657562  0.83520926
## Peces_est    0.14595043  0.14393496 -0.23038905 -0.03708317 -0.05862557
## Min_merc     0.24594093  0.43568870 -0.18643890 -0.52315245 -0.33715926
## Max_merc    -0.31539964 -0.58865459  0.39981950 -0.19248498 -0.29102414
## Est_conc    -0.01896973  0.21235311 -0.03264383  0.82000169 -0.30694627
## Edad_peces   0.08289019 -0.01571581  0.11766871  0.01975141  0.01188972
```



Cómo se puede observar en la tabla anterior,

El primer componente principal se ve afectado en su mayoria por Alcalinidad, PH, Calcio, Clorofila, la concentración de mercurio, el minimo de mercurio, el máximo de mercurio y la estimación de la concentración

El segundo componente principal se ve afectado principalmente por la variable que dice si un pez es joven o adulto

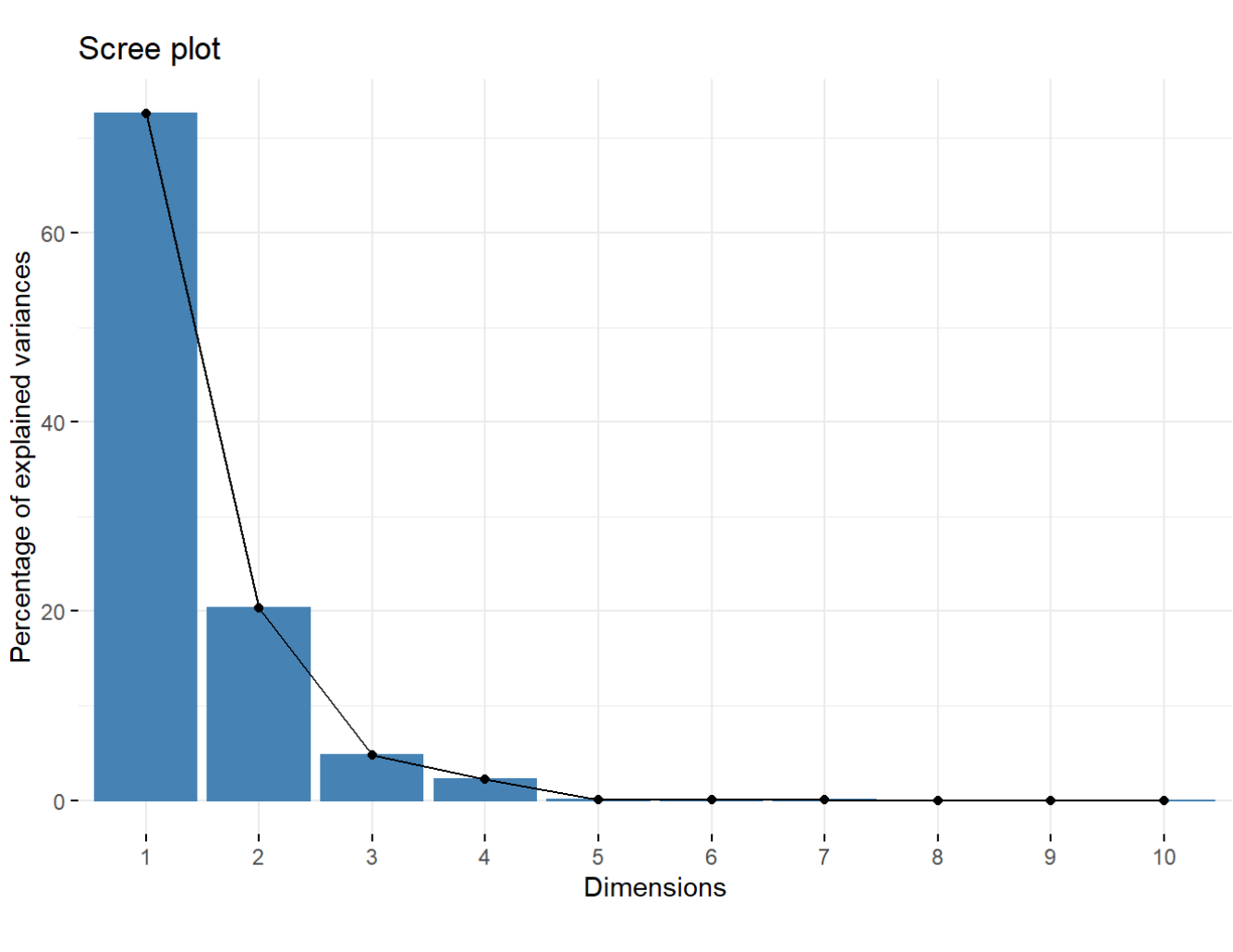
El tercer componente principal tiene mayor relación con Alcalinidad, Calcio, Concentración de mercurio, el minimo de mercurio, el máximo de mercurio y la estimación de la concentración

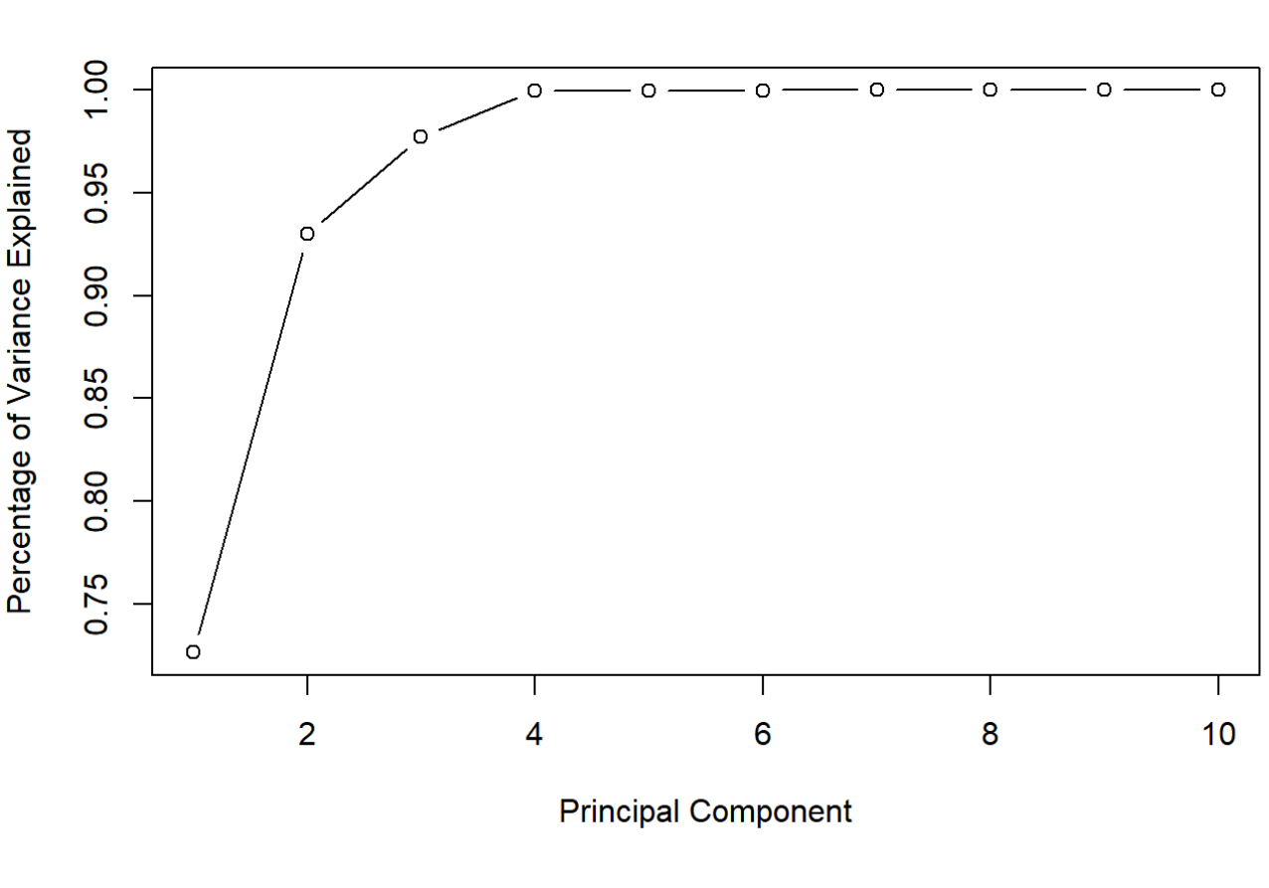
El cuarto componente principal se ve afectado en su mayoría por la cantidad de peces estudiados.

Esto es algo que se puede ver gráficamente en el loading plot anterior. Se puede ver como es que esas variables se van agrupando.

Por su parte, haré el mismo análisis pero esta vez sin la normalización de los datos. Esto con objetivo de ver si se tienen resultados similares.

Ya que tenemos los datos normalizados, podemos aplicar el PCA para ver cuantos componentes principales representan la mayoría de la variabilidad de los datos y que variables son las que tienen mayor relevancia en esos componentes.

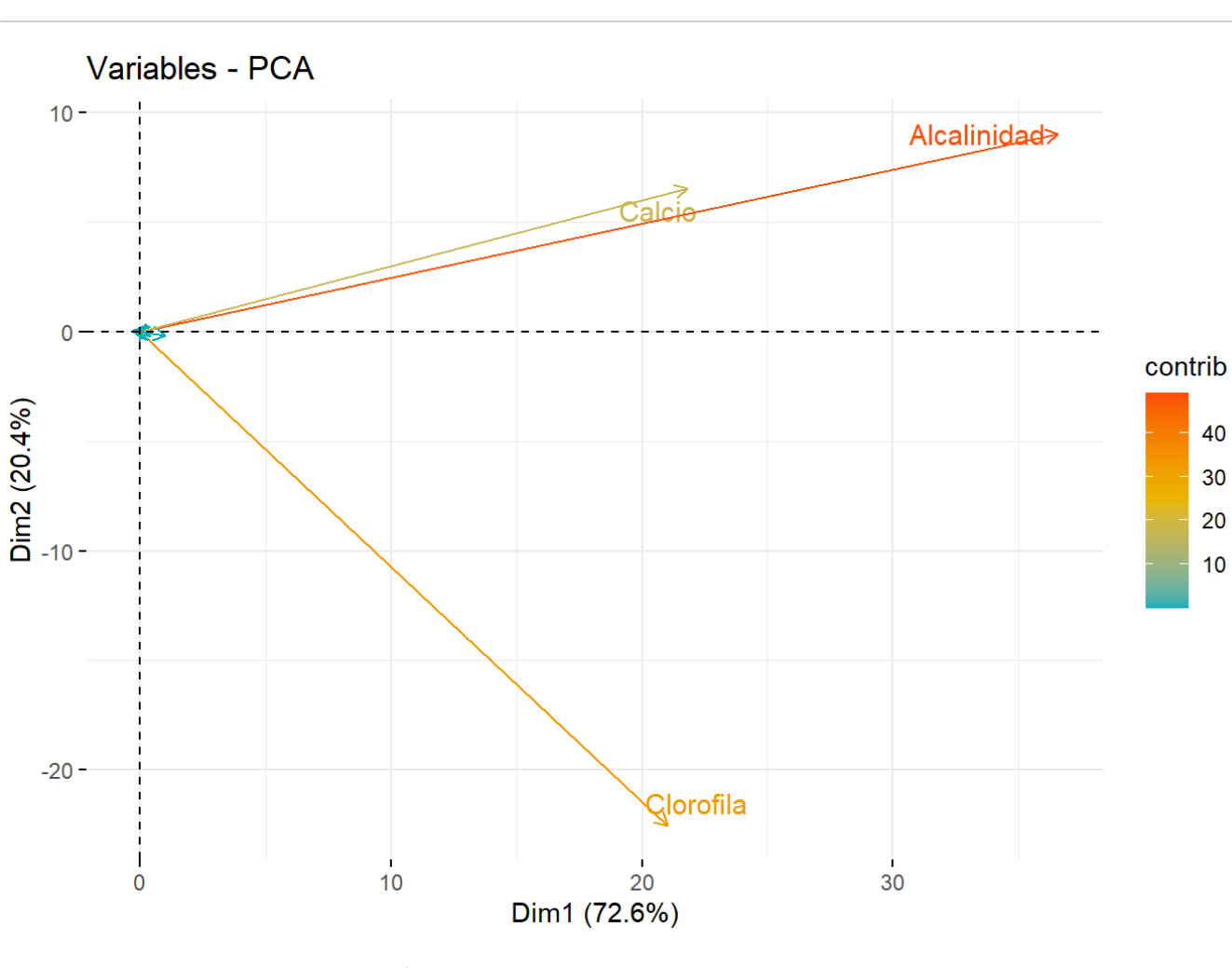




Cómo se puede observar con 4 componentes principales se tiene cerca del 90% de la variabilidad de los datos. Ahora solo falta observar esos componentes principales.

```
## Standard deviations (1, .., p=10):
##  [1] 47.50223254 25.15229597 12.14061049  8.29910593  0.82082273  0.50333372
##  [7]  0.33531760  0.16983359  0.06774799  0.04368155
##
## Rotation (n x k) = (10 x 10):
##                      PC1           PC2          PC3           PC4          PC5
## Alcalinidad  0.770051693  0.3595624308  0.512212889  1.211960e-01  0.022379095
## PH           0.020607395 -0.0064777176  0.013445136  7.815087e-03 -0.970898339
## Calcio       0.459103405  0.2606043150 -0.824602375 -2.029292e-01 -0.005136564
## Clorofila    0.442396618 -0.8959523487 -0.034896812 -8.682740e-03  0.014101339
## Con_Merc    -0.004349946  0.0015153985 -0.006280655 -6.263085e-03  0.070098841
## Peces_est   -0.003461490 -0.0017191525  0.236827940 -9.713807e-01 -0.003620863
## Min_merc    -0.002482186  0.0006039064 -0.004911645  6.911159e-05  0.061772060
## Max_merc    -0.006732314  0.0020103109 -0.009275841 -1.487778e-02  0.076192576
## Est_conc    -0.004611177  0.0012561924 -0.004970382 -3.212703e-03  0.079787785
## Edad_peces  -0.001169630  0.0040547980 -0.002920752 -1.197391e-02 -0.188987340
##                      PC6          PC7          PC8           PC9          PC10
## Alcalinidad  0.011296594  0.007055454 -0.002545190  1.581653e-05  0.0001951853
## PH           0.152551489 -0.173065657  0.057214145 -3.485511e-03 -0.0117752811
## Calcio      -0.009060983 -0.006678314  0.002454478  9.773667e-04 -0.0002013405
## Clorofila    0.003742677  0.006784090 -0.001424066  5.372560e-04  0.0004978075
## Con_Merc     0.470036339  0.066615441  0.279756491 -3.236927e-01  0.7658810638
## Peces_est   -0.009949302 -0.011307227  0.008472817 -9.632211e-04 -0.0020987371
## Min_merc     0.292267405  0.110635778  0.446372646 -5.796271e-01 -0.6027245954
## Max_merc     0.693362599 -0.035769112 -0.693360284  3.142457e-04 -0.1762334539
## Est_conc     0.432229792  0.082492810  0.471915643  7.474356e-01 -0.1363231520
## Edad_peces  -0.049765942  0.972120176 -0.125200363  2.410429e-02  0.0190692867
```

```
## Warning: ggrepel: 7 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Como se puede observar en las gráficas y tabla anteriores, parece que el pca si se ve muy sesgado por las variables que tienen mayores rangos, por lo que la normalizaión que realizamos fue un acierto.

Conclusiones

A partir del análisis de las correlaciones entre las variables y su significancia, además del análisis de componentes principales con los datos normalizados, se llega a la conclusión que los principales factores que influyen en el nivel de contaminación por mercurio en los peces de lagos de Florida son:

Alcalinidad, PH, Calcio, Clorofila, el mínimo de mercurio, el máximo de mercurio y la estimación de mercurio.

Algo que se debe de tomar en consideración es que el minimo de mercurio, el máximo de mercurio y la estimación de mercurio, por la forma en la que fueron obtenidas y calculadas, tienen bastante relación con la concentración de mercurio, por lo que considero que la Alcalinidad, el PH, el Calcio y la Clorofila son los factores más relevantes.

Referencias

Servicio Nacional de Salud Animal (SENASA) del Ministerio de Agricultura y Ganadería. (2008). RTCR 409: 2008 Reglamento de Límites Máximos Microbiológicos y de Residuos de Medicamentos y Contaminantes para los Productos y Subproductos de la Pesca y de la Acuicultura Destinados al Consumo Humano. Nº 34687-mag. Retrieved September 17, 2022, from http://www.pgrweb.go.cr/DOCS/NORMAS/1/VIGENTE/D/2000-2009/2005-2009/2008/F916/BE0D7.HTML