ANÁLISIS MULTIVARIANTE PARA COMPARAR EL ESTADO ECOLÓGICO DE LAS COMUNIDADES DE MACROINVERTEBRADOS EN RÍOS DEL PÁRAMO ECUATORIANO Y SU IMPACTO EN ÁREAS DE CONSERVACIÓN*

Maestría en Biodiversidad y Cambio Climático

Karina Hernández

16 de abril de 2024

*MAESTRÍA EN BIODIVERSIDAD Y CAMBIO CLIMÁTICO

En el presente Notebook de RMarkdown, se desarrolla todo el proceso estadístico multivariante que busca Analizar las Componentes Principales (ACP) y los Clúster que determinan la taxonomía de variables que identifican de mejor forma a las comunidades de macroinvertebrados en ríos del páramo ecuatoriano. Primero, se aplicará un ACP para reducir la dimensionalidad de los datos y examinar patrones de variación en la composición de las variables físico-quimicas y ecológicas entre diferentes ríos. Se realizará una evaluación de la adecuación del modelo mediante el estadístico de Kaiser-Meyer-Olkin (KMO) con su nivel significancia respecto al Test de Bartlett, el cual nos proporcionará información sobre la idoneidad de los datos para la aplicación del componentes principales. Además, se empleará un algoritmo de Clúster para agrupar variables en función de su similitud, esto permitirá identificar patrones para definir grupo de variables y xplorar posibles diferencias. Al integrar estos enfoques analíticos, este estudio busca proporcionar una comprensión más profunda del estado ecológico de los ríos del páramo ecuatoriano y de la efectividad de las áreas de conservación en la protección de las comunidades de macroinvertebrados. Los hallazgos de este análisis contribuirán a informar futuras estrategias de conservación y gestión de estos ecosistemas críticos.

CONTENIDOS

Listado de figuras

Listado de Tablas

```r
# Librerías y directorio de trabajo:
library(tidyverse)
library(DT)
library(highcharter)
library(psych)
library(gridExtra)
library(glue)
library(e1071)
library(reshape2)
library(factoextra)
library(FactoMineR)
library(corrplot)
library(knitr)
library(kableExtra)
library(RColorBrewer) # Librería para la paleta de colores

knitr::opts_knit$set(root.dir = "C:/Users/marcelochavez/Documents/TESIS/KARINA/")
```

1 Análisis Exploratorio de los Índices Físquico-Químicos y Ecológicos

Empezaremos con mostrar un Boxplot comparativo para visualizar si existen valores atípicos y como se distribuyen las variables en los años **1995 y 2021** según el formato o estructura de cada tabla con escalas de medición orignales en cada variable:

1.1 Dataset de los Índices Físico Químicos:

```
# Carga del DataSet y transformaciones ----------------------------------------

indices_fq <- readxl::read_excel("DATASETS/DatosFisicoQuimicosEcologicos.xlsx") %>%
  mutate_if(is.numeric, ~ round(., 2)) %>%
  rename_all(toupper)

df_largo <- indices_fq %>%
    pivot_longer(cols = -RIOS,
                 names_to = c(".value", "ANIO"),
                 names_pattern = "(.*)_(\\d{4})")

indices_fq_ddc <- df_largo %>%
    mutate(ID = glue("{RIOS}_{ANIO}")) %>%
    column_to_rownames(var = "ID") %>%
    select(-RIOS,
           -ANIO)

var_scale <- as.data.frame(lapply(indices_fq_ddc, function(x) if(is.numeric(x)) scale

df_fq_1995 <- pivot_longer(df_largo,
```

```
                                cols = c(ANCHO,
                                         PROFUNDIDAD,
                                         VELOCIDAD,
                                         DESCARGA,
                                         PH,
                                         TEMPERATURA,
                                         OXIGENO_DISUELTO,
                                         CONDUCTIVIDAD,
                                         RIQUEZA,
                                         ABUNDANCIA,
                                         DIVERSIDAD,
                                         ABI,
                                         BMWP,
                                         IBF),
                                names_to = "VARIABLE",
                                values_to = "VALOR") %>%
    filter(ANIO=="1995")

df_fq_2021 <- pivot_longer(df_largo,
                           cols = c(ANCHO,
                                    PROFUNDIDAD,
                                    VELOCIDAD,
                                    DESCARGA,
                                    PH,
                                    TEMPERATURA,
                                    OXIGENO_DISUELTO,
                                    CONDUCTIVIDAD,
                                    RIQUEZA,
                                    ABUNDANCIA,
                                    DIVERSIDAD,
                                    ABI,
                                    BMWP,
                                    IBF),
                           names_to = "VARIABLE",
                           values_to = "VALOR") %>%
```

```r
    filter(ANIO=="2021")


# Data para BoxPlot:

df_bx_95 <- data_to_boxplot(
    df_fq_1995,
    variable = VALOR,
    group_var = VARIABLE,
    # group_var2 = codigo,
    add_outliers = T
)

df_bx_21 <- data_to_boxplot(
    df_fq_2021,
    variable = VALOR,
    group_var = VARIABLE,
    # group_var2 = codigo,
    add_outliers = T
)

# Visualización del Dataset

DT::datatable(df_largo,
              class = 'cell-border stripe',
              # filter = 'top',
              caption = htmltools::tags$caption(
              style = 'caption-side: bottom; text-align: left;',
              'Tabla 1: ', htmltools::em('Indicadores Físico-Químicos y Ecológicos pa
              extensions = c('Buttons','Scroller'),
              options = list(scrollX = TRUE,
                             initComplete = JS(
    "function(settings, json) {",
    "$(this.api().table().header()).css({'background-color': '#000', 'color': '#fff'}
    "}"),
```

```
91      dom = 'Bfrtip',
92      buttons = c('excel'),
93      deferRender = TRUE,
94      scrollY = 500,
95      scroller = TRUE))
```

1.2 Boxplot de los indicadores físico-químicos y ecológicos:

```
highchart() %>%
    hc_xAxis(type = "category",
             labels = list(
                 rotation = 270
             )) %>%
    hc_add_series_list(df_bx_95) %>%
    hc_xAxis(title = list(text = "VARIABLES")) %>%
    hc_yAxis(title = list(text = "VALORES")) %>%
    hc_title(text = "BOXPLOT COMPARATIVO ENTRE ÍNDICES FÍSICO QUÍMICOS y ECOLÓGICOS -
    hc_subtitle(text = "AÑO 1995", align="left") %>%
    hc_caption(text = "ELABORADO POR: Karina Hernández") %>%
    hc_legend(enabled = F, title = list(text = "<b>ESTACIONES DE MONITOREO:</b>")) %>%
    hc_tooltip(formatter = JS("function() {
    var tooltip = '';
    if (this.point.low && this.point.q1 && this.point.median && this.point.q3 && this
        tooltip += '<b>Variable: </b>' + this.point.name + '<br/>' +
                   '<b>Mínimo: </b>' + this.point.low.toFixed(2) + '</b><br/>' +
                   '<b>Q1: </b>' + this.point.q1.toFixed(2) + '</b><br/>' +
                   '<b>Mediana: </b>' + this.point.median.toFixed(2) + '</b><br/>' +
                   '<b>Q3: </b>' + this.point.q3.toFixed(2) + '</b><br/>' +
                   '<b>Máximo: </b>' + this.point.high.toFixed(2) + '</b><br/>';
    }
    if (this.point.y && !(this.point.low && this.point.q1 && this.point.median && thi
        tooltip += '<b>Valor Atípico: </b>' + this.point.y.toFixed(2) + '<br/>';
    }
    return tooltip;}")) %>%
    hc_add_theme(hc_theme_gridlight())
```

La variable **Abundancia** presenta una alta dispersión para el año 1995 en relación a todas las variables de estudio

```
highchart() %>%
    hc_xAxis(type = "category",
             labels = list(
                 rotation = 270
             )) %>%
```

```
    hc_add_series_list(df_bx_21) %>%
    hc_xAxis(title = list(text = "VARIABLES")) %>%
    hc_yAxis(title = list(text = "VALORES")) %>%
    hc_title(text = "BOXPLOT COMPARATIVO ENTRE ÍNDICES FÍSICO QUÍMICOS y ECOLÓGICOS -
    hc_subtitle(text = "AÑO 2021", align="left") %>%
    hc_caption(text = "ELABORADO POR: Karina Hernández") %>%
    hc_legend(enabled = F, title = list(text = "<b>ESTACIONES DE MONITOREO:</b>")) %>%
    hc_tooltip(formatter = JS("function() {
    var tooltip = '';
    if (this.point.low && this.point.q1 && this.point.median && this.point.q3 && this
        tooltip += '<b>Variable: </b>' + this.point.name + '<br/>' +
                   '<b>Mínimo: </b>' + this.point.low.toFixed(2) + '</b><br/>' +
                   '<b>Q1: </b>' + this.point.q1.toFixed(2) + '</b><br/>' +
                   '<b>Mediana: </b>' + this.point.median.toFixed(2) + '</b><br/>' +
                   '<b>Q3: </b>' + this.point.q3.toFixed(2) + '</b><br/>' +
                   '<b>Máximo: </b>' + this.point.high.toFixed(2) + '</b><br/>';
    }
    if (this.point.y && !(this.point.low && this.point.q1 && this.point.median && thi
        tooltip += '<b>Valor Atípico: </b>' + this.point.y.toFixed(2) + '<br/>';
    }
    return tooltip;}")) %>%
    hc_add_theme(hc_theme_gridlight())
```

La variable **Diversidad** presenta una alta dispersión para el año 2021 en relación a todas las variables de estudio

1.3 Exploratorio de los datos:

```
exploratorio <- function(df) {
    tipos <- sapply(df, class)
    valores_min <- sapply(df, function(x) ifelse(is.numeric(x), min(x[complete.cases(
    valores_max <- sapply(df, function(x) ifelse(is.numeric(x), max(x[complete.cases(
    coeficientes_asimetria <- sapply(df, function(x) ifelse(is.numeric(x), round(e107
    curtosis <- sapply(df, function(x) ifelse(is.numeric(x), round(kurtosis(x[complet
    promedio <- sapply(df, function(x) ifelse(is.numeric(x), round(mean(x[complete.ca
```

```r
    medianas <- sapply(df, function(x) ifelse(is.numeric(x), round(median(x[complete.
    modas <- sapply(df, function(x) ifelse(is.numeric(x), {
        tab <- table(x[complete.cases(x)])
        as.numeric(names(tab)[tab == max(tab)])}, NA))
    rangos <- sapply(df, function(x) ifelse(is.numeric(x), round(max(x[complete.cases
    varianzas <- sapply(df, function(x) ifelse(is.numeric(x), round(var(x[complete.ca
    desviaciones <- sapply(df, function(x) ifelse(is.numeric(x), round(sd(x[complete.
    coeficientes_variacion <- sapply(df, function(x) ifelse(is.numeric(x), round(sd(x

    resumen <- data.frame(Variable = names(df),
                          Tipo = tipos,
                          Mínimo = valores_min,
                          Máximo = valores_max,
                          Rango = rangos,
                          Promedio = promedio,
                          Mediana = medianas,
                          Desviación_Estándar = desviaciones,
                          Coeficiente_Variación = coeficientes_variacion,
                          Moda = modas,
                          Varianza = varianzas,
                          Coeficiente_Asimetría = coeficientes_asimetria,
                          Curtosis = curtosis,
                          stringsAsFactors = FALSE)

    resumen$Tipo <- ifelse(resumen$Tipo == "factor", "Categórica", resumen$Tipo)
    resumen$Tipo <- ifelse(resumen$Tipo == "POSIXct", "Fecha", resumen$Tipo)
    resumen$Tipo <- ifelse(resumen$Tipo == "logical", "Booleana", resumen$Tipo)
    resumen$Tipo <- ifelse(resumen$Tipo == "numeric", "Numérica", resumen$Tipo)

    # Elimina los nombres de las etiquetas de las filas
    rownames(resumen) <- NULL

    return(resumen)
}

```

```r
DT::datatable(exploratorio(df_largo),
              class = 'cell-border stripe',
              # filter = 'top',
              caption = htmltools::tags$caption(
              style = 'caption-side: bottom; text-align: left;',
              'Tabla 2: ', htmltools::em('Estimadores de Tendencia Central, Dispersió
              extensions = c('Buttons','Scroller'),
              options = list(scrollX = TRUE,
                             initComplete = JS(
    "function(settings, json) {",
    "$(this.api().table().header()).css({'background-color': '#000', 'color': '#fff'}
    "}"),
    dom = 'Bfrtip',
    buttons = c('excel'),
    deferRender = TRUE,
    scrollY = 500,
    scroller = TRUE))
```

De las medidas del Análisis Exploratorio tanto para tendencia central, como de dispersión se puede deducir lo siguiente:

- Las variables que presentan mayor variabilidad son `Abundancia` y `Descarga` con **Coeficientes de Variación** por encima del 100%, esto lo puede corroborar la desviación estándar en comparación con el promedio de cada variable respectivamente.
- La variable `Diversidad` tiene un **Rango** $(ValorMaximo - ValorMnimo)$ considerable, esto nos puede alertar más adelante sobre el comportamiento de dicha variable.
- El **Coeficiente de Asimetría** en la variable `Temperatura` es cercano a **0** lo que podría sugerir que la variable se acerca a una **Distribución Normal Estándar**
- Respecto a la **Curtosis** (Mayor a 3) en las variables `Velocidad` y`Diversidad` son distribuciones con colas más largas, lo que puede sugerir la presencia de datos atípicos.
- En las variables `Descarga` y `Ancho` en **Curtosis** presentan valores cercanos a **0**, lo que puede sugerir variables con una Distribución cercana a la Normal.

- Es importante mencionar que el **Análisis Exploratorio** no es determinante para analizar la presencia de normalidad en los datos. Por consiguiente, es importante a través de un **Test Estadístico** realizar la comprobación correspondiente.

- Finalmente, en un contexto de análisis de datos previo a la aplicación de cualquier técnica estadística, una vez revisado los **estimadores en conjunto con los Boxplot** podemos confirmar que es necesario **realizar un escalamiento a la matriz de datos**, para evitar los sesgos que pueden causar los valores atípicos y su unidad de medida, en el conjunto completo de los indicadores físico-químicos y ecológicos.

2 Análisis de la normalidad de los datos:

Con base a los resultados obtenidos del Análisis Exploratorio, a continuación analizaremos el tipo de distribución de cada variable, es evidente que si consideramos a un año de estudio (*1995 o 2021*) con 11 estaciones de monitoreo, no podemos hablar de datos cercanos a una **Distribución Normal o Gaussiana** por el principio de la **Ley de los Grandes Números**. Por lo cual se ha considerado los datos de los dos años.

En cada variable se aplica el *Test de Normalidad* y su signicancia estadística para determinar si obedece o no a muestra proveniente de una Distribución Normal Estándar $N \approx (\bar{x} = 0, \sigma^2 = constante)$.

El **Test de Shapiro-Wilk (*W*)** (Gutiérrez, 2008), es uno de las pruebas estadísticas más robustas, y se obtiene de la siguiente forma:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2}$$

Donde:

- n es el tamaño de la muestra.
- x_i es el $i - simo$ valor más pequeño de la muestra ordenada.
- $\bar{x}$ es la media de la muestra.
- a_i son los coeficientes que se utilizan para calcular el estadístico de Shapiro-Wilks, y se calculan de la siguiente forma:
  $a_i = \frac{m_i}{\sum_{i=1}^{n} m_i^2}$
  $m_i = \sum_{j=1}^{n}(x_{(j)} - \bar{x})s_{(j)}$
  $s_{(j)} = \sum_{k=1}^{j}(x_{(k)} - \bar{x})^2$

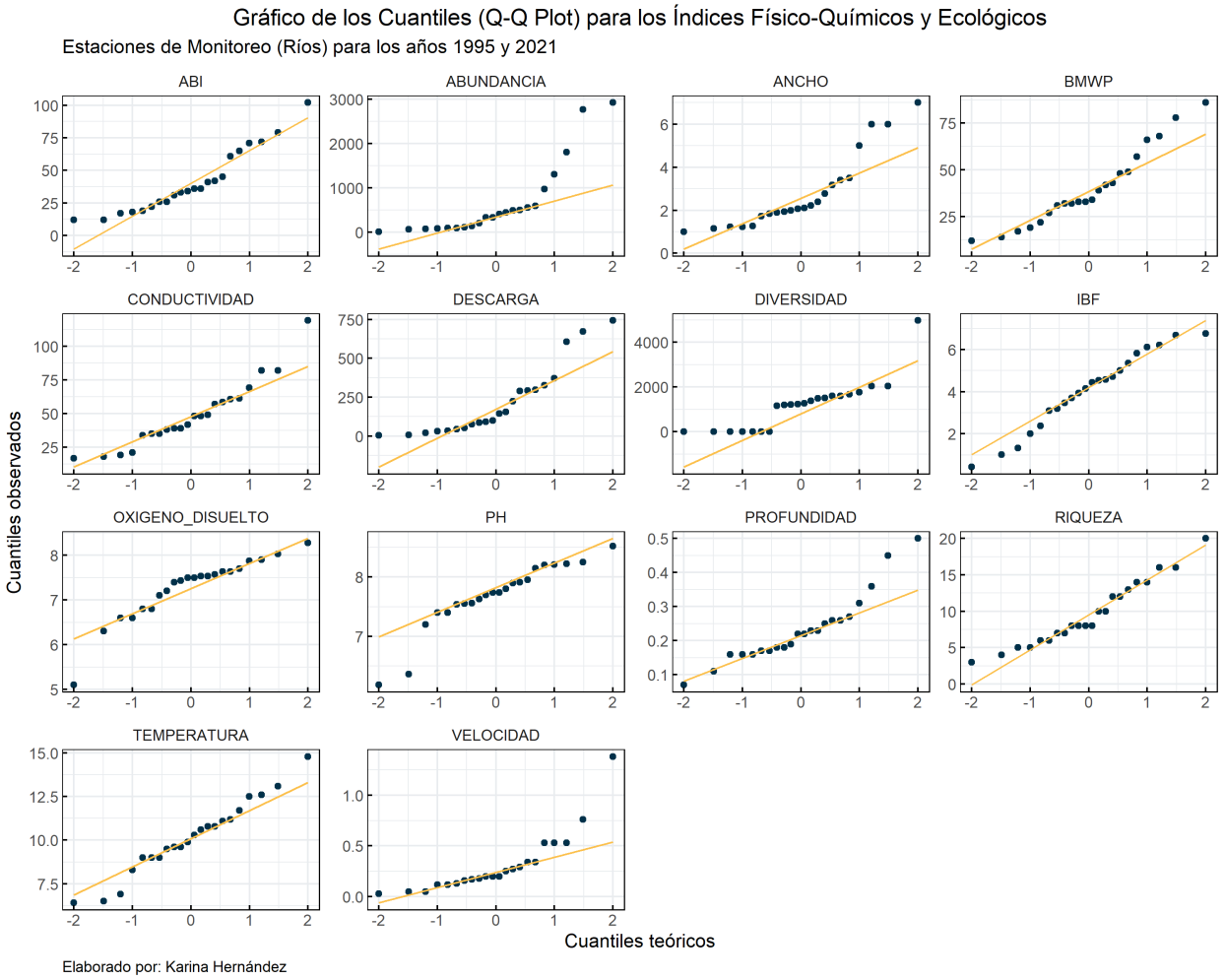
Los coeficientes a_i y los estadísticos auxiliares m_i y x_j se calculan a partir de las observaciones de la muestra y se utilizan en la fórmula del estadístico de Shapiro-Wilk. Estos coeficientes y estadísticos se definen matemáticamente de la siguiente manera:

Donde:

- m_i es la media de los elementos de las siguientes dos sumas:
- La suma en el numerador se calcula sobre todos los n valores de la muestra.
- La suma en el denominador se calcula sobre todos los n valores de la muestra.
- s_j representa la suma de cuadrados de las diferencias entre cada observación y la media, considerando solo las primeras j observaciones en orden ascendente. Esta cantidad se utiliza para calcular m_i, que a su vez se utiliza para calcular a_i, y finalmente en el cálculo del estadístico W.

Nota técnica: El valor del estadístico de Shapiro-Wilk (*W*) se compara con los valores críticos de la distribución para determinar si la muestra sigue una distribución normal. Si el valor de W es significativamente menor que el valor crítico correspondiente, entonces hay evidencia para rechazar la hipótesis nula de normalidad.

```r
numeric_vars <- indices_fq_ddc[sapply(indices_fq_ddc, is.numeric)]

# Convertir el dataframe en formato largo (long format) para usar facet_wrap
df_long <- tidyr::pivot_longer(numeric_vars, cols = everything())

# Crear el Q-Q plot con facet_wrap
ggplot(df_long, aes(sample = value)) +
    geom_qq(color = "#003049") +
    stat_qq_line(color = "#fcbf49") +
    labs(
        title = "Gráfico de los Cuantiles (Q-Q Plot) para los Índices Físico-Químicos
        subtitle = "Estaciones de Monitoreo (Ríos) para los años 1995 y 2021",
        x = "Cuantiles teóricos",
        y = "Cuantiles observados",
        caption = "Elaborado por: Karina Hernández"
    ) +
    facet_wrap(~ name, scales = "free") +
    theme_minimal() +
    theme(
        panel.grid = element_line(color = "#e9ecef", size = 0.5),  # Añadir grid
        panel.border = element_rect(color = "black", fill = NA),  # Añadir recuadro
        axis.ticks = element_line(color = "black"),  # Añadir ticks en los ejes
        axis.line = element_blank(),  # Eliminar líneas de ejes interiores
```




```
        plot.title = element_text(hjust = 0.5),  # Centrar el título
        plot.caption = element_text(hjust = 0),  # Alinear el footnote a la izquierda
        axis.ticks.length = unit(-0.1, "cm")  # Ajustar longitud de los ticks hacia a
    )
```

2.1 Test No paramétrico de Shapiro-Wilk para evaluar la normalidad en los Índices Físico-Químicos y Ecológicos:

```
# Definir una función para aplicar el test de Shapiro-Wilk a todas las variables numé
shapiro_test_all <- function(df) {

  # Filtrar solo las variables numéricas
```

```r
  numeric_vars <- df[sapply(df, is.numeric)]

  # Crear una lista para almacenar los resultados del test
  test_results <- list()

  # Iterar sobre cada variable numérica y aplicar el test de Shapiro-Wilk
  for (var_name in names(numeric_vars)) {

    result <- shapiro.test(numeric_vars[[var_name]])

    # Redondear el estadístico W y el valor p a dos decimales
    result$statistic <- round(result$statistic, 2)

    result$p.value <- round(result$p.value, 2)

    test_results[[var_name]] <- result
  }

  # Retornar los resultados como una lista
  return(test_results)
}

# Aplicar la función al dataframe indices_fq_ddc
shapiro_results <- shapiro_test_all(indices_fq_ddc)

# Convertir los resultados en un tibble
results_df <- tibble::enframe(shapiro_results,
                              name = "Variable",
                              value = "Resultados")

# Extraer los estadísticos W y los valores p en columnas separadas
results_df <- results_df %>%
  mutate(W = sapply(Resultados,
                    "[[",
                    "statistic"),
```

```
40
41          p_value = sapply(Resultados,
42                           "[[",
43                           "p.value"))
44
45  # Seleccionar las columnas relevantes
46  results_df <- results_df %>%
47    select(Variable, W, p_value)
48
49  # Definir los niveles de significancia
50  significance_levels <- c(0.05, 0.01, 0.10)
51
52  # Función para determinar si el valor p es mayor o menor que un nivel de significanci
53  check_significance <- function(p_value,
54                                 significance_level) {
55
56    if (p_value < significance_level) {
57      return("Menor al $\\alpha$")
58    } else {
59      return("Mayor o igual $\\alpha$")
60    }
61  }
62
63  # Aplicar la función a cada valor p y nivel de significancia
64  for (level in significance_levels) {
65
66    results_df <- results_df %>%
67      mutate(
68        !!paste0("Nivel de Significancia al ",
69                 level) := sapply(p_value,
70                                  check_significance,
71                                  significance_level = level),
72        !!paste0("Rechazas H0 al ",
73                 level) := ifelse(p_value < level,
74                                  "Sí",
```

```
                                    "No")
    )
}

# Mostrar la tabla actualizada
knitr::kable(results_df, format = "markdown")
```

Variable	W	p_value	Nivel de Significancia al 0.05	Rechazas H0 al 0.05	Nivel de Significancia al 0.01	Rechazas H0 al 0.01	Nivel de Significancia al 0.1	Rechazas H0 al 0.1
ANCHO	0.83	0.00	Menor al α	Sí	Menor al α	Sí	Menor al α	Sí
PROFUNDIDAD	0.94	0.03	Menor al α	Sí	Mayor o igual α	No	Menor al α	Sí
VELOCIDAD	0.75	0.00	Menor al α	Sí	Menor al α	Sí	Menor al α	Sí
DESCARGA	0.83	0.00	Menor al α	Sí	Menor al α	Sí	Menor al α	Sí
PH	0.88	0.01	Menor al α	Sí	Mayor o igual α	No	Menor al α	Sí
TEMPERATURA	0.98	0.84	Mayor o igual α	No	Mayor o igual α	No	Mayor o igual α	No
OXIGENO_DISUELTO	0.85	0.01	Menor al α	Sí	Mayor o igual α	No	Menor al α	Sí
CONDUCTIVIDAD	0.92	0.07	Mayor o igual α	No	Mayor o igual α	No	Menor al α	Sí
RIQUEZA	0.95	0.33	Mayor o igual α	No	Mayor o igual α	No	Mayor o igual α	No
ABUNDANCIA	0.71	0.00	Menor al α	Sí	Menor al α	Sí	Menor al α	Sí
DIVERSIDAD	0.79	0.00	Menor al α	Sí	Menor al α	Sí	Menor al α	Sí
ABI	0.91	0.04	Menor al α	Sí	Mayor o igual α	No	Menor al α	Sí
BMWP	0.94	0.16	Mayor o igual α	No	Mayor o igual α	No	Mayor o igual α	No
IBF	0.96	0.60	Mayor o igual α	No	Mayor o igual α	No	Mayor o igual α	No

2.2 Comprobación de hipótesis para análisis normalidad:

Para la comprobaciónd de normalidad se ha planteado la siguiente prueba de hipótesis:

H_0 : Las variables provienen de una Distribución Normal

H_1 : Las variables NO provienen de una Distribución Normal

Análisis: De igual forma en un contexto general los indicadores físico-químicos y ecológicos, con un nivel de significancia del 5%. De las 14 variables, 5 presentan normalidad y el resto de variables no lo son. Este resultado nos conduce a generalizar que para todo el conjunto de datos es necesario reducir la dimensionalidad y/o aplicar un **método No Paramétrico** para descubrir clústers o taxonomías entre las unidades de estudio.

3 Análisis Factorial por Componentes Principales:

El **análisis factorial** en un contexto de rigurosidad estadística presente los siguientes pasos:

- Matriz de correlaciones y su significancia estadística
- Determinante de la matriz de correlaciones
- Prueba de esfericidad de Bartlett
- Comprobación de hipótesis para el análisis factorial por componentes principales
- Gráfico de Sedimentación
- Proporción de Varianza Acumulada
- Componentes Principales
- Autovalores y Varianzas
- Círculo de las Correlaciones
- Variables versus Dimensiones
- Biplot (variables versus las estaciones de monitoreo)
- Índice Multivariante

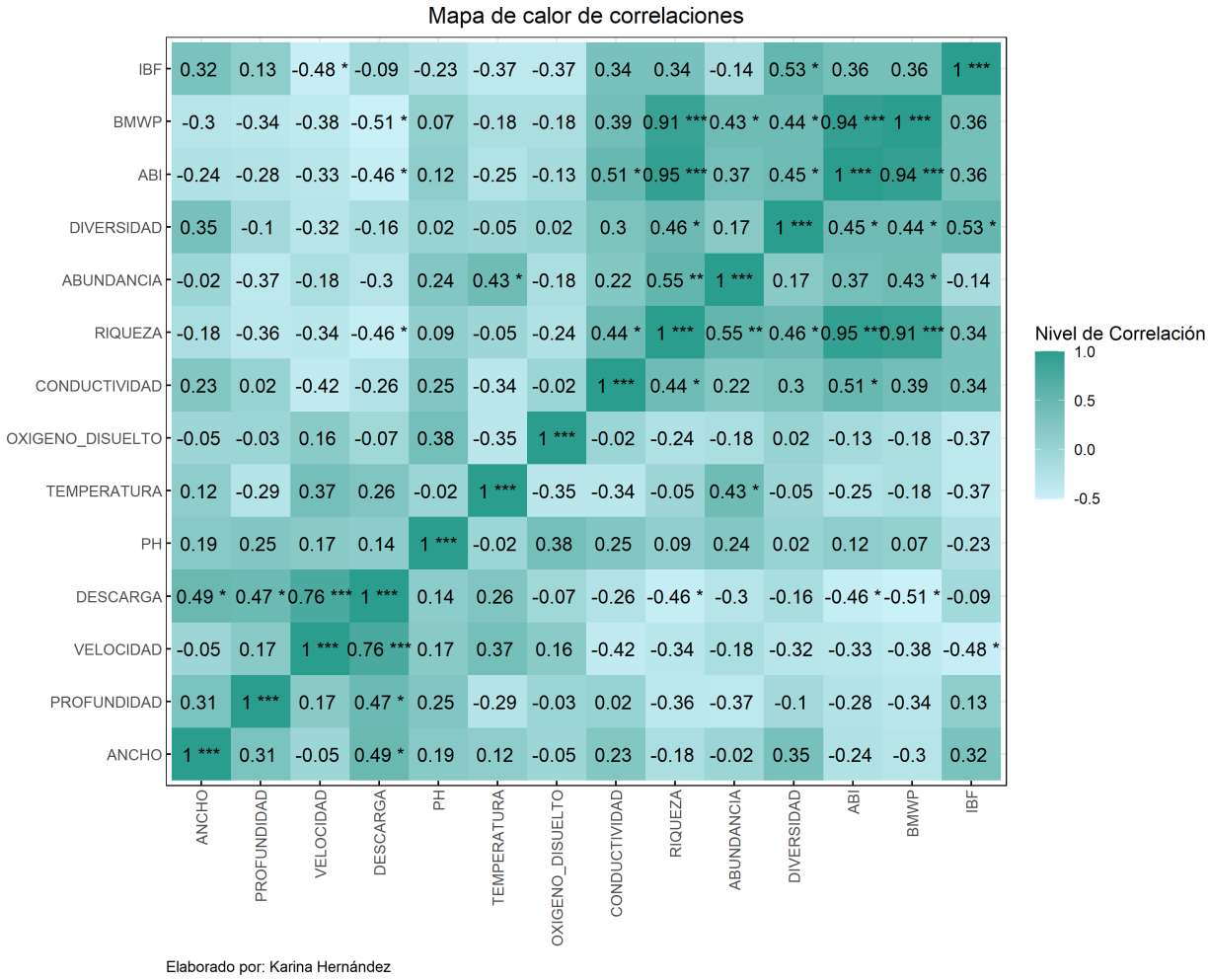
3.1 Matriz de Correlaciones:

```
# Calcular correlaciones y p-values
correlation_matrix <- cor(indices_fq_ddc)
p_values <- cor.mtest(indices_fq_ddc)$p

# Convertir p-values en asteriscos de acuerdo con el nivel de significancia
significant_level <- ifelse(p_values < 0.001, "***",
                            ifelse(p_values < 0.01, "**",
                                   ifelse(p_values < 0.05, "*", "")))
```

```r

# Crear gráfico
ggplot(data = melt(round(correlation_matrix, 2)), aes(x = Var1, y = Var2, fill = valu
    geom_tile() +
    geom_text(aes(label = paste(round(value, 2), significant_level)), color = "black"
    labs(
        title = "Mapa de calor de correlaciones",
        x = "",
        y = "",
        fill = "Nivel de Correlación",
        caption = "Elaborado por: Karina Hernández"
    ) +
    scale_fill_gradient(low = "#caf0f8", high = "#2a9d8f") +
    theme_minimal() +
    theme(
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_text(angle = 90),
        axis.title.y = element_text(angle = 0),
        panel.grid = element_line(color = "#e9ecef", linewidth = 0.5),  # Cambio a li
        panel.border = element_rect(color = "black", fill = NA),
        axis.ticks = element_line(color = "black"),
        axis.line = element_blank(),
        plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0),
        axis.ticks.length = unit(-0.1, "cm")
    )
```

Es importante conocer que el nivel de significancia de la matriz de correlaciones para definir variables que a más de tener un buen nivel de correlación, es significativo estadísticamente.

$p - value < 0.001(***)$: Es significativa la correlación al 99%
$p - value < 0.05(**)$: Es significativa la correlación al 95%
$p - value < 0.01(*)$: Es significativa la correlación al 90%



3.2 Determinante de la matriz de correlaciones

```
det(cor(indices_fq_ddc))
```

```
[1] 2.706342e-07
```

Si están correlacionadas, el determinantes de la matriz de correlaciones será próximo a cero. De lo contrario el **Análisis Factorial** no es viable.

3.3 Prueba de esfericidad de Bartlett:

```
KMO(indices_fq_ddc)
```

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = indices_fq_ddc)
```

```
Overall MSA =  0.44
MSA for each item =
           ANCHO      PROFUNDIDAD        VELOCIDAD         DESCARGA
            0.26             0.65             0.35             0.41
              PH      TEMPERATURA OXIGENO_DISUELTO    CONDUCTIVIDAD
            0.36             0.26             0.17             0.61
         RIQUEZA       ABUNDANCIA       DIVERSIDAD              ABI
            0.60             0.42             0.61             0.48
            BMWP              IBF
            0.58             0.55
```

```
1 cortest.bartlett(cor(indices_fq_ddc))
```

```
$chisq
[1] 1413.954

$p.value
[1] 3.472263e-236

$df
[1] 91
```

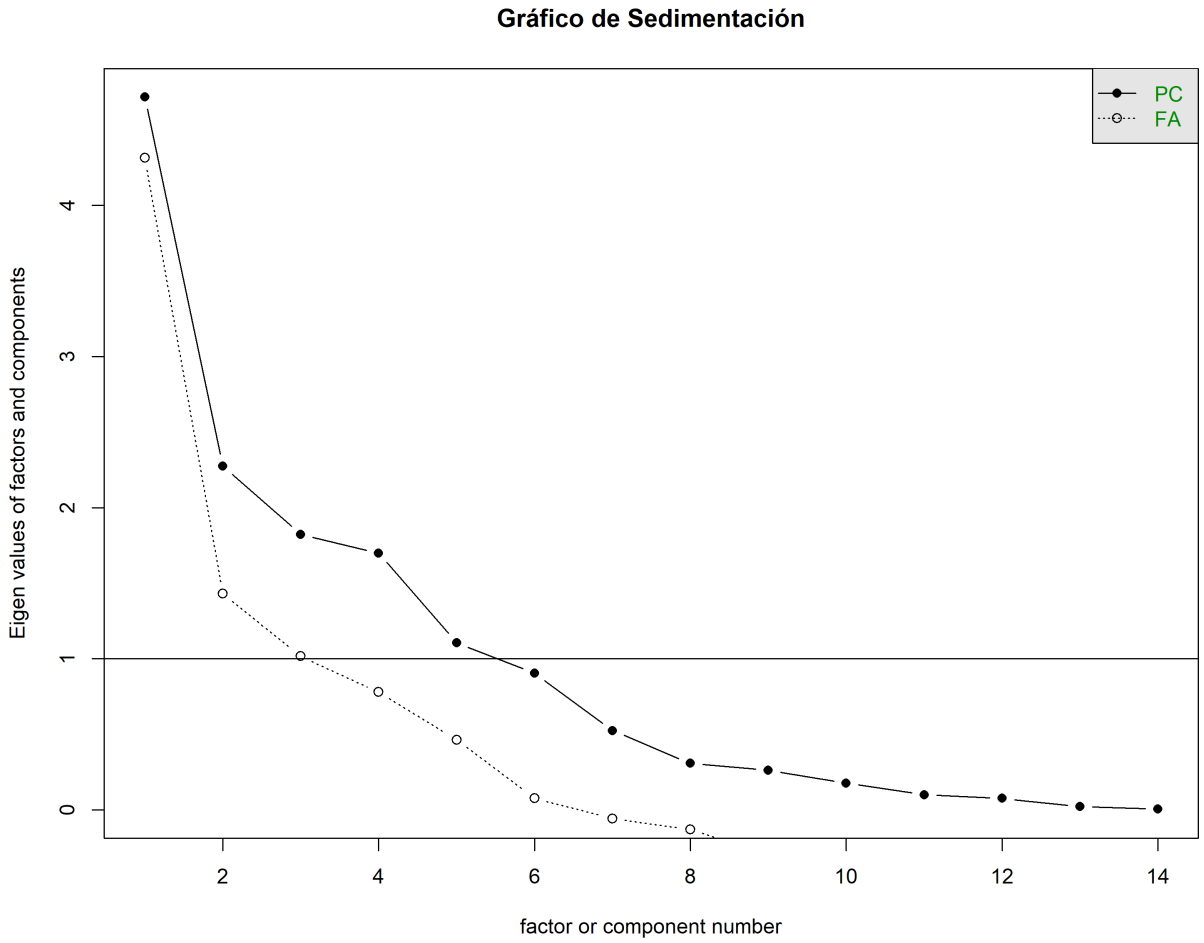
Un KMO de 0.44 es considerado bajo y podría sugerir que un análisis factorial puede no ser adecuado para estos datos. Sin embargo, la prueba de Bartlett tiene un p-valor menor al nivel de significancia del 5%, lo que indica que las variables están correlacionadas y que, por lo tanto, podría justificarse aplicar un método multivariante para reducción de dimensionalidad.

La hipótesis nula es que la matriz de correlaciones es una matriz identidad, lo que significa que no existirían correlaciones significativas, por lo que el modelo factorial no sería pertinente.

3.4 Comprobación de hipótesis para el análisis factorial por componentes principales

H_0 : Existe esfericidad

H_1 : No existe esfericidad



La esfericidad implica que todas las variables tienen la misma varianza y que no hay correlaciones entre ellas.

De aquí que el valor del *KMO* de 0.44 y su p-valor es menor al 95% de confianza. Por lo tanto: Rechazamos H_0, y concluimos que es factible estadísticamente realizar la reducción de dimensiones a través de un **Análisis Factorial por Componentes Principales**

3.5 Gráfico de Sedimentación:

```
scree(indices_fq_ddc,
      main ="Gráfico de Sedimentación")
```

3.6 Proporción de varianza acumulada:

```r
# Cálculo de los autovalores y autovectores:
value_pro<-eigen(cor(indices_fq_ddc))

Acumulado <- cumsum(value_pro$values)

Prop.acumulado <- Acumulado/sum(value_pro$values)

val.prop.porce <- data.frame(value_pro$values,
                             Acumulado,
                             Prop.acumulado)

row.names(val.prop.porce) = c(expression(lambda[1]),
                              expression(lambda[2]),
                              expression(lambda[3]),
                              expression(lambda[4]),
                              expression(lambda[5]),
                              expression(lambda[6]),
                              expression(lambda[7]),
                              expression(lambda[8]),
                              expression(lambda[9]),
                              expression(lambda[10]),
                              expression(lambda[11]),
                              expression(lambda[12]),
                              expression(lambda[13]),
                              expression(lambda[14]))

colnames(val.prop.porce) <- c("Valor Propio",
                              "Acumulado",
                              "Prop. Acumulado")
kable(val.prop.porce,
      caption = "Valores propios desde la matriz de correlaciones",
      digits = 2,
      format.args = list(decimal.mark =",")) %>%
  row_spec(row = 5,
```

```
            background="#a7c957",
            bold=T)
```

Tabla 3.1: Valores propios desde la matriz de correlaciones

	Valor Propio	Acumulado	Prop. Acumulado
lambda[1]	4,72	4,72	0,34
lambda[2]	2,27	6,99	0,50
lambda[3]	1,82	8,82	0,63
lambda[4]	1,70	10,51	0,75
lambda[5]	**1,11**	**11,62**	**0,83**
lambda[6]	0,90	12,52	0,89
lambda[7]	0,52	13,05	0,93
lambda[8]	0,31	13,36	0,95
lambda[9]	0,26	13,62	0,97
lambda[10]	0,18	13,80	0,99
lambda[11]	0,10	13,90	0,99
lambda[12]	0,08	13,97	1,00
lambda[13]	0,02	13,99	1,00
lambda[14]	0,01	14,00	1,00

El **Valor Propio** de la matriz de correlaciones cuando desciende hasta 1 sugiere que el número de componentes o factores que pueden explicar la variabilidad de todo el conjunto de datos está en un $\lambda = 5$ y lo corrobora el **Gráfico de Sedimentación**

3.7 Componentes Principales:

```
pca <- prcomp(indices_fq_ddc,
              scale=TRUE,
              center = TRUE,
              method = "varimax")
summary(pca)
```

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
```

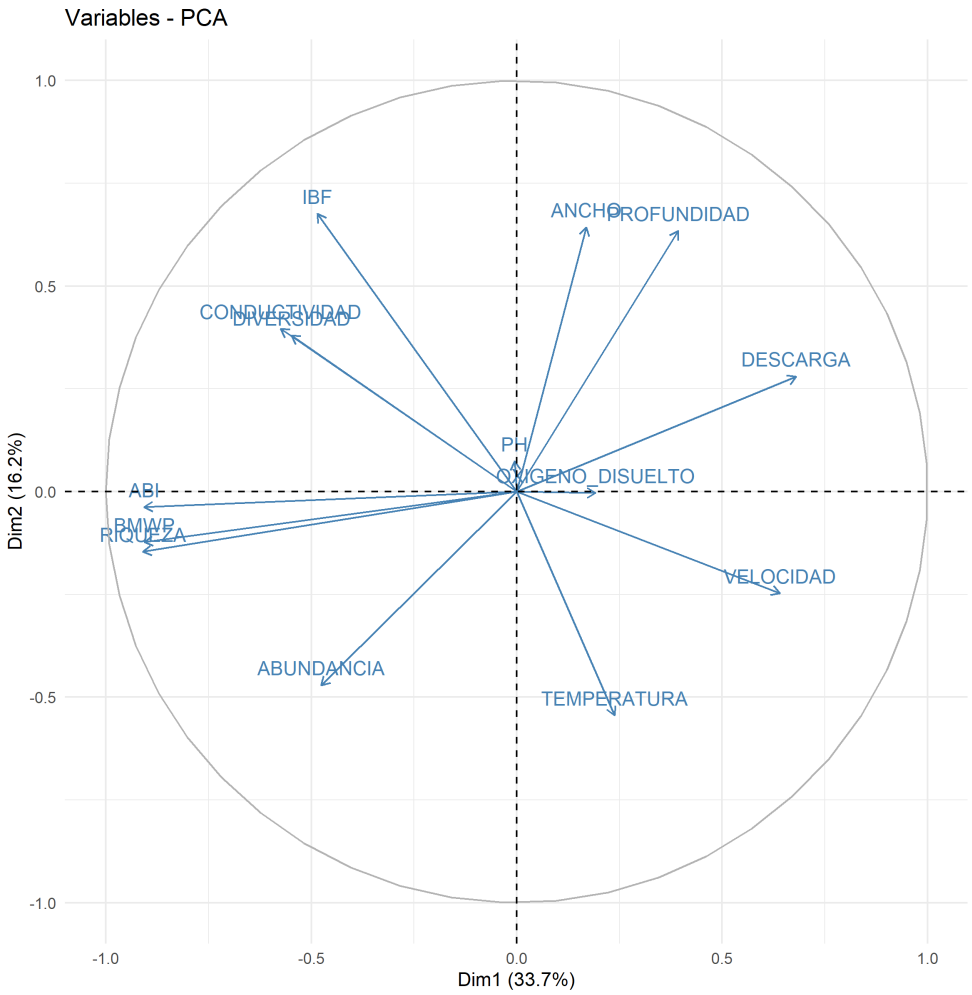
```
Standard deviation     2.172 1.5080 1.3503 1.3036 1.05141 0.95050 0.72437
Proportion of Variance 0.337 0.1624 0.1302 0.1214 0.07896 0.06453 0.03748
Cumulative Proportion  0.337 0.4994 0.6297 0.7511 0.83002 0.89456 0.93204
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.55491 0.51189 0.4216 0.3153 0.27573 0.14732 0.08124
Proportion of Variance 0.02199 0.01872 0.0127 0.0071 0.00543 0.00155 0.00047
Cumulative Proportion  0.95403 0.97275 0.9855 0.9926 0.99798 0.99953 1.00000
```

3.8 Autovalores y varianzas:

```
1 get_eig(pca)
```

```
        eigenvalue variance.percent cumulative.variance.percent
Dim.1  4.717836203      33.69883002                    33.69883
Dim.2  2.274213490      16.24438207                    49.94321
Dim.3  1.823386036      13.02418597                    62.96740
Dim.4  1.699435601      12.13882572                    75.10622
Dim.5  1.105467492       7.89619637                    83.00242
Dim.6  0.903448292       6.45320209                    89.45562
Dim.7  0.524706456       3.74790326                    93.20353
Dim.8  0.307925627       2.19946876                    95.40299
Dim.9  0.262032510       1.87166079                    97.27466
Dim.10 0.177786579       1.26990413                    98.54456
Dim.11 0.099431722       0.71022659                    99.25479
Dim.12 0.076026874       0.54304910                    99.79783
Dim.13 0.021703083       0.15502202                    99.95286
Dim.14 0.006600033       0.04714309                   100.00000
```

Al analizar los componentes principales podemos observar que en el **Componente 5** se explica el **83%** de la varianza de todas las dimensiones



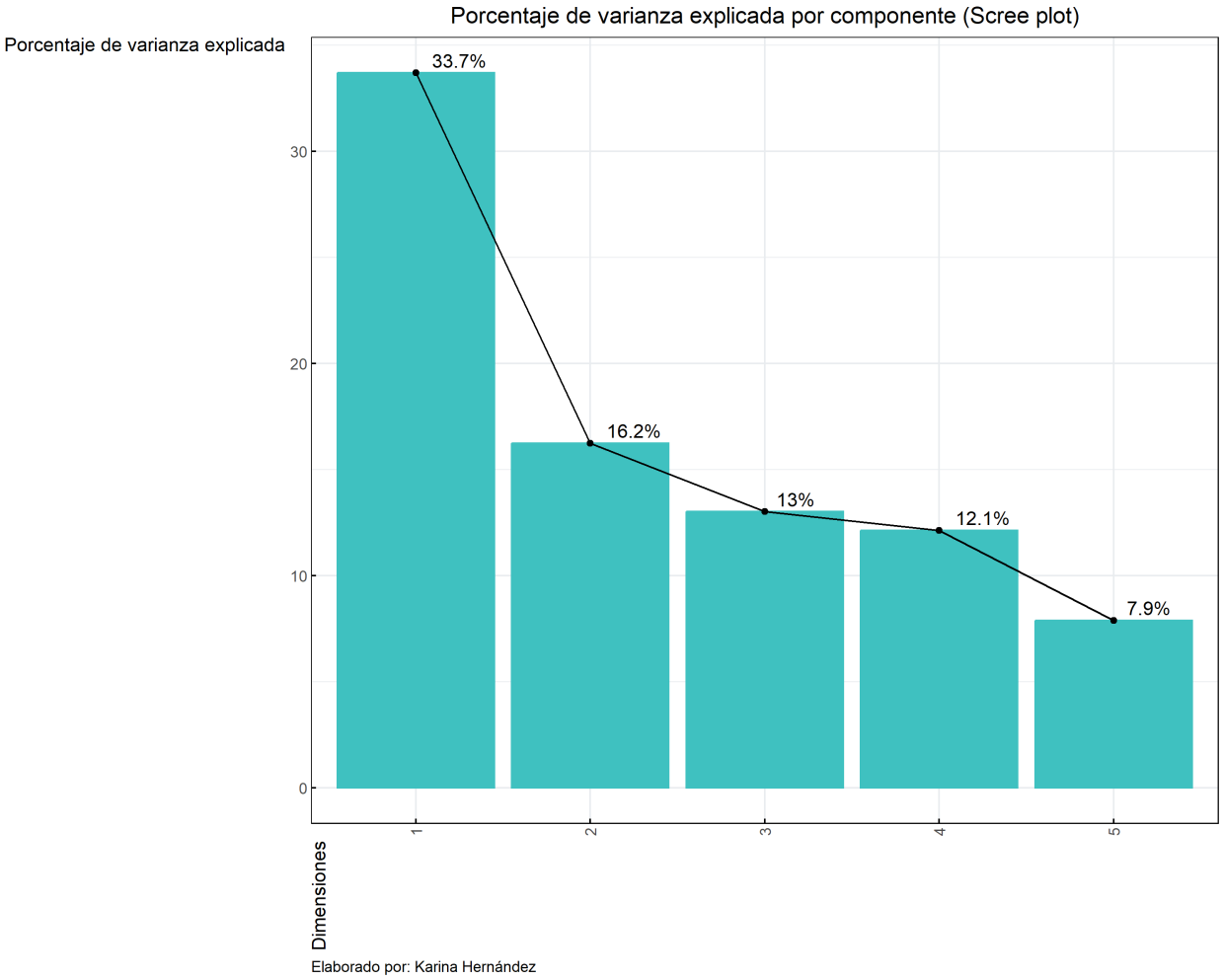
3.9 Círculo de correlaciones:

```
fviz_pca_var(pca,
             col.var="steelblue") +
  theme_minimal()
```

En el círculo de las correlaciones es una representación gráfica de las correlaciones entre las variables originales y los componentes principales extraídos durante el análisis factorial. Cada variable original se representa como un vector en el espacio de los componentes principales, donde la longitud del vector indica la correlación entre esa variable y el componente principal correspondiente, y la dirección del vector indica la dirección en la que esa variable contribuye al componente principal. Por ejemplo, al examinar el círculo de correlaciones, podemos observar que las variables **ABI, BMWP y RIQUEZA** están inversamente correlacionadas con la

Dimensión 2. Por otro lado, las variables **ANCHO y PROFUNDIDAD** muestran una correlación más fuerte con la Dimensión 1. En resumen, el círculo de correlaciones proporciona una representación visual que ayuda a entender cómo las variables originales están relacionadas y contribuyen a la estructura de los componentes principales identificados durante el análisis factorial

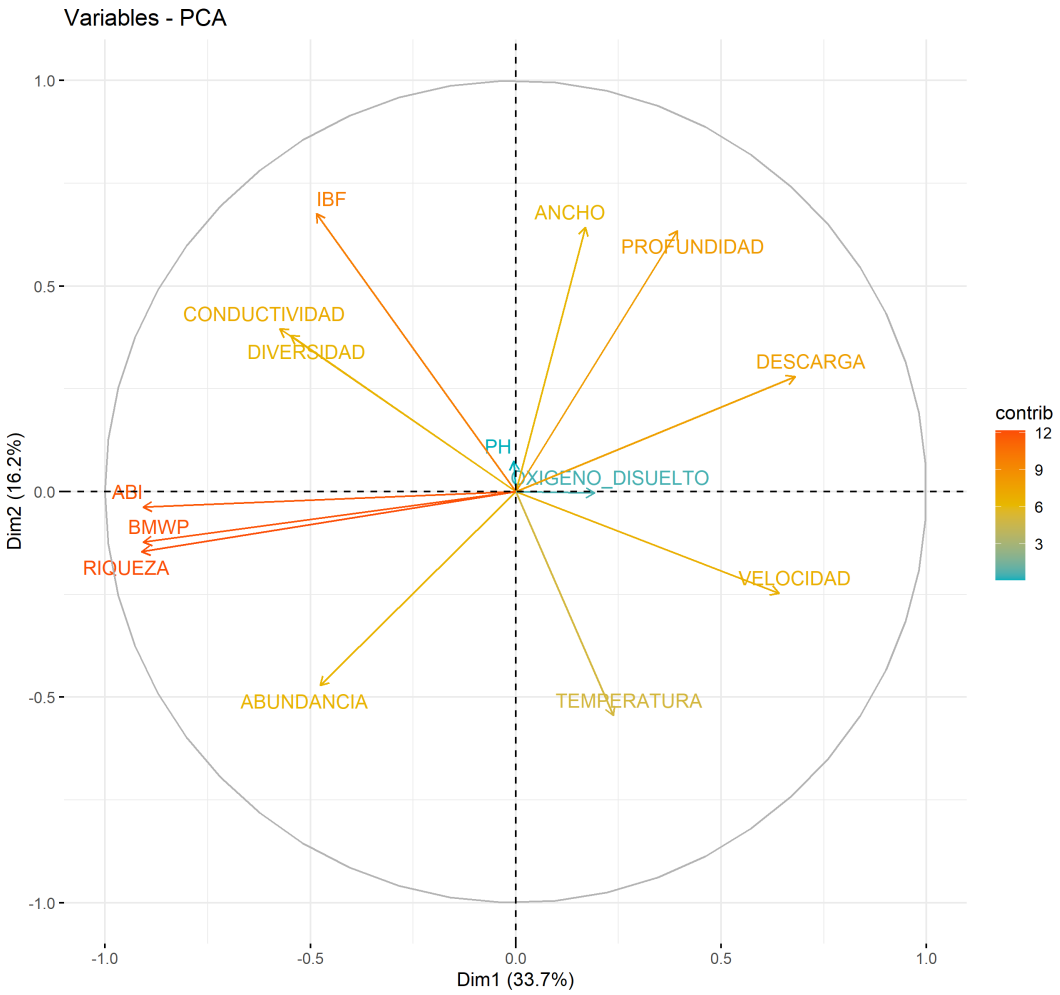
```
fviz_screeplot(pca,
               ncp = 5,
               barfill = "#3fc1c0",
               barcolor = "#3fc1c0",
               addlabels = TRUE,
               ylab = "Porcentaje de varianza explicada",
               xlab = "Dimensiones",
               main = "Porcentaje de varianza explicada por componente (Scree plot)",
               hjust = -0.3) +
  theme(plot.title = element_text(size = 14, face = "bold"),
        plot.caption = element_text(size = 8, face = "italic")) +
  labs(caption = "Elaborado por: Karina Hernández") +
  theme_minimal() +
    theme(
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_text(angle = 90),
        axis.title.y = element_text(angle = 0),
        panel.grid = element_line(color = "#e9ecef", linewidth = 0.5),  # Cambio a li
        panel.border = element_rect(color = "black", fill = NA),
        axis.ticks = element_line(color = "black"),
        axis.line = element_blank(),
        plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0),
        axis.ticks.length = unit(-0.1, "cm")
    )
```



3.10 Variables versus Dimensiones:

```
fviz_pca_var(pca,
             col.var="contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

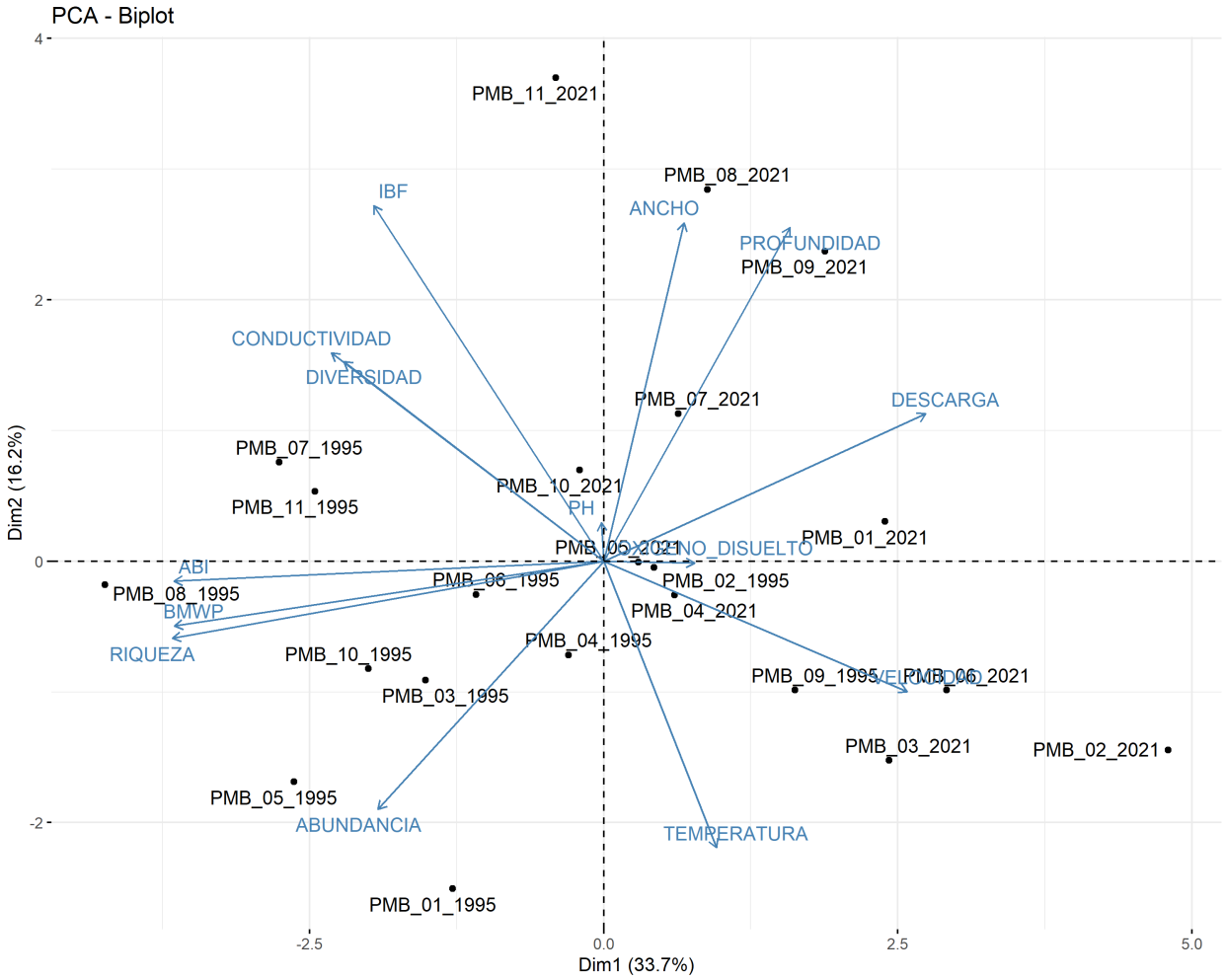
Es fundamental examinar el nivel de contribución de las variables a cada uno de los componentes principales. En este sentido, observamos que variables como **ABI, BMWP y RIQUEZA** destacan por su significativa contribución a la **Dimensión 2**. Sin embargo, al considerar **OXÍGENO y PH**, a pesar de ubicarse en los mismos componentes, la longitud de sus vectores indica una contribución mínima al modelo factorial en construcción.



3.11 Biplot (variables versus las estaciones de monitoreo)

```
# Biplot of individuals and variables
fviz_pca_biplot(pca, repel = TRUE)
```

Variables ecológicas como **ABI, BMWP y RIQUEZA** exhiben una alta similitud en la estación de monitoreo 8 durante el año 1995. De manera similar: **CONDUCTIVIDAD, DIVERSIDAD e IBF** comparten características destacadas en las estaciones de monitoreo 7 y 11, también en el año 1995. Por otro lado, un conjunto distinto de variables como: **ANCHO, PROFUNDIDAD y DESCARGA** para el año 2021, evidencian patrones similares en las estaciones de monitoreo 1, 7, 8 y 9. Sin embargo, variables como: **ABUNDANCIA, TEMPERATURA y VELOCIDAD**, a pesar de su comportamiento de no similaridad entre ellas, aportan de forma significativa al modelo factorial individualmente


PCA - Biplot
PMB_11_2021
PMB_08_2021
IBF
ANCHO
PROFUNDIDAD
PMB_09_2021
CONDUCTIVIDAD
DIVERSIDAD
PMB_07_2021
DESCARGA
PMB_07_1995
PMB_10_2021
PMB_11_1995
PH
PMB_01_2021
PMB_05_2021
OXIGENO_DISUELTO
ABI
PMB_08_1995
PMB_06_1995
PMB_02_1995
BMWP
PMB_04_2021
RIQUEZA
PMB_04_1995
PMB_10_1995
PMB_09_1995
VELOCIDAD
PMB_06_2021
PMB_03_1995
PMB_03_2021
PMB_02_2021
PMB_05_1995
ABUNDANCIA
TEMPERATURA
PMB_01_1995
Dim2 (16.2%)
Dim1 (33.7%)
4
2
0
-2
-2.5
0.0
2.5
5.0