Capítulo 4. Visualización gráfica

Javier Rodríguez Barrios

Introducción

Este capítulo es dedicado al análisis de graficas exploratorias, como un procedimiento casi obligatorio para cualquier tipo de investigación o de informe técnico que requiera del análisis estadístico para datos ambientales o ecológicos. Es importante resaltar que, de un buen análisis exploratorio, el investigador podrá obtener una idea más clara y estructurada de sus datos, para proceder al desarrollo de su diseño estadístico. Resulta imposible detallar en este documento la increíble variedad gráfica que ofrece el entorno de R, pues cada librería presenta un enorme número de opciones. Se detallará un pequeño grupo de posibilidades gráficas, que pueden servir a la hora de conocer el patrón de los datos.

Se incluye a algunas librerías disponibles, para el análisis exploratorio de datos univariados y multivariados, adicionalmente se brindan algunas opciones para la edición para las figuras y con ello, ofrecer la libertad al usuario de manipular sus figuras. Se detallan los procedimientos de rutina, como el diseño de gráficas para identificar valores atípicos, la detección de la media y el error estándar, entre otros estadísticos de tendencia central y de dispersión. Se analizarán diferentes paquetes gráficos entre las que se incluyen: *grid*, *lattice*, *ellipse*, *ggplot2*, entre otros.

De acuerdo con `(Paradais 2002)`, existen dos tipos de funciones gráficas: (1) funciones de alto nivel, que crean nuevas gráficas y (2) funciones de bajo nivel, que agregan elementos a una gráfica ya existente. Las últimas trabajan con parámetros gráficos que están definidos por defecto.

La exploración gráfica de variables, factores y de observaciones, permite dar respuesta a diferentes tipos de preguntas, como las siguientes:

1. ¿Se encuentran los datos centrados? ¿Cómo es su distribución? ¿Los datos son simétricos, asimétricos, o presentan alguna tendencia en particular?
2. ¿Hay datos o valores atípicos que puedan ser identificados?
3. ¿Las variables presentan una distribución normal o multinormal?
4. ¿Hay alguna relación entre las variables? ¿Las relaciones entre las variables son lineales? ¿Qué análisis exploratorio se debe aplicar para valorar esta relación?
5. ¿Las variables varían entre los diferentes niveles (grupos o muestras) de un factor?

4.1 Utilidad de R en la exploración de datos.

Una de las principales ventajas del trabajo en el programa R, es su versatilidad en el desarrollo de figuras básicas y avanzadas, para la exploración de una o más variables. En ese sentido, a continuación, se resumen algunas fuentes o herramientas virtuales, que permitirán realizar diferentes procedimientos numéricos y gráficos, con una o más variables en análisis.

Tutorial rápido de R. Esta es una guía que permite obtener información sobre las diferentes utilidades del programa R, tanto en su instalación, en sus componentes y el procesamiento numérico y gráfico de los datos `(Figura 5)`. https://www.statmethods.net/

Figure 1: Imagen de motor de búsqueda de utilidades de R.

4.2 Figuras exploratorias

El objetivo de esta temática consiste en explorar diferentes funciones gráficas que ofrece el R, para una exploración resumida de datos en los cuales se cuente con factores, variables cualitativas y cuantitativas. Se iniciará con figuras similares a las que pueden realizarse en Excel, como totas y columnas, que evalúan diferencias entre categorías o niveles de factores (ej. tramos, muestreos, etc). Posteriormente se realizará el análisis de otra base de datos, que incorpora variables ambientales y biológicas, en la cual se realizarán algunas figuras propias de R. Es importante aclarar, que el entorno gráfico del programa R, es una de sus principales fortalezas y por ende, se cuenta con numerosos textos, orientados exclusivamente a gráficas básicas y avanzadas, pero ese nivel de detalle no será el objeto del presente documento.

Ejemplo 1. GRÁFICAS EXPLORATORIAS

A continuación, se realizará un análisis exploratorio de una base de datos **"Datos1.csv"** la cual presenta observaciones (sitios) factores (cuencas), variables ambientales y biológicas. El script de R a revisar es "Insectos.explor.r". El objetivo de este ejercicio consistirá en visualizar y analizar diferentes opciones gráficas, para poder descifrar los patrones que subyacen de los datos.

```
#-------------------
# Librerías requeridas
library(lattice)
library(ellipse)
require(SciViews)
library(plotrix)
require(stats)
library(corrplot)
library(tidyverse)
library(ggplot2)
library(reshape2)
library(gridExtra)
library(gtable)
library(grid)
library(ggforce)

#-------------------
# Base de datos de insectos acuáticos
datos<-read.csv2("Datos1.csv",row.names=1)

# Organización de los datos
str(datos)      # Estructura de la base de datos
```

```
## 'data.frame':    20 obs. of  9 variables:
##  $ cuenca: chr  "cuen1" "cuen1" "cuen1" "cuen1" ...
##  $ pH    : num  6.8 7.3 5.6 6.3 5.6 6.3 7.5 7 7 5.7 ...
##  $ temp  : num  17.4 16.8 16 17.8 18.2 17 16.8 18.2 19.8 15.3 ...
##  $ Efem  : int  26 17 9 2 6 7 19 12 13 5 ...
##  $ Plec  : int  4 6 3 3 4 2 3 5 6 0 ...
##  $ Tric  : int  9 9 28 25 24 25 12 23 9 32 ...
##  $ Dipt  : int  30 25 24 21 12 10 12 9 0 11 ...
##  $ Cole  : int  3 1 3 6 13 1 3 4 15 8 ...
##  $ Ab    : int  72 58 67 57 59 45 49 53 43 56 ...
```

```
datos$cuenca=as.factor(datos$cuenca)  # Convertir cuenca a factor
str(datos)        # Nueva estructura de los datos
```

```
## 'data.frame':    20 obs. of  9 variables:
##  $ cuenca: Factor w/ 4 levels "cuen1","cuen2",..: 1 1 1 1 1 2 2 2 2 2 ...
##  $ pH    : num  6.8 7.3 5.6 6.3 5.6 6.3 7.5 7 7 5.7 ...
##  $ temp  : num  17.4 16.8 16 17.8 18.2 17 16.8 18.2 19.8 15.3 ...
##  $ Efem  : int  26 17 9 2 6 7 19 12 13 5 ...
##  $ Plec  : int  4 6 3 3 4 2 3 5 6 0 ...
##  $ Tric  : int  9 9 28 25 24 25 12 23 9 32 ...
##  $ Dipt  : int  30 25 24 21 12 10 12 9 0 11 ...
##  $ Cole  : int  3 1 3 6 13 1 3 4 15 8 ...
##  $ Ab    : int  72 58 67 57 59 45 49 53 43 56 ...
```

```
summary(datos[,3:9])  # Resumen estadístico
```

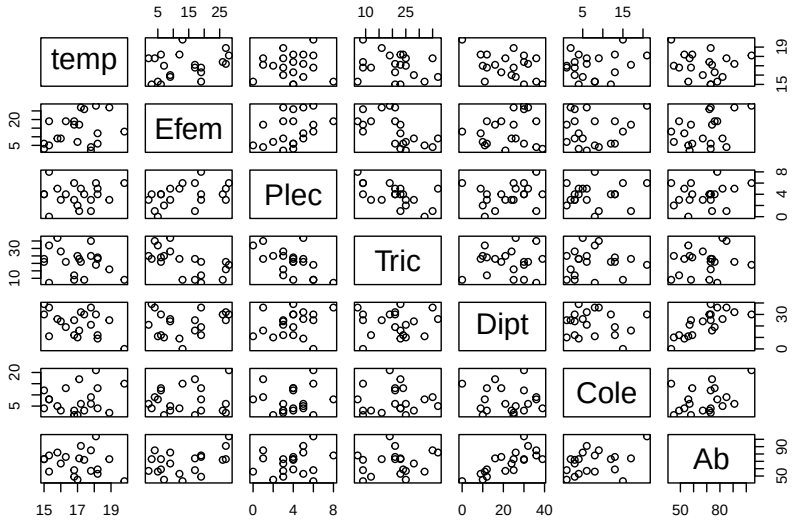
```
##       temp            Efem            Plec           Tric            Dipt
##  Min.   :15.00   Min.   : 2.00   Min.   :0.00   Min.   : 7.00   Min.   : 0.00
##  1st Qu.:15.95   1st Qu.: 6.00   1st Qu.:3.00   1st Qu.:15.00   1st Qu.:12.00
##  Median :17.05   Median :12.50   Median :4.00   Median :22.00   Median :24.50
##  Mean   :16.99   Mean   :13.75   Mean   :3.85   Mean   :20.95   Mean   :22.15
##  3rd Qu.:17.88   3rd Qu.:19.00   3rd Qu.:5.00   3rd Qu.:25.00   3rd Qu.:30.00
##  Max.   :19.80   Max.   :28.00   Max.   :8.00   Max.   :37.00   Max.   :39.00
##       Cole             Ab
##  Min.   : 1.00   Min.   : 43.00
##  1st Qu.: 3.00   1st Qu.: 56.75
##  Median : 6.00   Median : 72.50
##  Mean   : 7.70   Mean   : 68.40
##  3rd Qu.:12.25   3rd Qu.: 76.50
##  Max.   :21.00   Max.   :104.00
```

4.1 Graficas de pares (pairplot)

Permiten visualizar el nivel de relación de más de dos variables a través de un panel con una serie de diagramas de dispersión (uno para cada par de variables). Es apropiado para un máximo de 10 variables. Si el número de variables es mayor, se recomienda utilizar la función "ellipse". Las figuras de pares suelen utilizarse como exploraciones de relaciones lineales (con y sin transformaciones) que son requeridas en técnicas multivariadas como los componentes principales (PCA) y redundancias (RDA), y el resto que trabajen con la distancia euclídea (distancia métrica para relaciones lineales).

Estas figuras de pares pueden venir acompañadas de coeficientes de correlación, los cuales se muestran en la parte inferior del panel gráfico. Los valores de colinealidad pueden presentarse cuando el coeficiente de correlación sea cercano a uno (alta relación) en variables (independientes) que sean relacionadas (Zuur et al. 2007). De acuerdo es estos autores, las figuras de pares son útiles bajo tres tipos de relaciones entre parejas de variables: con variables respuesta, con variables explicativas y con variables respuesta versus variables explicativas.

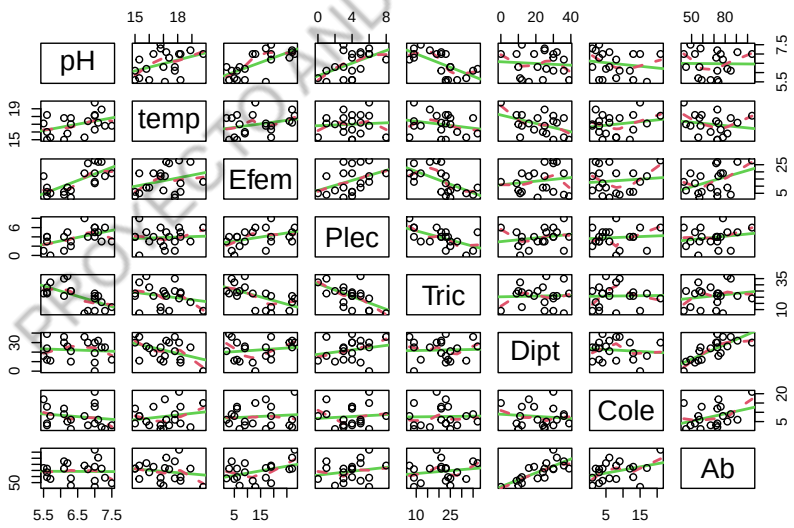
```
#------------------
# 1. Gráfica por pares
par(mar = c(4, 4, .2, .1))
pairs(datos[,3:9])          # [,3:9] relaciona a las columnas 2 a la 8.
```



Graficas de pares, con dispersión de los datos originales. Probar con `pairs(log10(datos[,3:9]))` en donde `log10` es la transformación logarítmica.

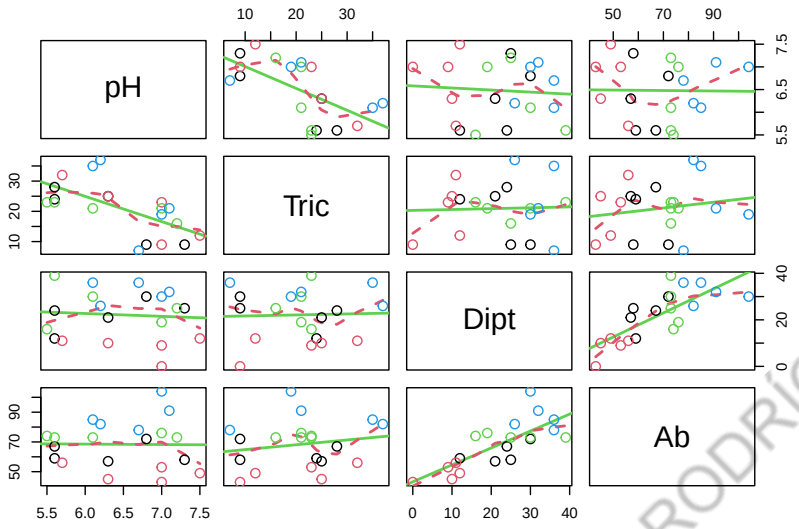
En este gráfico se incorporan dos tipos de líneas de ajuste. Las relaciones lineales las representa con las líneas verdes y las relaciones no lineales, las define con la línea suavizada roja, que se conoce como "loess" o "lowess", que sigue la tendencia más probable en la relación de las parejas de variables.

```
# 2. Figuras de pares con curvas de ajuste
pairs ((datos[,c(2:9)]),panel=function(x,y)
{abline(lsfit(x,y)$coef,lwd=2,col=3)
  lines(lowess(x,y),lty=2,lwd=2,col=2)
  points(x,y,cex=1)})
```



Grafica de pares, con líneas de ajuste lineal (líneas verdes) y no lineal o suavizada (líneas rojas). Los puntos corresponden a los valores de las variables en las quebradas

```
# 3. Pares con "cuenca" como un factor
pairs ((datos[,c(2,6,7,9)]),panel=function(x,y)
      {abline(lsfit(x,y)$coef,lwd=2,col=3)
      lines(lowess(x,y),lty=2,lwd=2,col=2)
      points(x,y,col=datos$cuenca, cex=1.4)})
```



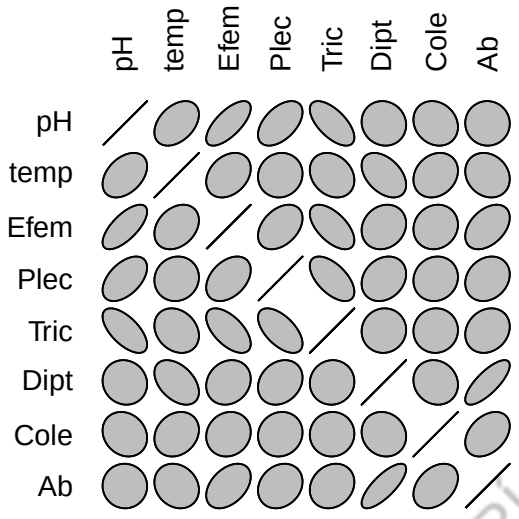
Grafica de pares, con líneas de ajuste lineal (líneas verdes) y no lineal o suavizada (líneas rojas). Los puntos corresponden a los valores de las variables en las quebradas. Los colores de los puntos relacionan a las diferentes cuencas o grupos evaluados.

En la siguiente figura, el panel superior relaciona a las relaciones suavizadas con los loess (líneas rojas), en la diagonal principal, se relaciona al patrón de distribución de frecuencias de cada variable (histograma) y en el panel inferior, a los coeficientes de correlación de Pearson, que indican si las relaciones en las parejas de variables son positivas (cercanas a 1) o negativas (cercanas a -1). Los asteriscos representan la significancia de las relaciones (* relaciones significativas, *** relaciones muy significativas).

4.2 Gráfica de elipses.

Estas figuras también permiten visualizar relaciones entre parejas de variables, dependiendo de la orientación de la elipse, así será el tipo de relación (positiva si hay una inclinación de la elipse hacia la derecha y negativa si la inclinación es hacia la izquierda), permiten visualizar un mayor número de relaciones que en las figuras de pares.

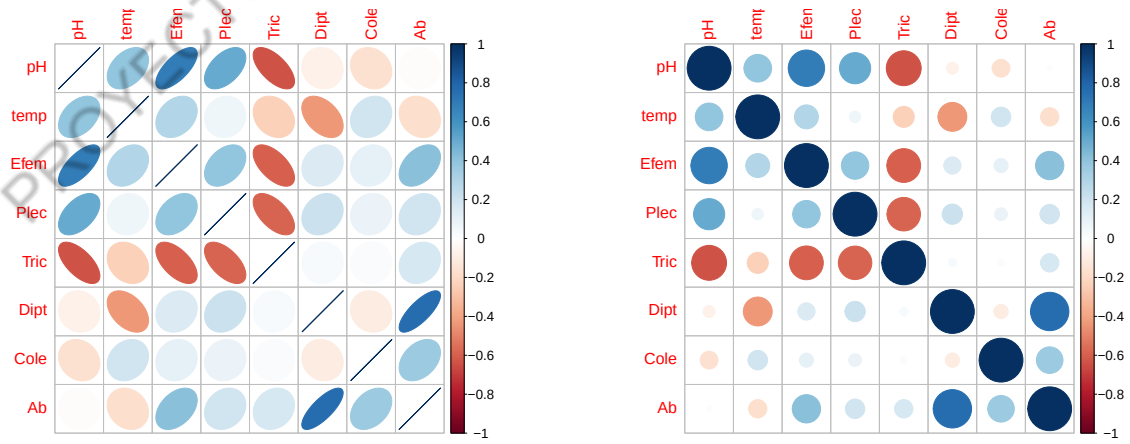
```
# 5.Figura elipses
plotcorr(cor(datos[,2:9]))
```



Graficas de elipses, que relacionan a parejas de variables.

En la figura de elipses, se puede visualizar relaciones positivas del pH con la abundancia de efemerópteros y de plecópteros. También se visualiza una relación negativa de esta variable ambiental con la abundancia de tricópteros. Hay otras relaciones entre órdenes de insectos, que no serán descritas en este documento.

```
# 6. Especies de insectos
M <- cor(datos[,2:9])                    # Matriz de Correlación (M)
x11()                                       # Panel gráfico adicional
par(mar = c(4, 4, .2, .1))
corrplot(M, method = "ellipse")      # Figura de correlaciones con elipses
corrplot(M, method = "circle")       # Figura de correlaciones con círculos
```



Graficas de elipses, que relacionan a parejas de variables. Colores azules, indican relaciones positivas y colores tojos, indican relaciones negativas. Se incluye a los coeficientes de correlación de Pearson.

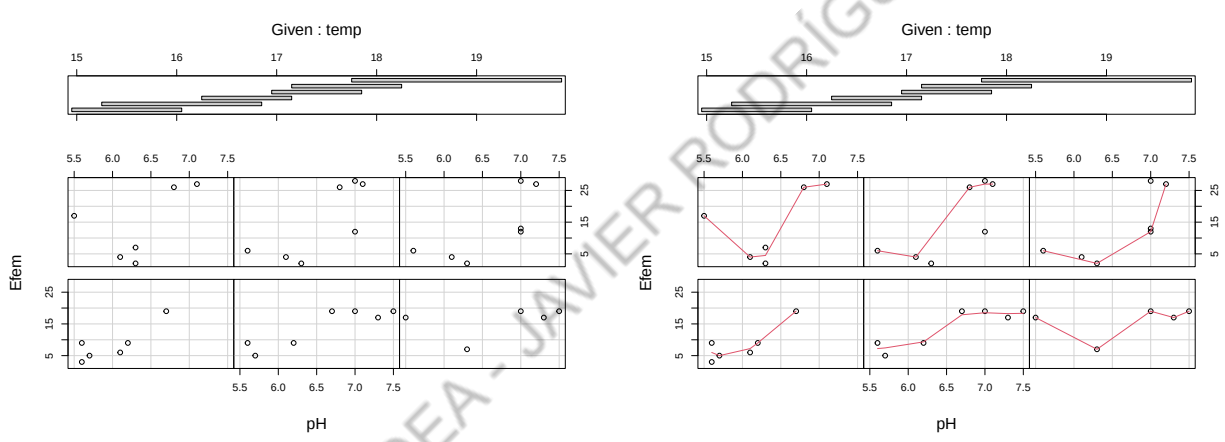
4.3 Figuras Coplot.

Son diagramas de dispersiones que diagnostican la asociación o relación entre dos variables continuas y una tercera categórica (o una continua que sea categorizada). También se puede incluir a una cuarta variable categórica en el análisis. Cuando la tercera variable (variable condicional) es nominal, no se presenta superposición de rangos. Cuando las variables son continuas que han sido discretizadas, suele presentarse cierto solapamiento. Cada panel en la relación de pH y Efemerópteros se asocia con las barras que representan los niveles de temperatura, partiendo de la izquierda.

```
par(mar = c(4, 4, .2, .1))

# 7. Figura con tres variables (Función: coplot)
with(datos,coplot(Efem~pH|temp))

# Coplot con líneas de ajuste suavizado (loess)
with(datos, {
    coplot(Efem~pH|temp,
         panel = panel.smooth)})
```



Coplot con suavizamiento (figura de la derecha). Se adicionan las líneas de suavizamiento o loess con rectas de color rojo (panel.smooth), que representan el tipo de relación más probable de las variables, para cada rango de temperatura.

Para las variables nominales, no hay superposición en los rangos de la variable condicional. Para las variables continúas acondicionado, se puede permitir un cierto solapamiento en los rangos de las variables condicionantes, y el número de gráficos, así como la cantidad de superposición se puede modificar. Cuando el tamaño de la muestra es similar en los diferentes paneles, suele utilizarse una línea de suavizamiento (loess).

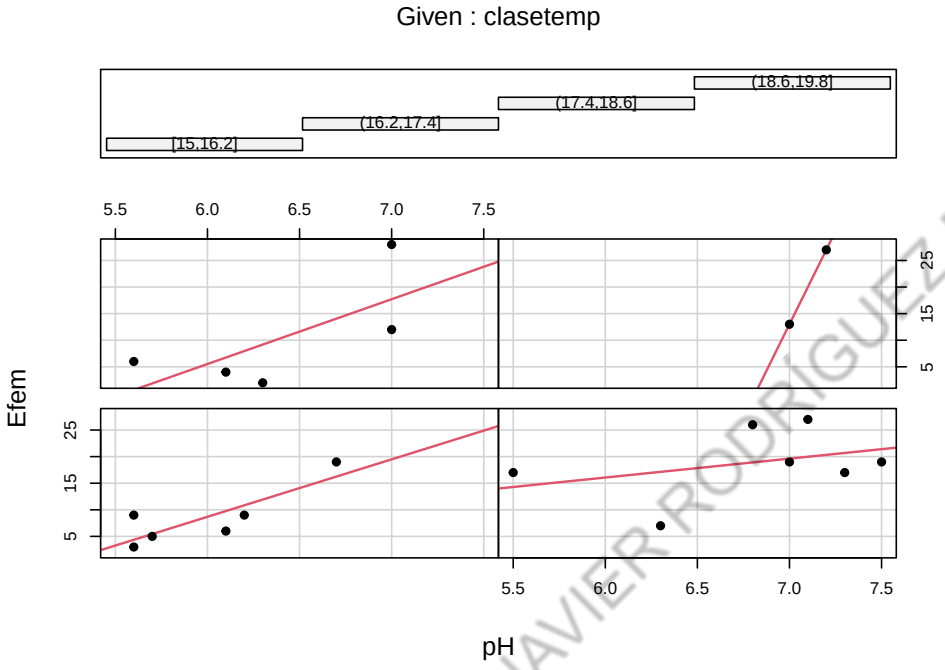
Coplot con variables categorizadas. Para este caso, se categoriza a una variable continua como la temperatura, con el fin de poder evaluar la relación entre las variables pH y Efemerópteros, con los rangos de temperatura. Los valores 15,20 son los niveles mínimos y máximos de temperatura, 1.2 son rangos de temperatura que se crean. Clasetemp, es el nombre de la variable temperatura discretizada en rangos de 1.2 oC.

```
# 8. Coplot con categórias
clasetemp<-cut(datos$temp,seq(15,20,1.2),include.lowest=T)
clasepH<-cut(datos$pH,seq(5,8,1,include.lowest=T))
```

```
## Warning: In seq.default(5, 8, 1, include.lowest = T) :
##  extra argument 'include.lowest' will be disregarded
```

```
panel.lm = function(x, y, ...) {
  tmp<-lm(y~x,na.action=na.omit)
  abline(tmp, lwd = 1.5, col= 2)
  points(x,y, ...)}

coplot(Efem~pH | clasetemp, pch=19, panel = panel.lm, data=datos)
```



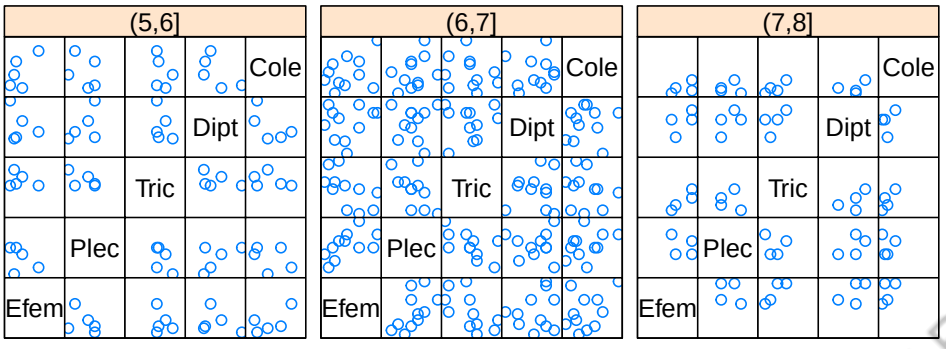
Graficas de coplot, con la variable temperatura discretizada en cuatro rangos. Los puntos corresponden a los valores de las variables en las quebradas, para cada rango de temperatura. Las líneas de ajuste no lineal o suavizado (líneas rojas), representan el tipo de relación entre las variables pH y Efemeropteros. Las barras corresponden a seis niveles de temperatura, que se asocian a cada panel inferior, partiendo del panel de la izquierda inferior.

4.4 Splom para variables categorizadas.

Con esta figura, se puede valorar las relaciones entre las variables biológicas (abundancia de los taxones), con rangos de las variables ambientales como la temperatura y el pH. Para este caso, las variables ambientales fueron categorizadas, utilizando el comando *"cut"* en *clasetemp* y *clasepH*.

El siguiente comando, permite ejecutar la siguiente figura, en la que se muestra la relación entre variables biológicas con tres rangos de pH (5-6, 6-7 y 7-8 unidades de pH).

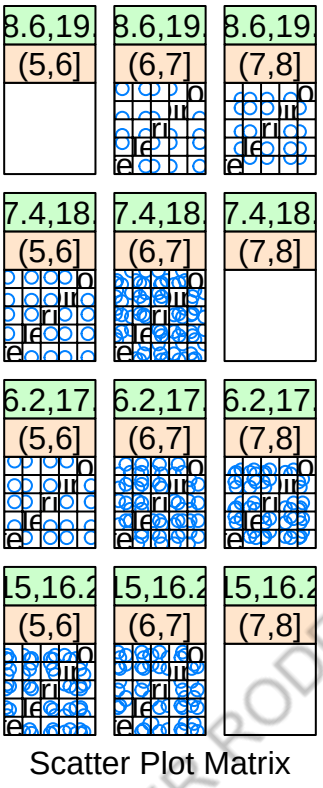
```
# 9. Splom para variables categorizadas
splom(~datos[,4:8]|clasepH,pscales=0)
```



Graficas splom, con la variable pH discretizada en tres rangos, que corresponden a los tres paneles visualizados. Los puntos corresponden a los valores de las variables en las quebradas.

El siguiente comando, permite ejecutar a la figura, en la cual se puede visualizar la relación entre las variables biológicas en seis paneles, definidos por categorías o rangos de pH y de temperatura.

```
splom(~datos[,4:8]|clasepH+clasetemp,pscales=0)
```



Graficas Splom, con las variables pH y temperatura, discretizadas en tres y dos rangos, respectivamente. Los puntos corresponden a los valores de las variables biológicas en las quebradas.

4.5 Histogramas de frecuencia.

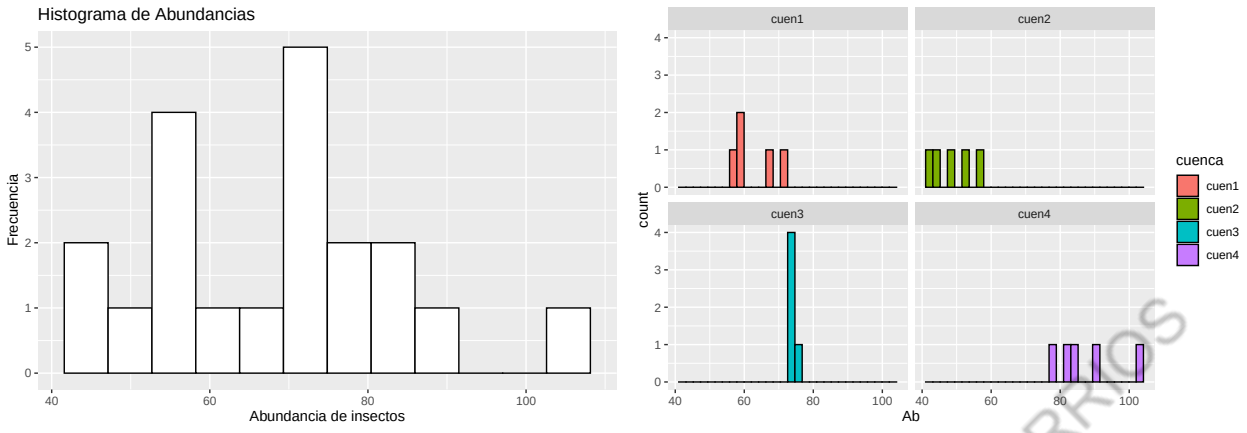
Muestran la simetría de las distribuciones de los datos en cada variable y brindan una idea de sus patrones de normalidad. Adicionalmente permiten evaluar el efecto de las transformaciones de las variables sobre su distribución. El gráfico de barras con los datos de la abundancia de invertebrados en las diferentes quebradas muestra cómo es su frecuencia general y por cada cuenca evaluada.

```
par(mar = c(4, 4, .2, .1))

ggplot(datos, aes(x=Ab)) +
      geom_histogram(bins = 12, color="black", fill="white") +
      labs( y="Frecuencia", x="Abundancia de insectos",
            title="Histograma de Abundancias")


# Histograma por tipos de cuencas
ggplot(datos, aes(x=Ab, fill=cuenca))+
  geom_histogram(color="black")+
  facet_wrap(~cuenca)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



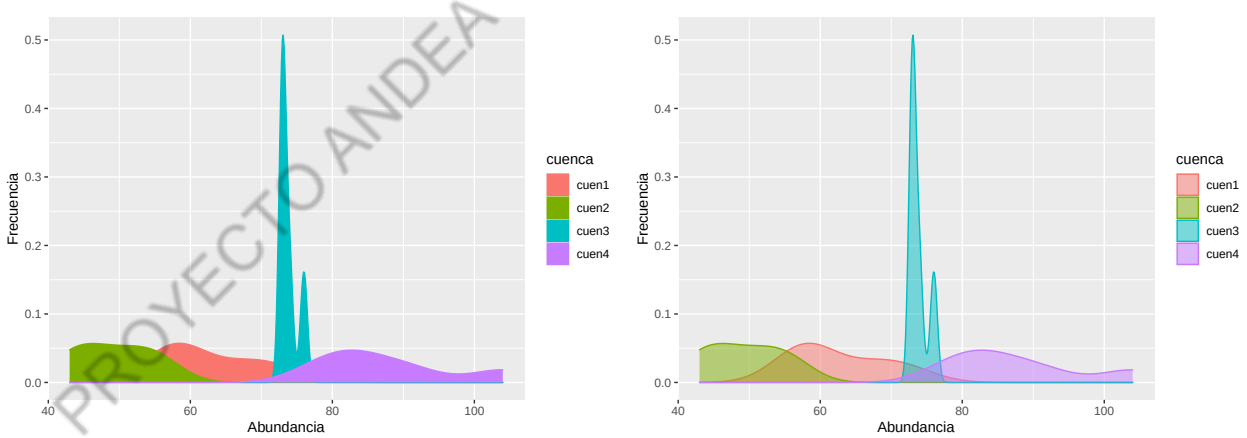
4.6 Histogramas de densidad.

Los siguientes histogramas, son similares a los anteriores, pero se realizan con el comando "densityplot", que permite visualizar las frecuencias en un gráfico de densidad

```
par(mar = c(4, 4, .2, .1))

# Figura de densidad por tipos de cuencas
ggplot(data = datos, aes(x = Ab, color = cuenca)) +
  geom_density(aes(fill = cuenca)) +
  labs( y="Frecuencia", x="Abundancia")

ggplot(data = datos, aes(x = Ab, color = cuenca)) +
  geom_density(aes(fill = cuenca), alpha = 0.5) +
  labs( y="Frecuencia", x="Abundancia")
```



Histogramas de frecuencias, con gráficas de densidad, relacionando a las abundancias de los invertebrados acuáticos (figura de la izquierda), por cada cuenca evaluada (figura de la derecha).

Ejercicios propuestos.

1. Realizar un análisis exploratorio gráfico de la base de datos de litios (género Iris), Seleccionando al menos cinco figuras que se ajusten mejor al patrón de los datos. La base de datos de "iris" se obtiene de la siguiente línea de comandos:

```
library(vegan) data(iris)

Con los datos tabulados, se requiere realizar y analizar las siguientes figuras (una de cada temática):
```

- Graficas de pares (pairplot)
- Figuras Coplot
- Histogramas
- Figuras quantil-quantil (QQ-plots)
- Diagrama de dispersión (plot y xyplot)
- Graficas de columnas o barras
- Figuras de Cajas (Boxplots)

2. Realizar el procedimiento del ejemplo 2, analizando a dos factores de la base de datos de variables ambientales medidas en dunas. De forma resumida se requiere explicar a los elementos de esta base y argumentar las variables (Horizonte A1) y a los factores seleccionados (Manejo y Uso). La base de datos de "dunas.env" se obtiene de la siguiente línea de comandos:

```
data(dune.env)
```

Con los datos tabulados, se requiere realizar y analizar las siguientes figuras:

- Figuras quantil-quantil (QQ-plots)
- Graficas de columnas o barras con desviaciones, para dos factores
- Figuras de Cajas (Boxplots)

NOTA: Para cada uno de los ejercicios, se requiere realizar la exploración gráfica, que incluya a los siguientes elementos:

1. Encabezado del taller, con un título, explicar de forma resumida a los elementos de cada base de datos, apoyado en el comando `help()`, el diseño de objetivos, una pregunta de análisis y redactar la hipótesis nula a explorar con el análisis gráfico.

2. Análisis de figuras exploratorias, similares a las realizadas los ejemplos de este capítulo, escogiendo las figuras que a su criterio sean las que mejor se ajusten a los datos analizados.

Nota: Los ejemplos 2 y 3 se encuentran en el libro Análisis de datos ecológicos y ambientales.