Análisis estadístico y visualización de datos. Informe de prácticas tema 4

# Ejercicio 1

El primer ejercicio se ha centrado en el análisis mediante PCA de un conjunto de datos sobre cinco especies de fitoplancton con el objetivo de conocer si es posible explicar las diferencias en la composición de especies a partir de los tres parámetros de región, lago y época. El set contiene una tabla con ocho columnas, representando los distintos parámetros observados, que son: Region, Lago, Epoca, Cianophyceae, Euglenophyceae, Chlorococales, Zygophyceae, Bacillariophyceae. En este caso los tres primeros parámetros (Región, Lago y Época) estarían representados por distintas categorías mientras que al resto de parámetros les correspondería una variable numérica en unidades desconocidas y que representarían el recuento de algas de cada una de las distintas especies para una determinada combinación de las tres primeras columnas.

Debo indicar que para este ejercicio hemos utilizado RStudio y las librerías Hmisc para el análisis por PCA y la librería ggfortify para la generación de alguna de las gráficas mostradas en este ejercicio, el resto del análisis se realizó a través de línea de comandos sin la utilización de R Commander.

Para el primer punto del análisis se nos pide comprobar la correlación entre las distintas variables (parámetros) mediante la generación de la matriz de correlación y la matriz de valores P. Estas se muestran a continuación:





En cuanto a la matriz de correlación, los valores próximos a ±1 indicarían correlación entre las dos variables, mientras que valores cercanos a 0 indicarían falta de correlación. Por otra parte, el valor P nos daría una idea de la significación de esa correlación. Cn nuestro caso, teniendo en cuenta el número de datos analizados, optamos por tomar 0,05 como valor umbral de significación, por debajo del cual consideramos como presencia de significación en los datos.

Partiendo de estas premisas podemos observar que nuestros datos no tienen una gran correlación para ninguna combinación de variables, ya que el mayor valor de correlación se encuentra entre las especies Euglenophyceae y Cianophyceae con un valor de correlación de 0.567 y un valor P muy inferior a 0,005.

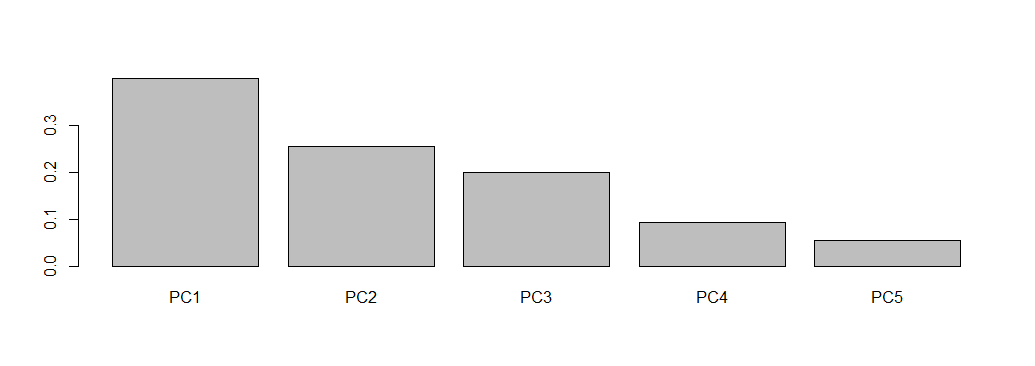
El siguiente paso de la práctica ha consistido en el análisis de componentes principales y la explicación de la elección de las dos componentes principales apoyándonos en la matriz de covarianzas. Los resultados obtenidos son los siguientes:



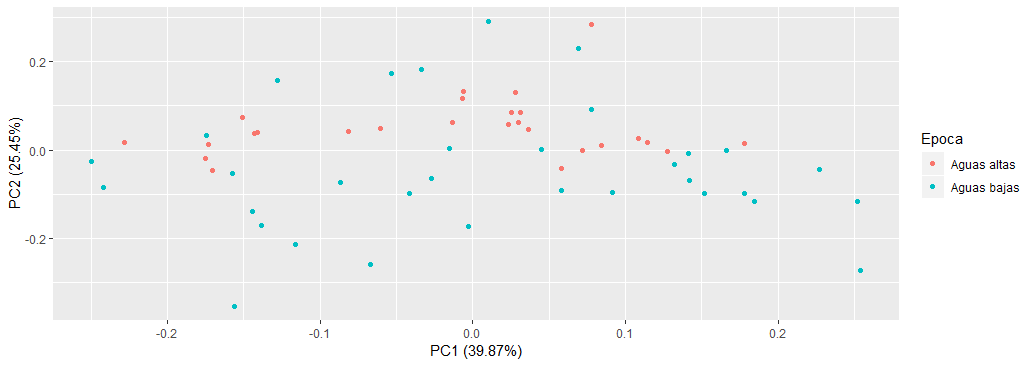
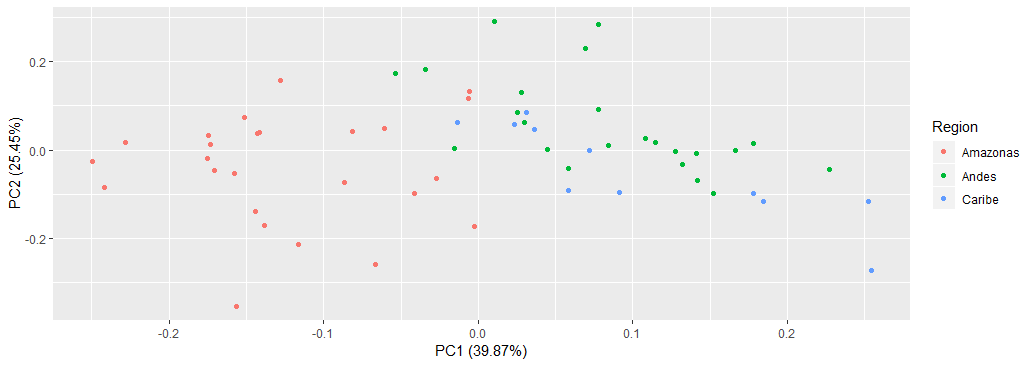


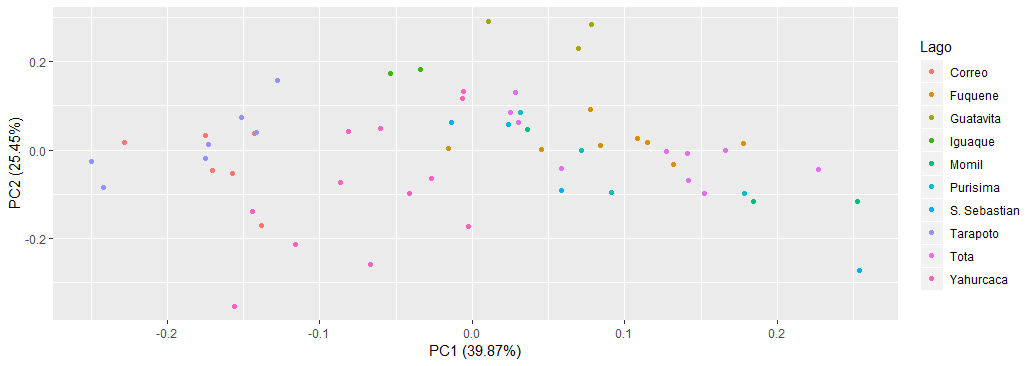
La matriz de rotación de componentes principales se forma a partir de la descomposición de los autovalores de la matriz de covarianzas y se encuentran ordenados por la cantidad de varianza que permiten explicar. El número de componentes principales que se suele escoger se relaciona con el porcentaje de varianza acumulado que pueden explicar en su conjunto, aceptándose como válido cuando este porcentaje ronda el 80-90%. En nuestro caso, como se verá más adelante, dos componentes principales no explican el porcentaje suficiente.

A continuación, el ejercicio pide construir un gráfico de sedimentación a partir del análisis de componentes principales. La gráfica obtenida es la siguiente:



Como se puede observar, la acumulación de varianza explicada por las primeras dos componentes no representa un porcentaje suficiente, representando un valor acumulado del 65% ya que la primera componente principal explica un 40% de la varianza mientras que la segunda componente explica un 25%. No sería hasta la tercera componente (explicando esta un 20%) donde la varianza explicada total se encontraría en valores superiores al 85%.

Por último, se pide construir gráficas de puntuación para los parámetros de Región, Época, y Lago. Estos gráficos se muestran a continuación. 



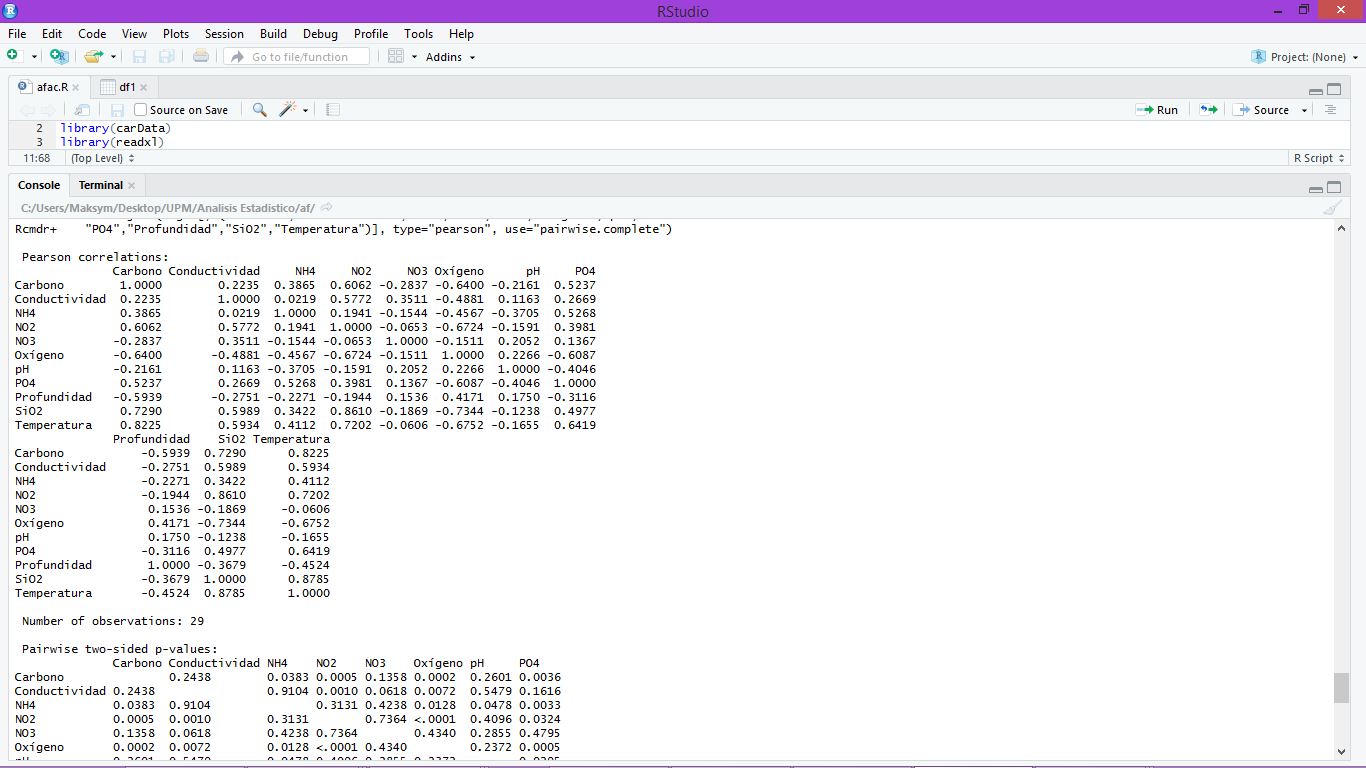
Como podemos ver, las gráficas de dispersión para las puntuaciones de los dos componentes principales seleccionados no permiten separar visualmente las distintas categorías, salvo en el caso de las regiones Amazonas frente a Caribe y Andes, y quizás entre aguas altas y bajas. Por ello, quizás convendría utilizar, como ya hemos mencionado anteriormente, una componente más, aunque la representación se volvería más complicada al incluir una dimensión más en la representación.

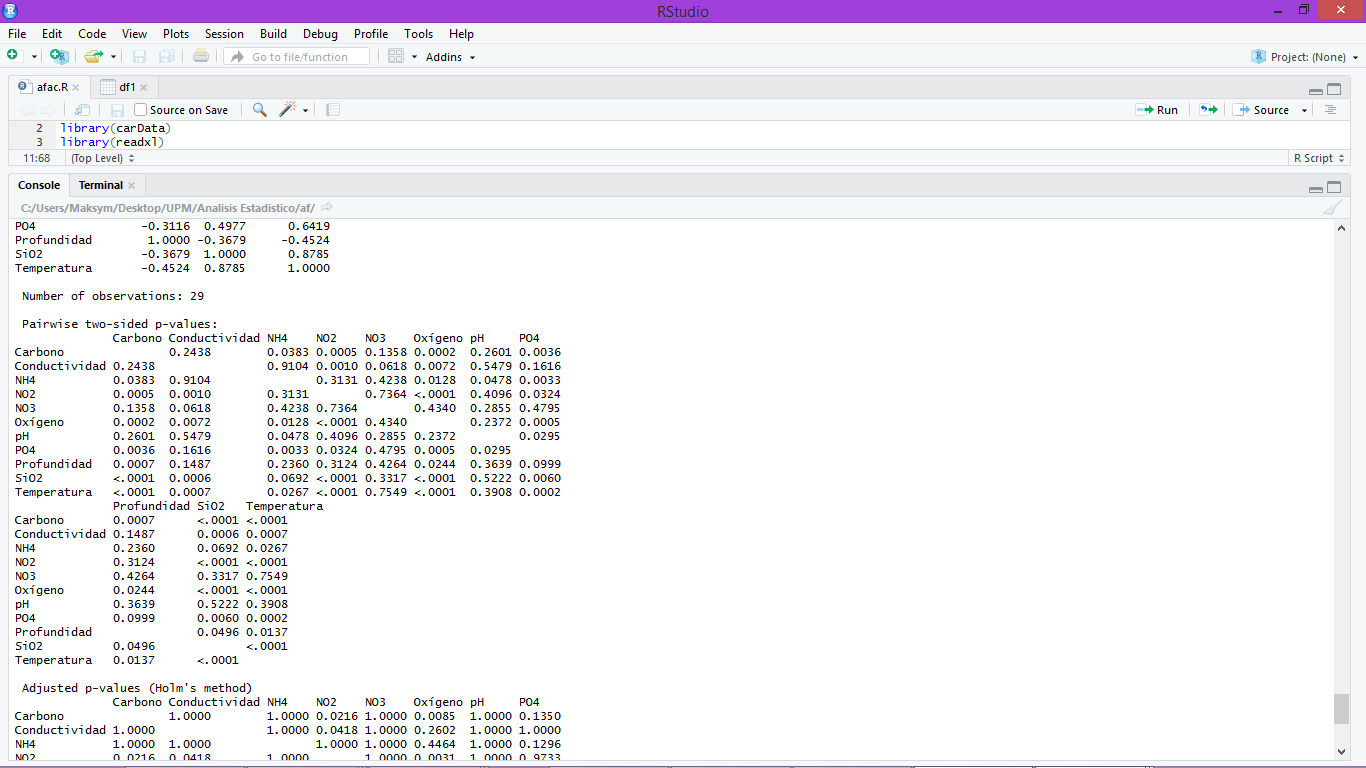
# Ejercicio 2

El segundo ejercicio se ha centrado en el análisis factorial de un conjunto de datos sobre variables limnológicas de varios lagos Neotropicales que indican el grado de productividad (concentración de nutrientes, carbono orgánico disuelto) y la adecuación del hábitat para la vida (profundidad, pH, conductividad, oxígeno disuelto, temperatura). Estos dos factores, productividad y adecuación, podrían explicar razonablemente la correlación observada entre las distintas variables. El objetivo del análisis factorial en este contexto es comprobar si este modelo es adecuado para explicar los datos.

1. Comprobar si las variables están correlacionadas.

Para comprobar la correlación entre variables, se ha en la matriz de correlaciones de R Commander la matriz de correlaciones de Pearson y también se ha mostrado la matriz de p-values.

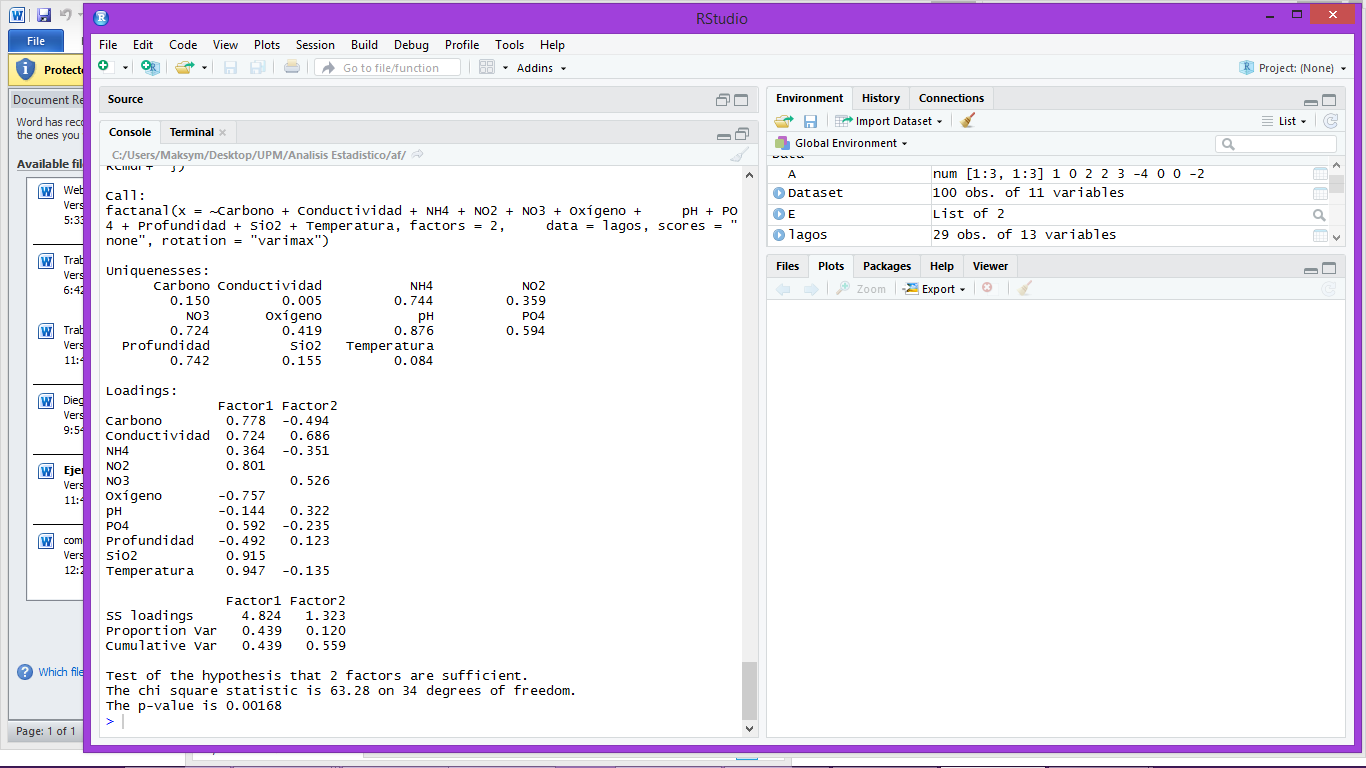




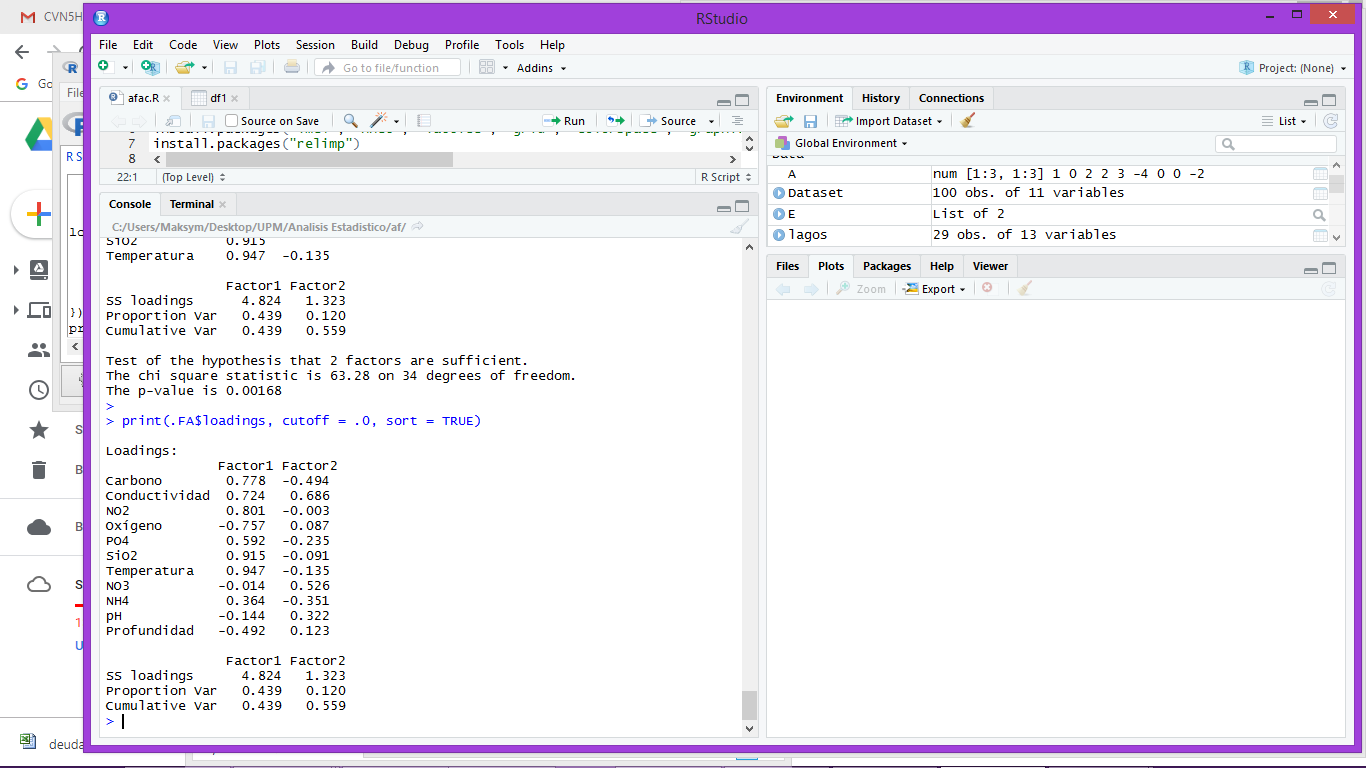
La matriz de correlaciones, como indica su nombre, nos da idea de cómo de correlacionadas están las variables. Los valores en diagonal serán siempre iguales a 1 ya que una variable siempre está relacionada consigo misma. En el caso de nuestro dataset, se puede decir que pocas variables están correlacionadas, con algunas excepciones (ej: NO2 y SiO2, Temperatura y Carbono, SiO2 y temperatura) que presentan valores absolutos >0.8 en la matriz de correlaciones. Teniendo eso en cuenta, a priori, predecimos que el análisis factorial va a ser de mala calidad, ya que este análisis depende directamente de la correlación entre variables.

1. Aplicar Análisis Factorial, con un modelo de dos factores, a todas las variables cuantitativas.

Todas las columnas del dataset excepto las dos primeras son numéricas. Mirando las “uniquenesses” de las variables, vemos que algunas variables están bien explicadas por el modelo ya que tienen una fracción baja de la variabilidad no explicada por ningún factor (conductividad y temperatura), pero las demás están mal explicadas.



Para interpretar nuestros factores, uno esperaría que ordenando las variables por el valor explicado por factores, los que están relacionados con el “Grado de Productividad” (concentración de nutrientes y carbono orgánico disuelto) tendrían un valor alto para el Factor 1 y los que están relacionados con el “Adecuación del hábitat para la vida” (profundidad, pH, conductividad, oxígeno disuelto, temperatura) tendrían un valor alto para el Factor 2, o vice-versa. Los valores de conductividad, oxigeno, profundidad y pH están en valores positivos en factor 2 y cerca del cero en el factor 1 (con la excepción de la conductividad), por lo que podemos inferir que el Factor 2 puede corresponder a la “Adecuación del hábitat para la vida”, aunque la variable “Temperatura” no está bien explicada por el factor (como sería de esperar porque ya intuíamos que el modelo de análisis factorial iba a ser malo, debido a altas uniquenesses y bajos valores de correlación entre variables). El Factor 1 corresponde por lo tanto al Grado de Productividad por lo que vemos que la variable “Carbono”, NO2 y NH4 (quizás relacionada con el metabolismo y por lo tanto la abundancia de nutrientes) está mucho mejor explicado por el Factor 1.



1. Calcular e interpretar el *índice de adecuación muestral KMO*

El indice de adecuacion muestral KMO para este modelo es igual a 0.416. Los valores menores de 0.5 se clasifican como “inacceptables”, por lo tanto, el modelo de analisis factoreial con 2 factores para este dataset es muy malo.

1. Interpretar los resultados: comunalidades, saturaciones, test de suficiencia del modelo con dos factores.

En un analisis factorial perfecto, el valor de saturaciones seria muy proximo de 1 para un factor y muy proximo de 0 para el otro, teniendo el factor especifico tambien muy proximo de 0. En nuestro dataset, para la mayoria de las variables, esto no ocurre ya que para algunas de ellas el valor de uniquenesses es alto y ninguno de los factores es capaz de explicar la variabilidad de la variable individualmente. Siedo que la comunalidad es igual a la suma de los cuandrados y es inversamente proporcional a la uniqueness, la comunalidad sera demasiado baja. Volviendo al grafico del apartado 2, vemos que este modelo, vemos que los dos factores juntos explican apenas cerca de 55% de la variabilidad y el test de suficiencia de factores da un valor de 0.00168, menor que 0.05, por lo que se puede rechazar la hipotesis de que 2 factores son suficientes para construir el modelo de analisis facctorial (tomando el 95% como nivel de significancia).

1. Construir e interpretar el gráfico de saturaciones. Discutir la validez global del modelo factorial para estos datos.



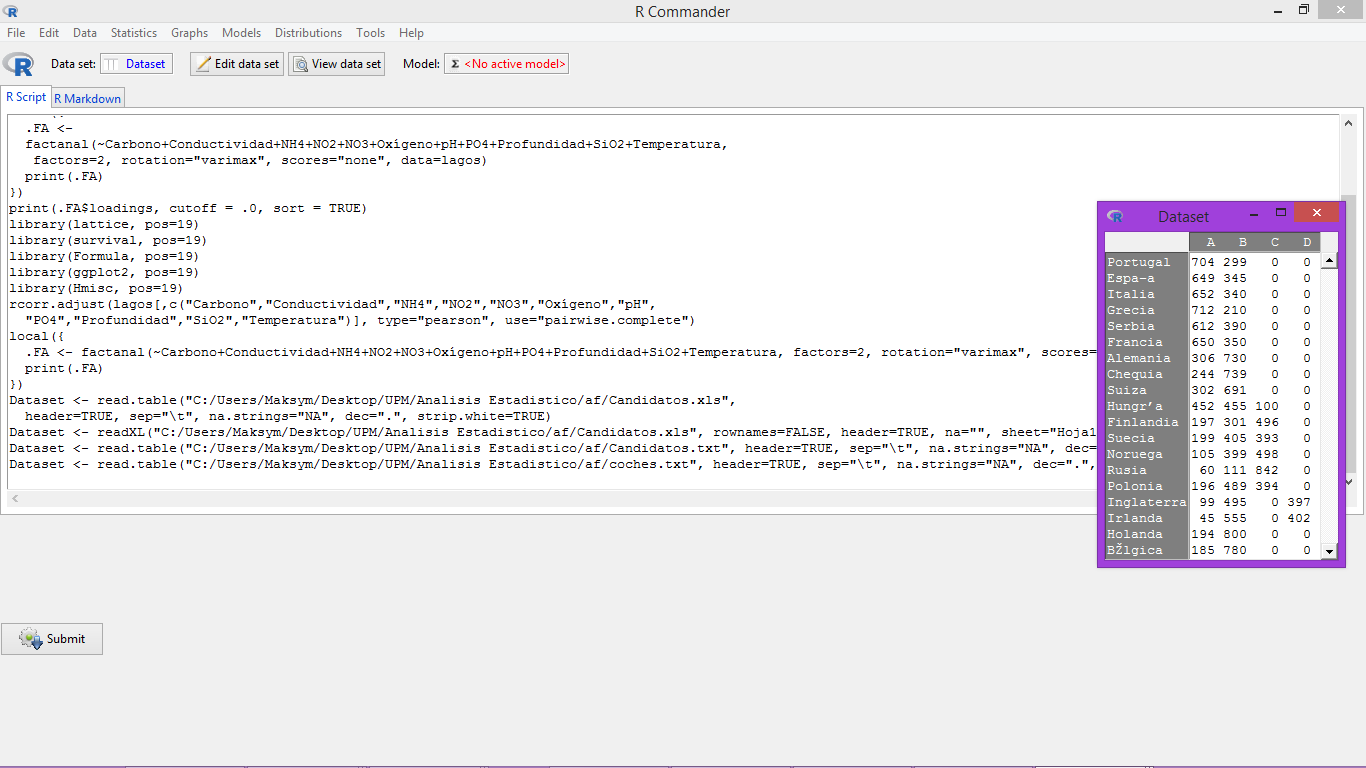
La interpretacion del modelo ha sido explicada ya en apartados anteriores. Aqui se puede ver de una manera mas visual la posicion de las variables en el plano cartesiano de factores. En un analisis factorial perfecto, las variables se agruparian cerca del (X,Y)=(0,1) o (X,Y)=(1,0). En nuestro caso eso no pasa, ya que no todas las variables que sospechariamos pertenecer al fator “Adecuacion al habitat” estan muy arriba en el eje del Y y cerca del 0 en el eje X.

Habria que redifinir las variables que componen los factores, o aumentar el numero de factores para construir un modelo de analisis factorial valido.

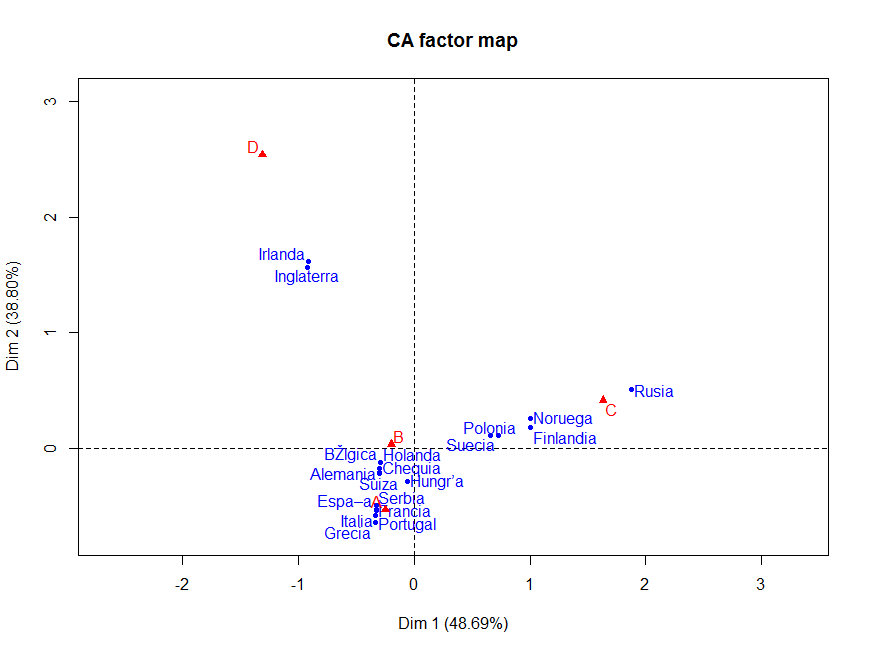
# Ejercicio 3

El segundo ejercicio consiste en realizar un análisis de correspondencia sobre datos tomados de muestras aleatorias en diferentes países europeos, de marcas de automóvil (identificadas como A, B, C y D). El objetivo consiste en determinar el grado de asociación entre marcas y países. Cada país tiene un determinado perfil de marcas, y cada marca una determinada distribución por países.

1. Leer desde R Commander el archivo Excel coches.xls, e inspeccionar los datos.



1. Realizar el Análisis de Correspondencias utilizando todos los países y marcas.



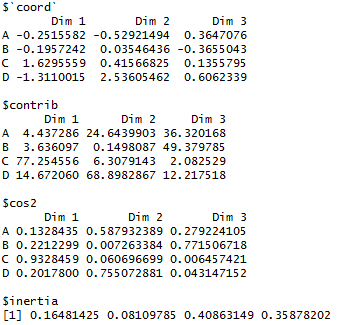
1. Interpretar el gráfico resultante.

Los coches C y D están en los extremos de los ejes de dimensión 1 y 2, respectivamente. Las marcas C y D son las que tienen una distribución de países que los compra más definida. Se nota claramente que los coches de marca D son comprados sobretodo en Irlanda e Inglaterra y los de marca C en Rusia, Noruega, Finlandia, Polonia y Suecia. Los demás coches son comprados por los restantes países en igual medida, siendo que la marca A tiende a tener más popularidad en los países de sur de Europa.

1. Interpretar las varianzas de los ejes, las contribuciones absolutas y relativas, y las inercias.



La dimensión 1 explica el 48.7% de la variabilidad de los datos y la segunda explica 38.8%, siendo que juntas, las dos dimensiones representadas en el grafico explican cerca de 87% de la variabilidad total de los datos. Con 3 dimensiones, es posible explicar el 100%.



Los datos de resultados de columnas nos dan casi la misma información que el grafico construido en el apartado 2. En primer lugar, en la tabla “coord”, vemos las mismas coordenadas que veíamos visualmente en a gráfica: la marca D se sitúa en el cuadrante superior izquierda ya que tiene un valor muy negativo en el eje X (Dimension 1) y muy positiva en Y (dimensión 2); la marca C se sitúa en el cuadrante superior derecho ya que tiene un valor muy positivo en el eje X positiva en Y. Si hubiéramos representado la dimensión 3, la marca D estaría también en el extremo del eje.

En según lugar, si miramos la tabla “contrib”, entendemos que las marcas C y D tienen los mayores valores en las 2 primeras dimensiones (principalmente la D), lo que significa que estas dos marcas de coche son las que nos permiten discriminar más la preferencia de los países en la compra de coches.

En tercer lugar, si miramos la tabla “cos2”, entendemos el peso de cada eje en la explicación de cada marca de coche. La elección del coche C está casi totalmente explicada por la dimensión 1, mientras que la elección de D y B esta explicada en más de 75% por la dimensión 2 y 3, respectivamente. La elección del coche A no está explicada en una proporción significativamente alta por ningún eje individual.

Finalmente, si miramos el apartado “inertia”, vemos que los coches C y D, como esperábamos, tienen el valor más elevado, por lo que tienen una mayor importancia relativa.





A continuación, se hizo el mismo análisis pero por países. Aquí vemos que Finlandia, Suecia, Noruega, Rusia, Irlanda e Inglaterra son los países que presentan una tendencia más acentuada de elegir una marca de coche particular.

Vemos también en la tabla de contribuciones, que Finlandia, Rusia y Noruega son los principales países que contribuyen para el eje 1 mientras que para el eje 2 son Inglaterra e Irlanda.

En resumen, en un análisis de correspondencia con 2 dimensiones, si nos basamos en la primera dimensión, veremos que los países que más contribuyen son Rusia, Finlandia, Noruega, Suecia y Polonia y que en estos países es más fácil encontrar coches de marca D, mientras que en caso del eje 2 son Inglaterra e Irlanda, siendo los coches de marca C más comprados. Los demás países y marcas contribuyen muy poco para las dos primeras dimensiones del análisis de correspondencia (sobre todo la marca B, cuyas coordenadas están muy cerca del 0). Sin embargo, se puede llegar a distinguir la tendencia de España, Portugal, Serbia, Italia y sobretodo Grecia en estar más abajo en el eje de la dimensión 2 y estar asociados a la compra de coches de marca A. Los coches de marca B tienen una distribución casi uniforme en todos los países estudiados.

#CODA

library(car)

library(carData)

library(readxl)

library(corrplot)

library(Rcmdr)

#CODA

#Factor Analysis

#Load data

lagos<- read.table("C:/Users/Maksym/Desktop/UPM/Analisis Estadistico/af/lagos.txt", header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)

#Correlation matixx

library(lattice, pos=19)

library(survival, pos=19)

library(Formula, pos=19)

library(ggplot2, pos=19)

library(Hmisc, pos=19)

rcorr.adjust(lagos[,c("Carbono","Conductividad","NH4","NO2","NO3","Oxígeno","pH",

+ "PO4","Profundidad","SiO2","Temperatura")], type="pearson", use="pairwise.complete")

#AF

.FA <- factanal(~Carbono+Conductividad+NH4+NO2+NO3+Oxígeno+pH+PO4+Profundidad+SiO2+Temperatura, factors=2, rotation="varimax", scores="none", data=lagos)

+ print(.FA)

factanal(x = ~Carbono + Conductividad + NH4 + NO2 + NO3 + Oxígeno + pH + PO4 + Profundidad + SiO2 + Temperatura, factors = 2, data = lagos, scores = "none", rotation = "varimax")

#Loadings sorting

print(.FA$loadings, cutoff = .0, sort = TRUE)

#Saturation Graph

.FA <-factanal(~Carbono+Conductividad+NO2+Oxígeno+PO4+SiO2+Temperatura+NO3+NH4+pH+Profundidad, factors=2, rotation="varimax", scores= "none", data=lagos)

plot(.FA$loadings, type="n")

text(.FA$loadings,labels=row.names(.FA$loadings),cex=.7)

#KMO calculation

R2 = cor(lagos[,3:10])^2 # matriz de correl. al cuadrado

R2.suma = sum(R2)-dim(R2)[1] # suma correl. al cuadrado

RP2 = partial.cor(lagos[,3:13])$R^2

RP2.suma = sum(RP2) # suma correl. parciales al cuadrado

kmo = R2.suma/(R2.suma + RP2.suma) # calcula KMO

kmo # escribe resultado KMO

----------------------------------------------------------------------

# Correspondence analysis

#Load Data

Dataset <- read.table("C:/Users/Maksym/Desktop/UPM/Analisis Estadistico/af/coches.txt", header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)

#Libraries

install.packages("FactoMineR")

library(FactoMineR) # carga el paquete FactoMineR

#Excluding 1st row

rownames(Dataset)=Dataset[,1]

Dataset <- Dataset[ ,2:dim(Dataset)[2]]

#Performing correspondence analysis

res <- CA(Dataset, ncp=5, graph = TRUE) # aplica AC DatoAC, creando el objeto “res”

#Looking at eigenvalues

res$eig

#Looking at columns coordinates

res$col

#Looking at row coordinates

res$row

#Rversion

#R version 3.5.1 (2018-07-02)