UNIVERSIDAD
NACIONAL DEL OESTE

Explotación de Datos

Actividad N^{o} 2

Análisis de Componentes Principales

Profesores:

Dejean, Gustavo
Españadero, Juan
Mendoza, Dante

Integrantes Grupo B:

Benitez, Nicolas
Garcia Ravlic, Ignacio Agustin
Rechimon, Pablo Hernan
Rodriguez, Miguel Angel

Fecha de entrega:

12 de septiembre de 2020

Resumen

A partir del dataset de ozono y el lenguaje R se deseaba analizar si es posible realizar un ACP, y realizarlo en ese caso, con el objetivo de encontrar la verdadera dimensión del problema y así poder realizar en un futuro un análisis. Como los test de KMO y Bartlett nos afirmaron que podíamos realizar tal ACP procedimos logrando así una reducción de 70 variables (de 73 a 3), manteniendo un 70% de la información más importante.

Palabras Clave: *Análisis de Componentes Principales - Reducción de la Dimensión - Ozono - Programación - Estudio de Correlación - Prueba de Bartlett - Indice de Kaiser-Meyer-Olkin*

Índice

Graficos

1 Introducción

1.1 Problemática

El dataset posee una gran dimensión (73 variables), además de que no conocemos si hay o no correlaciones entre estas, por lo cual el análisis de tal se vuelve muy dificultoso y engorroso, con un alto coste computacional.

1.2 Datos a utilizar

Utilizaremos el dataset Ozono, provisto por la cátedra, cuyos autores originales son:

- Kun Zhang, Department of Computer Science, Xavier University of Lousiana.
- Wei Fan, IBM T.J.Watson Research.
- XiaoJing Yuan, Engineering Technology Department, College of Technology, University of Houston

1.3 Objetivo

Por ende debemos corroborar que sea posible, y si lo es, realizar un Análisis de Componentes Principales para reducir la dimensión con la menor pérdida de información posible.

2 Desarrollo

2.1 Análisis de los datos

Los atributos Tn indican valores de Temperatura en el intervalo n[0,23] y los valores WSn indican los valores del viento en el intervalo n[0,23], ambos haciendo referencia a las horas del día.

- WSR_PK: continuo. velocidad del viento pico (máximo)
- WSR_AV: continuo. velocidad media del viento
- T_PK: continuo. Temperatura pico (máxima).
- T_AV: continuo. Temperatura promedio
- T85: continuo. Temperatura a un nivel de 850 hpa (o aproximadamente 1500 m de altura)
- RH85: continuo. Humedad relativa a 850 hpa
- U85: continuo. (Viento U - viento en dirección este-oeste a 850 hpa)
- V85: continuo. Viento V - Viento en dirección N-S a 850
- HT85: continuo. Altura geopotencial a 850 hpa, es aproximadamente la misma que la altura a baja altitud
- T70: continuo. T a un nivel de 700 hpa (aproximadamente 3100 m de altura)
- RH70: continuo.
- U70: continuo.
- V70: continuo.
- HT70: continuo.
- T50: continuo. T a nivel de 500 hpa (aproximadamente a 5500 m de altura)
- RH50: continuo.
- U50: continuo.
- V50: continuo.
- HT50: continuo.
- KI: continuo. Medida del potencial tormenta basado en gradiente vertical de temperatura, contenido de humedad de la atmósfera inferior y la extensión vertical de la capa húmeda.
- TT: continuo. Medida de Fuerza de Tormenta

- SLP: continuo. Presión a nivel del mar
- SLP_: continuo. Cambio de SLP desde el día anterior
- Precp: continuo. - Precipitaciónes

2.2 Preparación de los datos

Preparamos los datos en base a nuestros requerimientos. Para ello visualizamos y elminamos el numero de casos con valor null, hacemos un resumen de los nuevos datos y por ultimo creamos una matriz de correlaciones.

- sapply()
- na.omit()
- summary()
- cor()

Antes de aplicar el análisis de componentes principales, efectuamos el test de Barlett para saber si podemos factorizar las variables originales de forma eficiente:

```
> cortest.bartlett(cor_data)
$chisq
[1] 15655.81

$p.value
[1] 0

$df
[1] 2628
```

Fig. 1: Prueba de Bartlett

Como el p.value es 0 se rechaza la hipótesis nula que afirma que las variables no están correlacionadas, y se continúa con el ACP.

También revisamos el índice Kaiser-Meyer-Olkin (KMO) para comparar los valores de correlación de las variables y sus correlaciones parciales. Si el indice KMO es cercano a 1, significa que puede hacerse el análisis de componentes principales:

```
> KMO(data[, -1])
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = data[, -1])
Overall MSA =  0.91
```

Fig. 2: Indice Kaiser–Meyer–Olkin

Al ser 0.91, podemos proseguir.
Por último realizamos una visualización para la matriz de correlación:

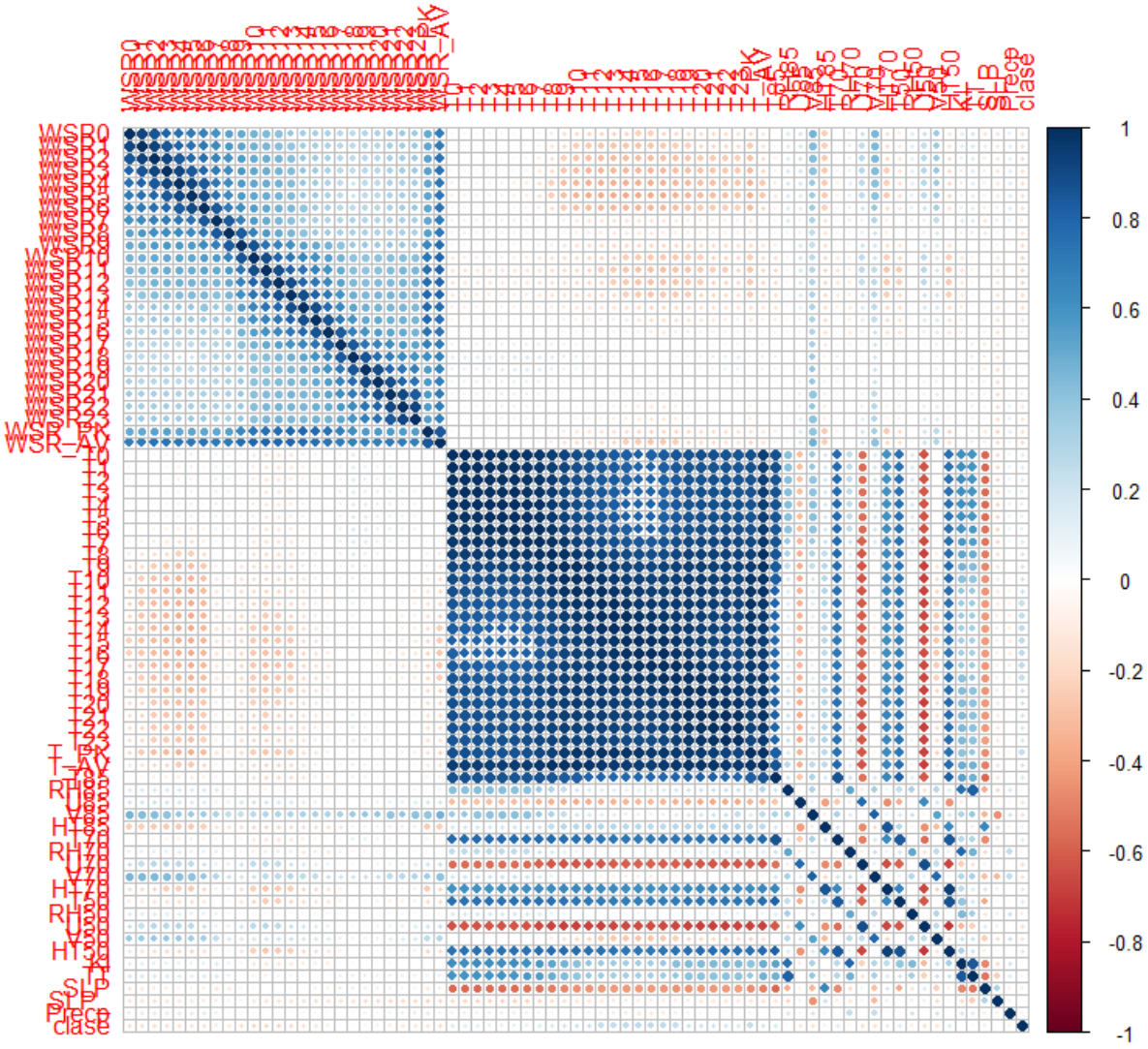

Fig. 3: Matriz de Correlacion

Como se puede observar, el análisis se vuelve muy dificultoso, aunque podemos rescatar que hay correlación positiva entre las variables del viento, también entre las variables de temperatura, y que entre viento y temperatura no hay ningún tipo de correlación.

2.3 Análisis de Componentes Principales

Realizamos el calculo de los componentes principales a travez de la funcion prcomp().

Hacemos un gráfico de dos dimensiones para explicar el porcentaje de varianza de cada componente principal, y ejecutamos un summary de las CP para obtener la proporción de Varianza acumulada y así definir a cuantas dimensiones reducir nuestro dataset.

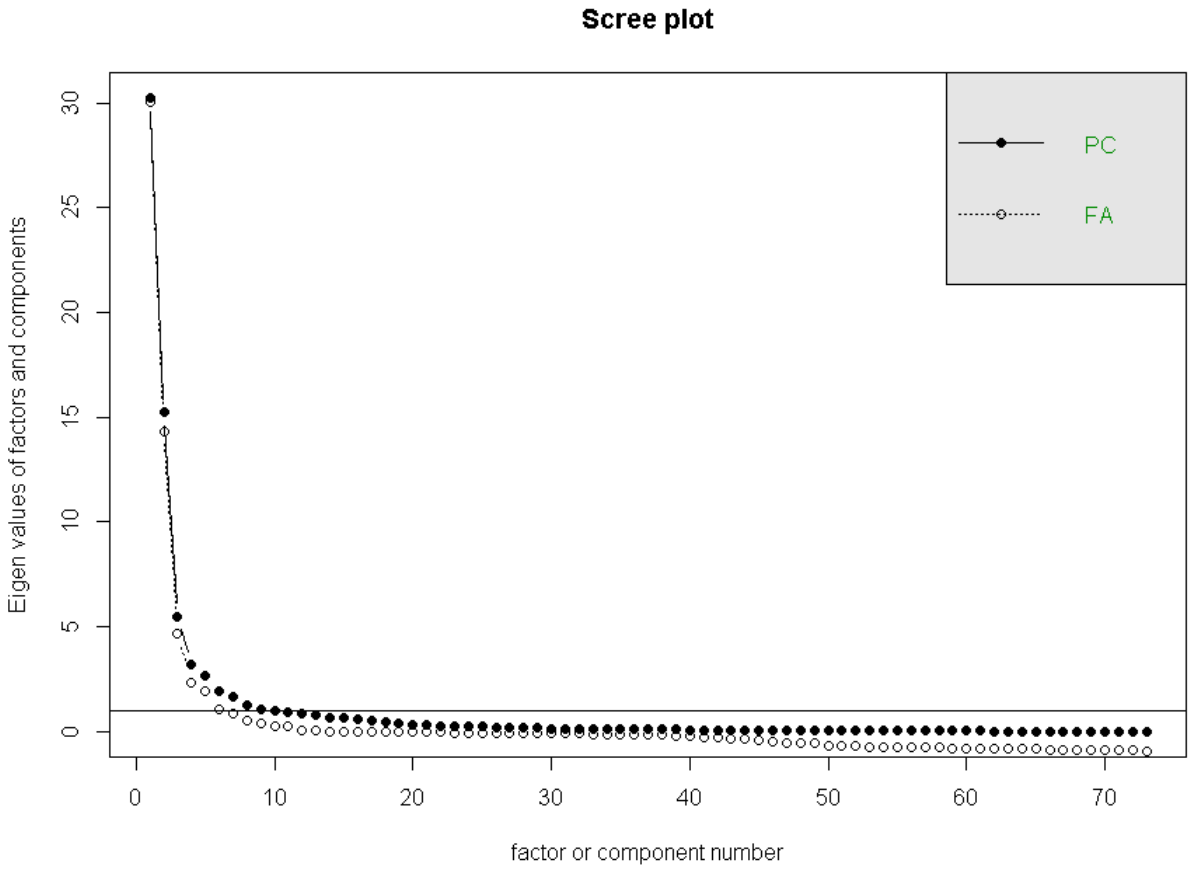

Fig. 4: Scree plot

```
> summary(cp)
Importance of components:
                          PC1    PC2     PC3
Standard deviation     5.4988 3.9020 2.34036
Proportion of Variance 0.4142 0.2086 0.07503
Cumulative Proportion  0.4142 0.6228 0.69780
```

Fig. 5: Proporción de varianza acumulada

Para este dataset, vamos a optar por tres componentes principales, ya que con estos podemos observar casí el 70% de los datos reduciendo el set de datos de 73 a 3 dimensiones.

2.4 Análisis de las visualizaciones

Realizamos unas pequeñas visualizaciones de los 3 componentes principales para la observacion de patrones y tendencias:

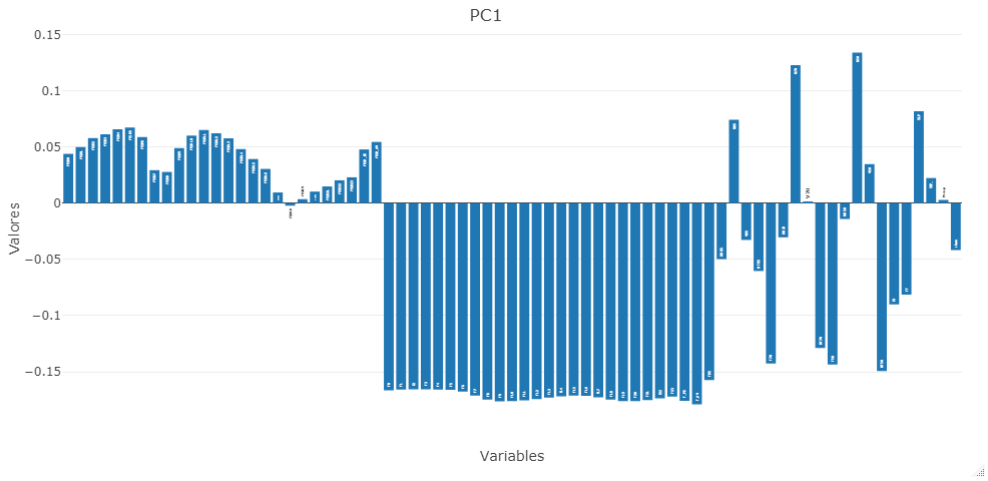

Fig. 6: Componente Principal 1

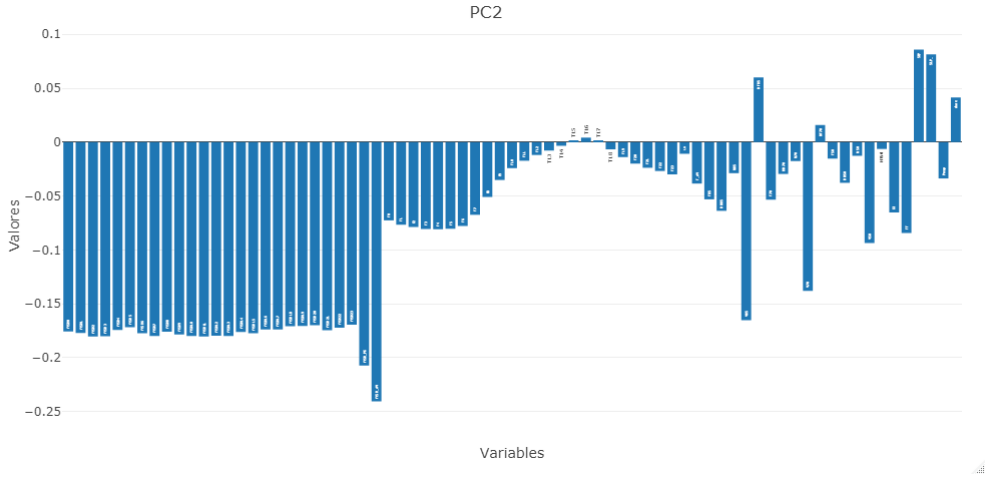

Fig. 7: Componente Principal 2

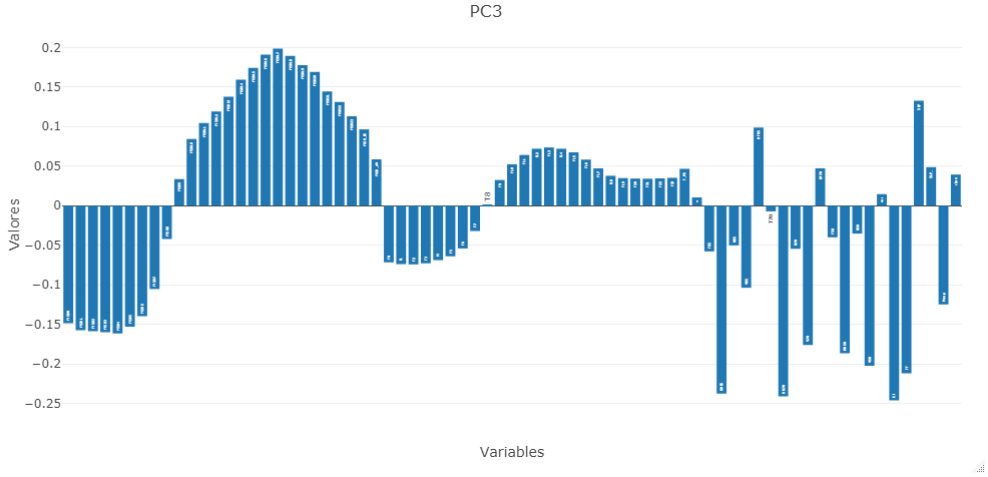

Fig. 8: Componente Principal 3

Y por último el biplot entre la Componente Principal 1 y 2, en el que se pueden apreciar dos grandes agrupamientos de variables, indicando correlación positiva en cada grupo, y que estos grupos están de forma perpendicular, indicando correlación nula

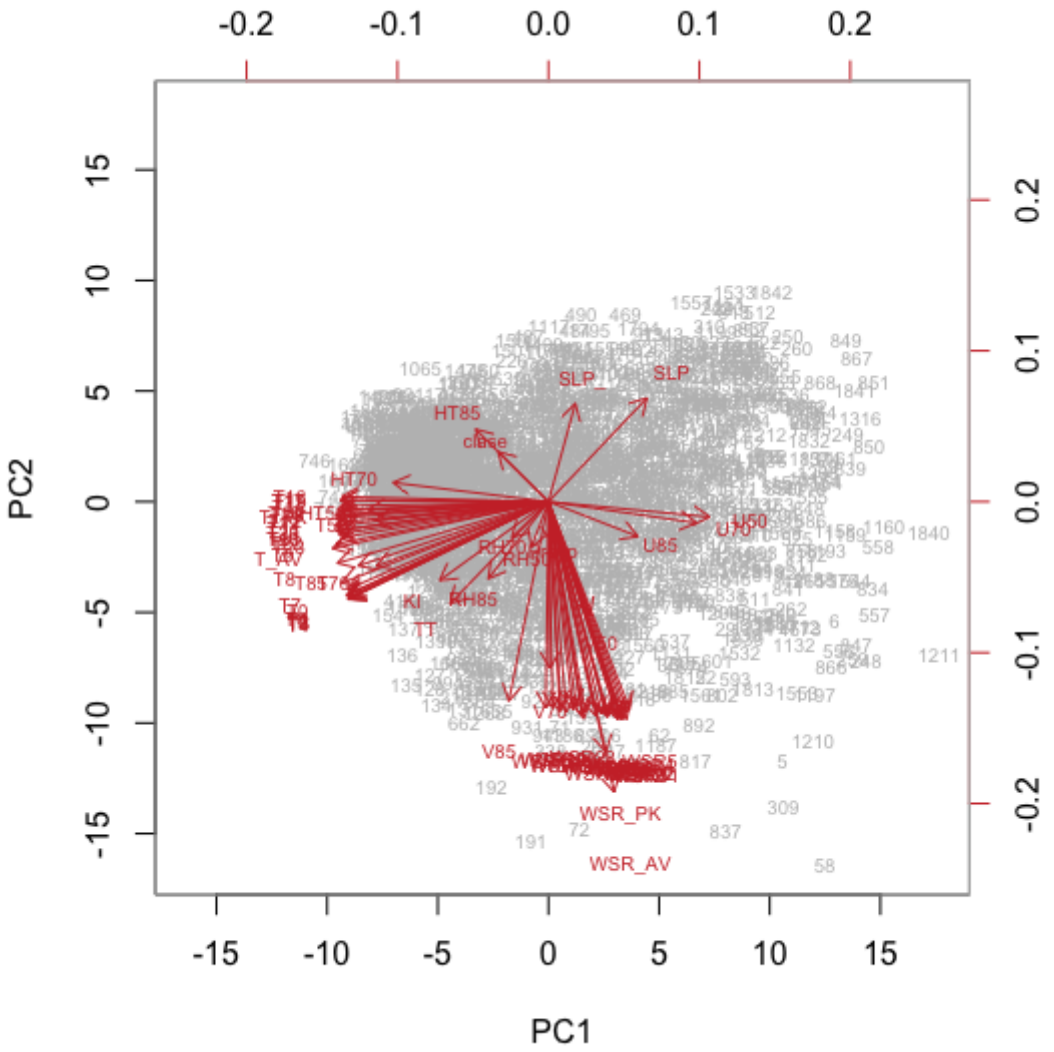

Fig. 9: Biplot

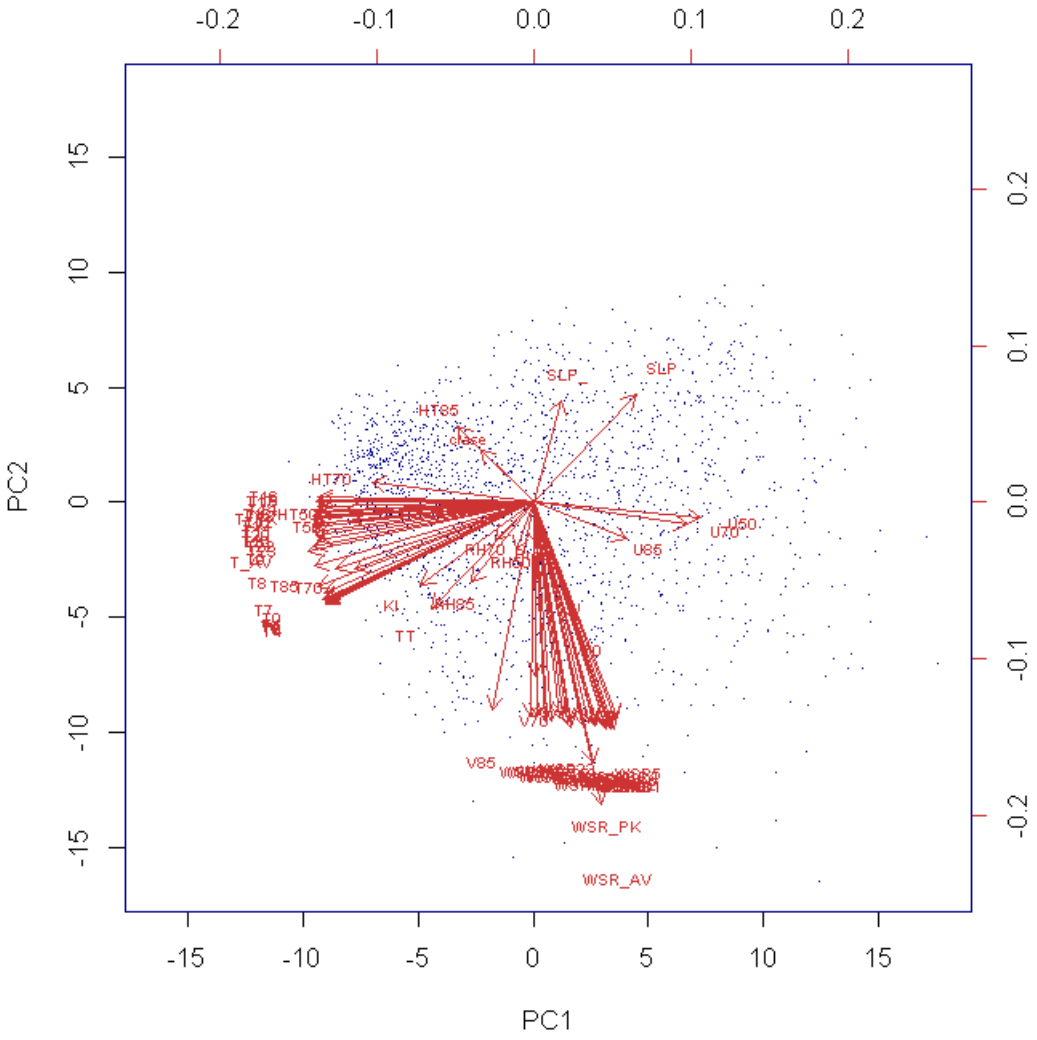

Fig. 10: Biplot con puntos

3 Conclusión

Una vez realizado el ACP obtuvimos una matriz en la que se agrupa la mayor porción de la información, en nuestro caso en sólo 3 variables logramos agrupar el 70% de la información, logrando así una gran reducción de la dimensión con una pequeña pérdida de información.

4 Anexo

4.1 Código en R

```
#################################################################
# Analisis de Componentes Principales del dataset de Ozono      #
#                                                               #
# Creado: 2020-09-06   v. 2020-09-10                            #
# Ultima Mod: plots                                             #
#                                                               #
# Grupo B: GAD, Benitez, Garcia, Rechimon, Rodriguez            #
#                                                               #
#################################################################

##### IMPORTAMOS LAS BIBLIOTECAS A USAR
library(readxl)
library(corrplot)
library(PerformanceAnalytics)
library(psych)
library(rela)
library(dplyr)
library(ggplot2)
library(plotly)

options(scipen = 6)



###### LEEMOS EL DATASET
data <- read_excel("ozono.xls", na = "?")

# Corroboramos que se haya leido bien
head(data, 5)

# Visualizamos el numero de variables y de casos
ncol(data)
nrow(data)



##### PREPARACION DE LOS DATOS
# Visualizamos el numero de casos con null
sapply(data, function(x) sum(is.na(x)))

#--- Preferimos mantener variables y eliminar casos NA ---#
data <- na.omit(data)

sapply(data, function(x) sum(is.na(x)))
nrow(data)

# Hacemos un resumen de los datos
summary(data)


```

```
# Matriz de correlaciones
cor_data <- cor(data[, -1])

#--- Azul cuando es cor+ y Rojo para cor-
corrplot(cor_data)

#chart.Correlation(data[, -1]) # para graficar correlciones y
    histogramas



###
# Test de barlett
cortest.bartlett(cor_data, n= 1847)

# Test de KMO
KMO(data[, -1])

# Grafico para ver con cuantas CP nos quedamos
scree(data[, -1])


# Componentes Principales
cp <- prcomp(data[,-1], scale = TRUE)

#Resumen de los Componentes Principales
summary(cp)
names(cp)
cp$center  # Media de cada variable
cp$scale   #  Desviacion estandard de cada variable
cp$sdev    # Varianza de cada componente
#cp$rotation
#cp$x


# para graficar:
biplot(x = cp, scale = 0, cex = 0.6, col = c("grey", "brown3"))

biplot(x = cp, scale = 0, cex = 0.6, xlabs=rep(".", nrow(data)),col
    = c("grey", "brown3"))

pc1 <- cp[[2]][,1]
pc2 <- cp[[2]][,2]
pc3 <- cp[[2]][,3]
pc4 <- cp[[2]][,4]
pc5 <- cp[[2]][,5]
pc6 <- cp[[2]][,6]
pc7 <- cp[[2]][,7]

pc <- data.frame(cbind(pc1, pc2, pc3, pc4, pc5, pc6, pc7))

rm(pc1, pc2, pc3, pc4, pc5, pc6, pc7)

fig <- plot_ly(x = ~pc[1,], y = ~pc$pc1, name = "PC1", type = "bar"
    , text = rownames(pc), textangle=-90,  textposition='auto')%>%
  layout(
```

```
    title = "PC1",
    xaxis = list(title = "Variables", showticklabels=FALSE),
    yaxis = list(title = "Valores")
  )
fig

fig1 <- plot_ly(x = ~pc[2,], y = ~pc$pc2, name = "PC2", type = "bar
    ", text = rownames(pc), textangle=-90,  textposition='auto')%>%
  layout(
    title = "PC2",
    xaxis = list(title = "Variables", showticklabels=FALSE),
    yaxis = list(title = "Valores")
  )
fig1

fig2 <- plot_ly(x = ~pc[2,], y = ~pc$pc3, name = "PC3", type = "bar
    ", text = rownames(pc), textangle=-90,  textposition='auto')%>%
  layout(
    title = "PC3",
    xaxis = list(title = "Variables", showticklabels=FALSE),
    yaxis = list(title = "Valores")
  )
fig2
```