1. Figuras exploratorias multivariadas

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?

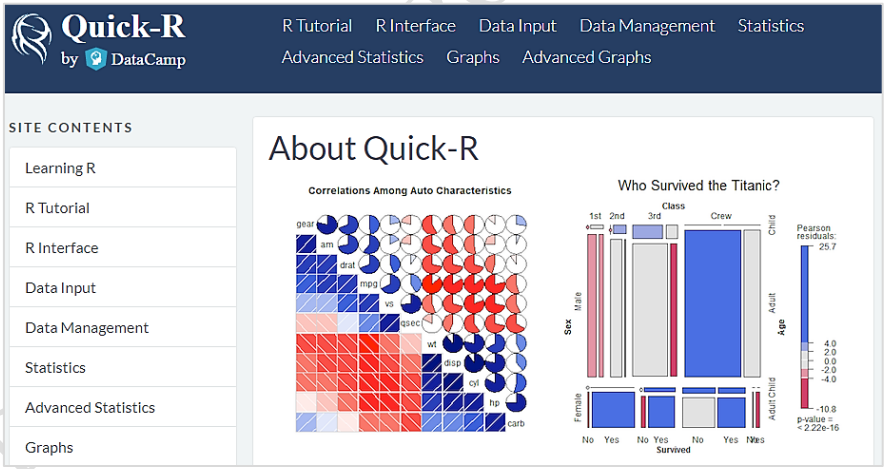
MANUAL ANDEA JAVIER RODRÍGUEZ B.

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.

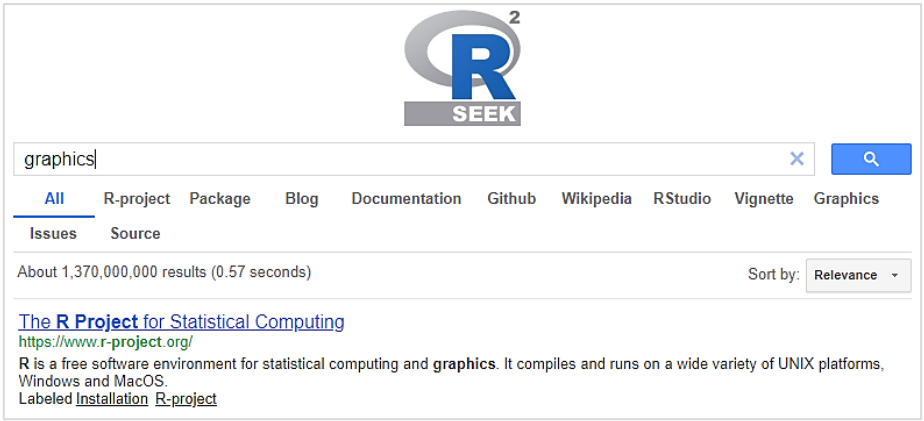
1. Utilidad de R en la exploración de datos.

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.

1.1 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/



1.2 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/



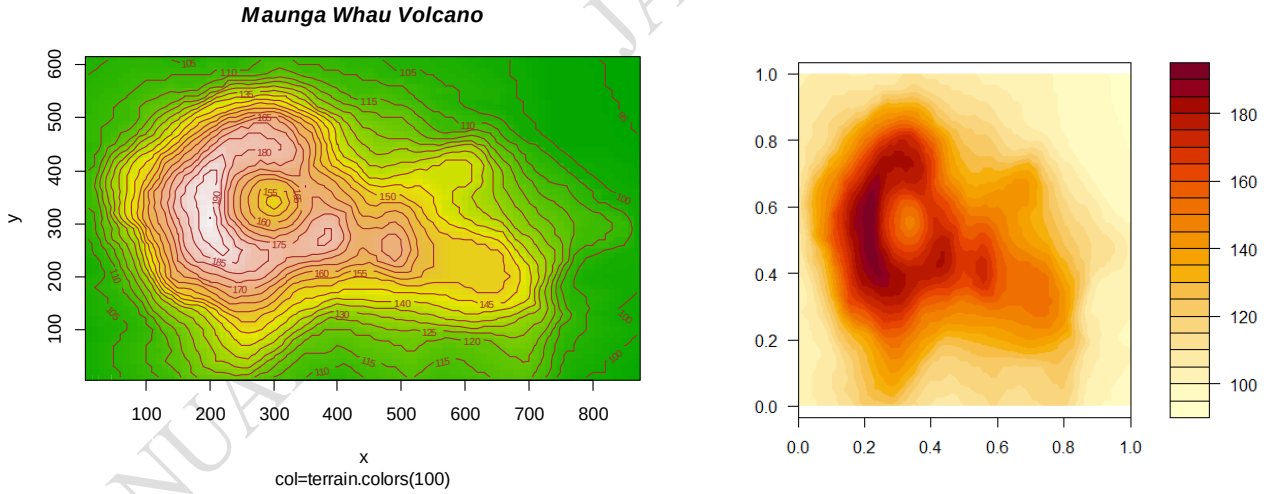
MANUAL ANDEA - J. RODRÍGUEZ B.

1.3 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/

1.4 Opciones gráficas

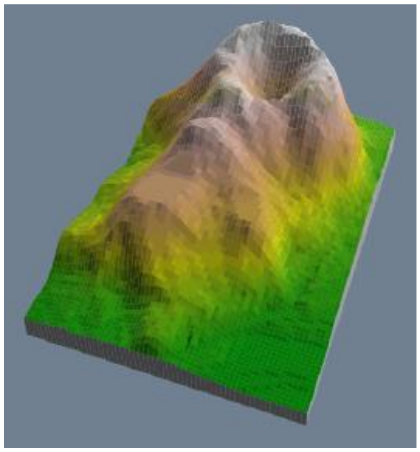
El siguiente demo se ejecuta en R, permite visualizar opciones gráficas
demo(image)

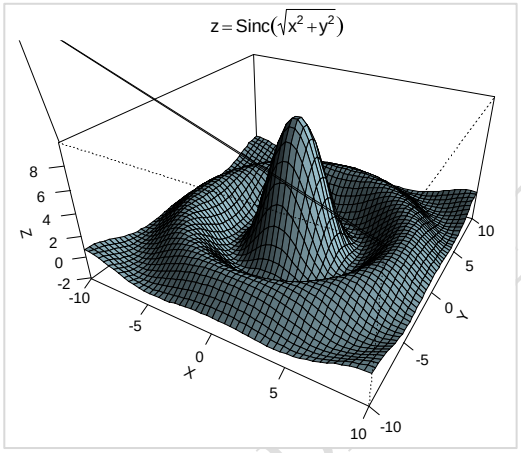


MANUAL JAVIER RODRÍGUEZ B.

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)

MANUAL ANDEA - JAVIER RODRÍGUEZ B.

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.

2. Figuras exploratorias

2.1 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

MANUAL ANDEA - JAVIER RODRÍGUEZ B.

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]))
```

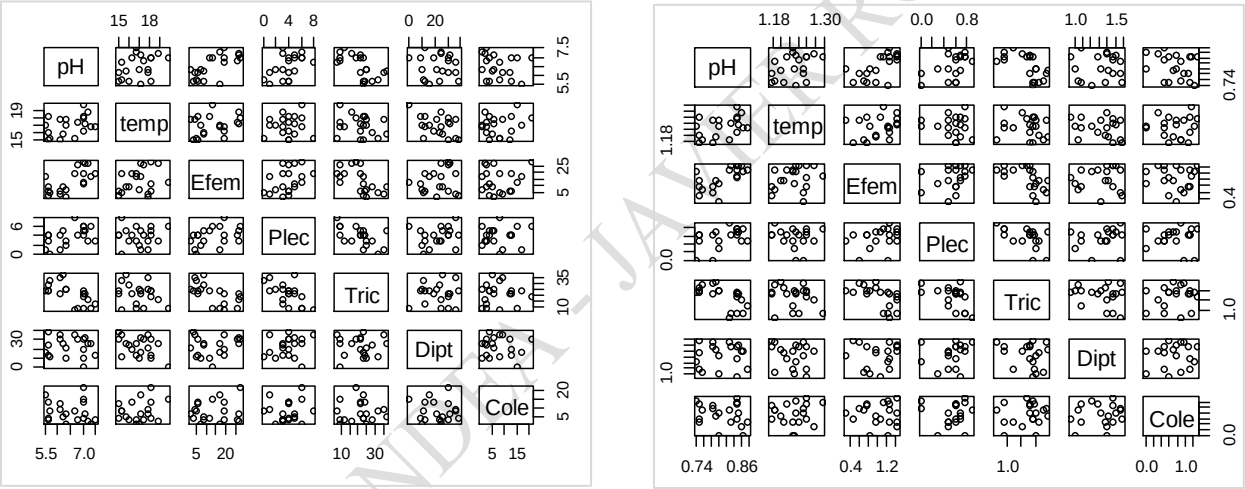


Figura 7. Graficas de pares, con dispersión de los datos originales (izquierda) y con transformación logarítmica base 10 (derecho).

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]))
```

MANUAL ANDEA - JAVIER RODRÍGUEZ B.

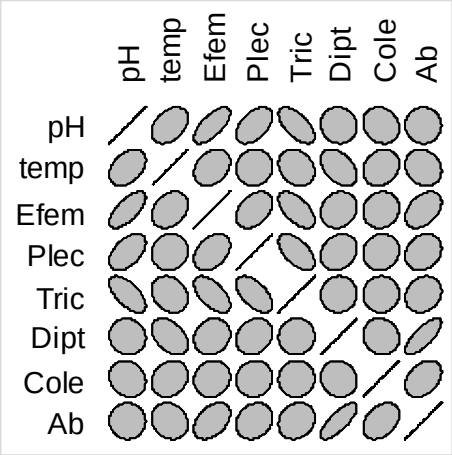


Figura 8. Graficas de elipses, que relacionan a parejas de variables.

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)})
```

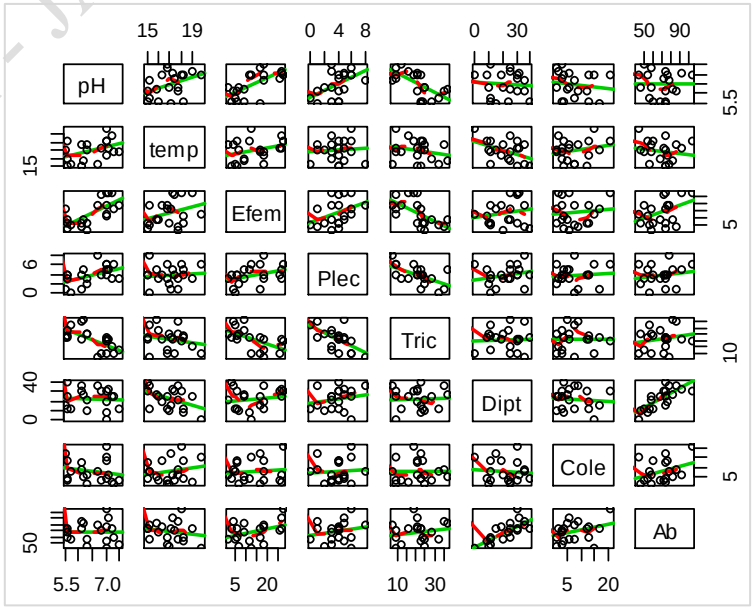


Figura 9. Graficas de pares, con líneas de ajuste lineal (líneas verdes) y no lineal o suavizada (líneas rojas). Los puntos corresponden a los valores de las variables en las quebradas.

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

MANUAL ANDEA - JAVIER RODRÍGUEZ B.

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)
```

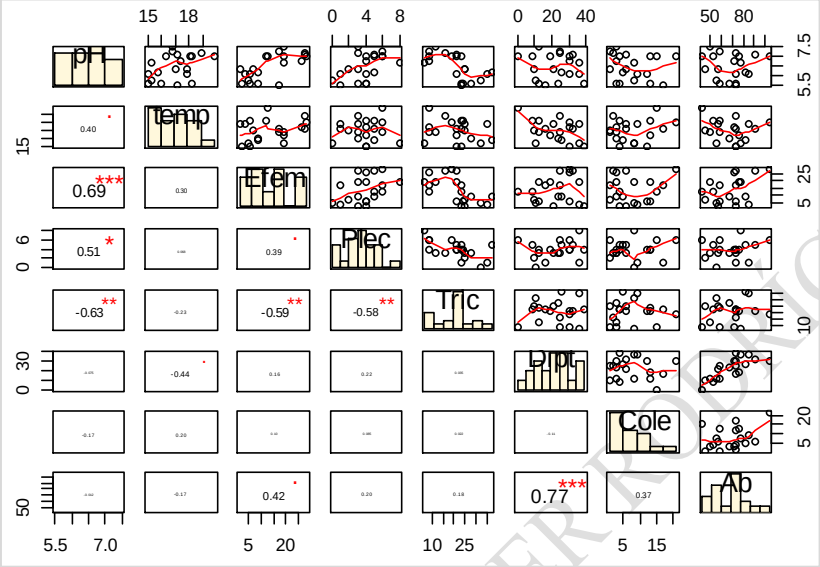


Figura 10. Graficas de pares, con líneas de ajuste no lineal o suavizado (líneas rojas). Los puntos corresponden a los valores de las variables en las quebradas. Los valores representan los coeficientes de correlación y los asteriscos, al nivel de significancia de las relaciones.

2.2 Figuras Coplot

2.3 Histogramas

2.4 Figuras quantil-quantil (QQ-plots)

2.5 Diagramas de dispersión (plot y xyplot)

2.6 Figuras circulares (Pie Chart)

2.7 Graficas de columnas o barras

2.8 Graficas de columnas o barras con desviaciones estándar

2.9 Gráficos de tiras

2.10 Figuras de Cajas (Boxplots)

Cuestionario propuesto

Nota: Se requiere contar con la versión completa del libro Análisis de Datos Ambientales y Ecológicos - ANDEA

MANUAL ANDEA - JAVIER RODRIGUEZ B.