OCE 313
TÉCNICAS DE ANÁLISIS NO PARAMÉTRICO

CLASE 13 – Análisis de componentes principales

Dr. José Gallardo

Junio 2021

Contenidos de la clase

- ¿Qué son los análisis de componentes principales?
- ACP con R para Oceanografía.
- Elaborar análisis de componentes principales con R

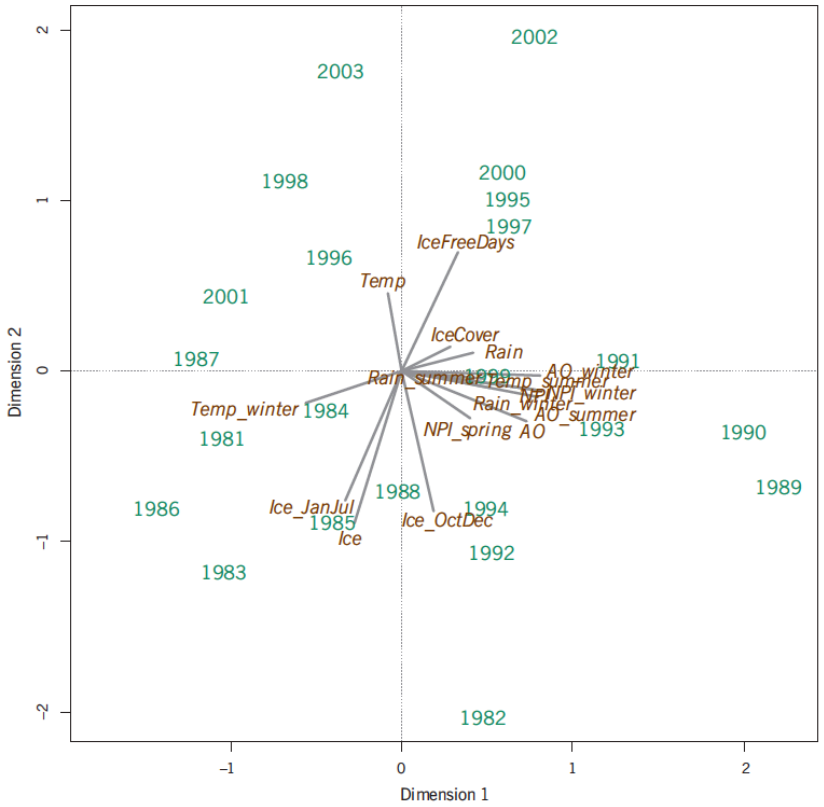
¿Qué son los análisis de componentes principales?

Análisis de componentes principales (ACP)

Gráficas biplot

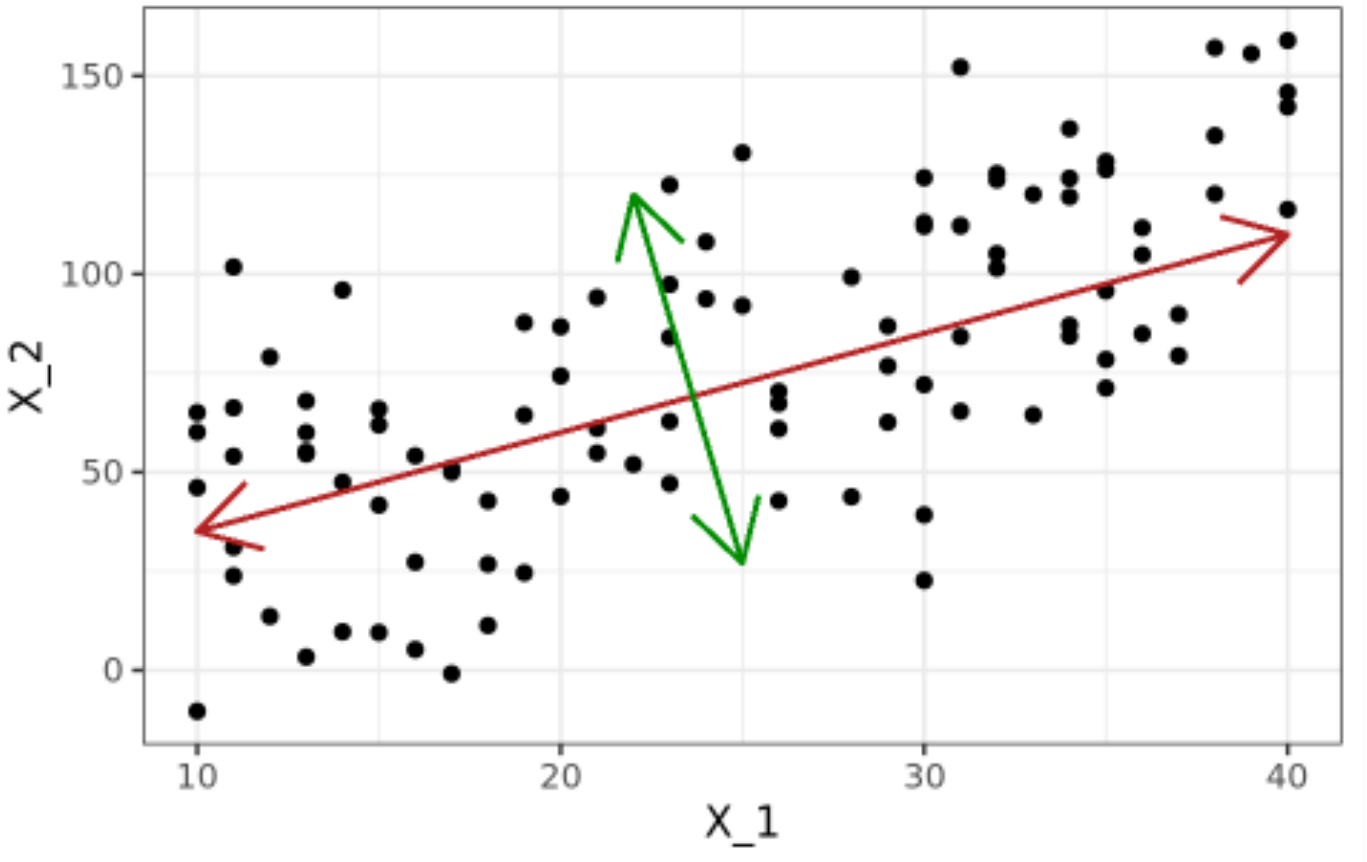
Es una herramienta utilizada para realizar análisis exploratorio de datos multivariantes y para construir modelos predictivos.

Permite reducir la dimensionalidad y encontrar patrones en un set de datos mediante el calculo de los "componentes principales".



¿Qué son los componentes principales?

CP: Combinación lineal de las variables originales no corr. entre si (perpendiculares / ortogonales).



Ejemplo
2 var.cor.
2 CP

Solución matemática para obtener los CP

Calcular los valores y vectores propios de la matriz de Varianza/covarianza de los datos.

Supuestos

Linealidad: Se asume que los datos observados son combinación lineal de una cierta base.

Normalidad: Los datos se distribuyen de manera gaussiana.

Eingenvalue y eigenvector

Cada eigenvector corresponde a un CP y la varianza explicada por cada CP se estima desde su eigenvalue

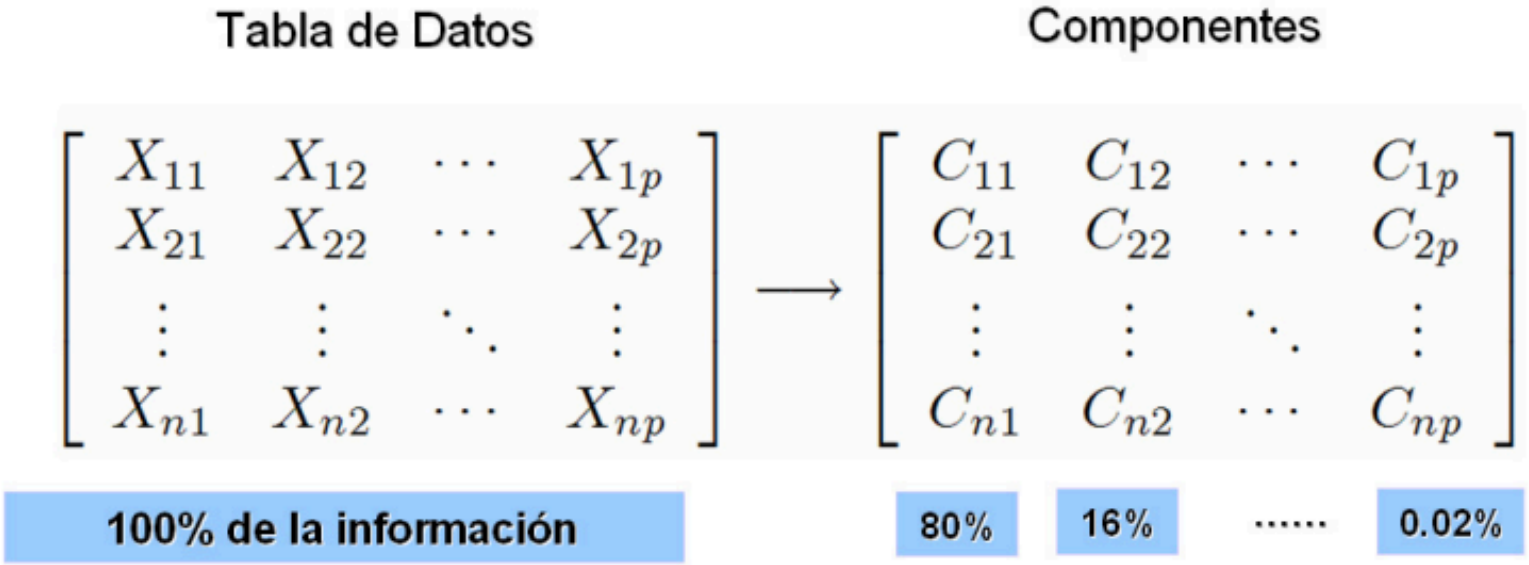


FIGURE 1. Transformación de las variables originales en componentes.

http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/cap2.143111051.pdf

ACP con R para Oceanografía

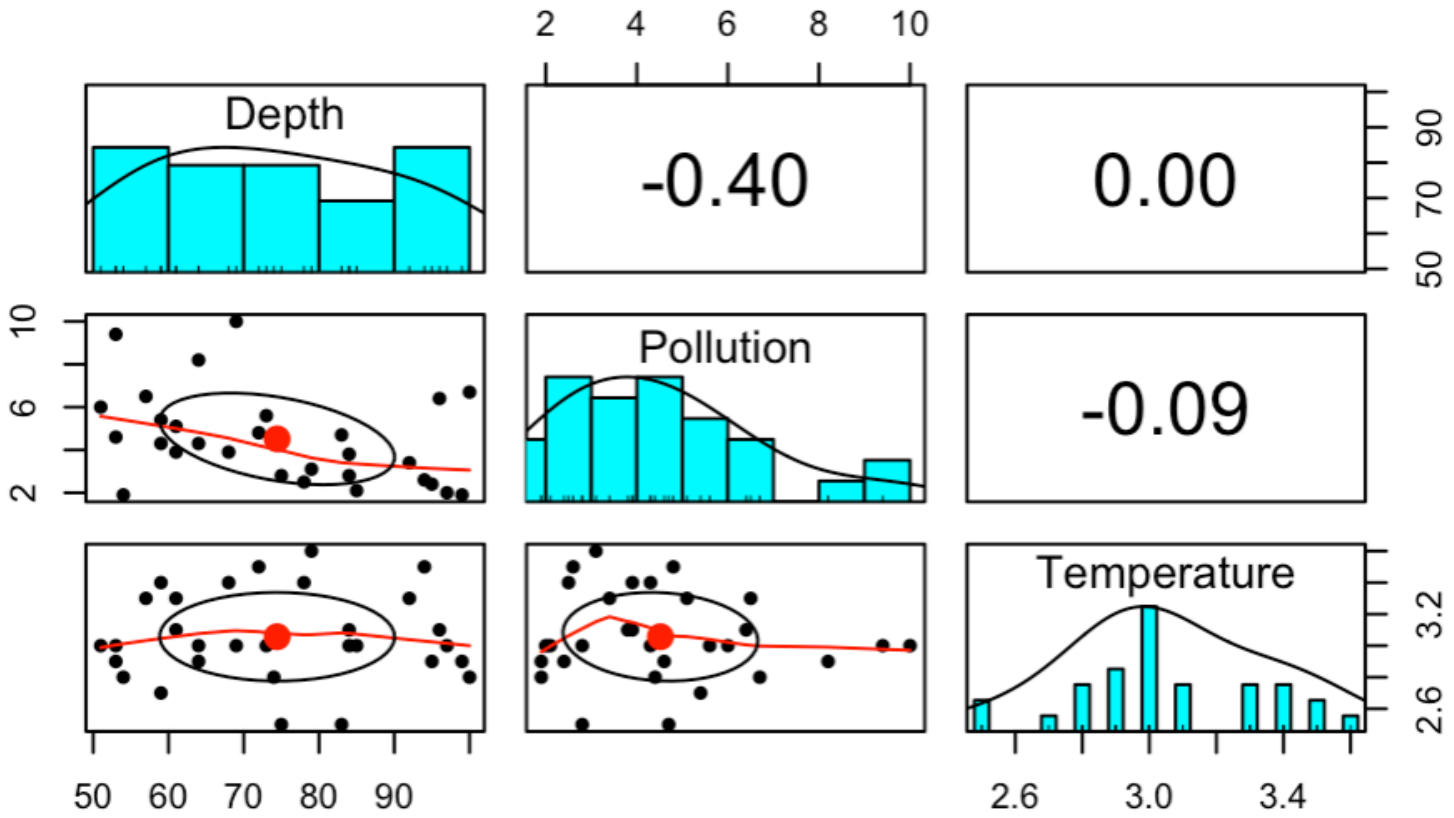
Datos multivariantes – Toy set

Sitio	*Depth*	*Pollution*	*Temperature*
s1	72	4,8	3,5
s2	75	2,8	2,5
s3	59	5,4	2,7
s4	64	8,2	2,9
s5	61	3,9	3,1
…	…	…	…
s26	78	2,5	3,4
s27	85	2,1	3,0
s28	92	3,4	3,3
s29	51	6,0	3,0
s30	99	1,9	2,9

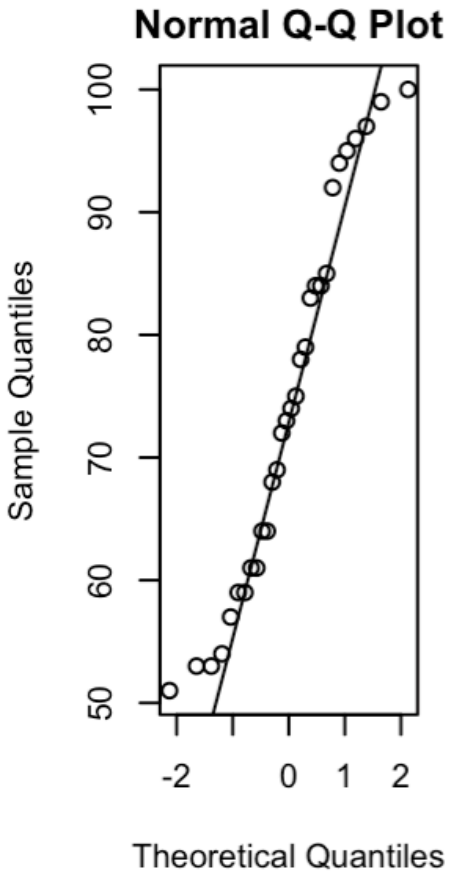
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

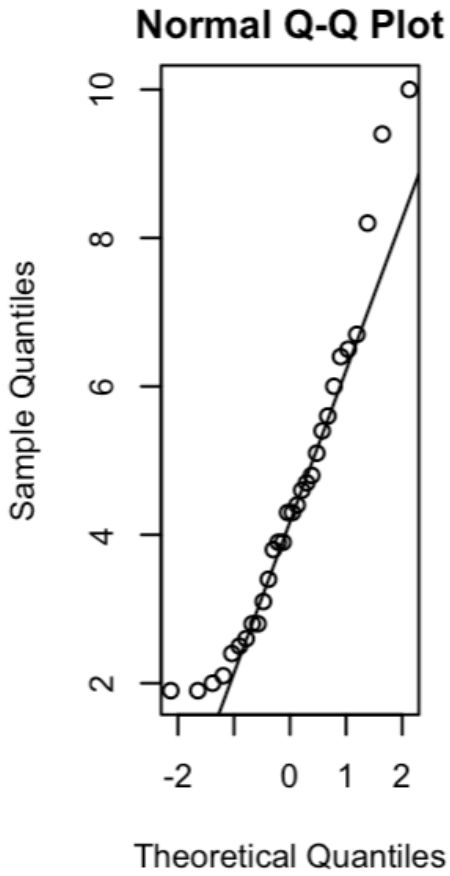
Etapa 1: Correlaciones

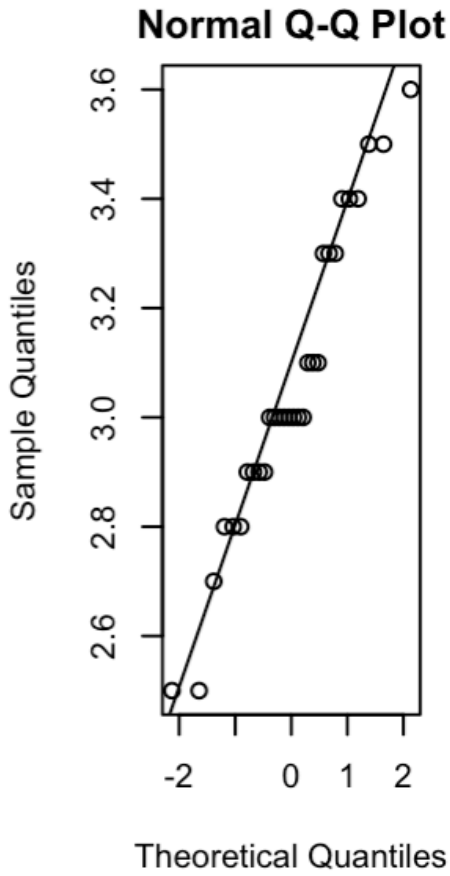
pairs.panels(bioenv[]) 7:9



Etapa 2: Normalidad qqplot







PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Etapa 2: Normalidad Shapiro test

Shapiro-Wilk normality test

data: bioenv$Depth
W = 0.93774, p-value = 0.080

data: bioenv$Pollution
W = 0.91871, p-value = 0.025

data: bioenv$Temperature
W = 0.95337, p-value = 0.21

Etapa 3 – Configuración de datos

- El nombre de los sitios debe ser incluido en el nombre de las filas.

	Depth <dbl>	Pollution <dbl>	Temperature <dbl>
s1	72	4.8	3.5
s2	75	2.8	2.5
s3	59	5.4	2.7
s4	64	8.2	2.9
s5	61	3.9	3.1
s6	94	2.6	3.5

Etapa 4 - ACP

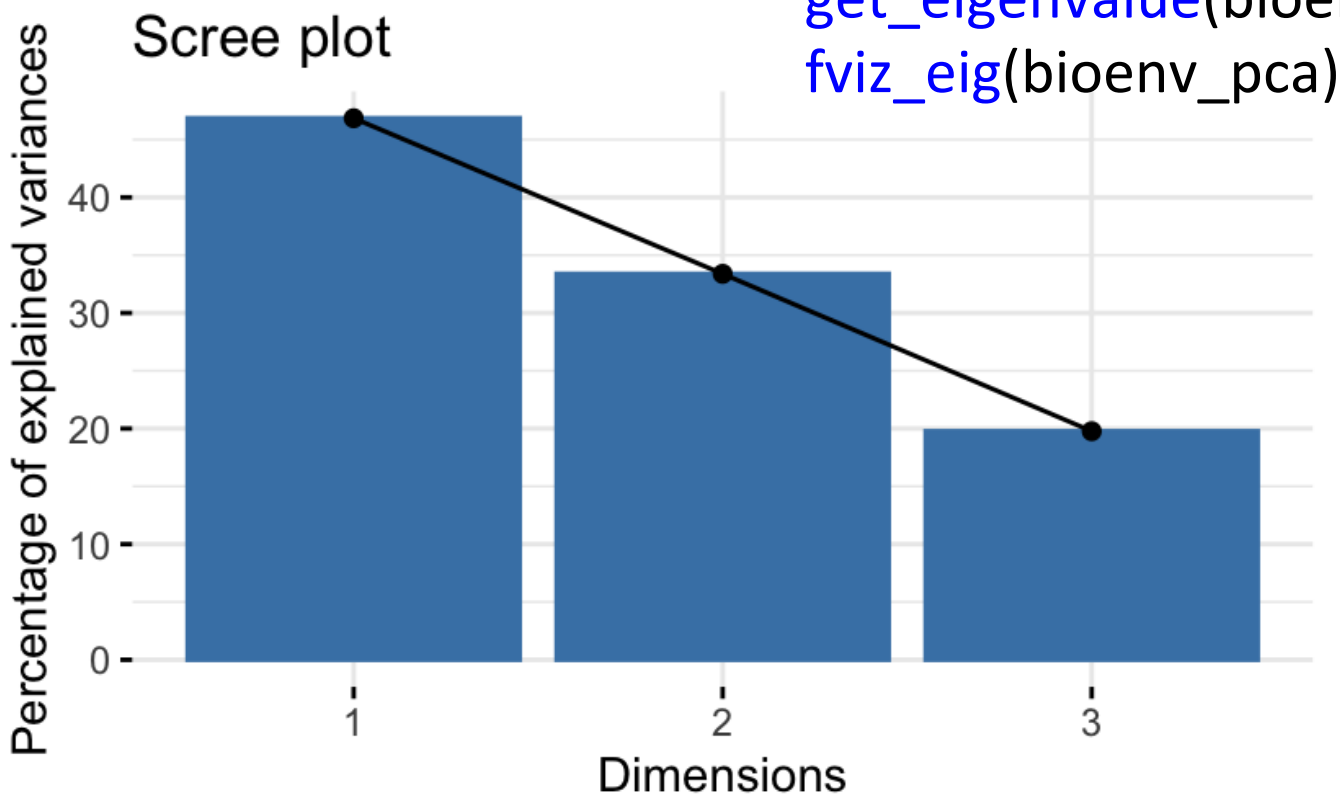
```
library(factoextra)
bioenv_pca <- prcomp(bioenv, scale = TRUE)
```

```
Standard deviations (1, .., p=3):
[1] 1.1854775 1.0007570 0.7701484

Rotation (n x k) = (3 x 3):
                    PC1          PC2       PC3
Depth         0.6892610 -0.226750181 0.6881160
Pollution    -0.7077454 -0.007574555 0.7064270
Temperature   0.1549703  0.973923499 0.1657022
```

Varianza explicada por cada CP

get_eigenvalue(bioenvpca)
fviz_eig(bioenv_pca)

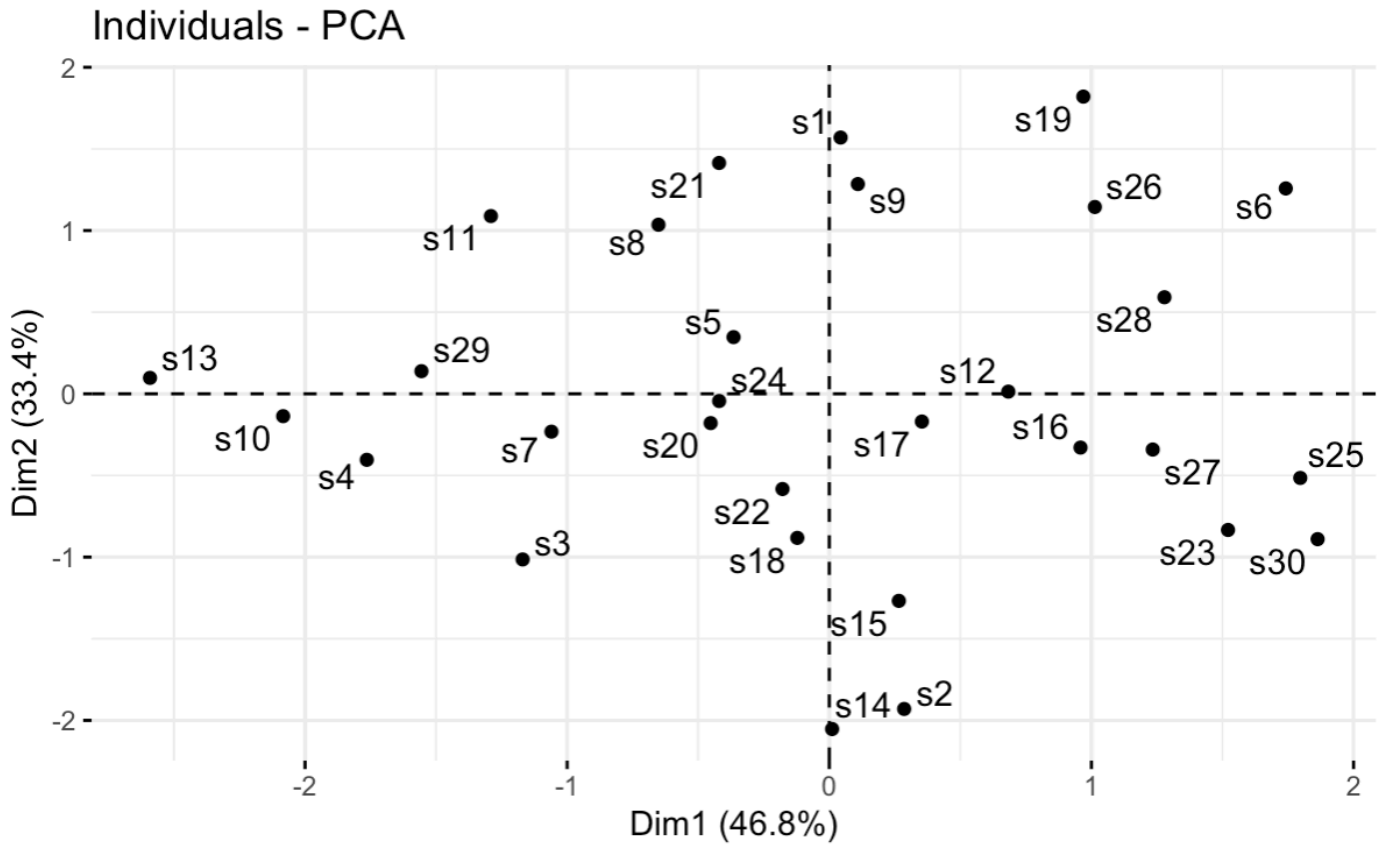


	eigenvalue <dbl>	variance.percent <dbl>	cumulative.variance.percent <dbl>
Dim.1	1.4053569	46.84523	46.84523
Dim.2	1.0015146	33.38382	80.22905
Dim.3	0.5931286	19.77095	100.00000

Gráfica de Sitios

Buscar patrones

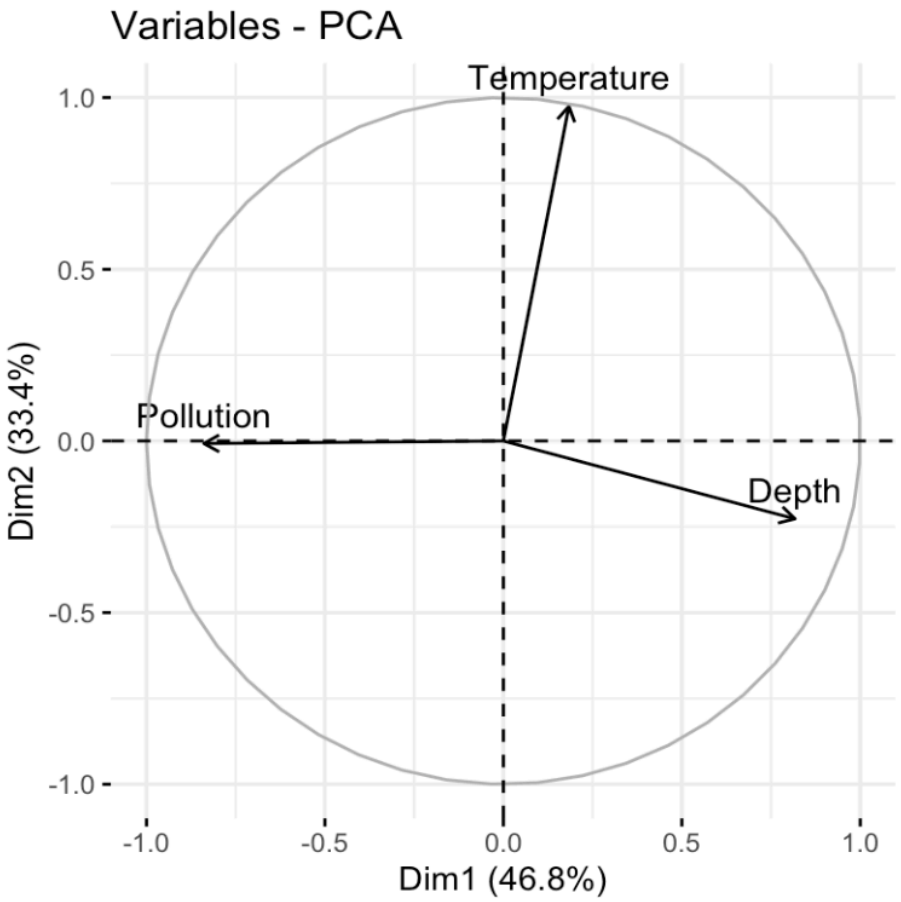
fviz_pca_ind(bioenvpca, repel = TRUE)



Gráfica de variables

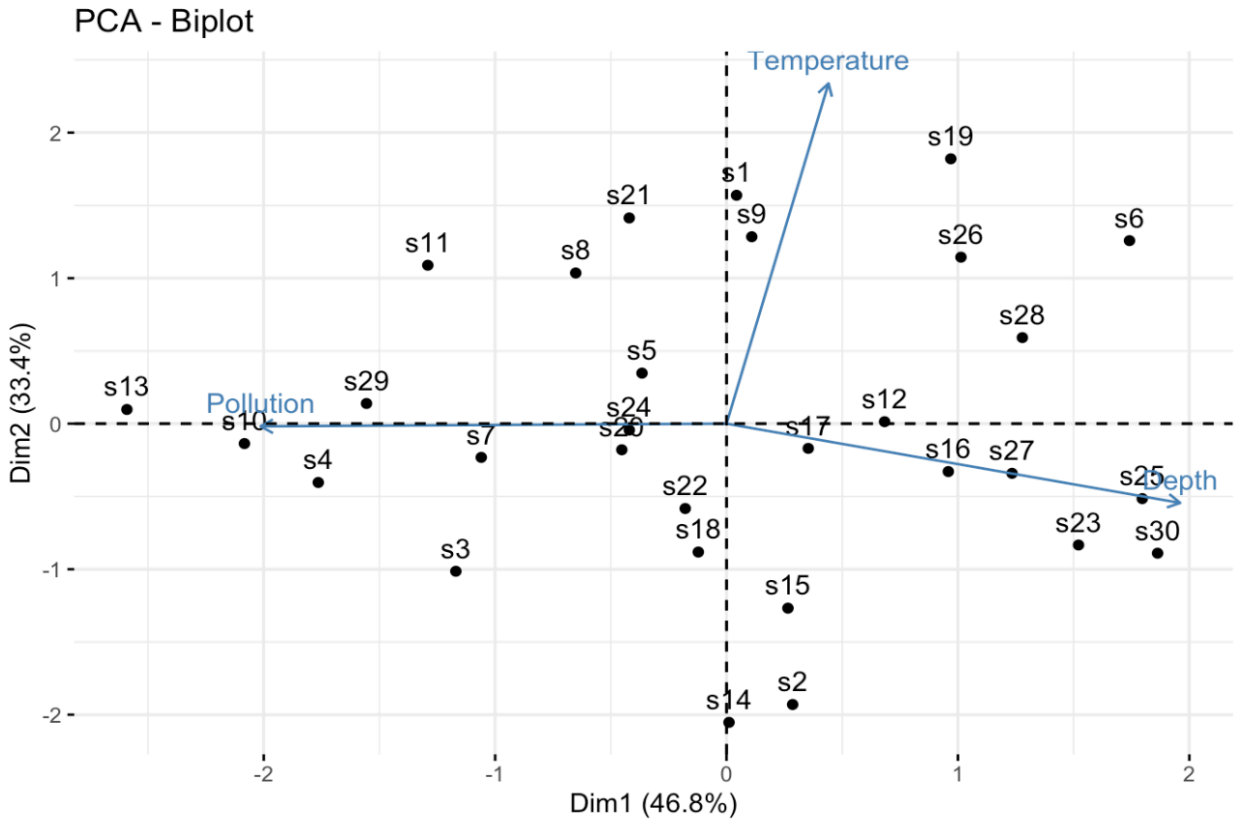
Vector de correlaciones

fviz_pca_var(bioenvpca)



Gráfica Biplot

fviz_pca_ind(bioenvpca, repel = TRUE)



Resumen de la clase

- Revisión de Análisis de componentes principales

- Práctica de análisis de componentes principales con R