1. Figuras exploratorias multivariadas

1. Utilidad de R en la exploración de datos.
1. Figuras exploratorias

Ejercicios propuestos (para trabajar en grupos de 3 estudiantes).

Javier Rodríguez-Barrios

Introducción

Este capítulo es dedicado al análisis de graficas exploratorias, como un procedimiento casi obligatorio para cualquier tipo de investigación que requiera del análisis estadístico. 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 de R, pues cada función gráfica de R tiene un enorme número de opciones, por su gran flexibilidad gráfica superior a la de cualquier otro paquete estadístico.

Se introducen las funciones gráficas más básicas para el análisis exploratorio de datos univariados y multivariados, adicionalmente se brindan algunas opciones para la edición de figuras, como una de las principales opciones del programa R para brindar libertad al usuario manipular sus figuras. Se detallan importantes procedimientos de rutina, como el diseño de gráficas para detectar valores atípicos, detección de la media y el error estándar. Se analizarán diferentes paquetes gráficos entre las que se incluyen: *grid*, *lattice* y *ellipse*.

De acuerdo a Paradais (2003), 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 u observaciones, permite dar respuesta a diferentes tipos de preguntas, como las siguientes:

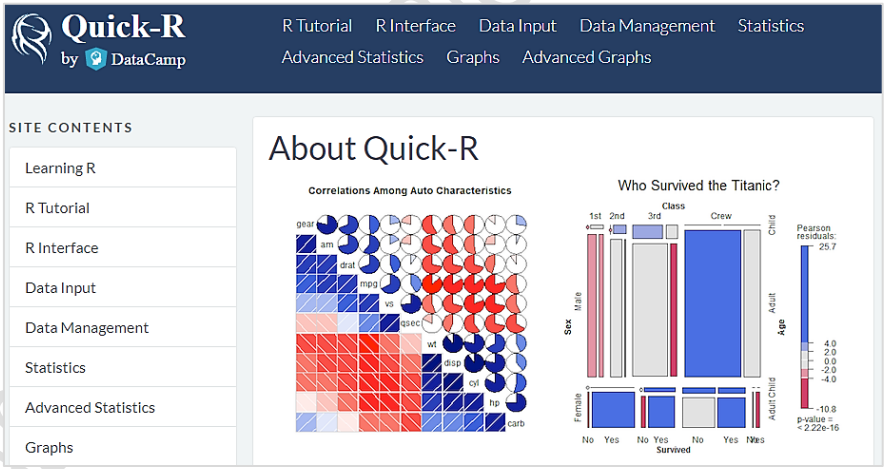
1. ¿Se encuentran los datos centrados? ¿Cómo es su distribución? ¿Los datos son simétricos, asimétricos, o presentan alguna tendencia particular?
2. ¿Hay datos o valores atípicos que puedan ser identificados?
3. ¿Las variables presentan una distribución normal o multinormal, homogénea u homocedástica?
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 requieren de una transformación o una estandarización?
6. ¿El esfuerzo de muestreo fue aproximadamente el mismo para cada observación o variable?

La tarea de todo investigador es hacer frente a estas preguntas, pues el paso a seguir consiste en revisar si los datos cumplen con varios supuestos antes de emitir cualquier conclusión. Por ejemplo, el análisis de componentes principales (ACP) depende de las relaciones lineales entre las variables. Los valores extremos o atípicos (outlying values) pueden causar las regresiones significativas, pero con errores en su análisis. En esta sección se analizan algunas herramientas de exploración principalmente gráfica, para intentar explicar cómo hacer para garantizar la validez de cualquier análisis posterior.

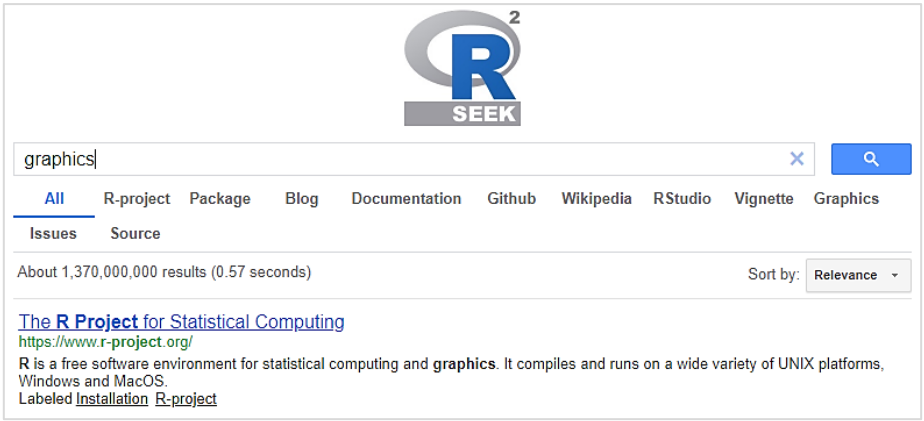
1. Utilidad de R en la exploración de datos.

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.

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



2. Motor de búsqueda de R (Rseek). Corresponde a una plataforma similar a la de google, en la que la comunidad que hace parte de la plataforma R, dispone de su información, para que sea de acceso gratuito.
https://rseek.org/

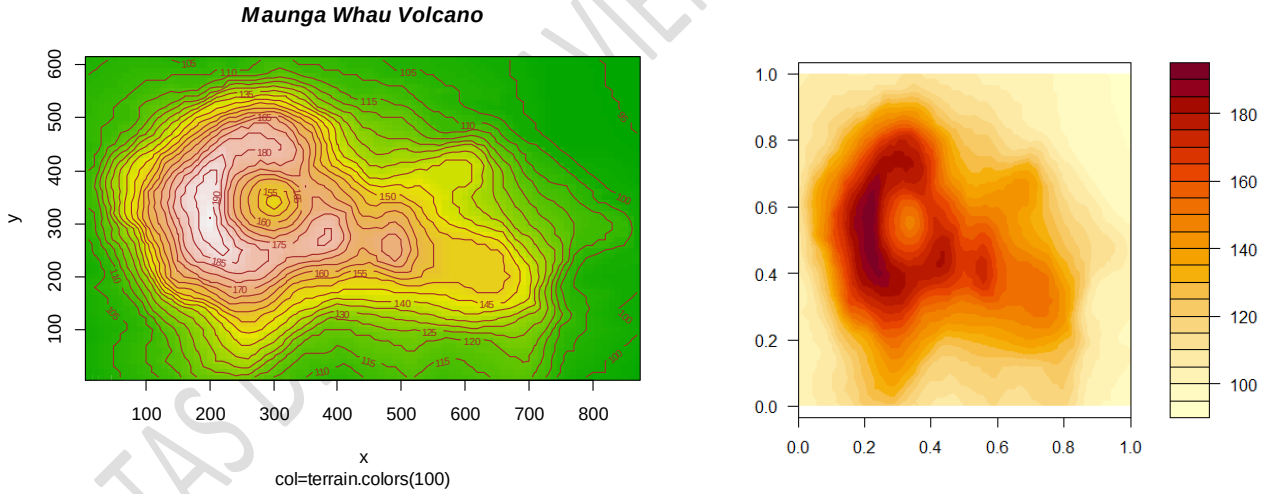


3. Otros sitios de utilidad

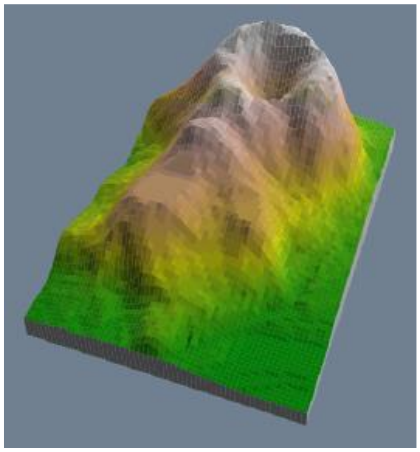
Sitio web de R, en el cual pueden descargarse las versiones más recientes de R.
http://www.r-project.org
http://www.cran.r-project.org
Enlace sobre opciones gráficas, manuales de soporte y paquetes disponibles en R
http://search.r-project.org/
Sitio web de R, en el cual se visualizan los tópicos generales de análisis que pueden realizase.
http://www.cran.r-project.org/web/views
Se cuenta con un espacio para realizar consultas sobre diferentes temas de análisis en R.
https://es.stackoverflow.com/questions/tagged/r
http://stats.stackexchange.com/questions/tagged/r
Se enumeran paquetes relacionados con análisis y modelos gráficos, son alrededor de 30 paquetes.
http://cran.r-project.org/web/views/gR.html
Enlace sobre respuestas a preguntas sobre temas de programación en R
http://stackoverfl ow.com/questions/tagged/r
R -bloggers: espacio informativo con más de 500 bloggers que proporcionan noticias y tutoriales sobre R.
https://www.r-bloggers.com/about/

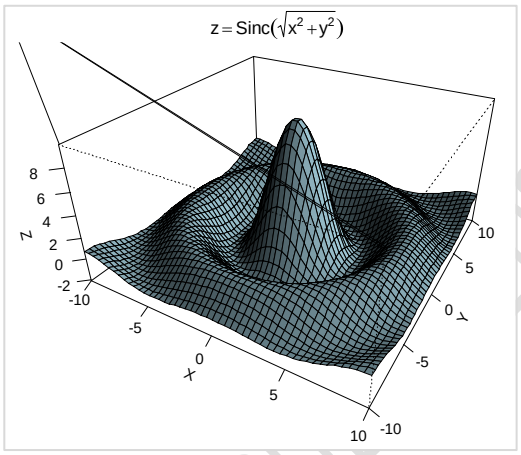
4. Opciones gráficas

El siguiente demo se ejecuta en R, permite visualizar opciones gráficas
demo(image)



La siguiente demostración, permite visualizar diferentes opciones de figuras tridimensionales, incluidos los modelos de elevación digital de terreno en un volcán (izquierda).
demo(persp)





A continuación, se relacionan otras demostraciones gráficas ofrecidas por R.
example(contour)
demo(graphics)
demo(plotmath)
demo(Hershey)

Las siguientes demostraciones gráficas, requieren del paquete gráfico “lattice”.
require("lattice")
demo(lattice)
example(wireframe)

Las siguientes demostraciones gráficas, requieren del paquete gráfico “rgl”.
require("rgl")
demo(rgl)
example(persp3d)

Ejemplo GRÁFICAS EXPLORATORIAS

El objetivo del presente ejercicio, 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 R, es una de sus principales fortalezas y por ende, se cuenta con numerosos textos, orientados exclusivamente a gráficas básicas y avanzadas, pero este no será el objeto del presente documento.

1. Figuras exploratorias

1.1 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 explicativas (independientes) que sean relacionadas (Zuur et al. 2007). De acuerdo es estos autores las figuras de pares son útiles bajo tres relaciones posibles: de variables respuesta, de variables explicativas y de respuesta versus explicativas.

***Ejemplo*.** A continuación, se realizará un análisis exploratorio de una base de datos "Insectos.csv", que incorpora variables ambientales y biológicas, las cuales caracterizan a diferentes cuencas. Se realizará una exploración de frecuencias y otros análisis que relacionen a las variables en las diferentes cuencas y quebradas.

Tabla 3. Representación de 10 de las 20 quebradas. Las variables fisicoquímicas corresponden al pH y Temperatura (Temp), las biológicas corresponden a órdenes de insectos acuáticos (Efem: Efemerópteros, Plec: Plecópteros, Tric: Tricópteros, Dipt: Dípteros, Cole: Coleópteros, Ab: Abundancia total de los insectos).

Quebrada	Cuenca	pH	Temp	Efem	Plec	Tric	Dipt	Cole	Ab
1	cuen1	6,8	17,4	26	4	9	30	3	72
4	cuen1	7,3	16,8	17	6	9	25	1	58
11	cuen1	5,6	16	9	3	28	24	3	67
13	cuen1	6,3	17,8	2	3	25	21	6	57
19	cuen1	5,6	18,2	6	4	24	12	13	59
3	cuen2	6,3	17	7	2	25	10	1	45
10	cuen2	7,5	16,8	19	3	12	12	3	49
15	cuen2	7	18,2	12	5	23	9	4	53
16	cuen2	7	19,8	13	6	9	0	15	43
17	cuen2	5,7	15,3	5	0	32	11	8	56

Lectura de la base de datos “Insecto.csv”
```
datos<-read.csv2("Insectos.csv",row.names=1)
```

Librerías requeridas
```
library(lattice)
library(ellipse)
require(SciViews)
require(stats)
```

1. Gráfica por pares |

Variables 2 a 8, corresponden a las dos ambientales y cinco biológicas. Aquí se comparan parejas de variables, buscando tendencias lineales (positivas o negativas), entre las variables, especialmente entre las biológicas y las ambientales. La transformación logarítmica es una opción para linealizar las relaciones.
```
pairs(datos[,2:8])
pairs(log10(datos[,2:8]))
```

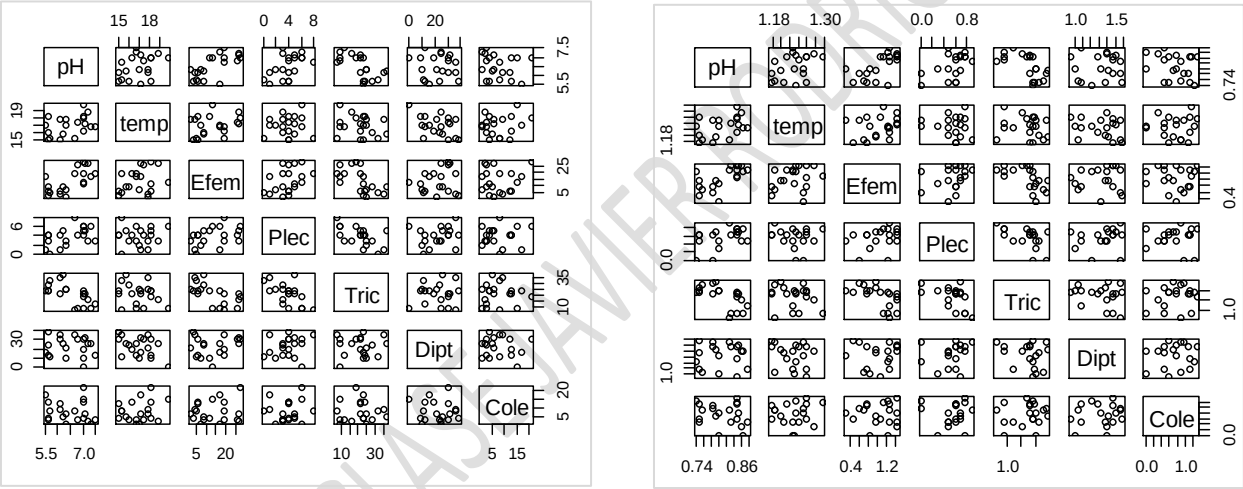


Figura 7. Graficas de pares, con dispersión de los datos originales (izquierda) y con transformación logarítmica base 10 (derecho).

2. Gráfica de elipses |

Estas figuras también permiten realizar relaciones entre parejas de variables, dependiendo de la orientación de la elipse, así será el tipo de relación (positiva si hay inclinación hacia la derecha y negativa, si la inclinación es hacia la izquierda).
```
plotcorr(cor(datos[,2:9]))
```

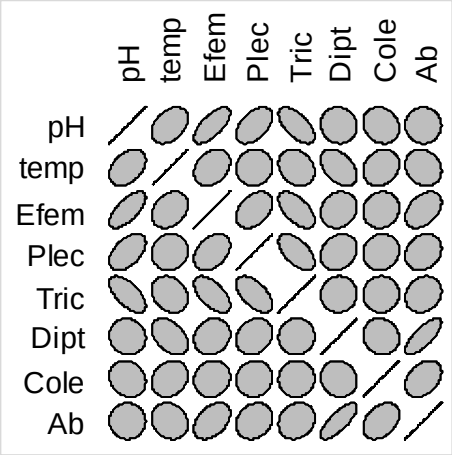


Figura 8. Graficas de elipses, que relacionan a parejas de variables.

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.

3. Otras graficas de pares

En este gráfico se incorporan dos tipos de líneas de ajuste. Las relaciones lineales las representa con las líeas 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.

```
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)})
```

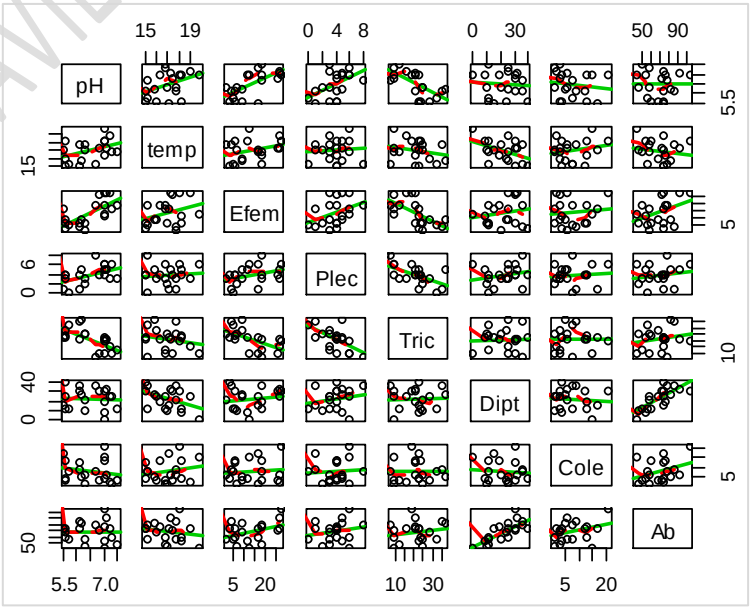


Figura 9. Graficas de pares, con líneas de ajuste lineal (líneas verdes) y no lineal o suavizada (líneas rojas). Los puntos corresponden a los valores de las variables en las quebradas.

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).

```
pairs    (datos[, 2:9], diag.panel = panel.hist,
          upper.panel = panel.smooth, lower.panel = panel.cor)
```

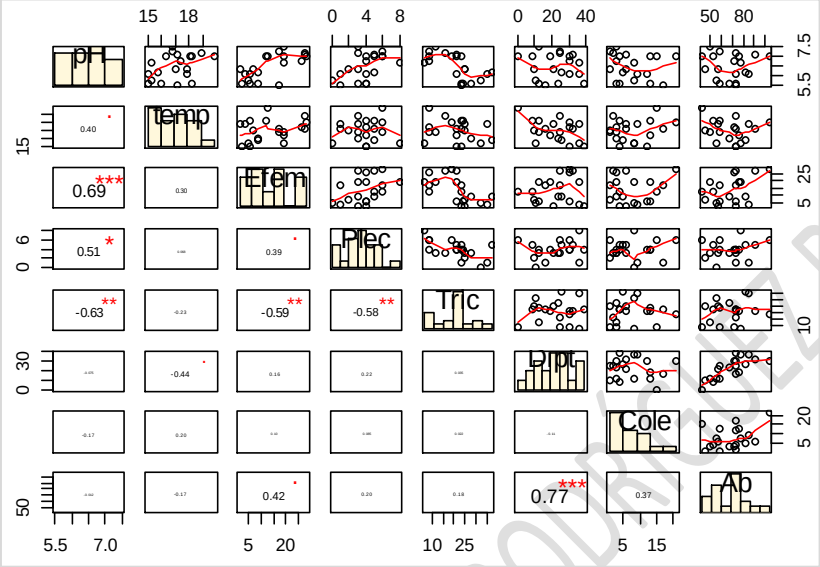


Figura 10. Graficas de pares, con líneas de ajuste no lineal o suavizado (líneas rojas). Los puntos corresponden a los valores de las variables en las quebradas. Los valores representan los coeficientes de correlación y los asteriscos, al nivel de significancia de las relaciones.

1.2 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). En ocasiones suele incluirse 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.

Para las variables nominales, como se muestra en las figuras 11 a 14, 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).

Ejemplo. Con los datos del ejercicio anterior, se analizará la relación entre dos variables (Efemeróteros y el pH), para niveles de otra variable (temperatura). Se utilizarán diferentes opciones de gráficas con la función “coplot”.

1. Coplot básico (figura de la izquierda).

Cada panel en la relación de pH y Efemerópteros, se asocia con las barras que representan los niveles de temperatura, partiendo de la izquierda.

```
with(datos,coplot(Efem~pH|temp))
```

2. 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.

```
with(datos, {
 coplot(Efem~pH|temp, number = 3,
      panel = function(x, y, ...) panel.smooth(x, y, span = .8, ...))
 coplot(Efem~pH|temp,
      panel = panel.smooth)
})
```

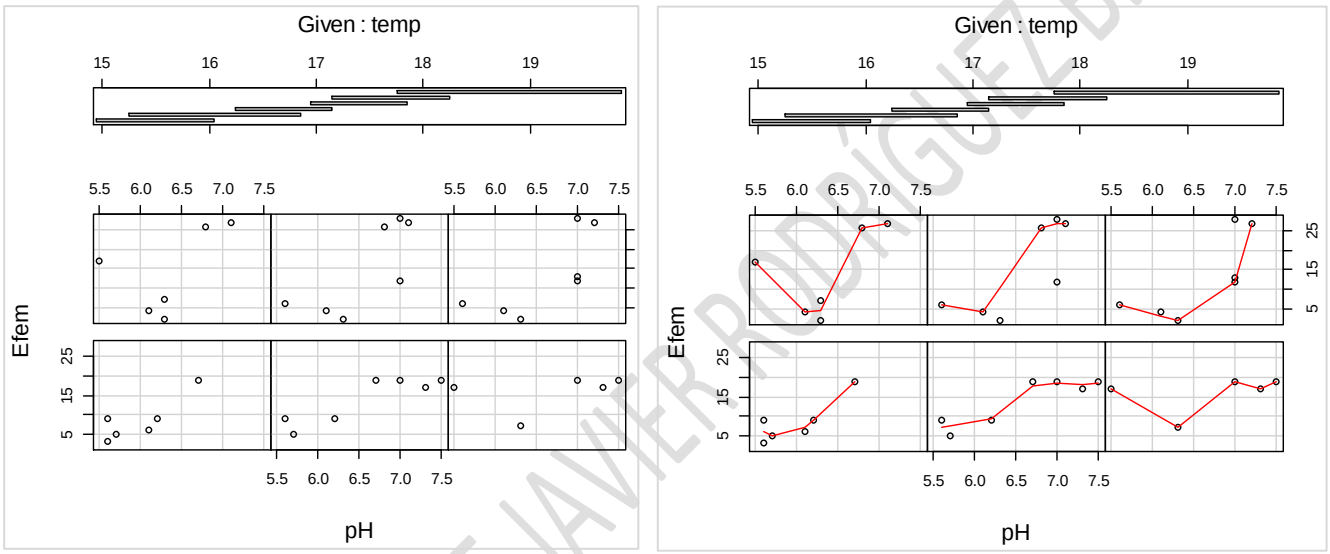


Figura 11. Graficas de coplot. Los puntos corresponden a los valores de las variables en las quebradas, para cada rango de temperatura. Las líneas de ajuste no lineal o suavizado (líneas rojas), representan el tipo de relación entre las variables pH y Efemeropteros. Las barras corresponden a seis niveles de temperatura, que se asocian a cada panel inferior, partiendo del panel de la izquierda inferior.

3. 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.

```
clasetemp<-cut(datos$temp,seq(15,20,1.2),include.lowest=T)
clasetemp
```

En el siguiente comando, se realiza la figura, con la temperatura categorizada.

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

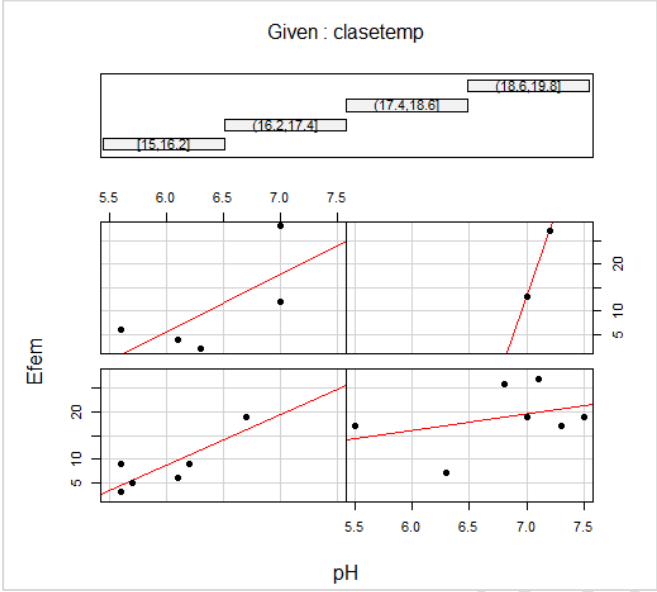


Figura 12. Graficas de coplot, con la variable temperatura discretizada en cuatro rangos. Los puntos corresponden a los valores de las variables en las quebradas, para cada rango de temperatura. Las líneas de ajuste no lineal o suavizado (líneas rojas), representan el tipo de relación entre las variables 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. Splom para variables categorizadas.

Con esta figura, se puede valorar las relaciones entre las variables biológicas, con rangos de las variables ambientales como la temperatura y el pH.

El siguiente comando, permite discretizar a la variable pH, similar a lo realizado con la temperatura.

```
clasepH<-cut(datos$pH,seq(5,8,1,include.lowest=T))
```

El siguiente comando, permite ejecutar la figura 13, 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).

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

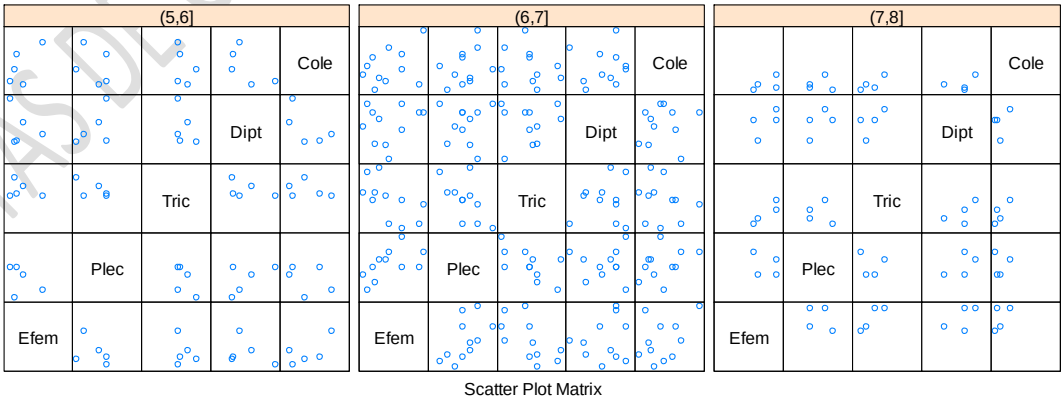


Figura 13. Graficas splom, con la variable pH discretizada en tres rangos, que corresponden a los tres paneles visualizados. Los puntos corresponden a los valores de las variables en las quebradas.

Categorización de las variables temperatura (clasetemp) y pH (clasepH), para realizar un gráfico que integre a las dos categorías.

```
clasetemp<-cut(datos$temp,seq(15,20,2),include.lowest=T)
clasepH<-cut(datos$pH,seq(5,8,1,include.lowest=T))
```

El siguiente comando, permite ejecutar a la figura 14, 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)
```

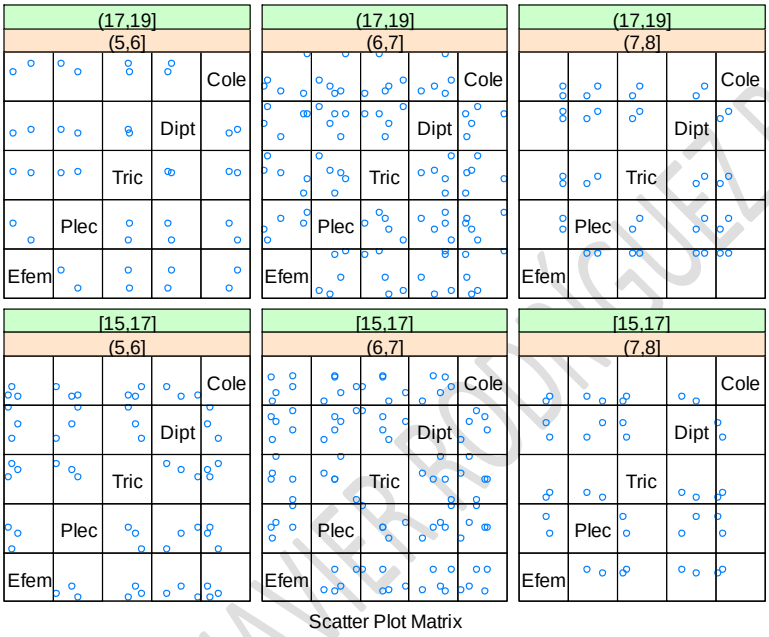


Figura 14. Graficas splom, con las variables pH y temperatura, discretizadas en tres y dos rangos, respectivamente. Los puntos corresponden a los valores de las variables biológicas en las quebradas.

5. xyplot

Con este comando, se puede realizar un entorno gráfico similar al realizado con el comendo "splom", para este caso, se realizar una relación entre el pH y la abundancia de Efemerópteros, en las cuatro cuencas evaluadas.

```
xyplot(Efem~pH|cuenca,data=datos)
```

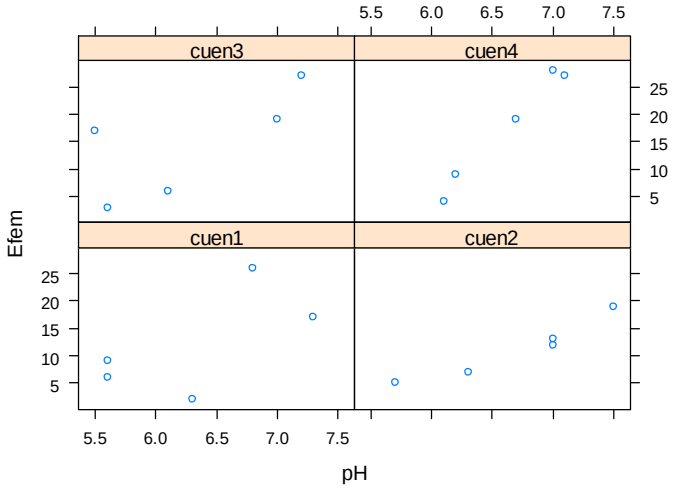


Figura 15. Graficas xyplot, con la relación entre el pH y los efemerópteros, para las cuatro cuencas evaluadas. Los puntos corresponden a los valores de las variables en las quebradas.

1.3 Histogramas

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.

Los dos comandos que se muestran a continuación, permiten realizar los histogramas de la figura 16.

```
histogram  (~Ab,data=datos, ylab="Porcentaje del Total", xlab="Abundancia de insectos")
histogram  (~Ab|cuenca,data=datos, ylab="Porcentaje del Total", xlab="Abundancia de insectos")
```

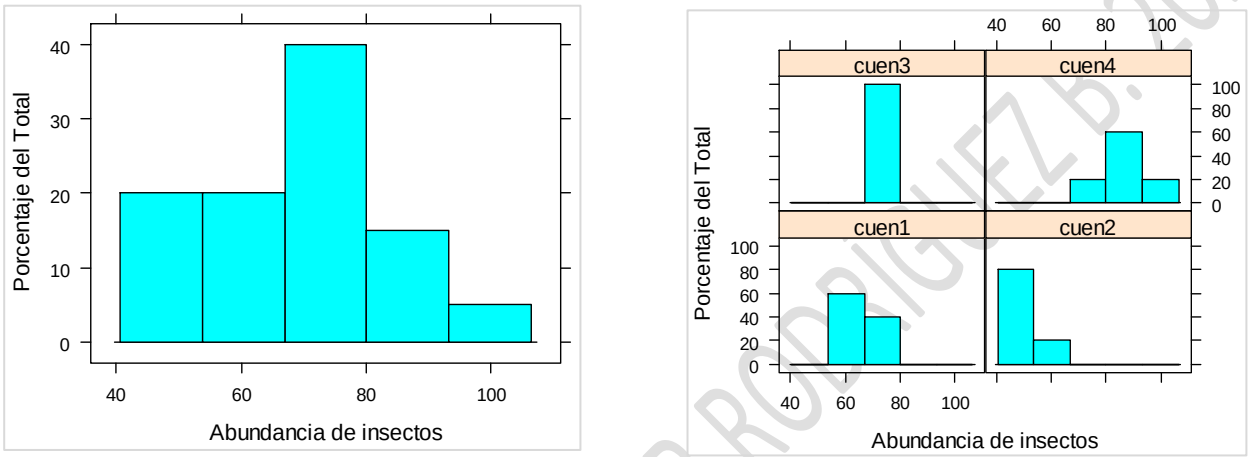


Figura 16. Histogramas de frecuencias de las abundancias de los invertebrados acuáticos (figura de la izquierda) y por cada cuenca evaluada (figura de la derecha).

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 (figura 17).

```
densityplot(~Ab,data=datos, ylab="Porcentaje del Total", xlab="Abundancia de insectos")
densityplot(~Ab|cuenca,data=datos, ylab="Porcentaje del Total", xlab="Abundancia de insectos")
```

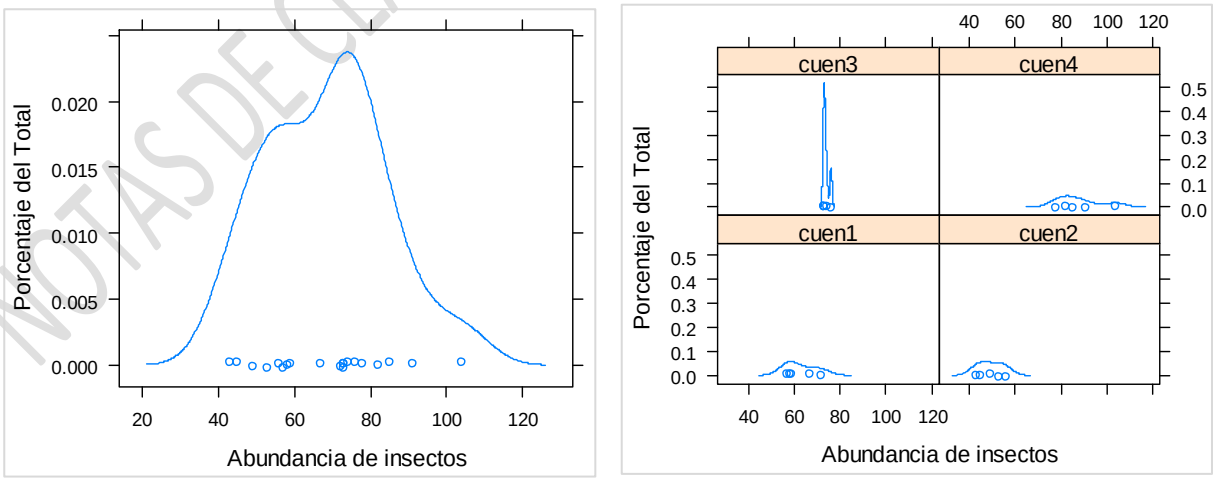


Figura 17. Histogramas de frecuencias, con gráficas de densidad, relacionando a las abundancias de los invertebrados acuáticos (figura de la izquierda), por cada cuenca evaluada (figura de la derecha).

1.4 Figuras quantil-quantil (QQ-plots)

Se utilizan para saber si los datos presentan una distribución particular (normal, logarítmica, etc.). La distribución gausiana o de normalidad en las variables es evaluada por figuras QQ-plot, en la medida que las distribuciones de los puntos se ubiquen en línea recta. Estas figuras pueden desarrollarse con base en los datos crudos o con diferentes transformaciones, que permitan conocer la distribución más probable de los datos de cada variable.

El siguiente comando, permite visualizar un panel que mostrará a dos figuras en simultánea.

```
panel<-par(mfrow=c(1,2), mar=c(4,3,3,2))
```

Los comandos "qqnorm" y "qqline" permiten realizar el gráfico que, para este caso, diagnostica el patrón de normalidad (ajuste de los datos a la recta), en la variable abundancia de insectos (Ab).

```
qqnorm (datos$Ab, main="Abundancia de Insectos", ylab="Cuantiles de la muestra", xlab="Cuantiles teóricos")
qqline(datos$Ab)
```

A continuación, se realiza una transformación logarítmica de los valores de abundancia.

```
Ab.log <- log10(datos$Ab+1)
```

Con la variable transformada se ejecuta el gráfico cuantil, para comparar el patrón de normalidad con el anterior.

```
qqnorm(Ab.log,main="Log de Abundancia de Insectos",ylab="Cuantiles de la muestra",xlab="Cuantiles teóricos")
qqline(Ab.log)
par(panel)
```

En la siguiente figura (figura 18) se logra visualizar, cierta tendencia al patrón de normalidad de la variable Abundancia de invertebrados acuáticos (figura de la izquierda), sin muchas diferencias con el patrón para la variable transformada (figura de la derecha).

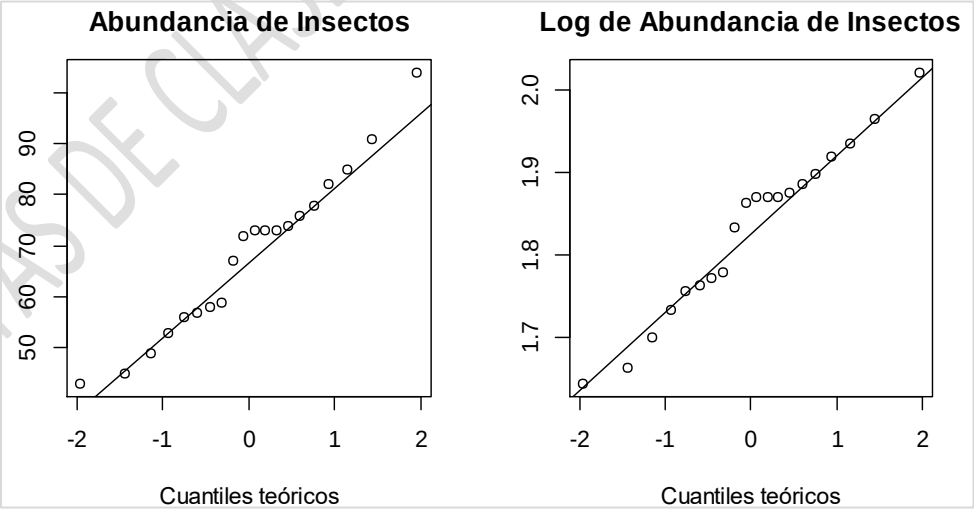


Figura 18. Gráficos qq o de cuantil-cuantil, para visualizar el patrón de normalidad de la variable abundancia de insectos acuáticos con datos crudos (izquierda) y con transformación logarítmica (derecha).

1.5 Diagramas de dispersión (plot y xyplot)

A diferencia de las figuras anteriores que diagnostican la distribución de los datos (valores atípicos, normalidad), la siguiente técnica permite diagnosticar relaciones entre las variables. Se destacan funciones como las figuras de pares de variables (plot, xyplot).

A continuación, se realiza un comparativo de dos figuras que relacionan la abundancia de dos grupos de insectos acuáticos (figura 19).

```
panel<-par(mfrow=c(1,2), mar=c(4,5,3,2))
with(datos,plot(Efem~Plec,type="p",ylab="Efemerópteros", xlab="Plecópteros"))
```

"lowess y lm", son los comandos utilizados para realizar las rectas de relación no lineal (azul) y lineal (roja), respectivamente.

```
lines(lowess(datos$Plec,datos$Efem),col=4)
lines(abline(lm(datos$Efem~datos$Plec),lwd=2,col=2, lty=2))
```

"legend", permite ubicar la leyenda de la figura de la derecha, en coordenadas x (0) e y (27), relacionando a las diferentes cuencas.

```
plot(Efem~Plec,col=as.integer(cuenca),data=datos,ylab="", xlab="Plecópteros")
legend(0,27,legend=levels(datos$cuenca),pch=19,col=1:4,cex=0.8)
lines(abline(lm(datos$Efem~datos$Plec),lwd=2,col=2, lty=2))
par(panel)
```

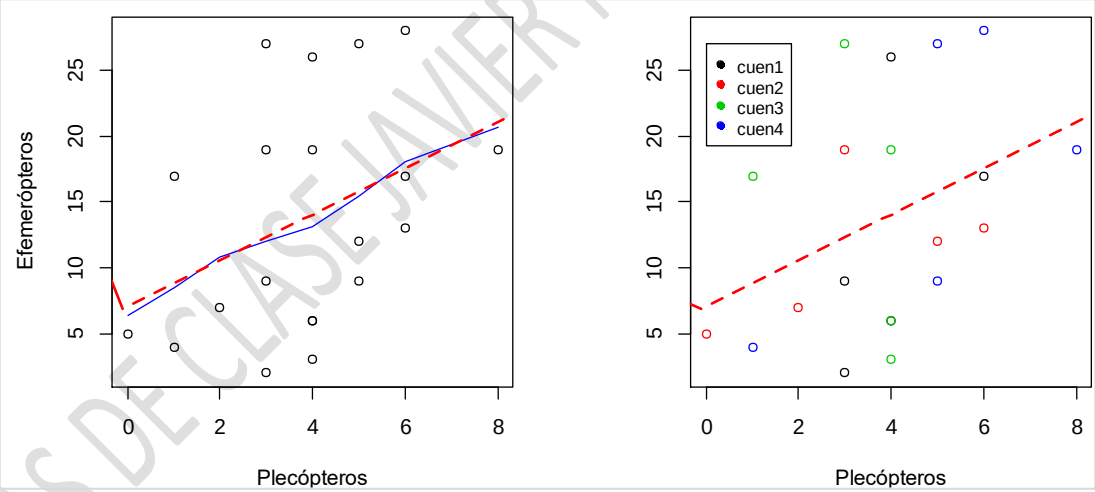


Figura 19. Relación entre la abundancia de plecópteros y de Efemerópteros. La línea azul, es de suavizamiento o loess, la línea roja punteada, corresponde a la relación lineal. Los colores de la figura de la derecha, representan a los valores de las variables en comparación, para las diferentes cuencas evaluadas.

El archivo que se revisará inicialmente, consiste en una base de datos, de invertebrados bentónicos, colectados en un río de la ciudad de Santa Marta, con el objeto de valorar las diferencias de biomasas en nos niveles de altura (A: Alto, B: Bajo) y entre microhábitatas (Unidades funcionales) (Rodríguez et al. 2011).

1. Lectura de base de datos de invertebrados bentónicos ("Datos1.csv")

```
datos<-read.csv2("Datos1.csv")
str(datos)
```

Tabla 1. Valores de biomasa (mg) de cinco grupos funcionales (GFA) macroinvertebrados acuáticos bentónicos (C-F: colectores – filtradores, C-R: colectores recolectores, R: raspadores, T: trituradores), presentes en dos tramos en un gradiente de altura del río Gaira – Santa Marta (A: alto, B: medio). BIOM-TOT: es la biomasa total.

COD	TRAMO	MUESTREO	LLUVIA	R	C-F	D	T	C-R	BIOM.TOT
AM1	A	M1	P1	46,2	48	152	553,7	20,8	820,7
AM2	A	M2	P1	66,1	111,9	372,9	559	105,6	1215,5
AM3	A	M3	P2	65,7	69	121,2	492,3	29,8	778
AM4	A	M4	P2	2,4	45,4	86,2	8,2	5,6	147,8
AM5	A	M5	P2	4,8	61	218,8	0,3	14,2	299,1
AM6	A	M6	P2	5,5	107,9	35,3	1895,7	30	2074,4
AM7	A	M7	P2	6	53,9	7,4	793,5	10,4	871,2
AM8	A	M8	P1	12,1	27,5	68,9	739,2	23,5	871,2
AM9	A	M9	P1	26,8	182,6	394,3	2824,4	43,1	3471,2
AM10	A	M10	P1	8	62,5	286,8	1041,1	28,9	1427,3
BM1	B	M1	P1	324,2	46,7	1226	306,2	36,9	1940
BM2	B	M2	P1	76,6	33,3	436	208,2	57,1	811,2
BM3	B	M3	P2	3,2	21	262,5	1708,9	17,4	2013
BM4	B	M4	P2	6	7,9	1146,9	1100,9	64,7	2326,4
BM5	B	M5	P2	0,4	7,9	552,8	0,8	30,4	592,3
BM6	B	M6	P2	4	74	559,6	425,9	11,1	1074,6
BM7	B	M7	P2	2	6,6	409,5	246,1	17,3	681,5
BM8	B	M8	P1	7,9	32	665,2	425,9	16	1147
BM9	B	M9	P1	178	69	1026	373,7	52,2	1698,9
BM10	B	M10	P1	12,5	9,9	834,6	191,9	17,6	1066,5

Estructura de la base de datos de la tabla 1.

```
str(datos)
```

```
data.frame      : 20 obs. of  10 variables:
 $ COD          : Factor w/ 20 levels "AM1","AM10","AM2",..: 1 3 4 5 6 7 8 9 10 2 ...
 $ TRAMO        : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
 $ MUESTREO     : Factor w/ 10 levels "M1","M10","M2",..: 1 3 4 5 6 7 8 9 10 2 ...
 $ LLUVIA       : Factor w/ 2 levels "P1","P2": 1 1 2 2 2 2 2 1 1 1 ...
 $ R            : num  46.2 66.1 65.7 2.4 4.8 5.5 6 12.1 26.8 8 ...
 $ C.F          : num  48 111.9 69 45.4 61 ...
 $ D            : num  152 372.9 121.2 86.2 218.8 ...
 $ T            : num  553.7 559 492.3 8.2 0.3 ...
 $ C.R          : num  20.8 105.6 29.8 5.6 14.2 ...
 $ BIOM.TOT     : num  821 1216 778 148 299 ...
```

La estructura de la base de datos, indica que se cuenta con 10 variables, de las cuales, la primera es el código o consecutivo, las tres siguientes son factores y las seis restantes son variables, asociadas a la biomasa de los invertebrados.

Librerías requeridas para el análisis grafico de todo el ejercicio.

```
library(lattice)
library(ellipse)
library(plotrix)
require(SciViews)
require(stats)
```

1.6 Figuras circulares (Pie Chart)

Estas figuras permiten realizar una comparación porcentual de variables categóricas o nominales, incluidos a los factores. La función *pie*, permite realizar figuras típicas de Excel, en relación a figuras de tortas.

Ejemplo. La figura 1, permite visualizar la comparación en los valores absolutos de abundancia en diferentes grupos funcionales alimenticios de macroinvertebrados acuáticos (GFA).

El comendo "colSums" permite sumar columnas, que para este caso serán las variables relacionadas a los grupos funcionales (variables 5 a 9).

```
datos1  <- colSums(datos[,5:9])
```

El comando "par" permite realizar un panel gráfico, en este caso, incorpora a dos figuras superiores y dos inferiores en el mismo gráfico (2,2).

```
par(mfrow = c(2,2), mar = c(3, 3, 2, 1))
```

El comando "pie", se orienta a la realización de gráficos circulares o de anillos, como se presentan a continuación.

```
pie(datos1 , main = "Figura Circular Ordinaria")
pie(datos1 , col = gray(seq(0.4,1.0,length=6)),
   clockwise=TRUE, main = "Escala de Grises", angle=45)
pie(datos1 , col = rainbow(6),clockwise = TRUE, main="Colores de Arcoiris")
pie3D(datos1 , labels = names(datos1), explode = 0.1,  main = "Figura Circular en 3D", labelcex=0.8)
```

El comando "par" para este caso, permite devolver el panel a una sola figura, por cada grafico a realizar.

```
par(bentos)
```

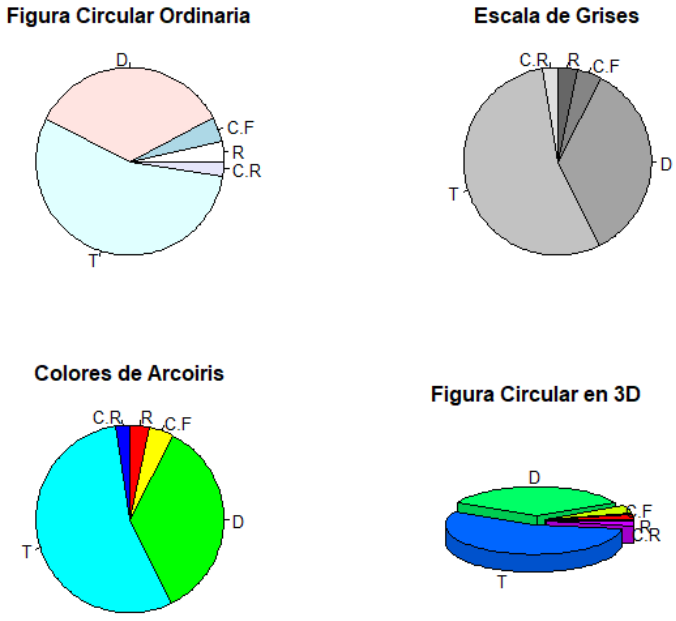


Figura 1. Figura circular o de totas, en la que se representa a los niveles de biomasa de diferentes grupos funcionales de insectos acuáticos. T= Trituradores, C-F= Colectores filtradores, C-R= Colectores recolectores, R= Raspadores y D= Depredadores (Tomado de Rodríguez et al. 2011).

1.7 Graficas de columnas o barras

Estas figuras permiten evaluar el patrón de las variables a través de diferentes categorías de otras variables, como lo es el tiempo, sitios, etc. La figura 2 presenta una forma de presentar la distribución de abundancia de los GFA en el tramo alto del río Gaira (Rodríguez 2011).

Ejemplo. Lectura de base de datos ("Datos1.csv"), se graficarán los datos de biomasa total de cada tramo.

```
datos<-read.csv2("Datos1.csv")
```

DatosA corresponde a los valores del tramo alto (Tramo A)

```
datosA=datos[1:10,]
datosA
```

Se extraen los datos de biomasa (BIOM.TOT) y se rotulan con los muestreos (datosA[,3])

```
tramoA=datosA$BIOM.TOT
tramoA
names(tramoA) <- datosA[,3]
tramoA
```

DatosB corresponde a los valores del tramo medio (Tramo B)

```
datosB=datos[11:20,]
datosB
```

Se extraen los datos de biomasa (BIOM.TOT) y se rotulan con los muestreos (datosA[,3])

```
tramoB=datosB$BIOM.TOT
tramoB
names(tramoB) <- datosB[,3]
tramoB
```

Se fusionan los datos de biomasa total de los dos tramos, en la base “tramos”

```
tramos <- cbind(tramoA, tramoB)
tramos
```

Se diseña el panel gráfico “par”, para cuatro figuras (mfrow = c(2,2)) y a continuación, se ejecutan las figuras.

```
par(mfrow = c(2,2), mar = c(3, 3, 2, 1))
barplot(tramoB , main = "Biomasas")
barplot(tramos)
barplot(t(tramos), col = gray(c(0.5,1)))
barplot(t(tramos), beside = TRUE)
par(mfrow = c(1,1))
```

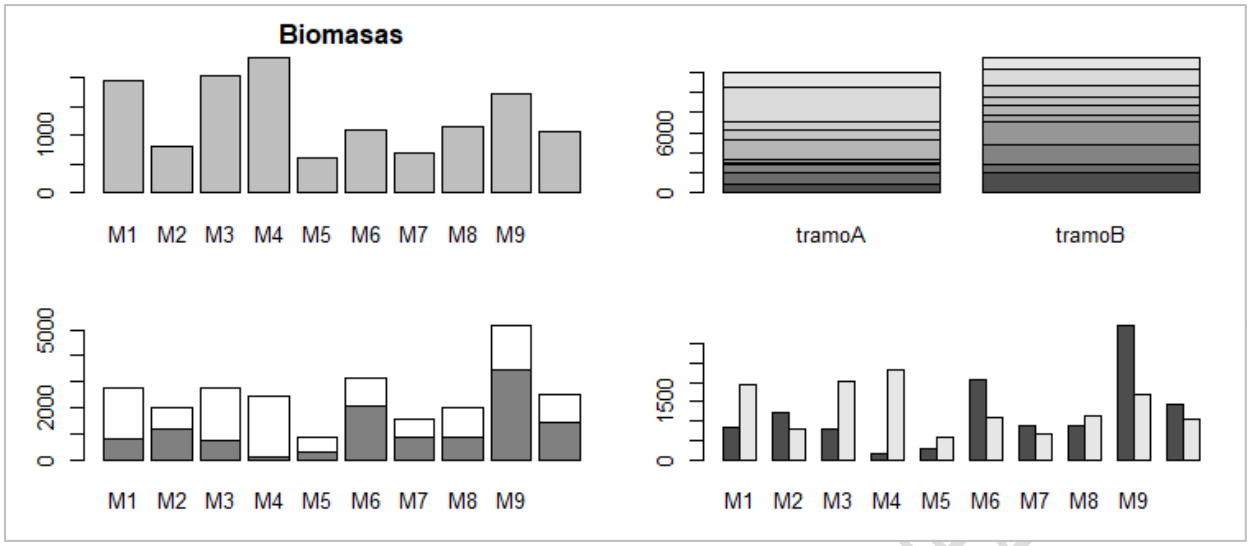


Figura 2. Figura de columnas, en la que se representa a los niveles de biomasa de diferentes grupos funcionales de insectos acuáticos, por muestreos y por tramos (Tomado de Rodríguez et al. 2011).

1.8 Graficas de columnas o barras con desviaciones estándar

Son figuras que en la parte superior se representan por barras o columnas que representan los promedios, con líneas acotadas que muestran las desviaciones estándar por encima y/o debajo del promedio. Tienen la desventaja de asumir que los datos presentan una distribución normal, lo cual puede generar limitantes, por lo que ciertos investigadores prefieren las figuras de cajas y bigotes. Sin embargo, los investigadores en ocasiones suelen encontrar útil a estas figuras (figuras de bajo nivel).

***Ejemplo*.** Figuras de columnas, con promedios y desviaciones estándar, para la abundancia de invertebrados acuáticos que derivan (se transportan en la columna de agua) del río Gaira de Santa Marta.

Lectura de la base de datos de invertebrados acuáticos “Datos2.csv”.

```
datos<-read.csv2("Datos2.csv")
datos
```

Tabla 2. Primeros 16 datos de la base de datos utilizada para las figuras de columnas y de cajas. M, representa a los 8 muestreos realizados, GF a los grupos funcionales, Lluvia y Caudal, al régimen hídrico del río (P1 y C1 corresponden a sequía y caudal bajo, P2 y C2 a lluvia y caudal alto). Ab, Biom y Largo, representan a la abundancia, la biomasa y el tamaño de los individuos, respectivamente.

No	M	GF	Lluvia	Caudal	Ab	Biom	Largo
1	M1	C-F	P1	C1	98	56,05	89,28
2	M2	C-F	P1	C1	198	52,66	236,35
3	M3	C-F	P2	C1	45	11,37	95,75
4	M4	C-F	P2	C2	51	25,26	101,4
5	M5	C-F	P2	C2	3	0,36	9
6	M6	C-F	P2	C2	69	23,59	57,34
7	M7	C-F	P2	C2	31	11,2	28,18
8	M8	C-F	P1	C2	17	2,42	24,48

No	M	GF	Lluvia	Caudal	Ab	Biom	Largo
9	M1	C-R	P1	C1	157	16,18	103,88
10	M2	C-R	P1	C1	385	68,55	632,08
11	M3	C-R	P2	C1	110	64,33	307,84
12	M4	C-R	P2	C2	399	290,51	389,78
13	M5	C-R	P2	C2	206	57,19	197,62
14	M6	C-R	P2	C2	48	14,46	113,33
15	M7	C-R	P2	C2	85	16,12	100,92
16	M8	C-R	P1	C2	72	10,26	123,06

Promedios y desviaciones para cada.

```
datos.m <- tapply(datos$Ab, INDEX=bentos$GF, FUN=mean)
datos.de <- tapply(bentos$Ab, INDEX=bentos$GF, FUN=sd)
```

Tabla de medias y desviaciones estándar por cada GFA

```
datos1<- cbind(Bent.m, Bent.de)
```

Dos opciones de figuras de barras con líneas acotadas que representan las desviaciones estándar.

```
par(mfrow = c(2,1), mar = c(3, 5, 2, 1))
barplot(Bent.m, xlab = "GFA", ylab = "Abuandancia de GFA", ylim=c(0,400))
arrows(bp, Bent.m, bp, Bent.m + Bent.de, lwd = 1.5,angle=90,length=0.1)

barplot(datos.m, xlab = "GFA",ylab = "Abundancias (Indv)", col=rainbow(9), ylim=c(0,700))
arrows(bp, datos.m, bp, datos.m + datos.de, lwd = 1.5, angle=90,length=0.1)
box()
par(mfrow = c(1,1))
```

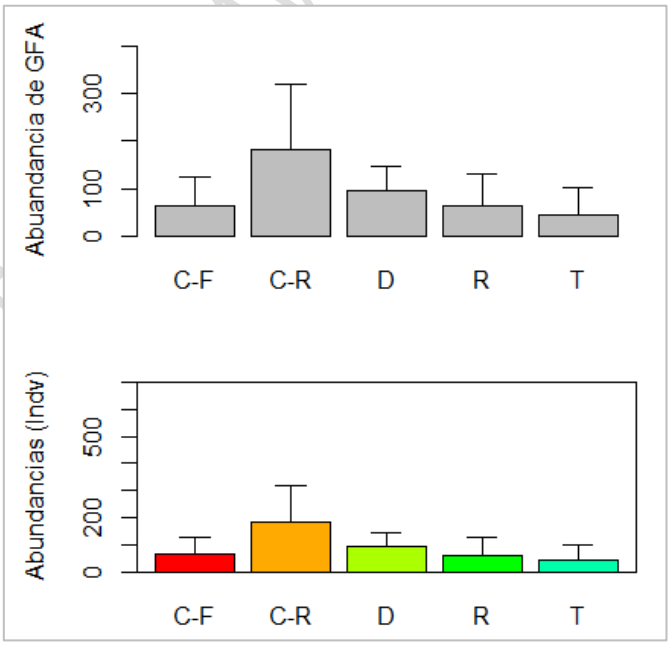


Figura 3. Figura de columnas, en la que se representa a los valores de abundancia de los diferentes grupos funcionales de insectos acuáticos. Las columnas representan los promedios y las líneas acotadas a una desviación estándar por encima de los promedios (Tomado de Rodríguez et al. 2011).

1.9 Gráficos de tiras

Conocidos también como gráficos de comparación de medias. Se representa a los datos crudos en forma de puntos vacíos, la media en puntos sólidos y líneas acotadas que representan a un error estándar encima y debajo de la media, obtenidos al dividir la desviación estándar sobre la raíz cuadrada del tamaño de la muestra. La función "*jitter*" permite distinguir los datos con el mismo valor (impide el solapamiento de puntos en la figura)

***Ejemplo*.** Comparación de medias del archivo "Datos2.csv", con errores estándar, para la abundancia de individuos de los diferentes tipos de grupos funcionales de invertebrados acuáticos del río Gaira de Santa Marta.

Con la base de datos anterior (datos1) y los promedios calculados, a continuación, se calculan los errores estándar (es), que representarán las líneas acotadas. El error estándar corresponde a la desviación estándar sobre la raíz cuadrada del tamaño de la muestra

Cálculo del tamaño de la muestra (tm), que representa el número de muestreos realizados.
```
datos.tm <- tapply(datos$Ab, INDEX=datos$GF, FUN=length)
```

Error estándar (ee) de los datos de abundancias
```
datos.ee <- Bent.de / sqrt(datos.tm)
```

Comando "*random jittering (variación)*" corrige el solapamiento de datos con valores de abundancia, iguales o similares. Puntos negros simbolizan la media por cada GFA
```
Stripchart      (datos$Biom ~ datos$GF, vert = TRUE,  pch=1, method = "jitter", jit = 0.05,
                xlab = "Grupos Funcionales",ylab = "Abundancias (Indv)")
points   (1:5,datos.m, pch = 16, cex = 1.5)
```

Líneas acotadas simbolizan los errores estándar (ee) por encima y debajo de los promedios (m)
```
arrows (1:8, datos.m,1:8, datos.m + datos.de, lwd = 1.5, angle=90, length=0.1)
arrows (1:8, datos.m,1:8, datos.m - datos.de, lwd = 1.5, angle=90, length=0.1)
```

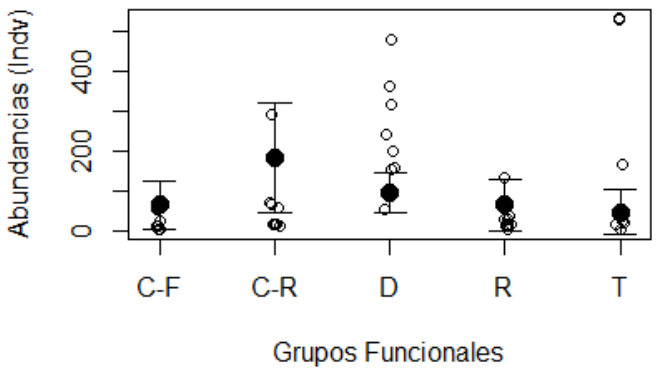


Figura 4. Comparación de medias de los valores de abundancia de los diferentes grupos funcionales de insectos acuáticos. Las líneas acotadas representan a los errores estándar por encima y por debajo de los promedios (puntos negros).

1.10 Figuras de Cajas (Boxplots)

También conocidos como figuras de cajas y bigotes, permiten visualizar estadísticos descriptivos de tendencia central y de dispersión, como la media (cuartil 2) y la dispersión de una o más variables. El punto (o línea) medio de una caja corresponde a la mediana, en ocasiones particulares puede ser la media. Los percentiles 25% y 75% (P25 y P75) definen el final de las cajas (rango intercuatílico), y la diferencia entre estos límites (bisagras) es la propagación. Las líneas acotadas (o bigotes) se generan de cada bisagra a 1,5 veces del valor más extremo de la extensión. Cualquier punto fuera de estos valores se suelen identificar como valores atípicos (*outliers*).

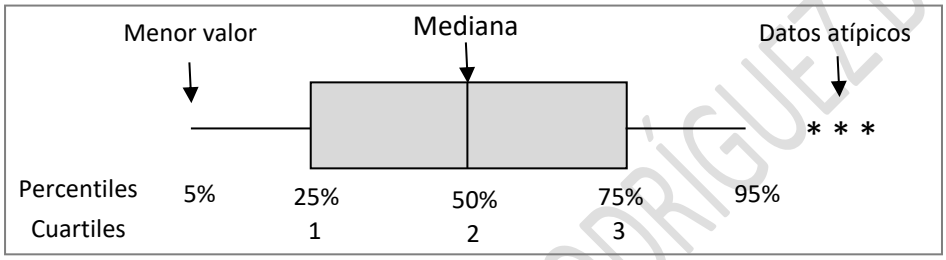


Figura 5. Diagrama de cajas y bigotes, representando la ubicación de los diferentes cuartiles y percentiles.

Los diagramas de cajas suelen ser herramientas útiles cuando se analizan variables continuas o discretas (dependientes) en respuesta a variables categóricas (independientes). Se destacan tres objetivos de estas figuras (1) detectar valores extremos, (2) heterogeneidad en la distribución de las variables (simetría de las cajas) y (3) visualizar el efecto de las variables explicativas (independientes).

Dependiendo del software, diagramas de caja se pueden modificar de varias maneras. Por ejemplo, las muescas (cinturas) se pueden extraer a cada lado de las cajas. Si las muescas de dos cajas no se superponen o solapan, entonces las medias son significativamente diferentes al nivel del 5% (Chambers et al. 1983). Suele ser más útil trazar diagramas de caja en sentido vertical, pues permiten visualizar mejor a las características de los datos originales.

Ejemplo 1. Figuras de cajas y bigotes, comparando a la abundancia (Ab) y biomasa (Biom) de invertebrados acuáticos de la deriva, por categorías de grupos funcionales (GF) y periodos de lluvia y de sequía (P1 y P2).

Inicialmente se llama a la base de datos de invertebrados de la deriva (Datos2.csv)

```
datos<-read.csv2("Datos2.csv")
```

Panel para graficar cuatro figuras de manera simultánea

```
par(mfrow = c(2,2), mar = c(3, 5, 2, 1))
```

Primera figura de cajas, que compara a los valores de abundancia por cada grupo funcional de invertebrados.

```
boxplot(Ab~GF, data = datos, ylab ="Abundancia (Indv)", cex.lab=1.3)
```

Figura de cajas con muescas o cinturas (intervalos de confianza alrededor de la mediana).

```
boxplot(Ab~GF,  notch=T, data = datos, ylab="")
```

Estas muescas permiten hacer comparaciones estadísticas entre los niveles de la variable x.

Figura de cajas para comparar la biomasa de los grupos funcionales por periodos climáticos de lluvia y de sequía (P1 y P2)

```
boxplot(Ab~GF * Lluvia, data = datos, ylab="Biomasa (Biom)", cex.lab=1.3)
```

Otra forma de comparar la biomasa de los grupos funcionales por periodos climáticos de lluvia y de sequía (P1 y P2)

```
boxplot(Ab~GF*Lluvia, names= c("P1/C-F","P2/C-F","P1/C-R","P2/C-R",
      "P1/D","P2/D","P1/R","P2/R","P1/T","P2/T"),data = datos, ylab ="")
par(mfrow = c(1,1))
```

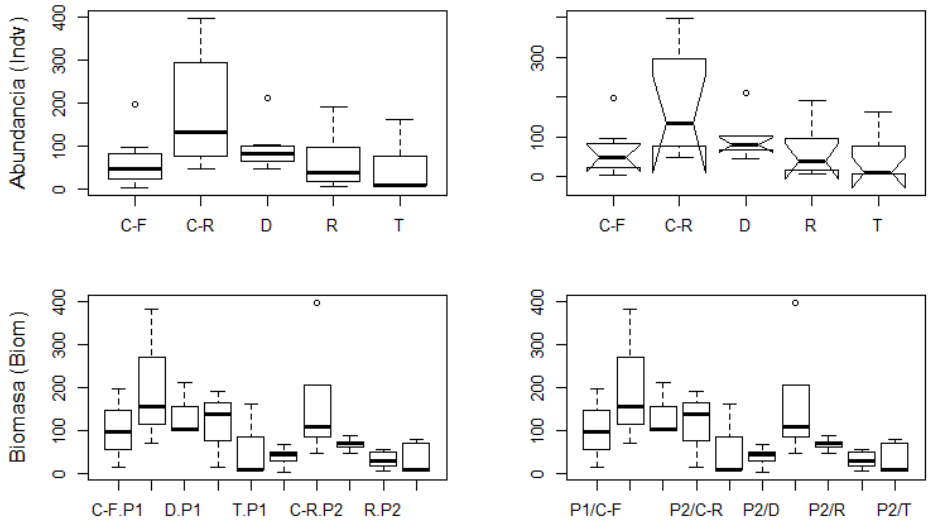


Figura 6. Diagrama de cajas y bigotes, representando el patrón de abundancias y de biomasas por grupos funcionales (C-F, C-R, D, R y T) y por periodos climáticos (P1 y P2).

Ejercicios propuestos (para trabajar en grupos de 3 estudiantes).

1. En la tabla 4, se resumen las variables ambientales que caracterizan a diferentes estaciones (ríos), presentes en 7 cuencas de la Sabana de Bogotá, de acuerdo al estudio realizado por (Ospina et al. 2002).

Tabla 4. Variables fisicoquímicas calculadas en diferentes cuencas de Bogotá.

Est	Cuencas	O2	cond	pH	NH4	Ptot	PO4	DBO5	DQO	sólidos	altitud	temp	conc-O2	sat-O2	Motot	MOP	vel	caudal	var-caudal	chorio
Av2	Av	1,7	32	6,6	0,08	0,46	0	1,3	9,5	0,11	3040	10,2	8	103	2,7	1,6	25	0,64	57	6
Av3	Av	3,7	44	6,8	0,1	0,15	0	1,4	17,5	0,03	2650	12,5	7,6	99	4	2,1	42	1,45	24	7
Fr1	Fr	5,2	79	6,6	0,23	1,34	0,02	2	20	0,14	2560	13,3	5,2	67	25,5	9,1	44	1,03	121	4
Fr2	Fr	1,7	7	4,8	0,05	0,09	0	1,4	16,5	0,07	3290	9,6	7,8	101	3,4	2,3	19	0,22	116	7
Fr3	Fr	3,2	21	6,1	0,18	0,61	0,05	1,3	15	0,17	3010	11,5	7,9	104	70,3	8,4	31	2,01	154	8
Ne1	Ne	4,9	215	6,8	0,28	1,07	0,05	1,8	18,5	0,27	2820	13,2	7,8	103	136,7	17	38	1,2	130	6
Ne2	Ne	3,4	6	6,1	0	0,2	0,02	1,3	14,5	0,04	3300	11,5	7,6	106	2,4	2,1	18	0,3	96	5
Ne3	Ne	2,5	23	7	0,06	0,13	0,02	1,4	12	0,03	2820	13,8	7,6	103	45,1	5	23	5,12	135	8
Si1	Si	5,9	37	6,9	0,3	0,43	0,02	3,4	12,5	0,17	2670	14,8	7,3	99	49,4	11,9	28	5,74	129	4
Si2	Si	8,5	176	7	0,05	0,31	0,06	1,6	16,5	0,1	2700	11,5	8,4	105	6	1,3	41	0,16	26	6
Si3	Si	5,3	134	6,5	0,25	0,95	0,02	2,3	22,5	1,37	2620	14,2	6,9	91	14,4	3,2	32	0,08	28	2
Su1	Su	5,9	246	6,9	2,18	0,82	0,06	3,1	32	0,6	2580	15,7	5,2	71	12,3	5,7	23	0,07	76	4
Su2	Su	5,5	48	6,8	0,18	0,14	0,01	2,3	14	0,04	2940	14,5	7,2	99	10,5	6,6	57	1,35	51	5
Su3	Su	1,7	50	7,3	0,16	0,41	0,01	2,1	10,5	0,68	2660	14,9	7,5	102	31,6	11,4	72	1,37	82	6
Teu1	Teu	4,4	53	7	0,08	0,23	0,01	1,6	14,5	0,08	2590	16,9	7,7	107	8,1	2,8	55	1,26	134	5
Teu2	Teu	0,5	15	6,5	0,03	0,38	0,02	1,4	8,5	0	2830	10,5	7,9	100	2,5	1,6	22	0,22	63	6
Teu3	Teu	2,1	22	6,7	0,1	0,31	0,01	1,7	34	0,04	2660	12,9	7,8	100	11,5	4,1	34	0,39	82	5
Tu1	Tu	3,3	28	6,8	0,18	0,67	0	2,3	40,5	0,17	2590	15,8	7,6	103	33,7	6,2	33	3,14	157	7
Tu2	Tu	1,8	18	6,8	0,03	0,2	0,01	1,3	6,5	0,01	2640	14	7,7	101	11,4	2,5	43	0,68	93	6
Tu3	Tu	4,4	21	6,8	0,08	0,28	0	1,1	9	0,04	2570	15,9	7,3	100	11,8	3,5	47	0,95	131	6

Con base a la información anterior, de sebe realizar un ejercicio de exploración gráfica, que incluya a los siguientes elementos:

a) Diseño de objetivos, elaborar una pregunta de investigación y redactar la hipótesis del diseño estadístico propuesto, que puedan ser visualizadas en el análisis gráfico. El enfoque de estos ítems, debe incluir la variación de algunas variables ambientales, en las cuencas y quebradas (Est.).

b) Análisis de algunas graficas exploratorias (en Excel y en R) que permitan interpretar en detalle los resultados (mínimo 5 figuras). Debajo de cada figura, se debe anotar su leyenda ya continuación, en un párrafo resumido analizar el patrón observado.

Incluir las figuras en los siguientes recuadros, debajo anotar su leyenda ya continuación, en un párrafo resumido analizar el patrón observado acorde con los objetivos propuestos.

2. Realizar un análisis exploratorio gráfico (Excel y en R) de la base de datos de iris (lirios), incluyendo por lo menos a cinco figuras exploratorias y una adicional que se obtengas de las ayudas que ofrece el R y que fueron incluidas al inicio de este capítulo.
   La base de datos de "iris" se obtiene de la siguiente línea de comandos:
   ```
   library(vegan)
   data(iris)
   ```
   Investigar cómo se puede guardar a la base de datos de "iris" como archivo de Excel "csv", a partir de este, se realiza el análisis correspondiente.

   Basado en información de la web, se reuiere iniciar el análisis con una contextualización de la base en análisis, a partir de este, se definirá una pregunta e hipótesis a probar y que debe relacionarse a las figuras a realizar.

   Realizar las figuras exploratorias en Excel y en R, de acuerdo a las pautas indicadas y analizar los resultados.

El plazo para enviar el taller, que incluye informe en Word, Archivos de Excel y el Script de RStudio realizado, será el miércoles 26 de marzo.