JAVIER RODRÍGUEZ-BARRIOS

ANÁLISIS DE DATOS ECOLÓGICOS Y AMBIENTALES

Aplicaciones con el programa R



DIAZ DE SANTOS

ANÁLISIS DE DATOS ECOLÓGICOS Y AMBIENTALES

APLICACIONES CON PROGRAMA R

JAVIER RODRÍGUEZ-BARRIOS

ANÁLISIS DE DATOS ECOLÓGICOS Y AMBIENTALES

APLICACIONES CON PROGRAMA R

DIAZ DE SANTOS
EDICIONES





El lector podrá descargar el material complementario de este libro en la siguiente dirección:
https://www.editdiazdesantos.com/libros/9788490524817

© Javier Rodríguez Barrios, 2023

Reservados todos los derechos.

«No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático,
ni la transmisión de ninguna forma o por cualquier medio,
ya sea electrónico, mecánico por fotocopia, por registro u otros métodos,
sin el permiso previo y por escrito de los titulares del Copyright.»

Ediciones Díaz de Santos
Internet: http//www.editdiazdesantos.com
E-mail: ediciones@editdiazdesantos.com

e-ISBN: 978-84-9052-485-5
Depósito Legal: M-26295-2022

Fotocomposición y diseño de cubiertas: P55 Servicios Culturales

Prefacio

Este libro surge como una iniciativa que tuve hace varios años, debido al interés para que mis estudiantes de pregrado y de posgrado valoraran la importancia del análisis de datos para su desempeño profesional, con la aplicación de lenguajes que en su momento eran poco convencionales para nuestra profesión, como lo es el R y el RStudio. Desde ese momento la experiencia ha sido muy gratificante, debido a que nos hemos podido adaptar a las nuevas exigencias, para poder mejorar la comprensión de la complejidad en los procesos ecológicos y ambientales, aplicando estrategias estadísticas y computacionales que en la actualidad ocupan un lugar muy importante para las ciencias básicas y aplicadas.

Los ejemplos de este documento se soportan en su mayoría por información obtenida a partir de experiencias investigativas que he tenido y las de algunos colegas con los que he interactuado, al igual que por experiencias enseñando el análisis de datos a nivel de pregrado y de posgrado. El libro se orienta a un público que inicia en el manejo de análisis ecológicos y ambientales con el uso de los programas R y RStudio. Para un abordaje más detallado sobre los temas descritos, he dejado algunas referencias en los capítulos, que permitirán profundizar en el entendimiento de cada procedimiento y técnica utilizada. Los códigos utilizados son generados por R Markdown en RStudio, para permitir que sean más amigables y didácticos.

Considero importante que las personas que inician en este proceso reconozcan que son muchas las ventajas que podrán obtener, iniciando por el enfoque filosófico que ofrece el lenguaje de comandos en R, dado que se apoya en un aprendizaje basado en cuestionamientos sobre la manera en la que se pueden tomar las mejores decisiones para que el análisis se ajuste a la realidad de los datos, soportados por preguntas de análisis, objetivos e hipótesis estadísticas. Se requiere de mucha concentración para minimizar las posibilidades de cometer errores en los procedimientos, con lo cual se podrá profundizar en el entendimiento de cada técnica, comparado a lo que se aprendería con programas que no se basan en lenguaje de programación.

Sobre el autor

Javier Rodríguez Barrios es profesor de Ecología, de Estadística Multivariada y de Análisis de Datos Biológicos e investigador en Ecología de la Universidad del Magdalena en Santa Marta, Colombia. Graduado como Biólogo en esa universidad y de magister y doctor en la Universidad Nacional de Colombia, sede Bogotá. Sus intereses de investigación se orientan a la ecología acuática y la estadística aplicada a las ciencias biológicas en general.

Índice

Prefacio ..... **V**
Sobre el autor ..... **VII**
1. Conceptos básicos **1**
Introducción 1
1.1. Etapas del ANDEA 2
1.2. Requisitos. 4
1.3. Objetivos 5
1.4. Tipos de datos. 5
1.5. Generalidades del análisis de datos. 6
Ejercicios propuestos 7
2. Generalidades de R y RStudio **9**
Introducción 9
2.1. Programa R y sus ventajas en el análisis de datos 9
2.2. Programa RStudio 10
2.3. Tutoriales de R Y RStudio 10
2.3.1. Tutorial rápido de R 10
2.3.2. Motor de búsqueda de R (Rseek). 10
2.3.3. Otros sitios de utilidad 11
2.3.4. Opciones gráficas 12
Ejemplo 1. NOCIONES DE RSTUDIO 14
Ejemplo 2. DATOS CON VARIABLES ALEATORIAS 29
Ejercicios propuestos 31
3. Conceptos básicos de matrices en R **33**
Introducción 33
3.1. Resumen estadístico de datos con múltiples variables 34
3.1.1. Medias 34
3.1.2. Varianzas 35
3.1.3. Covarianzas 35
3.1.4. Resumen estadístico 36
3.2. Correlaciones 36
3.3. Distancias 36
3.4. Valores y vectores propios 38
Ejemplo 1. APLICACIONES DE MATRICES EN RSTUDIO 49
Ejemplo 2. APLICACIONES DE MATRICES EN RSTUDIO 44
Ejemplo 3. APLICACIONES DE MATRICES EN RSTUDIO 52
Ejercicios propuestos 60
4. Figuras exploratorias multivariadas **63**
Introducción 63
4. Utilidad de R en la exploración de datos 63
4.2. Figuras exploratorias 64
Ejemplo 1. GRÁFICAS EXPLORATORIAS 65

4.1. Gráficas de pares (pairplot)..... 66
4.2. Gráfica de elipses..... 69
4.3. Figuras Coplot..... 70
4.4. Splom para variables categorizadas..... 72
4.5. xyplot para variables continuas y factores..... 73
4.6. Histogramas de frecuencia..... 81
4.7. Histogramas de densidad..... 75
4.8. Figuras quantil-quantil (QQ-plots)..... 76
4.9. Diagramas de dispersión (plot y xyplot)..... 77
4.10. Figuras de Cajas (Boxplots)..... 79
4.11. Figuras circulares (Pie Chart)..... 81
4.12. Gráficas de columnas o barras, con desviaciones estándar..... 82
Ejemplo 2. GRÁFICAS EXPLORATORIAS MULTIFACTORIALES..... 84
4.13. Gráficas de columnas o barras..... 84
4.14. Gráficos de tiras..... 86
Ejercicios propuestos..... 88
5. Transformaciones y Estandarizaciones..... 89
Introducción..... 89
Ejemplo 1. TRANSFORMACIONES GENERALES..... 90
5.1. Criterios generales en las transformaciones y las estandarizaciones..... 91
5.1.1. Transformaciones monotónicas o "simples"..... 92
5.1.2. Suavizamiento de Beals (beals smoothing)..... 93
5.1.3. Relativizaciones o estandarizaciones..... 93
5.1.4. Paquetes de R que integran transformaciones..... 94
Ejemplo 2. TRANSFORMACIONES Y ESTANDARIZACIONES..... 96
5.2. Regla de abultamiento de Mosteller y Tukey (1977)..... 101
Ejemplo 3. REGLA DE ABULTAMIENTO..... 102
5.3. Ley de potencias de Taylor (Taylor, 1961)..... 105
Ejemplo 4. LEY DE POTENCIAS DE TAYLOR..... 107
5.4. Transformación poder de Box-Cox (1964)..... 111
Ejemplo 5. LEY DE PODER DE BOX-COX..... 112
Ejemplo 7. LEY DE PODER DE BOX-COX..... 115
Ejercicios propuestos..... 117
6. Análisis de Ordenación..... 119
Introducción..... 119
6.1. Análisis de componentes principales (PCA o PCA)..... 119
6.1.1. Elementos principales de un PCA..... 120
Ejemplo 1. PCA CON VARIABLES AMBIENTALES..... 121
Ejemplo 2. PCA CON VARIABLES AMBIENTALES Y BIOLÓGICAS..... 128
Ejercicios propuestos..... 137
6.2. Análisis de Factores (AF)..... 138
Ejemplo 3. AF CON VARIABLES AMBIENTALES..... 139
Ejercicios propuestos..... 143
6.3. Análisis de escalamiento multidimensional (MDS y NMDS)..... 144
6.3.1. Escalamiento multidimensional métrico (MDS)..... 144

6.3.2. Escalamiento multidimensional no métrico (NMDS) ... 144
Ejemplo 4. MDS o PCoA CON VARIABLES BIOLÓGICAS ... 145
Ejemplo 5. NMDS CON DATOS DE PRESENCIA-AUSENCIA ... 147
Ejercicios propuestos ... 151
6.4. Análisis de correspondencia simple (CA) ... 153
6.5. Análisis de correspondencia dirigido (DCA) ... 153
6.6. Análisis de correspondencia múltiple (MCA) ... 155
6.7. El análisis factorial de datos mixtos (FAMD) ... 155
6.8. Análisis factorial múltiple (MFA) ... 155
Ejemplo 6. CA CON VARIABLES BIOLÓGICAS Y AMBIENTALES ... 156
Ejercicios propuestos ... 166
Ejemplo 7. MCA CON DATOS DE PECES ... 167
6.9. Análisis canónicos sin restricciones ... 178
6.9.1. Análisis de redundancia (RDA) ... 178
6.9.2. Análisis de Correspondencia Canónica (CCA) ... 178
6.9.3. Análisis de correlación canónica (CCoA) ... 179
EJEMPLO 8. RDA CON MICROALGAS ... 181
EJEMPLO 9. ACC CON MICROALGAS ... 194
Ejercicios propuestos ... 199
7. Análisis de Clasificación ... 201
Introducción ... 201
7.1. Medidas de asociación (modo Q y modo R). ... 201
7.1.1. Coeficientes de similitud “modo Q”. ... 202
7.1.2. Coeficientes de distancia “modo Q”. ... 203
7.1.3. Coeficientes de dependencia “Modo R”. ... 205
7.1.4. Escogencia del coeficiente de asociación o similitud. ... 205
Ejemplo 1. ANÁLISIS PARA DATOS DE PRESENCIA - AUSENCIA ... 207
Ejemplo 2. ANÁLISIS PARA DATOS DE ABUNDANCIA ... 218
Ejemplo 3. ANÁLISIS PARA DATOS AMBIENTALES ... 214
Ejercicios propuestos ... 216
7.2. Análisis de clúster jerárquico (CLA) ... 217
7.2.1. Tipología de los CLA. ... 217
(a) Técnicas exclusivas *vs.* No exclusivas ... 217
(b) Técnicas secuenciales *vs.* Simultáneas ... 217
(c) Jerárquicos *vs.* No jerárquicos ... 217
(d) Aglomerativos *vs.* Divisivos ... 218
(e) Politéticos *vs.* Monotéticos ... 218
(f) Probabilísticos *vs.* No probabilísticos ... 218
7.2.2. Selección del coeficiente de similitud o de disimilitud (distancias). ... 219
7.2.3. Métodos de agrupación jerárquica. ... 219
(1) Unión simple ... 219
(2) Unión completa ... 219
(3) Promedio aritmético no ponderado (UPGMA) ... 219
(4) Promedio aritmético ponderado (WPGMA) ... 219
(5) Centroide no ponderado (UPGMC) ... 219

(6) Centroide ponderado (WPGMC) 220
(7) Mínima varianza de Ward 220
7.2.4. Evaluación del mejor método de agrupación jerárquica 220
7.2.5. Clústeres no jerárquicos 220
(1) Partición por K-Medias 221
(2) Clúster de enlace completo 221
(3) Clústeres Probabilísticos 221
7.2.6. Clústeres basados en modelos 221
Ejemplo 4. CLÚSTER PARA DATOS DE ABUNDANCIA 222
Ejemplo 5. CLÚSTER PARA DATOS DE PRESENCIA - AUSENCIA 239
Ejercicios propuestos 247
7.3. Análisis Discriminante Lineal (LDA) 249
Ejemplo 6. ANÁLISIS DISCRIMINANTE LINEAL CON VARIABLES MORFOMÉTRICAS 250
7.4. Prueba T^2 de Hotelling (1931) 262
Ejemplo 7. T^2 DE HOTELLING CON VARIABLES MORFOMÉTRICAS 263
7.5. Análisis de Varianza Multivariado (MANOVA) 270
7.5.1. Matrices SSCP 270
7.5.2. Estadísticos de la MANOVA 270
7.5.3. Supuestos de la MANOVA. 270
Ejemplo 8. MANOVA CON VARABLES MORFOMÉTRICAS 271
Ejercicios propuestos 275
7.6. Cuándo utilizar cada técnica de clasificación 276
8. Diseños multivariados no paramétricos 277
Introducción 277
8.1. Análisis de disimilitud (MANTEL) 277
8.2. Variables ambientales con máxima correlación (BIOENV) 278
8.3. Análisis de permutación multirrespuesta (MRPP) 279
8.4. Análisis de similitud (ANOSIM) 280
8.5. Análisis de Varianza Multivariante Permutacional - PERMANOVA 281
8.6. Especies indicadoras – Esp.Ind 281
8.6.1. Utilidad del método. 282
8.6.2. Paquete Indicspecies 282
Ejemplo 1. MANTEL Y BIOENV CON DATOS DE MICROALGAS 283
Ejemplo 2. PERMUTACIONES CON VARIABLES MORFOMÉTRICAS 290
Ejemplo 3. ESPECIES INDICADORAS CON DATOS DE MICROALGAS 298
Ejercicios propuestos 301
Bibliografía. 303

Figuras

Figura 1. Imagen de la plataforma gráfica con comandos de R 10
Figura 2. Imagen de motor de búsqueda de utilidades de R 11
Figura 3. Representaciones de curvas de nivel en mapas realizados con R. 12
Figura 4. Representaciones tridimensionales en R 12
Figura 5. Imagen de motor de búsqueda de utilidades de R 64
Figura 6. Regla de Mosteller y Tukey. Las flechas indican el sentido en que debe realizar la transformación, en el caso de que la figura de dispersión apunte a uno de sus sentidos. . 101
Figura 7. (a) La dispersión original de los datos muestra un abultamiento hacia el cuadrante IV, que exige transformaciones de raíces en la variable *y* o potencias cuadradas y cúbicas en la variable *x*. (b) Variables linealizadas posterior a la transformación 101
Figura 8. Procedimiento resumido para hallar las componentes principales a partir de una tabla de datos de campo o matriz de datos crudos (tomado de Herrándo, 2005 – com. Pers.). 120
Figura 9. En el AF se construyen factores que presentan alta relación con las variables medidas (Xi), pero ninguna relación entre ellos. Además, se observa que cada variable presenta su factor único (residual). Tomado de Legendre & Legendre (1998). 138
Figura 10. (a) Ordenación normal CA de las observaciones en el que la distribución de los datos es unimodal invertida (efecto de herradura o de arco), (b) segmentación del eje 1 en cuatro partes, realizada en el DCA. Los escores de las observaciones son centrados en cero para cada segmento para visualizar un gradiente más claro a lo largo de este gradiente. Las flechas y las líneas punteadas representan al movimiento de cada segmento para reescalar las observaciones. Modificado de McCune *et al.* (2002). 154
Figura 11. Tabla de contingencia (frecuencias), mostrando los descriptores (presencia/ausencia = 0/1), usados para comparar a dos sitios (Objetos X_1 y X_2). Donde *a* es el número de taxones presentes en los dos objetos (1), *d* son los descriptores ausentes (0), b y c representan la presencia para cada objeto. *d* es utilizado solo en coeficientes simétricos (p= a + b + c + d). *p* es el número total de taxones. 202
Figura 12. La distancia entre los objetos (ej. sitios) x_1 y x_2, se calcula como la hipotenusa de un triángulo rectángulo. 203
Figura 13. Escogencia de una medida de asociación entre objetos (modo Q), usando bases de datos con especies y no especies (descriptores físicos, químicos, geológicos, etc.). Algunos coeficientes (S_i) y distancias (D_i) son detallados en el texto anterior y todos han sido tomados de Legendre y Legendre (1998). 206
Figura 14. Escogencia de una medida de asociación entre variables o descriptores (modo R), usando bases de datos con especies y no especies (descriptores físicos, químicos, geológicos, etc.). Algunos coeficientes (S_i) y distancias (D_i) son detallados en el texto anterior y todos han sido tomados de Legendre y Legendre (1998). 206
Figura 15. Clasificación de las técnicas de agrupación basadas en 5 propiedades. 218

Tablas

Tabla 1. Términos comunes en estadística, referidos a una investigación sobre el contenido de amoniaco en las excretas de serpientes de cascabel. 2

Tabla 2. Valores hipotéticos para datos multivariados. 6

Tabla 3. Datos hipotéticos de abundancia de órdenes de insectos en diferentes localidades ubicadas en dos tipos de ecosistemas. 34

Tabla 4. Transformaciones generales de variables. 91

Tabla 5. Transformaciones propuestas por McCune y Grace (2002) y comandos en R para ejecutarlas. Donde x_{ij} es la variable u observación sin transformar y b_{ij} es la variable u observación transformada. 95

Tabla 6. Aplicación de los métodos de ordenación comparación indirecta más utilizados 119

Tabla 7. Resumen de las diferencias entre las técnicas CCA y el RDA 179

1. Conceptos básicos

INTRODUCCIÓN

Una de las interrogantes iniciales conlleva a preguntar: ¿En qué consiste el análisis de datos ecológicos y ambientales (ANDEA)? En este sentido, el ANDEA suele aplicarse en áreas como la estadística, como una herramienta asociada a la recolección, análisis, representación e interpretación de datos y, por lo tanto, es fundamental para la mayoría de las actividades científicas en esta área del conocimiento. Para el profesional de las ciencias biológicas y ambientales, se convierte en una herramienta fundamental, debido a que ofrece elementos descriptivos e inferenciales de importancia en el diseño de experimentos y en la prueba de hipótesis, para cumplir con los requerimientos del método científico.

La jerarquía más elemental del análisis de datos se divide en poblaciones y muestras estadísticas, así como de variables. Las poblaciones estadísticas son referidas a todas las observaciones posibles, de las que puede extraerse una muestra o subconjunto sobre la cual se puede llegar a conclusiones. Estas poblaciones pueden representar unidades ecológicas naturales, como la población de serpientes de cascabel de los cerros en Santa Marta, aunque pueden dividirse en subunidades más artificiales si se limita a la población de hembras de estas serpientes.

Los parámetros son atributos o características sobre los cuales se pueden sacar conclusiones de una población. La generalidad es que no sea posible hacer inferencias sobre la población, por lo que los parámetros son estimados a partir de estadísticos que caracterizan partes de una población. Estos parámetros están relacionados con las muestras (ej. la media o varianza muestral), siempre y cuando estas representen adecuadamente a una población (tener un tamaño mínimo y representativo de la población). Por esta razón, es frecuente realizar diseños experimentales que asuman que una muestra haya sido obtenida de la población de forma aleatoria o por el azar (esto maximiza la posibilidad de que la muestra represente a la población), pero también hay casos en los que el interés se orienta a que la muestra sea sistemática, en la que su configuración sea definida por el investigador.

El objetivo del ANDEA se centrará en poder descifrar *patrones* o *señales* que brinden información ofrecida por una o múltiples variables, observaciones y en algunos casos su integración con factores que representan a los datos. Adicionalmente, en la actualidad existe una gran disponibilidad de técnicas y de programas estadísticos robustos que facilitan su aplicación e interpretación.

Otra de las interrogantes se describe así: ¿Por qué es necesario el análisis de datos en las ciencias naturales o ambientales? Y para dar respuesta, se debe ponderar el nivel de complejidad de la naturaleza, en donde sus patrones fluctúan considerablemente en espacio y en tiempo. Esto se puede corroborar al evaluar el comportamiento de los organismos y su respuesta a múltiples variables bióticas y abióticas, que pueden actuar de forma simultánea.

Uno de los aspectos a tener en cuenta en la pregunta anterior es la correlación que pueda existir entre las variables colectadas. Consiste en descifrar la información que subyace a la interacción de estos parámetros; algunos pueden descartarse dado a que generan ruido (variables colineales o

distorsionadas), aportan la misma información que otras (redundantes), por lo que existen procedimientos que permitirán descartarlas y dejar solo aquellas que muestren estructuras importantes en la correlación de los datos. Esta dinámica hace parte importante del "principio de la parsimonia" orientado a cómo lograr explicar patrones o estructuras de los datos, con el menor número posible de variables, resumiendo o simplificando la mayor cantidad de información expresada en las variables.

De acuerdo con lo anterior, se suele utilizar técnicas univariadas o multivariadas, que buscan identificar estructuras de los datos, basadas en un conjunto reducido de dimensiones. Por lo general se usan dos ejes construidos de forma matemática o computacional, en donde además se pueden visualizar las variables y/o las observaciones, orientado principalmente a la exploración de hipótesis construidas previamente. Si el objeto del análisis es hacia la prueba de hipótesis, se aplican otras técnicas, que permitan evaluar diferentes variables, conservando el nivel de significancia (alfa).

Tabla 1. Términos comunes en estadística, referidos a una investigación sobre el contenido de amoniaco en las excretas de serpientes de cascabel.

Término	Definición	Ejemplo
Medición	Dato único cuya información refleja una característica de interés (ej. longitud de un pez, número de individuos en un cuadrante, etc.)	Contenido de amoniaco en la excreta de una serpiente hembra.
Observación	Medida única o unidad experimental (ej. un individuo, un cuadrante, un sitio, un transepto, una visita de muestreo, etc.)	mg/L de amoniaco en una serpiente hembra.
Población	El total de observaciones posibles que pueden ser medidas de la unidad en la cual se busca sacar conclusiones.	Contenido de amoniaco presente en las excretas de todas las serpientes hembra de un lugar.
Muestra	Un subgrupo representativo de la población a evaluar.	mg/L de amoniaco en 20 serpientes hembra de un lugar.
Variable	Conjunto de mediciones del mismo tipo que componen la muestra. Las características medidas difieren (varían) de una observación a otra.	Contenido de amoniaco de cada serpiente hembra.
Factor	Conjunto de observaciones tomadas en una población y que son clasificadas por algún atributo particular (gradientes de profundidad o altura, épocas climáticas, etc.).	Amoniaco en excreta de serpientes en un gradiente de altura.

1.1. Etapas del ANDEA

El análisis de datos en el que interactúan diferentes variables suele realizarse en dos etapas generales, partiendo de las variables y de las observaciones (Tablas 1 y 2), para intentar identificar patrones generales y que puedan ser explicados desde el contexto en estudio, para poder validar hipótesis

propuestas previamente. De igual forma se busca descifrar patrones ocultos a la complejidad de bases de datos que pueden presentar muy pocas hasta cientos de variables, para lo cual el trabajo con el ANDEA suele presentar una potencia importante en su entorno gráfico.

En este sentido, el ANDEA suele dividirse en dos grandes grupos de técnicas: (1) las ***descriptivas*** que exploran la estructura de los datos, definidos por variables, observaciones y en algunos casos incluyen y/o construyen factores. Estas técnicas buscan identificar variables que presenten algún tipo de correlación o de combinación preferiblemente lineal para que facilite la generación de modelos con una buena base matemática (ver Capítulos 6 y 7) y la inferencia estadística que corresponde al segundo grupo de técnicas que se describen posteriormente.

Es importante aclarar que las técnicas descriptivas no necesariamente corresponden al ***análisis exploratorio***, el cual debería hacer parte inicial y de rutina para cualquier análisis de datos y está constituido principalmente por figuras univariadas y/o multivariadas, que permiten visualizar patrones generales en el comportamiento de los datos (ver Capítulo 4). Este componente suele ser subvalorado en muchos análisis, sin tener en cuenta que en algunos casos basta con un buen análisis exploratorio para identificar los patrones de nuestros datos.

(2) La ***inferencia multivariada*** es desarrollada por técnicas orientadas a la prueba de hipótesis y/o a la generación de modelos que pueden ser construidos a través de ecuaciones matemáticas y que exigen el cumplimiento de diferentes requisitos o supuestos numéricos (ver Secciones 7.4 y 7.5). Estas técnicas permiten a su vez elegir a las variables de mayor importancia en la explicación de un experimento realizado, independiente del número inicial que se analice de manera simultánea, debido a que son reguladas por un valor alfa, que corresponde al nivel de significancia, establecido previamente por el investigador (normalmente un alfa igual o menor de 0,05). Estas técnicas aportan control sobre la tasa de error de los diseños estadísticos.

La diferencia fundamental entre la inferencia estadística y el análisis descriptivo es que las últimas no requieren del cumplimiento de supuestos (homogeneidad de la matriz de varianza-covarianza, normalidad multivariada o la independencia, entre otros), o que las variables presenten una distribución esperada. Hace más de una década, la generalidad consistía en la aplicación de las técnicas descriptivas y exploratorias, pero con el desarrollo de la teoría estadística y de las herramientas computacionales, en la actualidad es mucho más frecuente la aplicación de la inferencia estadística, de pruebas que cumplen los supuestos (paramétricas), como de aquellas que no los exigen (no paramétricas o permutacionales).

Son diferentes los retos a los que se somete el investigador, al realizar el análisis e interpretación de bases de datos con diferentes variables. Entre las situaciones que pueden presentarse se destacan: (1) enfrentarse a numerosas variables y de diferente naturaleza, que han sido tomadas en diferentes localidades y periodos de muestreo, para lo cual se debe establecer una estrategia parsimoniosa que resuma y maximice la información ofrecida por los datos; (2) desarrollar modelos mentales que, previo al análisis, permitan establecer la mejor ruta en el ANDEA, dada la infinidad de técnicas que tiene a disposición; (3) en cuanto a la inferencia estadística, el tener la posibilidad de decidir si es necesario realizarla a nivel multivariado o si basta con pruebas univariadas para desarrollar el análisis requerido; (4) tener la capacidad de interactuar con diversas disciplinas, para abordar diseños multivariados que

integren variables de diversas tipologías, se convierte en uno de los retos más interesantes de este contexto.

Estas y otras situaciones son las que hacen del ANDEA un campo amplio y de mucha importancia para responder a problemáticas del entorno, que exigen el tratamiento de múltiples variables, en diferentes escalas espaciales y temporales, sin dejar a un lado la realización previa de un buen diseño experimental, que permita dar respuesta a hipótesis, controlando los tipos de error que se pueden presentar. Resumiendo lo anterior, el presente documento intentará proponer un esquema general para el ANDEA, que inicie con la exploración de los datos, a través de estadísticos básicos de tendencia central y de dispersión, así como de relaciones entre variables y muestras, para luego ser complementados con pruebas descriptivas e inferenciales con múltiples variables (en caso de ser necesario), para dar respuesta a preguntas e hipótesis sobre datos univariados y multivariados de naturaleza ecológica y ambiental.

1.2. Requisitos

El requisito inicial es el de contar en lo posible con múltiples variables (más de tres), para dar respuestas a preguntas de análisis, con muestras que han sido tomadas de manera apropiada y que representan a sus poblaciones estadísticas. Existen situaciones en que la aplicación de pruebas individuales, con cada una de las variables que caracterizan a la muestra en estudio, limita el poder revelar la estructura completa de los datos, dadas las interrelaciones que se pueden presentar entre variables.

Se requieren bases de conocimiento sobre aspectos de la estadística univariada, en especial de la bioestadística. Es importante tener conocimiento de la distribución normal, pruebas para comparar dos o más muestras, regresiones y análisis de varianza (paramétricos y no paramétricos), así como de los supuestos que deben cumplirse y las estandarizaciones o transformaciones que se pueden realizar a las variables. Las pruebas anteriores presentan un análogo multivariado, que será mucho más fácil de entender con las bases mencionadas.

Otro requerimiento importante es el conocimiento que se debe tener del álgebra lineal, especialmente del contexto matricial y vectorial que representa el insumo numérico de los análisis estadísticos multivariados (ver Capítulo 3). Vale la pena mencionar que el componente multivariado del ANDEA se soporta completamente sobre los elementos matemáticos mencionados, tanto en las operaciones realizadas como en los insumos numéricos y gráficos. Si bien este documento no se enfoca al contexto matemático de cada prueba, sí es necesario que el lector se familiarice con el origen y el significado de cada insumo matricial y de su importancia en cuanto al enfoque ecológico y ambiental.

Otro aspecto relevante es la experiencia previa que se tenga con el análisis y reporte de resultados estadísticos en las pruebas univariadas y multivariadas, como el registro de valores de significancia “p”, del estadístico o de los grados de libertad, elementos que son universales en el análisis de hipótesis estadísticas. Mucho más importante es la capacidad de análisis e interpretación de resultados numéricos y gráficos desde una perspectiva ecológica o ambiental, sin demeritar la importancia algebraica de las pruebas, aunque esto último no sea una prioridad para este documento.

1.3. Objetivos

Debido a su aplicación multidisciplinaria, el ANDEA presenta numerosos objetivos específicos, que, para el caso de este texto, pueden ser asignados a tres grupos generales, que se describen a continuación.

Un primer objetivo se orienta a brindar información, especialmente práctica, sobre los detalles y el significado ecológico o ambiental de diferentes técnicas multivariadas descriptivas e inferenciales, en cuanto a su contexto general, su aplicación, restricciones y su articulación con otras pruebas complementarias, priorizando en el componente práctico.

Como segundo objetivo, se busca promover la capacidad intuitiva para seleccionar las técnicas más apropiadas en el análisis de datos, dependiendo de la situación que se presente. Propendiendo porque cada técnica sea antecedida por un buen análisis exploratorio, a partir de gráficas que constituyen una de las fortalezas de la plataforma R.

El tercer objetivo consiste en la habilidad que se pueda adquirir en el desarrollo de cada técnica, por lo que cada una finaliza con un pequeño cuestionario, orientado al fortalecimiento conceptual e intuitivo, aunque para algunos casos, será necesario continuar entrenando para controlar cada una de las situaciones que se puedan presentar con estos análisis.

Estos objetivos pueden ser alcanzados si se cuenta con el planteamiento de preguntas análisis e hipótesis estadísticas de forma adecuada, para brindar facilidad al desarrollo e interpretación de las técnicas aplicadas. Esta habilidad se adquiere con el aprendizaje constante, especialmente si se trabaja con herramientas de cierto nivel de complejidad como la plataforma R, que se aplicará en capítulos posteriores. En resumen, *¡con la práctica se hacen buenos maestros para el análisis de datos!*

1.4. Tipos de datos

La forma común para presentar los datos es a través de matrices, en donde las filas agrupan muestras y/u observaciones y las columnas relacionan a las variables medidas para cada observación. Las observaciones pueden representar a individuos, lugares, periodos, etc. (Tabla 2).

Las variables no necesariamente deben ser del mismo tipo y pueden ser agrupadas en cuatro grandes grupos: (1) Las nominales, que suelen ser categóricas (ej. colores, sexos, etc.). (2) Las ordinales, definen un orden, aunque no mantengan una magnitud proporcional (ej. razas, niveles de contaminación, etc.). (3) Intervalos, presentan una magnitud proporcional a lo largo de una escala y la ubicación de su origen (cero) es arbitraria (ej. escalas de temperatura en grados Celsius o Fahrenheit). (4) Razón o proporción, corresponde a magnitudes relativas de las variables (ej. la edad, el peso, el tamaño, etc.).
En la Tabla 2 se presentan diferentes tipos de muestras y variables que pueden haber sido tomadas en estudios ambientales. (1) *Sitios* corresponden a las observaciones, para este caso son los lugares visitados, es normal que puedan ser numerados o codificados. (2) *Zonas y Regiones* corresponden a las muestras o factores que, a diferencia de los factores, cada una de sus categorías o niveles permiten presentar varios valores o réplicas por cada nivel. (3) *Textura* corresponde a una variable nominal o categórica, que representa a texturas del suelo presente en los sitios evaluados. (4) *Usos* representa a una variable de tipo ordinal, en la que se definen dos categorías de usos del suelo. (5) *K, Ca, Mg y Arcillas*, son variables continuas (con decimales), de tipo proporción, que corresponden a parámetros

físicos y químicos del suelo. (6) *Cyod, Dgbi* y *Lapae* son variables discretas (conteos de individuos), que representan la abundancia de tres especies vegetales, cuyos nombres han sido codificados.

Tabla 2. Valores hipotéticos para datos multivariados.

Sitios	Zonas	Regiones	Textura	Usos	T	K	Ca	Mg	Arcillas	Cyod	Dgbi	Lapae
1	Zona 1	Alta	G	1	27.2	19,54	98,2	126,64	10,75	10	0	3
2	Zona 1	Alta	G	1	27.5	21,94	63,8	113,49	21,35	12	8	4
3	Zona 1	Alta	G	1	26	23,89	63,3	118,42	28,61	11	9	7
4	Zona 1	Baja	SG	1	29	18,06	77,4	123,36	23,12	*df	3	9
5	Zona 1	Baja	SG	2	27.5	13,98	115,6	93,75	26,48	7	7	12
6	Zona 1	Baja	SG	2	28.5	13,69	98,1	100,33	32,67	8	9	12
7	Zona 1	Media	G	2	29	13,94	111,5	105,26	12,55	6	0	15
8	Zona 1	Media	G	2	*df	13,96	105,2	97,04	20,99	11	0	14
9	Zona 2	Media	G	1	28	4,33	15,1	15,62	19,67	11	0	13
10	Zona 2	Norte	F	1	25	5,76	17,8	16,45	21,71	13	5	1
11	Zona 2	Norte	F	1	24	1,5	7,3	2,63	30,15	15	0	7
12	Zona 2	Norte	F	1	25	1,24	5,8	1,97	19,67	17	0	0

Estas variables pueden ser evaluadas inicialmente a partir de estadística descriptiva univariada, como medidas de tendencia central (media, mediana, …), de dispersión (desviaciones, varianzas, …) o de posición (cuartiles, percentiles, …). Posteriormente pueden aplicarse técnicas multivariadas, para visualizar patrones con la integración simultánea de todas las variables.

Se puede presentar que falten datos para algunas variables, como se muestra en la Tabla 2 (**df*), lo cual podría ocurrir por varias razones, ya sea porque no fueron tomados o a que su valor es atípico por un posible problema en el instrumento de medida, lo cual indujo a su eliminación. Estos vacíos de información o de datos faltantes se pueden corregir con algunos métodos de imputación, como el realizado en técnicas de simulación con Monte Carlo en donde los valores faltantes son reemplazados por $m > 1$, en donde m es el número de simulaciones (3 a 10). Esto debe ser contrastado con intervalos de confianza, para incorporar cierto nivel de incertidumbre a los datos requeridos, sin perder de vista que estos datos no son reales sino estimados.

1.5. Generalidades del análisis de datos

En el ANDEA es común que nos enfrentemos a diferentes situaciones, en las que se encuentren distribuidos los datos, para llevarnos a pensar que deben ser analizados en un contexto multivariado. La práctica y la intuición son las herramientas fundamentales para poderlos categorizar en los siguientes tipos de casos:

1. Una muestra con diferentes variables (más de tres), en cada observación o muestra a evaluar.
2. Una muestra con dos grupos de diferentes variables medidas en cada observación.
3. Dos o más muestras (dos niveles de un factor) con distintas variables para cada observación.
4. Diferentes muestras en diferentes periodos (factores muestras y periodos), con distintas variables en cada observación.

Para el caso (1) se pueden presentar diferentes tipos de análisis a realizar:

- Probar la relación de las medias de cada variable en la muestra evaluada (ver Capítulo 4).
- Encontrar algunas dimensiones que sean combinación lineal de las variables, que definan alguna estructura de los datos y permitan explorar modelos lineales (ver Sección 6.1).
- Determinar algunas dimensiones que caractericen a las variables y a sus intercorrelaciones (ver Sección 6.2).
- Clasificar a las observaciones, de acuerdo con su nivel de similitud determinada por las variables (ver Sección 7.2).

Para el caso (2) también se pueden presentar diferentes tipos de análisis:

- Determinar el nivel de correlación en dos grupos de datos multivariados, evaluados sobre la misma muestra (ver Sección 8.1).
- Relacionar los dos conjuntos de variables, en donde uno depende del otro (ver Secciones 6.9.1 y 6.9.2).
- Seleccionar el subconjunto de variables de un grupo (explicativas) que presente la máxima relación con el otro grupo de variables (respuesta) (ver Secciones 8.2 y 6.9.1).

Para el caso (3) se describen los siguientes análisis a realizar:

- Ordenar a las observaciones y las variables en pocas dimensiones, de acuerdo con su similitud, y comparar gráficamente a las muestras (ver Secciones 6.1 - PCA, 6.3 y 6.4).
- Comparar dos muestras de acuerdo con los promedios de sus variables (ver Secciones 7.4, 7.5 y el Capítulo 8).
- Comparar a más de dos muestras de acuerdo con los promedios de sus variables (ver sección 7.5 y el Capítulo 8).
- Comparar a más de dos muestras de acuerdo con los promedios de sus variables, mediante métodos no paramétricos o que no requieran supuestos del MANOVA (ver Capítulo 8).
- Encontrar la combinación lineal de las variables que clasifique mejor a las muestras y verificar la membresía de cada observación a su muestra correspondiente (ver Sección 7.3).
- Determinar un método que permita validar la clasificación realizada de las observaciones o individuos a sus muestras, a partir de una escala de similitud o de distancia (ver Sección 7.2).

Para el caso (4) se describen los siguientes análisis a realizar:

- Ordenar a las observaciones y las variables en pocas dimensiones, de acuerdo con su nivel de similitud, y comparar gráficamente a las muestras y/o a periodos (ver Capítulo 6).
- Comparar a más de dos muestras y a periodos de acuerdo con los promedios de sus variables (ver Sección 7.5, Capítulo 8).
- Comparar a más de dos muestras y periodos de acuerdo con los promedios de sus variables, mediante métodos no paramétricos o que no requieran supuestos del MANOVA (ver Capítulo 8).

EJERCICIOS PROPUESTOS

1. Apoyándose en este capítulo de Conceptos Básicos y en revisión de literatura complementaria, realice un mapa conceptual en una página de Power Point, de Word o una imagen bien elaborada a mano, y resuma las etapas, los requisitos y los objetivos del análisis de datos ecológicos y ambientales - ANDEA.

2. Realice una breve descripción para relacionar a los aspectos que debe tener en cuenta un investigador que requiera en el procesamiento de datos - ANDEA.

3. Mediante una revisión exhaustiva, explique la importancia del álgebra lineal en la realización de análisis de datos, especialmente multivariados. Incluir las citas revisadas.

2. Generalidades de R y RStudio

INTRODUCCIÓN

2.1. Programa R y sus ventajas en el análisis de datos

R es un lenguaje y un entorno estadístico, desarrollado en 1995 por Ross Ihaka y Robert Gentleman - Universidad de Auckland, Nueva Zelanda (R Core Team, 2021). R se puede ejecutar en diferentes sistemas operativos como Windows, Mac y Linux. Una de las principales ventajas del trabajo en el entorno R, es su versatilidad en el desarrollo de figuras básicas y avanzadas, para el análisis de una o más variables, de igual forma, R permite desarrollar procedimientos univariados y multivariados de una forma eficiente. El programa R se puede descargar de una nube compuesta de servidores en diferentes lugares del mundo: https://cloud.r-project.org. Una vez al año se genera una versión principal de R y dos o tres versiones menores, por lo que es recomendable realizar las actualizaciones.

Las ventajas que ofrece R se resumen de la siguiente manera: es un programa gratuito, de código abierto, compatible con diferentes plataformas (Windows, Linus o MAC OSX). Presenta una gran variedad de pruebas estadísticas, permite reproducir procedimientos de forma sencilla, mediante la generación de archivos que contienen las rutinas, procedimientos o secuencias de comandos. Cuenta con una comunidad a nivel mundial, que constantemente realiza aportes para ofrecer más opciones de procedimientos. De esta manera, los usuarios tienen la cultura de compartir sus códigos, sus experiencias y de apoyarse entre sí. Cuenta con miles de aplicaciones, llamadas paquetes o librerías, que cuentan con su respectiva documentación y permiten realizar múltiples procedimientos estadísticos, entre otros.

La instalación de R para Windows se realiza con el siguiente procedimiento:

1. Descargar la versión más actual de R en https://cran.r-project.org/bin/windows/base/
2. Haga doble clic en el archivo que acaba de descargar para instalar R.
3. Hacer clic en Aceptar y dar “siguiente” en todos los procedimientos requeridos.

Para la versión de R, para MAC OSX, se realiza el siguiente procedimiento:

1. Ingresar a https://cran.r-project.org/bin/macosx/
2. Hacer doble clic en el archivo que acaba de descargar para instalar R.
3. Hacer clic en Aceptar y dar “siguiente” en todos los procedimientos requeridos.

Estos procedimientos se realizan para ir renovando las versiones que periódicamente son publicadas, debido a que en algunos casos necesitaremos de las mejoras que estas presentan, o porque algunos complementos o paquetes que requerimos instalar no son compatibles con versiones antiguas. El procedimiento no se puede realizar en RStudio, para ello se requiere abrir la consola de R (normalmente hay un acceso directo en el escritorio del computador). El procedimiento se describe a continuación:

1. Es necesario haber instalado el paquete *installr* y cargarlo con library(*installr*).
2. Posteriormente aplicamos el comando *updateR ().* Se realizarán una serie de preguntas orientadas a la actualización del R sin necesidad de desinstalar la versión actual. Recomiendo aceptar las opciones que el programa vaya sugiriendo.

El comando *updateR ()* permite realizar lo siguiente: encontrar la última versión R, descargarla, ejecutar su instalador, eliminar el archivo de instalación de la versión anterior, copiar y actualizar paquetes antiguos a la nueva instalación R.
Para más información, se puede ingresar al tutorial del paquete *instalr* desde: https://cran.r-project.org/web/packages/installr/index.html

2.2. Programa RStudio

RStudio es un programa de entorno integrado de desarrollo (IDE por sus siglas en inglés). En términos generales, este programa genera un espacio amigable, para desarrollar y ejecutar los comandos y códigos de R de una forma eficiente. Cuenta con herramientas para revisar los códigos realizados, ejecutar figuras, entre otros procedimientos requeridos por los analistas de datos. Las ventajas que ofrece RStudio se pueden resumir en lo siguiente: es un IDE de código abierto y por ello se actualiza con relativa frecuencia, su entorno es más amigable y completo que el de la plataforma de R. También permite compartir los códigos realizados entre comunidades de usuarios relacionadas. Esta plataforma se actualiza un par de veces al año. La versión que se instala en el computador, para sistemas de Windows, MAC y Linux, se consigue en:
https://www.rstudio.com/products/rstudio/download/
Adicional a la versión instalada, se cuenta con otra opción en línea (RStudio Cloud), a la que se puede acceder por este enlace: https://rstudio.cloud/content/
En caso de requerir actualizar la versión de R, es necesario actualizar también a la versión de RStudio, para que haya compatibilidad, de lo contrario, no se podrá realizar la interfaz entre estos programas.

2.3. Tutoriales de R Y RStudio

2.3.1. Tutorial rápido de R

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

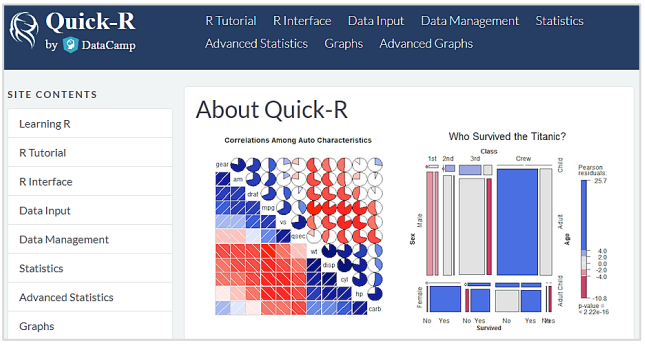


Figura 1. Imagen de la plataforma gráfica con comandos de R.

2.3.2. 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 (Figura 2). Se puede acceder a ella en el siguiente enlace: https://rseek.org/

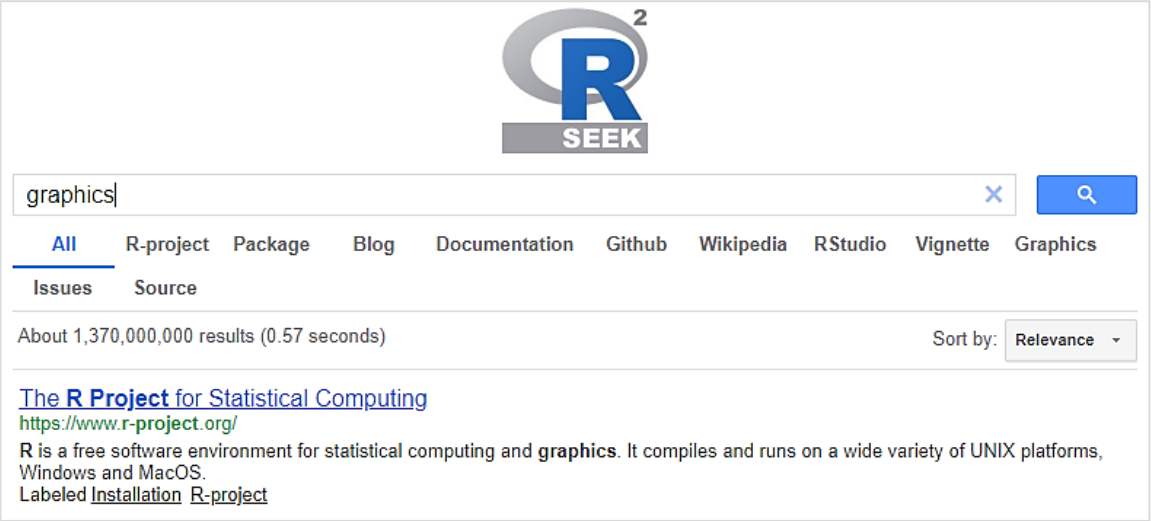


Figura 2. Imagen de motor de búsqueda de utilidades de R.

2.3.3. 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/
Hojas de trucos de RStudio, cuenta con diferentes guías rápidas para ejecutar procedimientos de ejecución y de programación en R:
https://www.rstudio.com/resources/cheatsheets/
En estos enlaces se puede visualizar algunos de los trucos de R, como: Introducción al R, Introducción al RStudio, Visualización de datos en ggplot2, entre otros.
Actualmente se cuenta con un conjunto de paquetes de R, denominados “Tidyverse”, diseñados para apoyar el proceso de importar, transformar, visualizar, modelar y comunicar la información en procesos de “ciencia de datos”. Han tomado mucho auge y son los que dominan los procedimientos de análisis de datos en diferentes ámbitos:
https://soka.gitlab.io/blog/post/2019-10-06-r-tidyverse-intro/

2.3.4. Opciones gráficas

El siguiente demo se ejecuta en R, permite visualizar opciones gráficas (Figura 3).
demo(image)

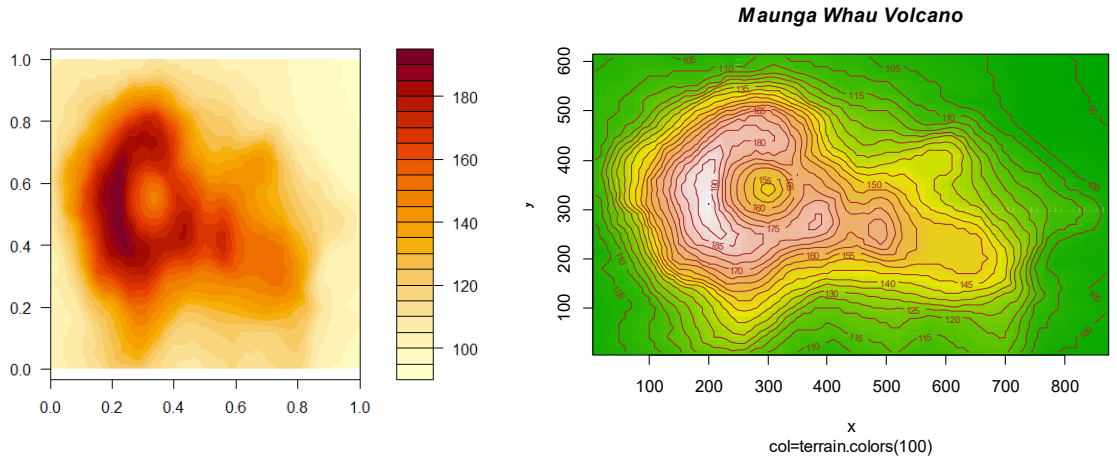


Figura 3. Representaciones de curvas de nivel en mapas realizados con R.

La siguiente demostración permite visualizar diferentes opciones de figuras tridimensionales, incluidos los modelos de elevación digital de terreno en un volcán (izquierda) (Figura 4).
demo(persp)

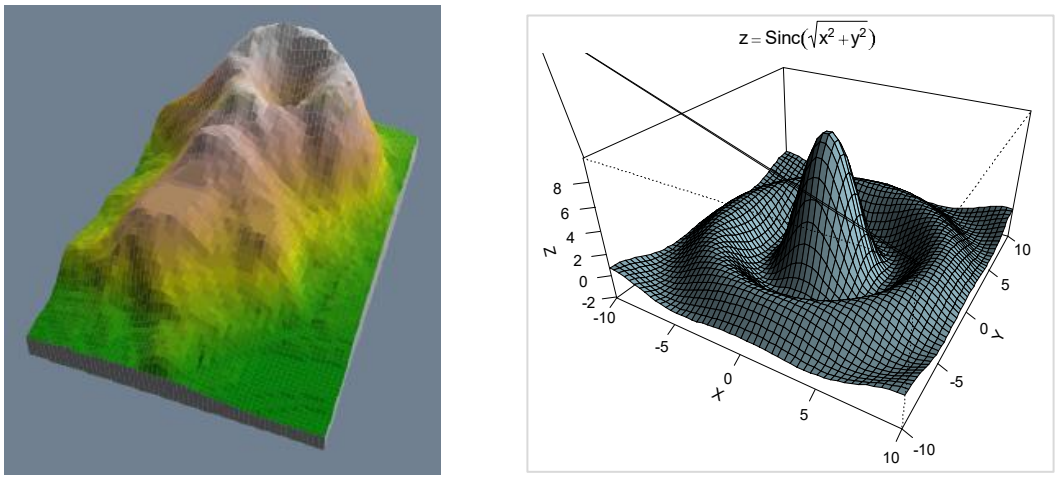


Figura 4. Representaciones tridimensionales en R.

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 1. NOCIONES DE RSTUDIO

A continuación, se describen algunos conceptos generales, relacionados al entorno de comandos de R, los cuales consisten en operaciones y funciones aritméticas básicas, así como en la creación de bases de datos (dataframes) y subconjuntos de datos.

```
#--------------
# 1) Operaciones aritméticas básicas

# Suma
5 + 7
## [1] 12
# Resta
5 - 3
## [1] 2
# Multiplicación
5 * 7
## [1] 35
# División
5/3
## [1] 1.666667
# Exponentes
2^3
## [1] 8
# Devolver el resto de la división de 5/3
5 %% 3
## [1] 2
# Logaritmos y exponenciales :
x = 5/3
log2(x)     # Logarítmo en base 2 de x
## [1] 0.7369656
log10(x)    # Logarítmo en base 10 de x
## [1] 0.2218487
exp(x)      # Exponencial de x
## [1] 5.29449
# Funciones trigonométricas :
cos(x)    # Coseno de x
## [1] -0.09572355
sin(x)    # Seno de x
## [1] 0.995408
tan(x)    # Tangente de x
## [1] -10.39878
acos(x/100)   # Arcoseno de x
## [1] 1.554129
asin(x/100)   # Arcseno de x
## [1] 0.01666744
atan(x/100)   # Arco tangente de x
## [1] 0.01666512
```

Ejemplo 1. NOCIONES DE RSTUDIO (*CONT.*)

```
abs(x)    # Valor absoluto de x
## [1] 1.666667
sqrt(x)   # Raíz cuadrada de x
## [1] 1.290994

#--------------
# 2) Asignación de valores a objetos o a variables
sitios <- 2         # Número de sitios = 2
sitios = 2          # Otra forma
n.sitios <- "dos"   # Número de sitios como un caracter
dos.sitios <- TRUE  # Objeto lógico

# Imprimir los valores asignados
sitios         # Imprime el número de sitios
## [1] 2
print(sitios)  # Otra forma
## [1] 2
n.sitios       # Sitios asignados como caracter
## [1] "dos"
dos.sitios     # Sentencia lógica
## [1] TRUE
# Identificar el tipo de asignación
class(sitios)         # Tipología de los sitios
## [1] "numeric"
class(n.sitios)       # Tipología del número de sitios
## [1] "character"
class(dos.sitios)     # Tipología de la sentencia lógica
## [1] "logical"
is.numeric(sitios)    # Identificador: is.numeric(), is.character(), is.logical()
## [1] TRUE
as.character(sitios)  # Cambiar: as.numeric(), as.character(), as.logical()
## [1] "2"
# Operaciones con las asignaciones
3 * sitios     # Multiplicar 3 por el número de sitios
## [1] 6
# Área de un rectángulo
altura <- 20      # Altura del rectángulo
ancho <- 10       # Ancho
área <- altura*ancho   # Cálculo de área
área              # Imprimir el área
## [1] 200
print(área)       # Otra forma de imprimir el área
## [1] 200
ls()        # Permite visualizar a los objetos creados
## [1] "altura"     "ancho"      "área"       "dos.sitios" "n.sitios"
## [6] "sitios"     "x"
```

Ejemplo 1. NOCIONES DE RSTUDIO (*CONT.*)

```
#--------------
# 3) Vectores
sitios <- c(2, 3, 2, 3)   # Vector sitios
sitios    # Imprimir el vector
## [1] 2 3 2 3
sitios <- c("dos", "tres", "dos", "dos")    # Vector como caracter
sitios
## [1] "dos"  "tres" "dos"  "dos"
abundancia <- c(TRUE, FALSE, TRUE, TRUE)    # vector con elementos lógicos
abundancia
## [1]  TRUE FALSE  TRUE  TRUE
sitios <- c(2, 3, 2, 3)   # Vector sitios
names(sitios) <- c("dos", "tres", "dos", "dos")   # Nombres de los elementos del vector
sitiossitios
sitios <- c(dos= 2, tres= 3, dos= 2, dos= 2)      # Otra forma
sitios
##  dos tres  dos  dos
##    2    3    2    2
length(sitios)    # Elementos del vector sitios
## [1] 4
sitios[2]         # Segundo elemento del vector sitios
## tres
##    3
sitios[c(1,4)]    # Primer y cuarto elemento del vector
## dos dos
##   2   2
sitios [1:3]      # Tres primeros elementos del vector sitios
##  dos tres  dos
##    2    3    2
sitios [-1]       # Eliminar el primer elemento del vector
## tres  dos  dos
##    3    2    2
sitios[abundancia == TRUE]  # Abundancias con valor verdadero
## dos dos dos
##   2   2   2
sitios[sitios >= 2]  # Sitios con abundancias con valores mayor o igual a 2
##  dos tres  dos  dos
##    2    3    2    2
sitios[sitios != 3]  # Sitios con abundancias diferentes a 3
## dos dos dos
##   2   2   2
#--------------
# 4) Operaciones con escalares y vectores

sitios.1 <- c(A = 2, B= 3, C= 2, D= 3)   # Vector sitios en muestreo 1
sitios.1
```

Ejemplo 1. NOCIONES DE RSTUDIO (*CONT.*)

```
## A B C D
## 2 3 2 3
sitios.2 <- c(A = 4, B= 6, C= 4, D= 5)   # Vector sitios en muestreo 2
sitios.2
## A B C D
## 4 6 4 5
sitios * 3            # Producto de los sitios x 2
##  dos tres  dos  dos
##    6    9    6    6
sitios.1 * sitios.2   # Producto de los sitios 1 y 2
##  A  B  C  D
##  8 18  8 15
sitios.1 <- c(2, 3, 2, 3)   # Vector sitios en muestreo 1
sitios.1
## [1] 2 3 2 3
max(sitios.1)         # Valor máximo
## [1] 3
min(sitios.1)         # Valor mínimo
## [1] 2
range(sitios.1)       # Rango de los datos del vector
## [1] 2 3
sum(sitios.1)         # Suma de los datos del vector
## [1] 10
prod(sitios.1)        # Producto de los datos del vector
## [1] 36
mean(sitios.1)        # Promedio de los datos del vector
## [1] 2.5
sd(sitios.1)          # Desviación estándar
## [1] 0.5773503
var(sitios.1)         # Varianza de los datos del vector
## [1] 0.3333333
sort(sitios.1)        # Organizar elementos de forma ascendente
## [1] 2 2 3 3
#--------------
# 5) Matrices
Matriz <- matrix(1:15, nrow = 3, ncol = 5)    # nrow y nrow: Número de filas y columnas
Matriz
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    4    7   10   13
## [2,]    2    5    8   11   14
## [3,]    3    6    9   12   15
Matriz <- matrix(1:15, nrow = 5, ncol = 3)    # 1:15: Secuencia de valores de 1 a 15
Matriz
```

Ejemplo 1. NOCIONES DE RSTUDIO (*CONT.*)

```
##      [,1] [,2] [,3]
## [1,]    1    6   11
## [2,]    2    7   12
## [3,]    3    8   13
## [4,]    4    9   14
## [5,]    5   10   15
Matriz <- matrix(c(1:15),5,3, byrow= FALSE)   # 5,3: Número de filas y columnas
Matriz
##      [,1] [,2] [,3]
## [1,]    1    6   11
## [2,]    2    7   12
## [3,]    3    8   13
## [4,]    4    9   14
## [5,]    5   10   15
# Primera opción para crear una Matriz
Matriz <- matrix(c(5, 6, 7, 8, 9,    # Elementos de la Matriz
                   2, 4, 5, 9, 8,
                   7, 3, 4, 8, 7),   # 15 elementos
                   5,3, byrow=F)     # 5 columnas y 3 filas
Matriz
##      [,1] [,2] [,3]
## [1,]    5    2    7
## [2,]    6    4    3
## [3,]    7    5    4
## [4,]    8    9    8
## [5,]    9    8    7
# Otra forma con las funciones cbind() o rbind()
col1 <- c(5, 6, 7, 8, 9)        # Col 1 a 3, son las tres columnas de la Matriz
col2 <- c(2, 4, 5, 9, 8)
col3 <- c(7, 3, 4, 8, 7)
Matriz <- cbind(col1, col2, col3)           # Nombres de las columnas
rownames(Matriz) <- c("A","B","C","D","E")  # Nombres de las filasMatriz
Matriz
##   col1 col2 col3
## A    5    2    7
## B    6    4    3
## C    7    5    4
## D    8    9    8
## E    9    8    7
# Otra forma con comandos "colnames" y "dimnames"
Matriz <- matrix(
          data = c(5, 6, 7, 8, 9,    # Elementos de la Matriz
                   2, 4, 5, 9, 8,
                   7, 3, 4, 8, 7),
          nrow = 5, byrow = FALSE,                  # nrow= columnas.
          dimnames = list(c("A","B","C","D","E"),    # dimnames= lista con vector fila y
columna
                     c("col1", "col2", "col3")))
Matriz
```